Donghong Ji
Guozheng Xiao (Eds.)

LNAI 7717

Chinese Lexical Semantics

13th Workshop, CLSW 2012
Wuhan, China, July 2012
Revised Selected Papers

Springer

Lecture Notes in Artificial Intelligence 7717

Subseries of Lecture Notes in Computer Science

Donghong Ji Guozheng Xiao (Eds.)

Chinese Lexical Semantics

13th Workshop, CLSW 2012
Wuhan, China, July 6-8, 2012
Revised Selected Papers

Springer

Volume Editors

Donghong Ji
Wuhan University
Computer School
Wuhan 430072, China
E-mail: dhji@whu.edu.cn

Guozheng Xiao
Wuhan University
College of Chinese Language and Literature
Wuhan 430072, China
E-mail: gzxiao@foxmail.com

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-36336-8 e-ISBN 978-3-642-36337-5
DOI 10.1007/978-3-642-36337-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012956180

CR Subject Classification (1998): I.2.7, I.2.6, H.2.4, H.3.6

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Since its initiation in 2000, the Chinese Lexical Semantics Workshop (CLSW) has become one of the most important forums for Chinese lexical semantics and related fields, including theoretical linguistics, computational lexicography, and natural language processing applications. So far, the series has been held in different cities in the Asia-Pacific region.

The 13th Chinese Lexical Semantics Workshop (CLSW 2012) was held in Wuhan, China, in July 2012. This workshop was organized by Wuhan University (China). Topics of interest involved theoretical, computational, and other related issues of lexical semantics research.

The Program Committee received a total of 169 papers representing work by academics and practitioners, and we would like to thank all of them for their work. The committee used a double-blind reviewing process and as a result 67 articles (39.6%) were accepted as full papers and 17 (10%) as short (poster) papers.

We would like to thank the invited and plenary speakers: Yukio Tono (Tokyo University of Foreign Studies), on "Lexical Semantics and Pedagogical Lexicography"; Lee Jong-Hyeok (Pohang University of Science and Engineering), on "Lexical Semantics in Machine Translation"; Guodong Zhou (Soochow University), on "Semantic Analysis in NLP"; Guozheng Xiao (Wuhan University), on "Case and Corpus Study of Lexical Concepts."

The success of this conference was only possible with the support of the Program Committee members. We would like to acknowledge Yanxiang He (Wuhan University) and Benjamin Tsou (Hong Kong Institute of Education), the Conference Chairs, for their generous support. We are also most grateful to Chu-Ren Huang (Hong Kong Polytechnic University) and the Advisory Committee for their help in promoting the workshop.

We are also grateful to the paper reviewers who devoted their time and energy in reviewing the papers selected for the proceedings.

October 2012

Donghong Ji
Guozheng Xiao

Organization

The 13th Chinese Lexical Semantics Workshop (CLSW 2012) was held in Wuhan, China, in July 2012. This workshop was organized by the faculty of the Computer School, Wuhan University (China).

Conference Chairs

Yanxiang He	Wuhan University
Benjamin Tsou	Hong Kong Institute of Education

Advisory Committee

Chu-Ren Huang	Hong Kong Polytechnic University
Donghong Ji	Wuhan University
Kim-Teng Lua	COLIPS
Mei-chun Liu	Taiwan Chiao Tung University
Xinchun Su	Xiamen University
Shiwen Yu	Beijing University
Shu-Kai Hsieh	Taiwan University
Chin-Chuan Cheng	National Taiwan Normal University
Benjamin Tsou	Hong Kong Institute of Education

Program Committee

Program Chairs

Jie Xu	University of Macau
Donghong Ji	Wuhan University

Members

Chao-Jan Chen	Taiwan Chi Nan University
Qunxiu Chen	Beijing University
Minghui Dong	Institute for Infocomm Research, Singapore
Guohong Fu	Heilongjiang University
Hong Gao	Nanyang Technical University
Tingting He	Central China Normal University
Degen Huang	Dalian University of Technology
Xuanjing Huang	Fudan University
Minghu Jiang	Tsinghua University
Kathleen Ahrens	Hong Kong Baptist University
Shiyong Kang	Ludong University

Kwong Olivia	City University of Hong Kong
Hui-ling Lai	Taiwan Cheng Chi University
Hongfei Lin	Dalian University of Technology
Mei-chun Liu	Taiwan Chiao Tung University
Qun Liu	Chinese Academy of Sciences
Yao Meng	Fujitsu Research Center, China
Juanzi Li	Tsinghua University
Haihua Pan	City University of Hong Kong
Chengqing Zong	Chinese Academy of Sciences
Le Sun	Chinese Academy of Sciences
Maosong Sun	Tsinghua University
Zhifang Sui	Beijing University
Chong Teng	Wuhan University
Weixin Tian	China Three Gorges University
Houfeng Wang	Beijing University
Hui Wang	National University of Singapore
Yunfang Wu	Beijing University
Shu-Kai Hsieh	Taiwan University
Endong Xun	Beijing Language and Culture University
Erhong Yang	Beijing Language and Culture University
Hua Yang	Guizhou Normal University
Zhengdao Ye	Australian National University
Hao Yu	Fujitsu Research Center, China
Yulin Yuan	Beijing University
Hongying Zan	Zhengzhou University
Min Zhang	Institute for Infocomm Research, Singapore
Yangsen Zhang	Beijing Information Science and Technology University
Zezhi Zheng	Xiamen University
Siaw-Fong Chung	National Chengchi University
Guodong Zhou	Soochow University
Qiang Zhou	Tsinghua University

Organizing Committee

Litang Liu	Wuhan University
Hongmiao Wu	Wuhan University
Chong Teng	Wuhan University
Han Ren	Wuhan University
Mingyao Zhang	Wuhan University

Publication Chair

Pengyuan Liu	Beijing University

Table of Contents

Applications on Natural Language Processing

Corpus Linguistics

Lexical Computation

Lexical Resources

Lexical Semantics

New Methods for Lexical Semantics

Other Topics

MT-Oriented and Computer-Based Subject Restoration for Chinese Empty-Subject Sentences

Guo-Nian Wang[1,2], Yi Qin[2], Min Jiang[2], and Qiu-Rong Zhao[2]

[1] College of Chinese Language and Literature, Wuhan University, Wuhan 430072, China
touchbobby@gmail.com
[2] School of Foreign Languages, China University of Geosciences, Wuhan 430074, China
{touchbobby,qinyi18,qiurong.zhao}@gmail.com,
jmyyb@163.com

Abstract. The Chinese language is rich in sentences with covert subject. The complicated formation mechanisms of empty-subject sentences (ESS) impose great barriers to Natural Language Processing (NLP) as far as the subject tagging and processing is concerned. This paper examines three ways in which Chinese empty-subject sentences are formed. A computer program blueprint based on discourse analysis is subsequently drawn so as to automatically restore, tag, and process ESS empty subjects, a step significant to the accuracy of Machine Translation (MT).

Keywords: empty subject, economy principle, discourse analysis, subject restoration, NLP, MT.

1 Introduction

Different from null-subject sentences (also known as "absolute sentences" in Chinese), which possess no overt syntactic subjects, and which could by no means have their "subject" retrieved or restored, the empty-subject sentences (ESS) under discussion refer to those whose syntactic subjects—retrievable and restorable, if need be—are left out for various reasons. These covert subjects (called "empty subjects" hereinafter) fall into the "Empty Category" proposed by Chomsky [1]. They are visible in the syntactic and semantic frameworks, but are phonetically absent.

Among the 60+ languages supported in Google's Web-based Translator, Chinese (used as source language) is believed to be the one that generates the worst-quality target languages, be it Japanese or any Indo-European language. The reason for this awkward situation, as Google CEO puts it, is form-related. Most European languages share similarities in word-formation and grammar rules, which are robust in forms. Compared with Indo-European family, Chinese lacks rich forms of words and grammatical markers; there is even no clear-cut state between words and non-words, or a standardized classification of parts of speech. Empty subjects of various kinds, among other linguistic ellipses, add to the barriers that render Chinese-related Machine Translation (MT) unsatisfactory.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 1–10, 2013.

Lin argues that the linguistic errors in target language generated from MT with Chinese as source language are derived from erroneous analyses, inter-lingua, generational rules, or dictionaries, etc. [2] All these possible errors are more or less related to the studies and rule abstraction of Chinese, corpus design or programming.

We attempt to approach—by way of stereotyped expression, economy principle and discourse features in Chinese—the formation mechanism of empty subjects and the programmed restoration of such subjects. Our small amounts of experimental samples have testified the fact that complete syntactic structures (S-V and S-V-O) in Chinese are capable of generating far more accurate and readable MT target languages than are those with incomplete grammatical elements.

As a trial study of the possible approaches to automatic Chinese discourse analysis and subject restoration and tagging, it defines or modifies some terms that may have been established or formularized in the Chinese language. It is not our intention to challenge any authorities or norms. Instead, these re-definitions are for shear computational purposes. They include:

- "Empty-Subject Sentence" (ESS). The term "empty subject", as part of Chomsky's "Empty Category", is generally known and agreed upon, while "ESS" is seldom employed to refer to such sentences as have phonetically intangible "empty subjects." In this paper, ESS is narrowly and seriously defined as Type 1 sentences (shown in Table 1) with omitted subject (termed as "empty subject" hereinafter). Types 2—4 are not our concern but some will assist us with the programming blueprint.

Table 1. Categorization of traditional "non-subject" sentences

Types	Names	Illustrations	Notes
1	Omitted Subject	—你吃晚饭了吗？—(e)吃了。 *—Ni chi wanfan le ma?* —(e) *Chi le*. (Chinese *pinyin*, literally) —Did you have dinner? —(e) Did. (Translated, literally)	The subject ellipsis (Chinese) renders it hard for MT. The covert/empty subject (e) that is left out in discourse for economy purpose needs restoring and tagging for better MT result. Type 1 sentences are referred to in this paper as "Empty-Subject Sentences". Types 2-4 are not the concern of the study.
		—他昨天干嘛去了？—(e)看房子。 *—Ta zuotian gan ma qu le?* —(e) *Kan fangzi.* —What did he do outside yesterday? —(e) Searched for a house.	
		(e)连续看了两天电视剧，我眼睛又酸又疼。 (e) *Lianxu kan le liangtian dianshiju, wo yanjing you suan you teng.* (e) Have watched TV series for two successive days, (so) I have sore and aching eyes.	
		我在工作中不断摸索，(e)在摸索中不断进步 *Wo zai gongzuo zhong buduan mosuo,* (e) *zai mosuo zhong buduan jinbu.* I keep learning the ropes in my job, (and) (e) keep progressing in the learning.	
2	Implied Logical Subject	他打算明天(e)出差一趟。 *Ta dasuan mingtian* (e) *chuchai yitang.* He plans to (e) go out on business tomorrow.	These sentences possess syntactical subjects. It is unnecessary to restore the implied logical subjects in the infinitive phrases.
		我们劝他(e)早点离开。 *Women quan ta* (e) *zaodian likai.* We persuaded him to (e) leave early.	

Table 1. (*continued*)

3	**NP-Traced Subject**	(e_t)疼得她(t)差点晕过去。(SS)→她疼得差点晕过去。(DS) (e_t) *Teng de ta* (t) *chadian yun guoqu.* →*Ta teng de chadian yun guoqu.* (e_t) So painful (that) she (t) almost fainted. →She was so painful that she almost pained. (e_t)饿得小贝(t)嗷嗷叫。(SS)→小贝饿得嗷嗷叫。(DS) (e_t) *E de Xiaobei* (t) *aoao jiao*. →*Xiaobei e de aoao jiao.* (e_t) So hungry (that) the little baby (t) cried out loud. →The little baby was so hungry that he cried out loud.	The Chinese deep structure (DD) belongs to the typical S-V category. It is unnecessary to restore the NP-traced subject (e_t) in the surface structure (SS).
4	**Null Subject**	下雨了。 *Xiayu le.* (It's) Raining. 禁止吸烟。 *Jinzhi xiyan.* No smoking.	These (in Chinese) are called "absolute null-subject sentences", whose null "subjects" are by no means restorable.

- "Programmed Simple Sentence" (PSS) and "Programmed Complex Sentence" (PCS). They are no match for or equivalent to traditionally-defined grammatical simple sentences or complex sentences. It is assumed that the colon, semi-colon, period, question mark and exclamation mark in Chinese share the similar or same syntactic function, then a PCS is the cluster of texts between any such two marks (including themselves). A PSS is usually an ESS, a null-subject sentence, a regular S-V or S-V-O sentence, or even an idiom or any other phrase—any of which could stand alone as a PSS (similar to a grammatical simple sentence), or stay together with another one or more than one PSS to form a PCS (similar to a grammatical complex sentence). See 3.2 for details.

2 Mechanisms for Subject Ellipsis

So far, the categorization of non-subject sentences is approaching a consensus [3—6], as is depicted in Table 1 above. The mechanisms by which these covert subjects come into being, however, are still at dispute. [3][7—8] We believe the stereotyped expressions, economy principle and discourse features typical of Chinese combine to restrain the uses of too many syntactic subjects.

2.1 Stereotyped Expressions

Subject ellipsis occurs in Indo-European languages very occasionally, but it thrives in the Chinese language. Lv stated that "run-on sentences" (流水句, *liushui ju*, running-water sentences) are very particular but commonly seen in Chinese, and that "minor sentences" (小句, *xiaoju*, simple sentences or clauses) could even stand together—without any connectives—as a syntactically-legal (complex) sentence. [9] Such

run-on sentences and minor sentences mostly lack subject markers but are capable of composing a complex sentence, dialogue or paragraph. They differ a lot from the real "clauses" in, for example, English.

(1) 我从北京到徐州，(e_1)打算跟着父亲奔丧回家。(e_2)到徐州见着父亲，(e_3)看着满院狼藉的东西，(e_4)又想起祖母，(e_5)不禁簌簌地流下眼泪。(朱自清《背影》[10])

Wo cong Beijing dao Xuzhou, (e_1) *dasuan gen zhe Fuqin bensang huijia.* (e_2) *Dao Xuzhou jian zhe Fuqin,* (e_3) *kan zhe man yuan langjide dongxi,* (e_4) *you xiangqi Zumu,* (e_5) *bujin susude liuxia yanlei.* (*Zhu Ziqing*《*Beiying*》)

I went from Beijing to Xuzhou, (e_1) (I) planned to attend a funeral home with Father. (e_2) (I) saw Father in Xuzhou, (e_3) (I) watched the whole lot of mess in the yard, (e_4) and (I) thought of Grandma, (e_5) (I) could not hold back (my) tears. (Excerpted from Zhu's essay "The Sight of Father's Back")

This paragraph consists of 6 run-on sentences, of which only one is equipped with a subject, and the rest 5 belong to ESS in question. MT of the 5 ESS sentences has witnessed inaccurate and unreadable target language. With their empty subjects "I" (e_1—e_5) restored, however, the target quality improves to an accurate and readable level.

Besides the stereotyped thinking mode, expressions and constructions, the economy principle also contributes to the loss of Chinese subjects in most cases.

2.2 Principle of Economy

The Principle of Economy proposed by French linguist André Martinet is usually employed to address a wide variety of ellipses in language performances. It is "the principle of least effort, which makes him restrict his output of energy, both mental and physical, to the minimum compatible with achieving his ends."[11] The Chinese language users seem to have taken the greatest advantage of this principle, resulting in huge amounts of ellipses of syntactic elements such as the losses of subjects (e_1 and e_3) and object (e_2) in the following dialogue.

(2) a. — (e_1)吃(e_2)了吗?
a. — (e_1) *Chi* (e_2) *le ma*?
a. — (e_1) Ate (e_2)? (Did you have your meal?)
b. — (e_3)吃(e_2)了。
b. — (e_3) *Chi* (e_2) *le.*
b. — (e_3) Ate (e_2). (Yes, I did.)

No subjects ("you" and "I", respectively) are phonetically existent in this dialogue, nor is the object ("meal"). The two sentences, however, are grammatically legal and self-sufficient for interactive purpose in a Chinese phatic context. As Huang argues, "Ellipsis (in this manner) is a logically sound and linguistically explicit way of communication."[12] It is socially accepted and habitually practical. A supplement of the lost subjects, however, is most likely to make the two parties in communication feel unnatural and unpleasant; in the meantime, it offends the principle of economy, as is shown in Case (3).

(3) a. —? 你吃(饭)了吗?
a. —? *Ni chi* (*fan*) *le ma*?
a. —? Did you eat/have (meal)?
b. —? 我吃(饭)了。
b. —? *Wo chi* (*fan*) *le*.
b. —? I ate/did (it).

Whether *fan* (饭, meal/it) appears or disappear in this discourse, the forcible use of personal pronouns *Ni* (你, You) and *Wo* (我, I) as syntactic subjects render the dialogue ineffective and uneconomical. Ellipsis—if used appropriately, as is mostly the case in Chinese—will help achieve such multiple effects as coherence and economy, among others. [13]

2.3 Discourse Features

As is emphasized by Functionalists, ellipsis, especially ellipsis of empty categories, could not be analyzed from mere syntactic perspective. But rather, a pragmatic approach should also be adopted [7], for it is a human behavior relying on contexts and the locale of speech.

The empty subject, as one of ellipsis instruments, bears and shares the general features of ellipsis. The nature of Chinese subject ellipsis, until very recently, has seldom been identified. We define it as a typical pragmatic performance. Subject ellipsis occurs in language uses, and thus is only perceivable and restorable in greater linguistic contexts, i.e., discourse. Examine Case 1, and we will find the truth that some empty subject could only be correctly retrieved in discourse contexts, not at sentential level. Zoom in the third to sixth run-on sentences to compose Case 4:

(4) (e_1)到徐州见着父亲(t)，(e_2/e_t)看着满院狼藉的东西，(e_3)又想起祖母……
(e_1) *Dao Xuzhou jian zhe Fuqin* (t), (e_2/e_t) *kan zhe man yuan langjide dongxi*, (e_3) *you xiangqi Zumu…*
(e_1) (I) saw Father (t) in Xuzhou, (e_2/e_t) (I/Father) watched the whole lot of mess in the yard, (e_3) and (I) thought of (my) Grandma…

The first empty subject "I" (e_1) is easily perceivable in the first ESS itself, for the guy who saw Father (父亲, *Fuqin*) could not be Father himself, but the one who narrates the story. The problem, however, strikes at the second run-on ESS. The one who "watched the whole lot of mess in the yard" could be either "I" or "Father", and more often than not, the subsequent subject is identical to the preceding object. In this dilemma, analysis at sentential level fails to restore the real hidden subject (e_2). The term "Grandma" in the third run-on sentence helps to shoot the trouble: that is "my Grandma" in Chinese diction habit. That is to say, the third sentence has "I" as subject. It is rare to see, among a sequence of 3 or more run-on sentences, the first and third subjects are identical ("I") but the second in between is different ("Father"). The more likely and actually correct empty agent of "watched", therefore, is "I" rather than "Father." This is only perceivably right in a larger context than sentence level, i.e., discourse.

For all the above-mentioned factors that are characteristic of the Chinese language and of ESS empty subjects, the forthcoming chapter focuses on the design of a programming blueprint that is effective for automatic discourse analysis and subject restoration, a preparatory step for better MT production. In more technical terms, we aim at formalizing, on the basis of discourse scanning and analysis, the empty subjects of ESS sentences.

3 Restoration of Empty Subjects

To restore the empty subjects left out in a given discourse, it is necessary, above all, to determine the very texts that are empty-subject sentences. Two blueprints are designed for this critical stage: program-based ESS identification, and program-based empty subject restoration (retrieval and tagging).

3.1 ESS Identification

The following steps are programmed for computational operation in order to exclude non-ESS sentences. The remaining texts of a given discourse are mostly what we need for further computation, the ESS.

Exclusion of Null-Subject Sentences (NSS). The NSS as shown in Type 4, Table 1 are relatively fixed and set in scope and number. They are irregular in terms of syntax, but are regularly limited in terms of construction. Suppose a given discourse (referred to as "A" hereinafter) consists of all Programmed Simple Sentences (PSS). By scanning sequentially through all PSS in A and comparing with NSS corpus, the program will spot all possible NSS (referred to as "B" hereinafter) before excluding them. They are left intact or sent for MT processing with NSS parallel corpus. The remaining non-NSS (texts of (A-B)) are entered for a second-step computation.

Exclusion of Regular S-V / S-V-O Sentences (RS). The recognition of RS (referred to as "C" hereinafter) through the texts (A-B) is facilitated by the Corpus of Subject Rules. They are left intact or sent for gist translation (or more precise manual rendering, if needed). The remaining non-RS (texts of (A-B-C)) are entered for a third-step computation.

Exclusion of Special Constructions (SC). Most of the remaining non-RS texts generated by the above two steps of computation belong to the ESS defined in this study, namely, those sentences with discourse-based subject ellipsis. It is interestingly noticeable, as far as Chinese is concerned, that a small proportion of these near-target PSS "sentences" might be special constructions (SC, referred to as "C" hereinafter), including but not limited to nominal phrases of time, place and direction, prepositional phrases, adverbial phrases, idioms and slang, exclamatory phrases and modal words, etc. They are scanned, spotted and excluded in the third round of computation.

Identification and Verification of ESS. So far, the texts surviving all the three rounds of exclusion are determined as ESS under discussion (referred to as "E" hereinafter, E=A-B-C-D), which are to be dealt with in Phase 2, subject restoration. For better understanding, the program flowchart for identification of ESS is furnished in Fig. 1.:

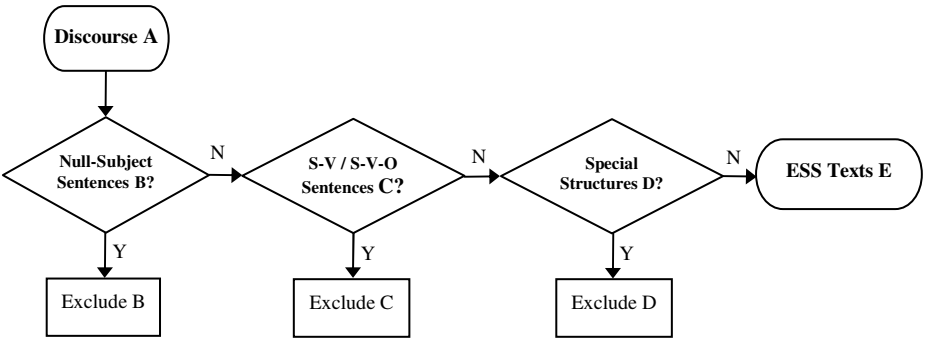


Fig. 1. Program flowchart for identification of Empty-Subject Sentences (ESS)

3.2 Restoration and Tagging of ESS Empty Subjects (ES)

As the syntactic subject, the empty subject of ESS sentences are hidden in the contexts of a discourse, it is, therefore, important to abstract the rules that govern the discourse-based covert ES. By working under the rules, the program will be capable of automatically scanning, identifying and restoring the real ES for subject tagging and/or MT processing.

Scanning within PCS. The empty subject (ES) of one or more than one ESS (also called a PSS as defined in the "Introduction" section, simplified as "S_e" hereinafter) could firstly and optimally be scanned within the PCS (defined in the "Introduction" section, simplified as "S" hereinafter) where the S_e stand(s):

- The first S_e within S will preferably take the forcibly overt subject preceding and neighboring S_e as its subject. If such overt subject exists, tag it as S_e subject;
- If such overt subject does not exist, for example, in the case that S_e has no neighboring PSS (S_e-1) preceding it (S_e actually initiates the PCS or directly follows a preceding PCS, sentence (S-1)), go to the next neighboring PSS, sentence (S_e+1) for the possible overt subject till the rear of S. If such overt subject exists, tag it as S_e subject.

Scanning Outside PCS. If the above two steps fail to restore the subject for S_e inside its residential S, then go to the preceding PCS, (S-1), for subject scanning.

- The most neighboring overt subject in (S-1) is preferably taken as S_e subject, with the most neighboring overt object less preferably if the subject is absent;
- If both overt subject and object fail to be identified, then go to the next neighboring PCS, (S+1), for subject scanning till the rear of (S+1). If such overt subject exists (most neighboring), tag it as S_e subject.

For better understanding, the program flowchart for empty subject restoration, as depicted in the above 4 steps, is furnished in Fig. 2.:

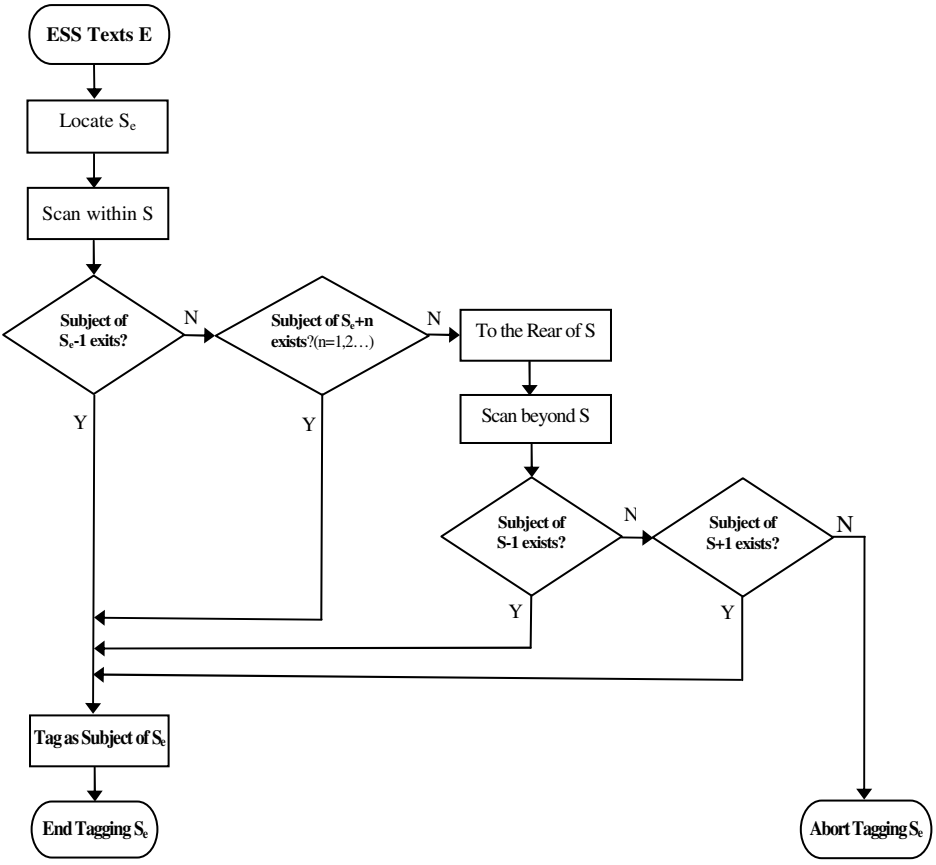


Fig. 2. Program flowchart for Empty-Subject Restoration

Manual Processing. If some step identifies any overt subject (or object), the program will end tagging S_e; if all these four steps fail to identify any subject (or preceding object) for S_e, the program will abort tagging S_e. The program will then loop computing for the next ESS, (S_e+1), until all ESS sentences are tried for possible subject restoration. If any ESS fails to have its "empty subject" spotted and restored, like in the latter case, the "ESS" here might be a fake. The fake ESS could be any of

irregular Construction B, C or D that has escaped programming exclusion (as detailed in 3.1). It is subject to manual identification or processing, if precise MT is required of the given discourse A. This step shall be otherwise ignored.

3.3 Case Study

Try the following discourse with the above two programs, the first to identify ESS sentences, and the second to restore their subjects:

(5) ‖(e_1)回家变卖典质(S_{e1})，父亲(t)还了亏空(C_1)；‖(e_2)又借钱办了丧事(S_{e2})。‖这些日子(D_1)，家中光景很是惨淡(C_2)，一半为了丧事(D_2)，一半为了父亲赋闲(D_3)。‖(朱自清《背影》)

‖ (e_1) *Huijia bianmai dianzhi* (S_{e1}), *Fuqin* (t) *huan le kuikong* (C_1);‖(e_2)*You jieqian ban le sangshi* (S_{e2}). ‖ *Zhexie rizi* (D_1), *jia zhong guangjing hen shi candan* (C_2), *yiban wei le sangshi* (D_2), *yiban wei le Fuqin fuxian* (D_3). ‖ (*Zhu Ziqing*《*Beiying*》)

‖ (e_1) Went home and pawned the valuables (S_{e1}), Father (t) liquidated (all our) debts (C_1);‖(e_2) And borrowed some money for the funeral (S_{e2}). ‖ These days (D_1), conditions at home were really gloomy (C_2), partially for the funeral (D_2), (and) partially for Father's unemployment (D_3). ‖ (Excerpted from Zhu's essay "The Sight of Father's Back")

Texts between two neighboring double-strokes (‖) constitute a Programmed Complex Sentence (PCS). Any construction of S_e, C, or D stands alone as a Programmed Simple Sentence (PSS). The two underlined PSS's are ESS sentences to be identified and restored. No Construction B (NSS) is spotted in this discourse.

The first program excludes all C's and D's, leaving only two S_e's (underlined texts) to be restored.

The second program attempts to restore the subject, with loop computation, for the two S_e's. Within the first PCS, the overt subject "Father" is spotted, and thus tagged as the subject for S_{e1}. Within the second PCS (also a PSS), the S_{e2} itself, no overt subject is identified; within the neighboring PCS preceding it, "Father" is spotted, and thus tagged as the subject for S_{e2}. The program then ends scanning.

With manual verification, both e_1 and e_2 can be restored with the NP-traced (t) "Father". Therefore, the computational results are identical to manual outcome.

4 Concluding Remarks

The factors that lead to the subject ellipsis are rather complicated, with habitual and stereotyped expression, the principle of economy and special traits of discourse contributing much. The recognition of ESS is based on the exclusion of constructions of, i.e., regular S-V and S-V-O, irregular NSS, and various types of phrases. The restoration of ESS subject is conducted in at most three successive CPS sentences.

For more precise subject tagging and MT results, the following resources are fundamentally necessary:

- NSS corpus and NSS rules corpus, with parallel translated texts cross-referenced. NSS are constant in construction and limited in number.
- Subject rules corpus for Chinese. Subjects in Chinese are rather different from those in other languages, but studies in this domain have reaped satisfactory results. Translated parallel subject corpus is advised to focus on the linguistic forms of subjects in Chinese, such as morpheme, person and number.
- Corpus of special constructions like idioms and phrases. They are various and numerous, but are relatively easy in the corpus building.
- Norms and standards of Chinese punctuation. Commonly used marks are fewer than 10 in number, which are vital to computer-based recognition of PSS and PCS, which in turn are responsible for the accuracy of subject restoration.
- Standardized use of Chinese punctuation marks, and accurate typing-in of Chinese texts.

Some of these domains are what this study is to further explore.

Acknowledgement. This work has been supported by the Research Project in Humanities and Social Sciences by the Ministry of Education (11YJC740154), the National Social Science Foundation Project (12CYY001), as well as the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUGW120227, CUGW120231, and CUGW100224). We also appreciate Prof. Lin He for her inspiration and encouragement in the academic writing. Two anonymous reviewers are acknowledged for their revision suggestions and comments.

References

1. Chomsky, N.: Lectures in Government and Binding. Foris, Dordrecht (1981)
2. Lin, X.G.: Lexical Semantics and Computational Linguistics. Language Publishing House, Beijing (1999)
3. Huang, Y.: On Empty Categories in Chinese. Zhongguo Yuwen 230, 383–394 (1992)
4. Wang, D.Z.: On the Classification and Referent of Chinese Null Subject. Journal of PLA Foreign Languages University 22, 29–32 (1999)
5. Zhu, L.H., He, Z.H.: On Null Subject in Chinese. Journal of Social Science of Hunan Normal University 39, 132–135 (2010)
6. Hua, H.Y.: The Ellipses of Subjects after Subjects. Journal of Yantai Normal University 18, 83–89 (2001)
7. Zhang, G.X.: On Implication. Zhongguo Yuwen 233, 126–133 (1993)
8. Xu, L.J.: Some Chinese Grammatical Facts Related to Empty Categories. Zhongguo Yuwen 242, 321–329 (1994)
9. Lv, S.X.: Issues on Chinese Grammatical Analyses. Commercial Press, Beijing (1979)
10. Zhu, Z.Q.: Classics of Zhu Ziqing's Essays. Beijing Publishing House, Beijing (2008)
11. Martinet, A.: A Functional View of Language. Clarendon Press, Oxford (1962)
12. Huang, N.S.: Ellipsis and Discourse. Linguistic Researches 62, 9–16 (1997)
13. Chen, W.Y.: Ellipsis and Economy. Journal of Zhejiang University (Humanities and Social Sciences) 35, 177–184 (2005)

Incorporating Lexical Semantic Similarity to Tree Kernel-Based Chinese Relation Extraction

Dandan Liu, Zhiwei Zhao, Yanan Hu, and Longhua Qian

Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu, 215006
School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, 215006
{20104227054,20104227044,20114227025,qianlonghua}@suda.edu.cn

Abstract. Lexical semantic information plays an important role in semantic relation extraction between named entities. This paper incorporates two kinds of lexical semantic similarity measures, thesaurus-based and corpus-based, into convolution tree kernels and systematically investigates their effects on Chinese relation extraction. The experiments on the ACE2005 Chinese corpus shows that the incorporation of lexical semantic similarity into tree kernel-based Chinese relation extraction can significantly improve the extraction performance when entity types are unknown, while in the case of known entity types, these lexical similarity measures also enhance the extraction performance for some person-related relationships. This demonstrates the usefulness of lexical semantic similarity in Chinese relation extraction.

Keywords: Lexical Semantic Similarity, Chinese Entity Relation Extraction, Convolution Tree Kernel, TongYiCi CiLin, HowNet.

1 Introduction

Relation extraction (RE) is an important information extraction task in natural language processing (NLP), with many practical applications, including learning by reading, automatic question answering, text summarization and so on. The goal of relation extraction is to detect and characterize semantic relationships between pairs of named entities in text. For example, a typical relation extraction system needs to extract a Person-Social relationship between the person entities "他" (person, PER) and "妻子" (person, PER) in the Chinese phrase "他 的 妻子" (ta de qi zi, his wife).

Generally, machine learning-based methods are adopted in relation extraction due to their high accuracy. In terms of the expression of learning examples (i.e., the relation instances) they can be divided into feature-based methods and kernel-based ones. The key issue of feature-based RE is how to extract various lexical, phrasal, syntactic, and semantic features[1-3], which are important for relation extraction, from the sentence involving two entities, while for kernel-based RE, the structured representation of relation instances, such as syntactic parse trees[4-9], becomes the central problem. In Chinese relation extraction, many studies focus on feature-based methods, such as

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 11–21, 2013.

[10-12] .while kernel-based methods, such as edit distance kernel[13], string kernel [14], convolution tree kernels over parse trees[15-17], have gained wide popularity.

It is widely held that lexical semantic information plays an important role in relation extraction between named entities, since two words, different in surface but similar in semantic, may represent the same relationship. For example, the two phrases "他 的 妻子" (ta de qi zi, his wife) and "她 的 丈夫" (ta de zhang fu, her husband) convey the same relationship PER-SOC.Family in the ACE terminology, though "他" (ta, he) and "她" (ta, she), "妻子" (qi zi, wife) and "丈夫" (zhang fu, husband) are two distinctive words. Therefore, different methods are proposed to exploit this lexical semantic information in Chinese relation extraction. [13]employ the Improved-Edit-Distance (IED) to calculate the similarity between two Chinese strings, and further considering lexical semantic similarity between words based on TongYiCi CiLin, their experiments show that the lexical semantic-embedded IED kernel method performs well for the person-affiliation relation extraction. [14] acquire lexical semantic similarity scores based on HowNet, and incorporate them into a string kernel, experiments in some ACE-defined minor relationships show promising results. However, the unresolved issues are that whether or not the widely adopted convolution tree kernel[18] can benefit from such lexical semantic information and which lexical semantic resources are more appropriate for relation extraction.

[19] propose a generalized framework for syntactic and semantic tree kernels which incorporate semantic information when computing structural similarity between two parse trees. They rely on the hierarchy of nouns in WordNet to calculate the term similarity and their methods can significantly improve the performance in Question Classification. Inspired by their work, we incorporate lexical semantic similarity measures based on two Chinese lexical resources, namely TongYiCi CiLin (abbreviated as CiLin, or more concisely as CL) and HowNet (abbreviated as HN), into convolution tree kernels for Chinese relation extraction, and further compare their effects on Chinese relation extraction. Additionally, we also calculate the lexical similarity based on corpus statistics and incorporate it into tree kernels to investigate its impact on Chinese relation extraction.

The rest of the paper is organized as follows. In Section 2, we elaborate the semantic convolution tree kernel over parse trees. Section 3 reports experimental results and analysis. Finally, Section 4 concludes the paper and points out the future directions.

2 Semantic Convolution Tree Kernel for RE

The key issues of tree kernel-based relation extraction are the representation of tree structure and the similarity calculation between trees. This section deals with the latter, i.e. the semantic convolution tree kernel, while for the tree structure in RE, we adopt the Unified Parse and Entity Semantic Tree (UPEST) which incorporates entity-related semantic information, such as entity types and subtypes, into Shortest Path-enclosed Tree (SPT). For details, please refer to [16].

2.1 Convolution Tree Kernel

The convolution tree kernel[18] counts the number of common sub-trees between two parse trees T_1 and T_2 as their similarity measure without explicitly considering the whole tree space. It can be computed as follows:

$$K_{CTK}(T_1,T_2)=\sum_{n_1\in N_1,n_2\in N_2}\Delta(n_1,n_2) \tag{1}$$

where N_1 and N_2 are the sets of nodes for T_1 and T_2 respectively, and $\triangle(n_1, n_2)$ evaluates the number of two common sub-trees rooted at n_1 and n_2. It can be computed recursively as follows:

1. If the productions at n_1 and n_2 are different then $\triangle(n_1,n_2)=0$; otherwise go to Step 2;
2. If both n_1 and n_2 are part of speech (POS) tags, then $\triangle(n_1,n_2)=\lambda$; otherwise go to Step 3;
3. Calculate recursively as follows:

$$\Delta(n_1,n_2)=\lambda\prod_{k=1}^{\#ch(n_1)}(1+\Delta(ch(n_1,k),ch(n_2,k))) \tag{2}$$

where *#ch(n)* is the number of children of the node n, *ch(n,k)* is the *k*-th child of the node *n,* and λ ($0<\lambda<1$) is a decay factor, which is used for preventing the similarity of sub-trees exceedingly depending on the size of sub-trees.

2.2 Semantic Convolution Tree Kernel

While convolution tree kernels exhibit promising results in the task of relation extraction[4,6,8,17], they disregard lexical semantic similarity between words in parse tree, which is critical for relation extraction in some scenarios. Motivated by the successful application of the syntactic and semantic convolution tree kernel[19] to the task of Question Classification (QC), we adopt a similar Semantic Convolution Tree Kernel (SCTK) to Chinese relation extraction with the lexical semantic similarity being calculated using Chinese lexical semantic resources.

The computation process for the SCTK is largely the same as that of the standard CTK except that in Step 1, one additional case should be considered as follows:

1. If the productions at n_1 and n_2 are the same, then go to Step 2; otherwise, if both n1 and n2 are the parents of entity headword nodes, then $\triangle(n_1,n_2)=\lambda * LexSim(hw_1, hw_2)$; otherwise $\triangle(n_1,n_2)=0$;

where hw_1 and hw_2 denote the headwords corresponding to two entities immediately under n_1 and n_2 respectively and $LexSim(hw_1, hw_2)$ denotes the lexical semantic similarity between these two headwords which can be calculated using lexical resources

such as *CiLin* or *HowNet*, or using corpus statistics. The details concerning the lexical similarity calculation will be introduced in Section 2.3.

In most cases, the entity headwords can be used directly to calculate their lexical similarity scores, e.g., in the relation instance "他 的 妻子" (ta de qi zi, his wife), both "他" (ta, he) and "妻子" (qi zi, wife) could be passed to the similarity calculation module since as common names they can be found in lexical resources. However, take the entity mention "大安森林公园" (da an sen lin gong yuan, DaAn Forest Park) as an example, since this headword is not a well-known proper noun and can not be found in CiLin or HowNet. Our solution to this problem is to segment the entity headword and then to take the rightmost word as the new headword. For example, the entity mention "大安森林公园" is segmented into "大安 森林 公园" and then the word "公园" is passed to the lexical similarity calculation module.

2.3 Lexical Semantic Similarity Calculation

The core element of SCTK is the lexical semantic similarity calculation between two entities, or more specifically speaking, between two words corresponding to the entity mentions. Here, we employ two kinds of lexical similarity measures, thesaurus-based one, which takes advantage of two commonly used Chinese lexical resources, namely, CiLin and HowNet, and corpus-based one, which derives distributional similarity between lexical items from corpus statistics.

2.3.1 Thesaurus–Based Similarity

TongYiCi CiLin[20] is a Chinese thesaurus where 77,492 words are organized as a tree structure according to their semantic classes. There are five layers in the hierarchy, including 12 major classes, 94 intermediate classes and 1428 minor classes, groups, and atomic groups.

HowNet[1] is a lexical knowledge base with rich semantic information, where a word is described as a group of sememes in a complicated multi-dimensional knowledge description language. Due to its richness in lexical semantics, it has been widely exploited in various NLP researches.

We adopt the software package by Liu and Li[21] [2] to calculate lexical semantic similarity scores based on both CiLin and HowNet. For *CL* they compute the lexical similarity score based on the distance between words in a semantic taxonomy tree. For *HN* the similarity score between content words is a linear interpolation of four different similarity scores, i.e. similarity between primary sememes, that between other sememes, that between sets, and that between feature structures.

2.3.2 Corpus-Based Similarity

We use the parallel Chinese-English Foreign Broadcast Information Service (FBIS) corpus to calculate lexical similarity, which gathered from the news domain. This

[1] http://www.keenage.com/html/c_index.html

[2] http://code.google.com/p/xsimilarity/downloads/list

bilingual corpus contains about 240k sentences, 6.9 million words in Chinese and 8.9 million words in English. However, here we only use the Chinese side of this corpus.

Each entity headword is represented by a feature vector, where each feature corresponds to a context word with which the headword co-occurs. Each feature in the vector is weighted by its point-wise mutual information (PMI) with the headword *hw*, defined as:

$$PMI(hw,c) = \log_2 \frac{N(hw,c) * N}{N(hw) * N(c)} \tag{3}$$

where $N(hw, c)$ is the co-occurrence frequency of the headword *hw* and its context word *c*, $N(hw)$ and $N(c)$ are the occurrence frequencies of the words *hw* and *c* respectively, *N* is the number of all word occurrences in the corpus. Since PMI is usually biased towards infrequent words, we multiplied it with a discounting factor as described in[22]:

$$\frac{N(hw,c)}{N(hw,c)+1} \times \frac{\min(N(hw),N(c))}{\min(N(hw),N(c))+1} \tag{4}$$

Then, the pair-wise similarity scores are computed between one entity headword hw_1 and another entity headword hw_2 using the cosine similarity measure as follows:

$$LexSim(hw_1,hw_2) = COS(HW_1,HW_2) = \frac{\sum_i HW_1^i \times HW_2^i}{\sqrt{\sum_i HW_1^i} \times \sqrt{\sum_i HW_2^i}} \tag{5}$$

where HW_1 and HW_2 are the feature vectors for headword hw_1 and headword hw_2 respectively, HW_1^i and HW_2^i are the discounted PMI values of the *i*th features f_1^i and f_2^i, s.t. $f_1^i = f_2^i$. This measure is usually called distributional similarity since it depends on the similarity of context distribution.

3 Experimentation

This section experimentally investigates the effects of lexical semantic resources on Chinese relation extraction.

3.1 Experimental Setting

The ACE RDC 2005 Chinese corpus is used as the experimental corpus for Chinese semantic relation extraction. The corpus defines 6 major relationships and 18 minor relationships. It contains 633 files, including 298 broadcast news, 38 newswires and 97 micro-blogs and others.

The corpus is first word-segmented using the ICTCLAS package3 based on the multi-layer HMM model, and then parsed using the state-of-the-art Charniak's parser [23] with the boundaries of all the entity mentions kept. Finally, relation instances are generated by iterating over all pairs of entity mentions occurring in the same sentence, extracting corresponding tree structures and incorporating optional entity-related information (e.g., entity types or subtypes). In total, we obtain 9,147 positive relation instances and 97,540 negative relation instances.

In our experimentations, SVMLight-TK[4] toolkit is adopted as our classifier. The package is modified to incorporate the lexical similarity calculation module. We apply the one vs. others strategy. Particularly, the SubSet Tree (SST) kernel is used since it yields the best performance, while the decay factor λ (tree kernel) is set to the default value (0.4).

In order to make full use of the corpus resources, and reduce the influence of experimental results caused by the corpus variance, the five-fold cross validation strategy is adopted for training and testing, and the averages of 5 runs are taken as the final performance scores. Evaluation metrics are commonly used Precise, Recall, F harmonic measure, which can be abbreviated as P/R/F1.

3.2 Experimental Results and Analysis

Since entity-related semantic information, i.e., entity types and subtypes, is also important for relation extraction[16] and may correlate with entity lexical semantic information, we conducted two series of experiments, with or without entity types respectively, in order to better investigate the effect of lexical semantic information and its correlation with entity types for Chinese relation extraction.

3.2.1 The Overall Effect of Lexical Similarity on Chinese Relation Extraction

We compare in Table 1 the P/R/F1 performance scores of major type and subtype relation extraction on the ACE 2005 Chinese corpus respectively. Among different systems, the baseline (BL) system only uses the basic SPT structure, considering neither lexical semantic information nor entity types, and "ET" denotes the UPEST structure augmented with entity type information, while "+CL" 、"+HN" and "+FBIS" denote that additional lexical semantic similarity measures corresponding to *CiLin*、*HowNet* and *FBIS* are incorporated into the tree kernel computation.

We can see from Table 1 that

- All three lexical semantic similarity measures significantly improve the performance scores over the baseline system without entity types. For example, the overall scores of ΔF reach 3.6, 3.8 and 3.3 units on major type extraction when considering lexical similarity measures based on *CiLin*, *HowNet* and *FBIS* respectively. This illustrates that the lexical semantic similarity is very helpful for Chinese relation extraction. Moreover, the improvements of F1 come from both precision and recall

[3] http://ictclas.org/ictclas_download.aspx
[4] http://ai-nlp.info.uniroma2.it/moschitti/

increases in similar degrees. This suggests that the incorporation of lexical semantic similarity makes the convolution tree kernel more precise and concise in capturing the essence of relationships.

- However, when the entity types are known, the performance increases brought about by lexical similarity are greatly diminished. For example, the ΔF scores decrease to 0.6, 0.6 and 0.1 units on major type extraction corresponding to *CiLin*, *HowNet* and *FBIS* respectively. This shows that in general the effect of lexical semantic information is severely dented when entity types are incorporated. The main reason is that entity type information, which is specifically designed for relation extraction, may have already contained most of semantic information embedded in three lexical semantic similarity measures.

Table 1. Overall performance comparison of incorporating various lexical similarity for Chinese relation extraction

Systems	Major type extraction			Subtype extraction		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Baseline(BL)	72.2	45.8	56.0	69.1	42.7	52.7
BL+CL	**75.9**	49.1	59.6	**74.4**	**47.5**	**58.0**
BL+HN	75.8	**49.4**	**59.8**	73.6	46.9	57.3
BL+FBIS	75.8	48.8	59.3	73.7	46.2	56.8
ET	80.4	56.5	66.4	77.1	54.3	63.7
ET+CL	**82.**0	56.6	67.0	**79.6**	54.5	**64.7**
ET+HN	81.9	**56.8**	**67.0**	79.2	**54.6**	64.6
ET+FBIS	81.8	56.0	66.5	79.2	54.1	64.3

- Generally, two thesaurus–based similarity measures are better than the corpus-based one, no matter whether the entity type information is provided or not, though the advantage is not significant, and there is almost no difference between the CiLin-based measure and the HowNet-based one. This shows that in assisting Chinese relation extraction this way, CiLin and HowNet perform comparably, while the corpus-based method underperforms them probably due to the noise and incompleteness of lexical features when calculating the lexical similarity from corpus statistics.

3.2.2 The Effect of Entity Types on Major Relation Types

Table 2 compares the performance scores of BL and ET (c.f. Table 1) on major relation types, where # is the number of instances for major types.

From Table 2 we can see that, as verified in previous studies on relation extraction [16], entity type information is very critical for relation extraction, as their incorporation drastically improves the performance, with the average increases of P/R/F1 reaching 8.2/10.7/10.4 units respectively.

Table 2. Performance comparison on baseline system and entity types system for Chinese relation extraction

Major Types	#	BL			ET		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
PHYS	1552	64.2	9.3	16.2	71.4	19.1	30.2
PART-WHOLE	2249	69.5	60.8	64.9	80.5	**72.3**	76.2
PER-SOC	652	74.0	32.4	44.9	79.4	35.4	48.9
ORG-AFF	2166	**79.8**	**61.3**	**69.3**	**87.0**	69.1	**77.0**
ART	623	62.3	14.4	23.4	73.8	32.6	45.1
GEN-AFF	1905	75.8	55.3	63.9	82.8	68.9	75.2
Average	9147	72.2	45.8	56.0	80.4	56.5	66.4

3.2.3 The Effect of Lexical Semantic Information on Major Relation Types with or without Entity Types

In Figure 1, we compare the F1 improvements on major relation types of different systems over the baseline one, without augmented entity types, where "(BL+CL)-BL", "(BL+HN)-BL", and "(BL+FBIS)-BL" denotes the F1 increases corresponding to three lexical similarity based on *CiLin*, *HowNet*, and *FBIS*.

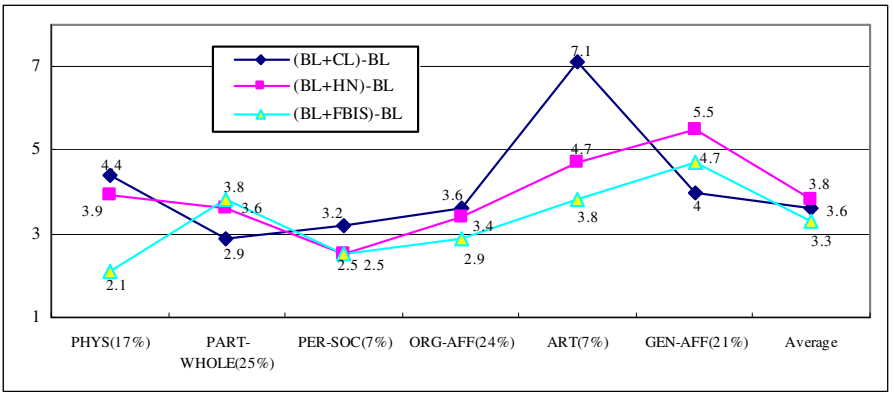


Fig. 1. The comparison of △F on major type extraction over the baseline system by incorporating lexical similarity

From Figure 1, we can see that the F1 score on every major relation types can be significantly improved by all three lexical similarity measures, though in different degrees. An interesting phenomenon is that, HowNet outperforms FBIS consistently in somewhat similar degree on most relationships except PART-WHOLE for which FBIS slightly outperforms HowNet. It seems that the HowNet-based similarity is

consistent with the corpus-based similarity from the viewpoint of relation extraction. However, HowNet and CiLin perform divergently on most relation types. This probably reflects the fact that HowNet embodies complicated semantic relationships between words than CiLin, which organizes lexical items in a clearly hierarchical way.

In Figure 2, we compare the F1 improvements on major relation types over the ET system with the entity type information augmented, where "(ET+CL)-ET", "(ET+HN)-ET" and "(ET+FBIS)-ET" denotes the increasements of F1 when incorporating lexical similarity measures based on CiLin, HowNet, and FBIS respectively.

The figure clearly shows that when entity types are known, additional consideration of lexical semantic similarity measures has a similar trend among themselves but a divergent effect on various relation types. That is, to these three measures, for three relation types (PER-SOC, ORG-AFF, and GEN-AFF) the performance is improved in varied degree, but for other three types (PHYS, PART-WHOLE and ART), the performance improvements are minor or even negative. This suggests that with the entity type information provided, lexical similarity is detrimental for some relationships, yet helpful for other relationships. Particularly interestingly, compared with Figure 1, we find that there are two kinds of relationships, namely, PER-SOC and ART, on the former lexical similarity improves the performance better when entity types are known than when they are unknown, while on the latter lexical similarity decreases the performance when entity types are known other than improves when entity types are unknown. This means that lexical similarity complements with entity types on extracting some specific relation types.

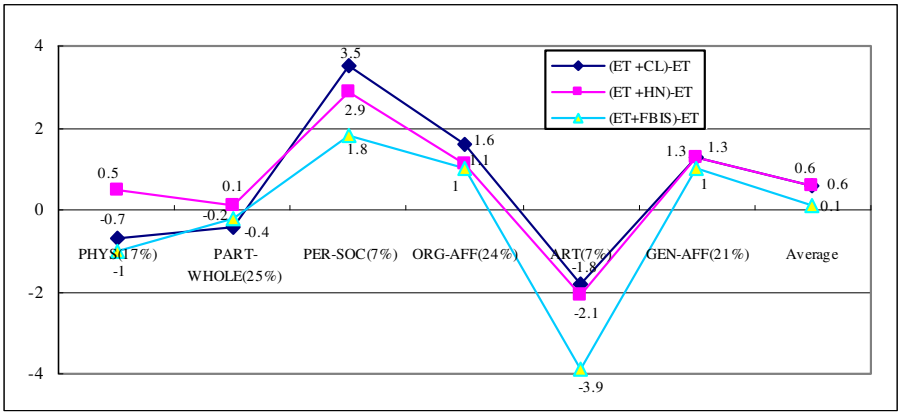


Fig. 2. The comparison of △F on major type extraction over the ET system by incorporating lexical similarity

Further experiments show that a few of minor relationships, largely person-related, such as PER-SOC.Business, PER-SOC.Family etc., significantly improve their performance when entity types are provided. This eventually reaches our conclusion that lexical semantic similarity are useful to Chinese relation extraction no matter whether entity types are provided or not, the distinctions are the degree to which and the kinds of relationships for which the performance scores of relation extraction are improved.

4 Conclusion and Future Work

In this paper, we first examine two methods of lexical semantic similarity measures, namely, the thesaurus-based measure and the corpus-based one, and then empirically demonstrate the impact of lexical semantic similarity on Chinese relation extraction. Specifically we incorporate these lexical semantic similarity measures into tree kernel-based Chinese relation extraction. A series of experiments on the ACE 2005 benchmark corpus indicate that these lexical similarity measures perform comparably and can improve the performance of Chinese relation extraction with or without entity type information, the only distinctions are that the degree of improvement is weakened and the kinds of relationships being improved are shrunk when entity type information are provided. Nevertheless, some person-related relationships always benefit from lexical semantic similarity measures.

For future work, we will extend our work to relation extraction from English texts using lexical resources such as WordNet and large corpus like Gigaword. As the research focus in relation extraction moves from close-domain and supervised learning towards open-domain and self-supervised learning, we will explore the role of lexical semantics in this new direction.

Acknowledgement. This work is funded by China Jiangsu NSF Grants BK2010219 and 11KJA520003.

References

1. Zhou, G.D., Su, J., Zhang, J., Zhang, M.: Exploring Various Knowledge in Relation Extraction. In: ACL 2005, pp. 427–434 (2005)
2. Zhou, G.D., Su, J., et al.: Modeling Commonality among Related Classes in Relation Extraction. In: COLING-ACL 2006, pp. 121–128 (2006)
3. Zhou, G.D., Zhang, M.: Extracting Relation Information from Text Documents by Exploring Various Types of Knowledge. Information Processing and Management 43, 969–982 (2007)
4. Zhang, M., Zhang, J., Su, J., Zhou, G.D.: A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. In: COLING-ACL 2006, pp. 825–832 (2006)
5. Zhang, M., Zhou, G.D.: Exploring Syntactic Structured Features over Parse Trees for Relation Extraction Using Kernel Methods. Information Processing and Management 44, 687–701 (2008)
6. Zhou, G.D., Zhang, M., Ji, D.H., Zhu, Q.M.: Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information. In: EMNLP/CoNLL 2007, pp. 728–736 (2007)
7. Zhou, G.D., Qian, L.H., Fan, J.X.: Tree kernel-based Semantic Relation Extraction with Rich Syntactic and Semantic Information. Information Sciences 18(8), 1313–1325 (2010)
8. Qian, L.H., Zhou, G.D., Kong, F., Zhu, Q.M., Qian, P.D.: Exploiting Constituent Dependen-cies for Tree Kernel-based Semantic Relation Extraction. In: COLING 2008, Manchester, pp. 697–704 (2008)

9. Zhuang, C.L., Qian, L.H., Zhou, G.D.: Research on Tree Kernel-Based Entity Semantic Relation Extraction. Journal of Chinese Information 23(1), 3–9 (2009) (in Chinese)
10. Che, W.X., Liu, T., Li, S.: Automatic Entity Relation Extraction 19(2), 1–6 (2005)
11. Dong, J., Sun, L., Feng, Y.Y., Huang, R.H.: Chinese Automatic Entity Relation Extraction. Journal of Chinese Information 21(4), 80–85, 91 (2007) (in Chinese)
12. Li, W.J., Zhang, P., Wei, F.R., Hou, Y.X., Lu, Q.: A Novel Feature-based Approach to Chinese Entity Relation Extraction. In: ACL 2008, pp. 89–92 (2008)
13. Che, W.X., Jiang, J., Su, Z., Pan, Y., Liu, T.: Improved-Edit-Distance Kernel for Chinese Relation Extraction. In: IJCNLP 2005, pp. 132–137 (2005)
14. Liu, K.B., Li, F., Liu, L., Han, Y.: Implementation of a Kernel-Based Chinese Relation Extraction System. Computer Research and Development 44(8), 1406–1411 (2007) (in Chinese)
15. Huang, R.H., Sun, L., Feng, Y.Y., Huang, Y.P.: A Study on Kernel-based Chinese Relation Extraction. Journal of Chinese Information 22(5), 102–108 (2008) (in Chinese)
16. Yu, H.H., Qian, L.H., Zhou, G.D., Zhu, Q.M.: Chinese Semantic Relation Extraction Based on Unified Syntactic and Entity Semantic Tree. Journal of Chinese Information 24(5), 17–23 (2010) (in Chinese)
17. Qian, L.H., Zhou, G.D., Zhu, Q.M.: Employing Constituent Dependency Information for Tree Kernel-based Semantic Relation Extraction between Named Entities. ACM Transaction on Asian Language Information Processing 10(3), Article 15, 24 (2011)
18. Collins, M., Duffy, N.: Covolution Tree Kernels for Natural Language. In: NIPS 2001, pp. 625–632 (2001)
19. Bloehdorn, S., Moschitti, A.: Exploiting Structure and Semantics for Expressive Text Kernels. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, Lisbon, Portugal (2007)
20. Mei, J.J., Zhu, Y.M., Gao, Y.Q., Yin, H.X.: TongYiCi CiLin, 2nd edn. Shanghai Lexicographic Publishing House, Shanghai (1996) (in Chinese)
21. Liu, Q., Li, S.J.: Word Similarity Computing Based on How-net. Computational Linguistics. Chinese Information Processing, 59–76 (2002)
22. Lin, D.K., Pantel, P.: Concept Discovery from Text. In: COLING 2002, pp. 42–48 (2002)
23. Charniak, E.: Immediate-head Parsing for Language Models. In: ACL 2001 (2001)

Research on Chinese Sentence Compression for the Title Generation

Yonglei Zhang, Cheng Peng, and Hongling Wang

Natural Language Processing Lab, Soochow University
1 Shizi Street, Suzhou, China 215006
{20104227009,20104227048,hlwang}@suda.edu.cn

Abstract. Automatic Title Generation means generating a title which can show the central information of the original text via natural language processing technologies. One method is by extracting a sentence which represents the original text's central information and then compressing it to a short sentence as the title, in which the core technology is the sentence compression. But the research of Chinese sentence compression has not carried out, it is mainly facing the following difficulties: lacking of the corpus, suffering from the poor performance of Chinese word segmentation and parsing, and having no unified automatic evaluation metric. This paper realizes a Chinese sentence compression method through simply shorting a sentence by deleting words or constituents which is main practice is by learning a subtree from the source parsing tree of a sentence, and then uses the manual and automatic evaluations to evaluate the sentence compression performance. The experimental results show that the method and evaluation metrics used in this paper are valid and effective.

Keywords: Title Generation, Chinese Sentence Compression, Syntactic Tree, Automatic Evaluation.

1 Introduction

With the development of the network technology, collection of information has become a topic which has got more and more attention. In real life, people select articles through the titles, the abstracts and the keywords of the article. However, due to the complexity of the Chinese, vocabulary with a literary flavor is usually consciously added to the title. In many cases, the title of the article can not be intuitive demonstration of central idea of the article. In addition, a large number of documents and document fragments without a title. Therefore, generating a title which represents central information for document or document fragments is a meaningful issue.

Automatic title generation is a task to generate a title which represents original central information by using natural language processing techniques. One of the methods for the automatic title generation is to extract a sentence on behalf of the central information of the article, and compress it to a short sentence or phrase as the title. The work can be divided into the following two parts: the center sentence extraction and the sentence compression. The authors [1] were the first using the

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 22–31, 2013.

sentence compression into the Automatic Title Generation. They generate a title for a dialogue by removing the redundant and non-important phrase in the sentence and retaining the main topic of argument. In this paper, our work is focus on the sentence compression part. The sentence compression for the Automatic Title Generation main difference with the usual sentence compression is the high compression ratio requirements.

We define the sentence compression task as follows: given an input sentence, to produce a sentence which is shorter and retains the important information from the original, and also it is grammatical. In previous work, the researchers mainly use additional operations such as deletion, substitution, reordering, and insertion in the sentence compression task. Most prior work has focused on a specific instantiation of sentence compression, namely word deletion, such as [2-4]. In our paper, we also consider using the word deletion approach. Here, sentence compression aims to shorten a sentence $x=l_1,l_2,\ldots\ldots,l_n$ into a substring $y^*=c_1,c_2,\ldots\ldots c_m$, where $c_i \in \{ l_1,l_2,\ldots\ldots,l_n \}$. We define the function $F(c_i) \in \{1, \ldots, n\}$ that maps word c_i in the compression to the index of the word in the original sentence. Then, we include the constraint $F(c_i)< F(c_{i+1})$, which forces each word in x to occur at most once in the compression y^*, so in the compression process we don't change the word's order. This paper implements a Chinese sentence compression system by learning a subtree from the source parsing tree of a sentence (Figure 1).

Example(Figure 1):

Original Sentence:	据 法新社 报道 ， 有 目击者 称 ， 以军 23日 空袭 加沙 地带 中部 ， 目前 尚 无 伤亡 报告 。
Pinyin:	ju faxinshe baodao , you mujizhe cheng , yijun 23ri kongxi jiasha didai zhongbu , muqian shang wu shangwang baogao .
Translation:	According to the report of the Agence France-Presse, witnesses said the Israel army on the 23rd air strikes the central Gaza Strip, there were no casualties were reported.
Target Sentence:	目击者 称 以军 空袭 加沙 地带 中部
Pinyin:	mujizhe cheng yijun kongxi jiasha didai zhongbu
Translation:	Witnesses said the Israel army air strikes the central Gaza Strip

The remainder of this paper is structured as follows. Section 2 provides an overview of related work of the sentence compression and we analysis previous models benefits and defects. And then we conclude the challenges of our work for the Chinese sentence compression. In Section 3 we present the way of the corpus we constructed, we also show the features of our corpus. Section 4 describes the framework of our Chinese sentence compression system, and we describe the feature set, the decoding method, the loss function and the evaluation methods. In Section 5 we present the experimental evaluation of our system under different max length of the target sentence and the final results of the system. At the last section of this paper discusses future work.

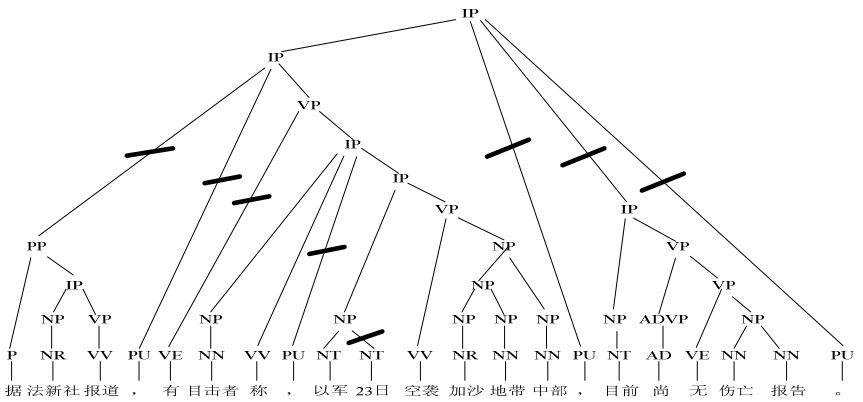


Fig. 1. Compression example, the coarse slashes means in the compressed target sentence the edges are deleted

2 Previous Work

The authors [2] firstly introduce a generative noisy-channel model and a discriminative tree-to-tree decision tree model for the task of sentence compression. And since then, their evaluation indexes are widely used. They mainly used the following evaluation indexes: Importance (How much of the important information is retained from the original), Grammaticality (Grammatical degree), Compression Ratio (How much compression took place).

Though the performance of the noisy-channel model is quite well, they do have some shortcomings, such as the source model which represent the probability of compressed sentences, but it is trained on uncompressed sentences, and the channel model requires aligned parse trees for both compressed and uncompressed sentences in the training set in order to calculate probability estimates. These parse trees with many mistakes for both the original and compressed versions will make alignment difficult and the channel probability estimates unreliable as a result.

There is another model included in [2], which they call the decision tree model. The decision tree method learns a decision tree to incrementally convert between original parse trees and compressed parse trees. This model avoid the unreliable of the tree alignment, but their model features encode properties related to including or dropping constituents from the tree with no encoding of bigram or trigram surface features to promote grammaticality. As a result, the model will generate some short and ungrammatical targets.

The author [4] used max-margin leaning algorithm to study the feature weight, then rank the subtrees, and finally select the tree with the highest score as the optimal target sentence. McDonald's work had achieved a well performance by using the manual evaluations. But manual evaluations have some disadvantages, such as heavy workload, strength subjectivity, and so on.

The research for Chinese sentence compression just started, and the research facing many difficulties such as the lack of corpus, Chinese word segmentation and syntactic analysis's performance aren't well. In this paper, we extract corpus form the WangYi news, and then we use the manual and automatic evaluations to evaluate the sentence compression performance. The manual evaluations come from [2]. The automatic evaluations include compression ratio and Bleu score which used in our English sentence compression evaluation system [5].

3 Corpora

In our experiments, we use the WangYi news as our corpora. After some processing (Figure 2), we extract 1308 pair sentences. We use ICTCLAS5[1] kit to do Chinese word segmentation. In the figure 2, the process of manual selection and modifying mainly include the following operations: (1) checking whether the sentence group expressing the same information; (2) modifying the sentence group which does not match the words to meet the methodological framework based on simple word deletion.

Here, it should be noted that the compression ratio is very high since the title is matched with the article in the corpus, which meets the requirement of the automatic title generation task. The target sentences come from the titles which have no punctuation, and we also do not add punctuation in the experiments, In addition, by the performance limitation of Chinese word segmentation, there are many segmentation errors after word segmentation. We ignore these errors in our experiments.

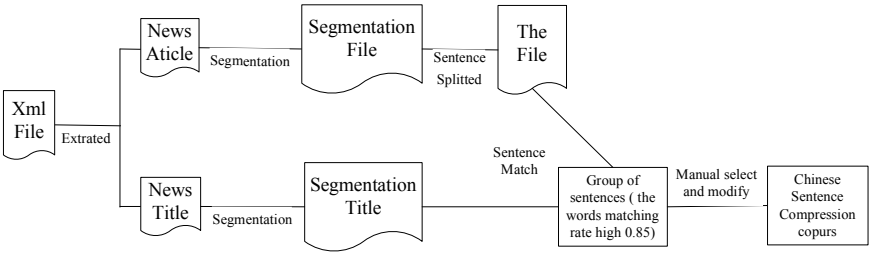


Fig. 2. The Flowchart of the Construction Corpora

4 Our Work

The framework of the Chinese sentence compression method is shown in Figure 3. The tool of words alignment used in the experiment is developed by our own. Although other open-source kits, such as Giza++, Berkeley Aligner etc, can also be used, we don't use them for their poor performance in the same language. The parser

[1] http://ictclas.org/ictclas_download.aspx

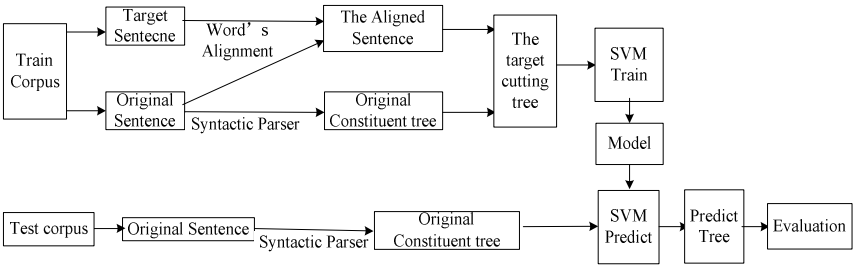


Fig. 3. The Framework of Chinese Sentence Compression

we used is the Stanford University's open-source tool Stanford Parser[2]. And we use Structured SVM [6] which supports structured outputs to solve structured learning problem. The tool we used is an open-source tool SVMstruct[3].

After preprocessing for the corpus, the system extracts the features which generated during the source syntax tree transformed into the target tree, and then uses the SVM to train feature weights, finally selects the highest score as the best target tree.

4.1 Features

1. Word/POS Features

Due to the small size of our corpora, which will lead to sparse and over-fitting, we mainly extract the feature of the word's POS and rarely include the word itself. In this paper, following features are used: the remaining word's POS, the remaining word's context POS (PosBigram (目击者 称) = NN&VV[4], PosTrigram (目击者 称 以军) = NN&VV&NT), whether each dropped word is a stop word (IsStop (据) = 1), whether each dropped word is the headword of the source sentence, the number of remaining words. In our experiments, we find that if those features are not extracted, the results would be shorter and ungrammatical. So the features of word/POS are necessary.

2. Syntax Features

Although the features of word/POS are very important, they can't reflect the status of the word in the sentence. The syntax features can reflect the status of the word in the sentence, so we include the following syntax features: the parent-children relationship of the cutting edge (del-Edge (PP) = IP-PP), the framework of the remaining edge (Parent-and-Child (IP) = IP (IP PU-D IP-D PU-D)), and the number of the cutting edge etc. In our early experiment, we find the target sentence remained the most importance information, but the convergence between words is inconsistent, so we add some

[2] http://nlp.stanford.edu/software/lex-parser.shtml
[3] http://download.joachims.org/svm_struct/current/svm_struct.tar.gz
[4] In Figure 1, for example (the same below).

dependency features in our system. The dependency features include the following: the dropped word's POS with its dependence word's POS (dep_link (，) = PU-VMOD-VV), whether the dependence tree's root is deleted (del_ROOT (无) = 1), and whether each dropped word is a leaf of the dependence tree (del_Leaf (法新社) = 1).

4.2 Decoding

In this paper, the system is based on simply shorting a sentence by deleting words or constituents. Assuming the source sentence x has n words, and then the target set has 2^n elements. With the increasing number of the word the source sentence has, decoding set exponential growth. Finding the best target sentence in such a large decoding space, time complexity is very large. So we use McDonald's simplify method to decode.

Supposing that $f(i , l)$ means the optimal target sentence which length is l and end at the i-th word. And we defined a two-dimensional array $A[n][n]$.The element $A[0][0]$ save the score of the optimal sentence which length is 0 and it is 0. $A[i][l]$ save the score of the optimal target sentence which length is l and end at the i-th word, then $A[i][l]=max\ Score_{0<l-1<j<i}(f (j , l-1)+l_i)$.Then, the time complexity of finding the best target sentence is $O(n^3)$. And we also through set the upper length of the target sentence to decrease the decoding space. Finally, the complexity of our decoding algorithm is $O(n^2*l)$, and l is the upper length we set.

Decoding Algorithm

```
1:  Input: x = l_1, l_2 ....., l_n ;
2:  t[n][0].string= "";  t[0][0].score=<w，f("",x)>;
    maxid=0 ;  maxlen=0 ;  maxscore=t[0][0]
3:  for  i=1,....,n do        /*  t[i][0] hasn't been used */
4:      t[i][1].string= l_i ;  t[i][1].score=<w，f( l_i ,x)>
5:      if  maxscore<t[i][1].score
6:        maxid=i;   maxlen=1
7:      for len=2,....,i    do
8:          for  j=len-1,....i-1  do
9:              if  <w，f(t[j][len-1].string+ l_i ,x)>  >  t[i][len].score
10:                 t[i][len].string=t[j][len-1].strig+ l_i ;
                    t[i][len].score=<w,f(x,t[j][len-1]+ l_i )>
11:                 if       maxscore < t[i][len].score
12                        maxid= i ;   maxlen=len;  maxscore=t[i][len].score
13:                 end if
14:             end if
15:         end for
16:     end for
17: end for
18: y=t[maxid][maxlen].string
```

4.3 Loss Function

In our work, loss function means the difference between the predict sentence and the golden target sentence. In our experiments, we use bigram loss ratio (Formula (1)) as the loss function. And we also find that its performance is better than using word loss ratio which is used in McDonald's work.

$$l(y, y^*) = \frac{|B(y)| - |B(y) \cap B(y^*)| + Max\{(|B(y^*)| - |B(y)|), 0\}}{|B(y^*)|} \quad (1)$$

4.4 Evaluation

In previous work, Importance, Grammaticality, and Compression ratio are used as the main evaluations. In our work, we also use these evaluations to evaluate our system. In addition, we also use sentence similarity as evaluation metric. Here we introduce BLEU score as similarity evaluation to compare the n-gram difference between the predict sentence and the golden target sentence.

1. Manual Evaluation Metrics

The metric of Importance means how much center information of the original sentence the target sentence contains, and Grammaticality means the grammatical degree of the target sentence. We evaluate each compression on a scale of 1-5 for these two metrics.

2. Automatic Evaluation Metrics

The automatic evaluation metrics we used are compression ratio and Bleu. The compression ratio means the percentage of words in the original sentence that are left in the compressed sentence. This metric has been used in all of the previous compression tasks. In our experiments, we find that 30% words of the original sentence contain the most important information. Bleu which is often used in Machine Translations means the similarity of the predict sentence with the golden target sentence. We find when two system give a similar compression rate, which's bleu is higher than another also contains more information than another. So, we can use Bleu to evaluate the information the predict sentence contains. In our experiments, we calculate to 4-gram.

5 Experiments

In our experiments, we use 1200 pair sentences as training set, and use 100 pair sentences as test set. Due to the heavy workload of manual evaluation, we use fewer test set to reduce the workload. In the training process, we add the target sentence and itself as a pair into the training set, so the training set has 2400 pair sentence. In the following experiments, we find this method dramatically improved the performance of our system.

5.1 Experimental Set-up

In the experiments, we can reduce the decoding space by setting the upper length l of the predict sentence. Figure 4 shows the results of different upper length on system. In the figure 4, the horizontal axis represents the different values of the upper length *l*, the vertical axis represents the value of the compression rate and the similarity for the corresponding upper length *l*. We can find that Bleu score is best when the upper length is 13. So in the following experiments the upper length is set to 13 in this paper.

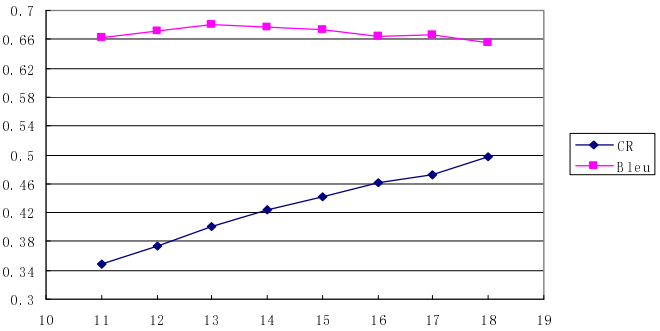


Fig. 4. The results of difference upper length on system (CR: compression ratio, Bleu: sentence similarity)

5.2 Results

Table 1 shows the evaluation results of our system. The item of OurWork is the result of the system when the upper length is 13. And OurWork_1 didn't add golden target sentence and itself as a pair into the training set, OurWork_2 added golden target sentence and itself as a pair into the training set.

Table 1. Compression Results

	CR	Importance	Grammaticality	Bleu
Golden	0.291	4.335±0.265	4.977±0.077	
OurWork_1	0.367	3.879±0.739	4.159±0.976	0.650±0.183
OurWork_2	0.401	4.200±0.562	4.444±0.776	0.686±0.160

From Table 1, we can conclude the following results:

- When we add golden target sentence and itself as a pair into the training set, the result is better
- In the case of ensuring a better compression rate, the results are grammatical and contain the center information.

Table 2 shows some test examples. From the example 2 and 3, we can conclude that the Chinese word segmentation performance affect our system results. As the error of the word segmentation generated, our system will generate some ungrammatical target sentence.

Table 2. Example compressions for the evaluation data

Original Sentence	2009年 2月 20日 清晨 ， 沈 阳 夏宫 成功 爆破 。
Golden	沈 阳 夏宫 成功 爆破
OurWork_1	沈 阳 夏宫 成功 爆破
OurWork_2	沈 阳 夏宫 成功 爆破
Original Sentence	据 云南 当地 媒体 消息 ， 目前 “ **躲 猫 猫** ” 事件 调查 委员会 正 在 封闭 完成 调查 报告 ， 报告 内容 即 将 公布 。
Golden	**躲 猫 猫** 事件 调查 委员会 正 封闭 完成 调查 报告
OurWork_1	事件 调查 委员会 正 在 封闭 完成 调查 报告 报告 内容 公布
OurWork_2	**猫 猫** 事件 调查 委员会 正 在 封闭 完成 调查 报告
Original Sentence	据 英国 广播 公司 1日 报道 ， 流亡 海外 的 泰国 前 总理 **他 信** 取消 在 香港 的 演讲 。
Golden	泰国 前 总理 **他 信** 取消 在 香港 演讲
OurWork_1	泰国 **信** 取消 在 香港 的 演讲
OurWork_2	泰国 前 总理 **他 信** 取消 在 香港 的 演讲

6 Conclusion

In this paper, Chinese sentence compression is studied for the task of title generation under the framework of deleting words. The method of adding the target sentence and itself as a pair into the training set dramatically improved the performance of our system. And our method can generate target sentence which is grammatical and contains the center information of the original sentence in the case of ensuring a better compression rate. But our system's performance is limited by the performances of Chinese word segmentation and syntactic parsing.

The future work mainly contains the followings:

- Using SRL system [7-10] add the SRL as a kind of features into our system;
- Building an automatic title generation corpus, and using our compression system into the automatic title generation task;
- Using our system into Multi-document summarization task [11-12].

References

1. Vandeghinste, V., Pan, Y.: Sentence compression for automated subtitling: A hybrid approach. In: Marie-Francine Moens, S.S. (ed.) Text Summarization Branches Out: Proceedings of the ACL Workshop, Barcelona, Spain, pp. 89–95 (2004)

2. Knight, K., Marcu, D.: Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artificial Intelligence, 91–107 (2000)
3. Riezler, S., King, T.H., Crouch, R., Zaenen, A.: Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In: Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics, Edmonton, Canada, pp. 118–125 (2003)
4. McDonald, R.: Discriminative sentence compression with soft syntactic constraints. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, pp. 297–309 (2006)
5. Zhang, Y.L., Wang, H.L., Zhou, G.D.: Sentence Compression Based on Structured Learning. Journal of Chinese Information Processing (2012)
6. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 1453–1484 (2005)
7. Zhou, G.D., Li, J.H., Fan, J.X., Zhu, Q.M.: Tree kernel-based semantic role labeling with enriched parse tree structure. Information Processing and Management, 349–362 (2011)
8. Li, J.H., Zhou, G.D., Zhao, H., Zhu, Q.M., Qian, P.D.: Improving Nominal SRL in Chinese Language with Verbal SRL and Automatic Predicate Recognition. In: Proceedings of the Empirical Methods for Natural Language Processing, Singapore, pp. 1280–1288 (2009)
9. Li, J.H., Zhou, G.D.: Unified Semantic Role Labeling for Verbal and Nominal Predicates in Chinese Language. ACM Transaction on Asian Language Information Processing, 13–21 (2011)
10. Li, J.H., Zhou, G.D., Zhu, Q.M., Qian, P.D.: Semantic Role Labeling in Chinese Language for Nominal Predicates. Journal of Software, 1725–1737 (2011)
11. Wang, H.L., Zhou, G.D.: Topic-driven Multi-Document Summarization. In: International Conference on Asian Language Processing (2010)
12. Wang, H.L., Zhou, G.D.: Toward a unified framework for standard and update multi-document summarization. ACM Transaction on Asian Language Information Processing (2012)

Event Argument Extraction Based on CRF

Libin Hou, Peifeng Li, Qiaoming Zhu, and Yuan Cao

Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu, 215006
School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, 215006
{20104227033,pfli,qmzhu,20114227022}@suda.edu.cn

Abstract. Event argument extraction is an important component of event extraction which plays a decisive role in whether event extraction can be applied to the actual. This paper proposes a method of event argument extraction based on Conditional Random Fields (CRFs). After employing frequently used features, we summarize all the features into five categories, i.e., lexical, semantic, dependency, syntactic and relative-position. More importantly, we propose using semantic role as a specific feature. Great efforts have been made to evaluate the performance by exploring various features and their combination. Experimental results show that semantic role is a good indicator for event argument extraction.

Keywords: event argument extraction, Conditional Random Fields (CRFs), Semantic role labeling, event extraction.

1 Introduction

With the development and wide spread of the internet, large amounts of information make it more and more difficult for people to obtain the message that they need. It is a desiderate problem that how to extract interesting event information from the mass of unordered, chaotic and not-structured data. Event extraction technology is just a powerful tool for solving such problem which belongs to the field of information extraction (IE). There are three main programs that focus on the relative research on event extraction. They are MUC (Message Understanding Conference), TimeML and ACE (Automatic Content Extraction). There are two main extraction targets supported by ACE: event detection and recognition, arguments (roles) extraction. The ACE 2005 tasks are more fully described in (ACE, 2005). In this paper, we focus on the latter task. Each event type has a set of possible arguments, which may be filled by entities, values, or timexs. The event arguments extraction task is to detect and recognize these event arguments from the entities, values and timexs. ACE event arguments extraction task does not offer any template in an open field, this increases the difficulty of arguments extraction to a certain extent. Firstly, we introduce some relevant glossaries of ACE Evaluation :

- Event mention: a phrase or sentence within which an event is described, including a trigger and its arguments;
- Trigger: the main word that most clearly expresses the occurrence of an event, so recognizing an event can be recast as identifying a corresponding trigger;

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 32–39, 2013.

- Trigger type/Event type: The type of an event;
- Argument: the entity mentions are involved in an event;
- Argument role: the relation of an argument to an event where it participates. There are 35 roles types. Every event type has its own set of arguments, as can be seen in Table 1.

Table 1. Event Type and its Corresponding Arguments

Event Type	Arguments
Injure	Agent Victim Instrument Time Place
Marry	Person Time Place
Start-Org	Agent Org Time Place
Attack	Agent Target Instrument Time Place
……	……

Take the following sentence as an example:

(E1) Chinese:她的丈夫和女儿在日治时期过世。
Pinyin: tāde zhàngfū hé nǚér zài rìzhì shíqī guòshì.
English:Her husband and daughter pasted away during the japanese invasion.

In this event mention , the trigger is "过世" (guòshì, past away), the event type is "Die" and arguments are "她的丈夫" (tāde zhàngfū, Her husband) and "日治时期" (rìzhì shíqī, the japanese invasion) with their corresponding roles "Victim" and "Time-Within". The event argument extraction is to determine whether the annotated entities, times and values are arguments, if so, assign them to roles. It is based on the first sub-task of the event extraction, that is, the detection and recognition of event trigger and type.

The rest of this paper is structured as follows: we discuss the relevant work in section 2, we explain our research motivation in section 3, we describe our approach in section 4, we discuss the results in section 5, and finally, we draw a conclusion in section 6.

2 Related Work

There are more relevant researches of event argument extraction which are carried on English corpus. Ahn [1] regarded the event argument extraction as a classification task and use Maximum Entropy (ME) model to realize it. It considered every entity, time expression and value as instances, thus, introduced large quantity of negative examples and led to the serious imbalance of positive and negative examples. Hong [2] regarded entity type consistency as a key feature to predict event mentions and adopted this inference method to improve the traditional event extraction system. Finkel [3] obtained a much larger extent of structure and content using CRFs models in information extraction.

In Chinese corpus, Zhao [4], Chen [5] et al. extracted event arguments using the ME model. Tan [6] explained the relation of arguments and triggers by constructing models, and also researched the event argument extraction methods based on multiple-layer module and CRFs model.

3 Motivation

Event arguments consist of event participants and relevant event attributes (e.g., time, place, etc), it is classified as different roles according to the semantic relations of argument and event type. The task of event arguments extraction is described as : Given the event's type, let R be the set of the corresponding roles of this event type where R= {r1，r2，…rn，None} and let E be "entity" set (this "entity" set includes entities, times and values as was just mentioned) where E= {etv1，etv2，…, etvm} .The task of argument extraction is to mark every etvi by its corresponding type rj and then build a map between collection E and collection R. In the following sections.

In recent years, the CRFs[7] (Conditional Random Fields) model is gradually accepted in information extraction field and obtains favorable results. E.g., Peng [8] applied the CRFs to IE field; Jakob [9] applied this model to evaluate object extraction. The task of event argument extraction is to annotate the entities in the event sentence, that is, to assign a role or none(non-argument) to every entities. In this paper, we will use the CRFs model to event argument extraction because of its good effect in sequence labeling field.

At present, most realization is based on the probability model, such as ME. But, the ME classifier can not fully consider the context features as well as it can only get a local optimal result and bring about the label bias because of its independent assumption. In contrast, the CRFs model can select any features and all features are normalized in the overall situation, so the real optimal result can be obtained. In CRFs model, the features of every entity are not analyzed independently and we can make better use of the context information. For example, in a "Transport" event mention, if an entity has been identified as the *Origin*-ARG, the probability of a nearby entity being *Destination*-ARG is increased.

We also use the features from semantic role labeling (SRL) to improve the performance of argument extraction. The SRL is annotating the predication's corresponding semantic composition in a given sentence and making semantic marks (e.g., agent, target, instrument or adjunct etc.). It may be observed that the SRL and event argument extraction are similar in some respects. The former reflects the semantic relation between the prediction and relevant phases (entities), while the latter reflects the relation between trigger and relevant entities. Take the following sentence as an example:

(E2) Chinese: [那位疑似小偷的男人 Arg1] 已被[村民 Arg0]活活打死。

Pinyin: [nàwèi yísì xiǎotōu de nánrén Arg1] yǐbèi [cūnmín Arg0] huóhuó dǎ sǐ

English: [The man seemed to be a thief Arg1] has been beat to death by [villager Arg0].

The sentence E2 contains an Attack event with the trigger “打” (dǎ,beat). In the event arguments extraction, “那位疑似小偷的男人” (nàwèi yísì xiǎotōu de nánrén, The man seemed to be a thief) is tagged as the role Victim and “村民” (cūnmín, vil-lager) is tagged as the role Agent. In SRL “那位疑似小偷的男人” is tagged as the Arg1 (target) and “村民” is tagged as the Arg0 (Agent). So there is an intimate rela-tionship between them. In this paper, we propose the ACE argument-SRL feature extraction algorithm in which the SRL role features [10-14] are merged into current features of event argument extraction to improve the latter’s performance.

4 Solution

4.1 Feature Selection

In this paper, we combine the features used in Chen [5] and Tan [6] , and introduce new SRL features into them. Table 2 gives the detail description of these features which use E2 as an example. In our experiment, the first three features are called basic features, and the others are called expanded features.

Table 2. Feature Selection and Description

Features	Feature Description	Examples
Basic features	Morphology and parts-of-speech of the trigger	Trigger:打(dǎ,beat); parts of speech: verb
	Type of event	Type: Attack
	Type of entity; head of entity	Entity type: PER; head:村民(cūnmín, villager)
Around words’ features	Morphology and parts-of-speech of the word before entity	Morphology: 被(bèi, been);parts of speech: preposition
	Morphology and parts-of-speech of the word next entity	Morphology: 活 活 ;parts of speech: adverb
	Morphology and parts-of-speech of the word before trigger	Morphology: 活活; parts of speech: adverb
	Morphology and parts-of-speech of the word next trigger	Morphology: 死 ; parts of speech: noun
Dependency features	Dependency path between trigger and current entity(or entity head noun)	Nsubj (打-13, 村民-11)
Syntactic features	Shortest path between the trigger and entity	NN↑NP↑IP↓VP↓VP↓VV
	Difference between the depth of entity and the depth of trigger in the parsing tree	Discrepancy: 1
Relative position features	Position of entity relative to the trigger, former or latter	Relative position: former
Semantic role Features	Entity’s role in this event	Role ： Arg0

Algorithm SRL Feature Extration
Require the entity list and SRL result tables
Output the corresponding semantic role for the entity
Procedure Combine the annotations of several possible predictions in SRL result, and eliminate the nest structures between multiple predictions, and reserve the smallest semantic role unit.
foreach entity
IF there is a semantic role unit R corresponding to the entity boundary
set the semantic role of R to the entity
ELSE IF there is a senmatic role unit R containing the entity
set the semantic role of R to the entity
ELSE IF there is a semantic role unit R containing the head noun of the entity
set the semantic role of R to the entity
ELSE set the feature to NULL
Repeat
End

Fig. 1. SRL Feature Extraction Method

Procedure of algorithm is as follows:

(a) Input the entity list and SRL result tables in the given event sentence.
(b) Combine the annotations of several possible predictions in SRL result, and eliminate the nest structures between multiple predictions, and reserve the smallest semantic role unit.
(c) Finally, we should give entities the role feature of semantic role unit according to some certainly rules, if there is no corresponding semantic role unit, then set the feature to NULL.

5 Experiments

5.1 Experiment Settings

We use the ACE 2005 Chinese event corpus, which contains 633 annotated documents. The material statistics is shown in the tables 3 below. In the experiment, 30 documents are chosen as the development data, and 533 documents are chosen as the training data. For evaluation, we conduct a test on a set of 70 documents.

Table 3. Material Statistics

the number of documents	633
the number of sentences	6325
the number of sentences containing events	3241
the number of arguments	8013

The event argument extraction is based on the first stage of event extraction in which the event type has been detected and the entity is standard entity in annotation document.

What we use in our experiment is a SRL tool developed by ourselves. It is a SRL system based on phrase structure syntactic parsing and feature vector. It makes a score of 78.75 points in F1 when it is tested on WSJ of CoNLL 2005 SRL shared task.

5.2 Results

1. Feature Combination and SRL Feature Analysis

In the experiment, several groups of experiments are designed. We first conduct the experiments on features and their combinations, then we compare the performance before or after introducing SRL features to reflect the influence of SRL. The results are shown in table 4 below.

Table 4. Feature Combination and Result Performance

Features	P	R	F
Basic Features	44.7	32.8	37.8
Basic + adjoin features	52.4	34.6	41.7
Basic + Syntax features	47.9	33.8	39.1
Basic + dependency features	49.1	33.4	39.8
Basic + Position features	45.7	32.9	38.3
Basic + adjoin + dependency	57.3	36.2	44.4
Basic + adjoin + dependency + Syntax	61.7	37.0	46.3
All	62.6	37.3	**46.7**
All + SRL	63.7	39.7	**48.9**

Can be seen from table 4, for single feature, when we combine the basic feature with the adjoin feature (basic+adjoin), system performance is increasing fastest by 3.9 points. For example, when the phrase "前往" comes before an entity, the entity is often an event argument and act as the Destination. The next best feature is dependency feature, improving the performance by 2.0 points. For example, the entity that has a NSUBJ relation with the trigger is often an argument and act as the Agent in a Die event. The performance increases with the increase of features involved, finally, when we count all the features in, the performance peaks. By comparing the performances of experiments before and after adding SRL features, it could be seen that new semantic role features are the good indicator to event arguments. The reason for the improvement of system performance after introducing semantic argument features is that semantic argument feature can better reflect the relationship between trigger and entities.

2. The Influence of Classifier

To confirm the CRFs model is effective and make a comparison with other classifier, we introduce the SVM classifier to do the task of extracting event arguments. There

Table 5. Results Based on Classification Model

Method	P	R	F
SVM	60.9	35.7	45.0
SVM + SRLFeatures	61.5	36.5	45.8
CRFs	62.6	37.3	46.7
CRFs + SRL Features	63.7	39.7	48.9

were four groups: among them, "+SRL features" represents the introduction of semantic features. The results are shown in the Table5 below:

Table 5 indicates that the performance of SVM classifier and CRFs model are both definitely improved after introducing SRL features. And CRFs model works better than SVM classifier with the same features. System performance peaks in the condition of "CRFs+SRL features".

3. Comparison with Other Relevant Systems

The event extraction system is a modular system, and there is a pipeline structure between these modules; results from previous module are used as the input for the next module. This means errors existing in the event type detection phase will be added up and magnified; A sharp reduction in the experiment performance is result. In this paper, we compare our system with Chen[7] which is also used pipeline structure. As can be seen in Table 6:

Table 6. The Contrast Results

System	P	R	F
Chen	60.6	34.3	43.8
Our	63.7	39.7	48.9

5.3 Problems Existing in the Experiments

Although some progress has been achieved in event extraction experiment, but the performance is pretty limited. The reason is :(1)some foundation tasks in NLP, such as parsing, dependency parsing and semantic role analysis are not very practical. (2) the size of training set is too small and there is some certain error in these data.(3)The extraction system is in a pipeline structure ,error is added up and magnified. For example, when the trigger is not detected, arguments in event mention can not be identified either. Thus, the R value of system is very low.

6 Conclusion

In this paper, we introduce a novel way of event argument extraction based on CRFs model, investigate the contribution degree of different features， especially, introduce new semantic role arguments, and propose entity-semantic extraction algorithm.

Experiments show that new proposed semantic role features are help to improve system performance.

In addition, we use SRL features in the system. Although it has a better indicative effect on event argument extraction, but its function and effect is not elaborated. Therefore, how to apply SRL framework into event argument extraction is worthy of studying.

References

1. David, A.: The stages of event extraction. In: Proceedings of the Workshop on Annotations and Reasoning about Time and Events, pp. 1–8 (2006)
2. Hong, Y., Jianfeng, Z., Bin, M., Jianmin, Y., Guodong, Z.: Using Cross-Entity Inference to Improve Event Extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 1127–1136 (2011)
3. Jenny, R.F., Trond, G., Christopher, M.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proc. 43rd Annual Meeting of the Association for Computational Linguistics, pp. 363–370 (2005)
4. Yanyan, Z., Bing, Q., Wanxiang, C., Ting, L.: Research on Chinese Event Extraction. Journal of Chinese Information Processing, 3–8 (2007)
5. Zheng, C., Heng, J.: Language specific issue and feature exploration in Chinese event extraction. In: Proceedings of NAACL, pp. 209–212 (2009)
6. Hongye, T.: Research on event extraction. Harbin Institute of Technology (2008)
7. Lafferty, J., Mccallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: The Proceedings of ICML, pp. 282–289 (2001)
8. Fuchun, P., Andrew, M.: Information extraction from research papers using conditional random fields. Information Processing and Management, 963–979 (2006)
9. Jakob, N., Gurevych, I.: Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields. In: Proceedings of EMNLP, pp. 1035–1045 (2001)
10. Junhui, L., Guodong, Z.: Improving Nominal SRL in Chinese Language with Verbal SRL Information and Automatic Predicate Recognition. In: Proceedings of EMNLP, pp. 1280–1288 (2009)
11. Junhui, L., Guodong, Z.: Unified Semantic Role Labeling for Verbal and Nominal Predicates in Chinese Language. ACM Transaction on Asian Language Information Processing (2011)
12. Li, J., Zhou, G., Zhu, Q.-M., Qian, P.: Syntactic Parsing with Hierarchical Modeling. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 561–566. Springer, Heidelberg (2008)
13. Min, Z., Wanxiang, C., Aw, A.T., Tan, C.L., Guodong, Z., Ting, L., Sheng, L.: A grammar-driven convolution tree kernel for semantic role classification. In: Proceedings of ACL, pp. 200–207 (2007)
14. Min, Z., Wanxiang, C., Guodong, Z., Aw, A.T., Tan, C.L., Ting, L., Sheng, L.: Semantic role labeling using a grammar-driven convolution tree kernel. IEEE Transaction on Audio, Speech and Language Processing, 1315–1329 (2008)

Fuzzy Matching for N-Gram-Based MT Evaluation

Liangyou Li and Zhengxian Gong

School of Computer Science & Technology, Soochow University, Suzhou, China
{20104227013,zhxgong}@suda.edu.cn

Abstract. N-gram-based metrics have been used widely in automatic evaluation of machine translation. However, most of them also lose merits due to the strict policy of matching of n-grams. Especially, the policy of exact matching leads to take synonyms as totally different words and thus give unreasonable estimation. This paper introduces fuzzy matching for n-grams, which refers to a semantic similarity function based on WordNet. And it is used to find a match with the highest similarity when incorporated into BLEU, the representative of n-gram-based evaluation metrics. Since WordNet can contribute more to high-order n-grams and fuzzy matching can perform well even with fewer references, experiments on MTC Part 2 (LDC2003T17) show our proposed method can greatly improve correlation between BLEU and human evaluation both at segment-level and document-level. Furthermore, BLEU incorporating fuzzy matching achieves more significant improvement at document-level evaluation.

Keywords: fuzzy matching, WordNet, machine translation, automatic evaluation.

1 Introduction

In recent years, machine translation (MT) has made great progress, partly because of introduction and development of MT automatic evaluation. Compared with manual evaluation, automatic evaluation can quickly display the quality of translation. Although the accuracy of automatic evaluation needs to be further improved, it has an important significance for frequent updates of MT system because of its quickness and objectivity.

Among all automatic evaluation metrics, n-gram-based methods are most widely used. A basic part of n-gram-based evaluation is to estimate whether the collection of n-grams from candidate translation can match with those of reference translation or not. Thus similarity between n-grams in the candidate and reference translations is crucial for matching. During evaluation, high-order n-grams can play a role in the measurement of local word-order problem. Such method has an underlying assumption: a candidate translation with higher quality should have closer words and word-order with reference translation. Obviously this assumption is not always true since translations which are semantically equivalent may have different manifestations. The simplest case is synonyms which are not in the same lexical form but have a strong semantic relation.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 40–48, 2013.

Therefore, this paper proposes to view lexical semantic similarity, which is based on existing linguistic resource, WordNet, as a way of fuzzy matching and use it during the process of matching n-grams. The method takes synonym into consideration when matching two n-grams. This paper also describes how to incorporate such matching into existing metrics, such as BLEU [1], representative of n-gram-based metrics, which is also test bed for our method.

2 Related Works

Automatic evaluation has experienced wealth of researches and changes. Nowadays, numerous evaluation metrics have been developed varying from methods based on edit distance, such as TER [2], to n-gram-based evaluation metrics, such as BLEU [1], METEOR [3], MAXSIM [4], to linguistic-based methods, such as STM [5], MEANT [6] etc. In all of these metrics, those based on n-grams are most widely used because of their simplicity and good correlation with human evaluation. In this section, we will talk about BLEU and some related works on fuzzy matching.

2.1 BLEU

BLEU first calculates the precision of n-grams of candidate translation, which is the ratio of n-grams from candidates that also appear in reference translations. Then geometric mean of precisions on different length of n-grams can be obtained. Particularly, when match of one n-gram happens successfully, BLEU clips the number of the n-gram with its count in reference translations, that is, the effective number of a n-gram in candidate is not greater than maximum of the same n-gram in reference. Formula (1) shows the way of calculating BLEU score.

$$\mathrm{BLEU} = \mathrm{BP} \bullet \exp\left(\sum_{n=1}^{N} \frac{\log p_n}{N} \right), \tag{1}$$

where p_n is precision of n-grams and BP is penalty factor which prevents BLEU from favoring short segments because of lacking direct consideration of recall measurement:

$$\mathrm{BP} = e^{\min(1-r/c,0)}, \tag{2}$$

where c denotes the number of words in candidate and r is the length of reference which is closest to c.

2.2 Fuzzy Matching

Studies have pointed out some pitfalls existing in BLEU metric, one of which is exact matching between n-grams and thus lack of considering synonyms. Therefore, following BLEU, some other n-gram-based evaluation metrics try to handle this defect, such

as METEOR and MAXSIM both of which consider synonym between candidate and reference with the help of WordNet during evaluation. However, both do not differentiate synonyms with different degree of semantic relations, that is, similarities in this paper.

The authors [7] propose using LCCSR (Longest Continuous Common String Ratio) to calculate the similarity of words, experimental results show that their method can effectively improve the correlation between BLEU and human evaluation. Yet their method is based on an assumption -- the more same continuous characters two words have, the more similar they are – which is not always true, so their method adopts some limitations during calculation. Moreover, calculation based on an alignment matrix also to some extent increases computational complexity and limits the scope of application of their method.

3 Fuzzy Matching for N-Grams

This paper uses existing linguistic resource, WordNet, which is more accurate and reliable compared with automatically building a synonym matching, to determine synonym and calculate semantic similarity of two words. Thus n-gram similarity can be built on similarities of their corresponding words. In this paper, the assumption that any two words can match with a certain similarity can maximize the possibility of matching.

3.1 WordNet-Based Lexical Similarity

In WordNet, nouns, verbs, adjectives and adverbs are organized into a synonym network and each synonym set (synset) represents a basic semantic concept and synsets are connected to each other with various relationships. For calculating the similarity between two words in WordNet, synonym relationships are useful information. This paper uses the method proposed by [8] to calculate lexical semantic similarity in WordNet:

$$\mathrm{sim}_{\mathrm{lin}}\left(c_1, c_2\right) = \frac{2\log P\left(c_0\right)}{\log P\left(c_1\right) + \log P\left(c_2\right)}, \tag{3}$$

where c_1 and c_2 are two synsets in WordNet, c_0 is the lowest level of word set which subsumes c_1 and c_2, and $P(c)$ is the probability of an object belonging to the synset c.

Since a polysemous word appears in more than one synsets, the similarity of two such words would meet a number of different values. In this paper, we select the maximum of all lexical similarities:

$$\mathrm{sim}_{\mathrm{wn}}\left(w_2, w_2\right) = \max\left\{\mathrm{sim}_{\mathrm{lin}}\left(c_1, c_2\right) \mid c_1 \in \mathrm{synsets}\left(w_1\right), c_2 \in \mathrm{synsets}\left(w_2\right)\right\}. \tag{4}$$

Note that our calculation in (4) ignores the context of synonym, but this method is simple, and in some applications asking for evaluation speed, this approach could be a tolerable compromise.

3.2 N-Gram Similarity

This paper derives n-gram similarity from similarities of corresponding words between two n-grams. For two n-grams $ng =< w_1 \cdots w_N >$ and $ng' =< w'_1 \cdots w'_N >$, the similarity of them is defined as follows:

$$\mathrm{sim}_{\mathrm{ngram}}(ng, ng') = \frac{1}{N}\sum_{i=1}^{N} \mathrm{sim}_{\mathrm{w}}(w_i, w'_i) \quad , \tag{5}$$

where $\mathrm{sim}_{\mathrm{w}}(w_1, w_2)$ is similarity between word w_1 and w_2, which is defined as:

$$\mathrm{sim}_{\mathrm{w}}(w_1, w_2) = \begin{cases} 1 & \text{if } w_1 = w_2 \\ \mathrm{sim}_{\mathrm{wn}}(w_1, w_2) & \text{else} \end{cases} \quad . \tag{6}$$

Equation (6) means that when lexical similarity is calculated, exact matching is first used to determine whether two words are in the same lexical form. If it does, then the similarity is 1; otherwise WordNet is used to calculate their semantic similarity. Since WordNet contains only nouns, verbs, adjectives and adverbs, the exact matching can prevent words not in WordNet from being ignored. At the same time, (5) defines n-gram similarity as average rather than minimum or maximum of lexical similarities to ensure fairness to some extent. In fact, other matching, such as those based on stem or lemma and so on, can also be added into (6) to capture more WordNet-ignored words.

3.3 Implementation Details of Fuzzy Matching in BLEU

In this paper, BLEU is our test bed. As an alternative to exact matching in BLEU, fuzzy matching would play two roles: consideration of synonyms and smoothing scores of high-order n-grams to reduce the possibility of a zero score caused by data sparsity.

The upper part above dotted line in Algorithm 1 shows the process of matching in BLEU, which uses exact matching. BLEU always keep a counter during matching. If a n-gram ng_t in candidate appears in reference, then the counter get an increment of an integer value, which is the smaller one between amounts of the n-gram in candidate and reference; otherwise, the counter remains unchanged. In Algorithm 1, $\mathrm{C}(ng, s)$ represents the number of n-gram ng in translation s, for multiple references, which is the maximum among all numbers in each reference.

The algorithm of fuzzy matching is in the lower part of Algorithm 1. If a n-gram ng_t in candidate cannot be matched exactly, then fuzzy matching will calculate similarities between ng_t and all n-grams in reference, and then select one with the maximum similarity as final matching; after that a product value between the similarity and the clipped number of n-gram is added to the counter.

Algorithm 1. Fuzzy matching for BLEU

for each n-gram ng_t in candidate t

if ng_t appears in reference r

add $\min\{C(ng_t,t),C(ng_r,r)\}$ to counter

---

else

$maxsim = 0$, $addcount = 0$

for each n-gram ng_r in r

if $\text{sim}_{\text{ngram}}(ng_t,ng_r) > maxsim$

$maxsim = \text{sim}_{\text{ngram}}(ng_t,ng_r)$

$addcount = \min\{C(ng_t,t),C(ng_r,r)\}$

add $addcount$ to counter

Obviously, the similarity of fuzzy matching is on behalf of weight of the matching. During fuzzy matching, all n-grams in reference are considered, so this method reduces the possibility of zero value of high-order n-grams, and thus can be seen as a smoothing technique.

4 Experiment

Experiments are conducted on MTC Part 2 (LDC2003T17), which contains 100 source documents in Chinese, a total of 878 segments. For each source document this dataset provides four manual reference translations and seven machine translations in English.

Translations of three machines are manually evaluated in terms of adequacy and fluency scores for each segment. All manual scores are normalized by method of [9]. Final adequacy and fluency scores for one segment is the average score of all manual scores on this segment when there are more than one human evaluating the same segment. For a document, its score is the average of scores of its segments.

Before automatic evaluation, all candidate and reference translations are tokenized and lowercased. This paper uses Pearson correlation coefficient between automatic and human evaluation scores to indicate the accuracy of automatic evaluation metric. Our experiments use WordNet 3.0[1] and a free tool[2] to compute lexical semantic similarity.

4.1 Evaluation with Fuzzy Matching

Table 1 lists the Pearson correlation coefficient between automatic evaluation and manual evaluation scores. Obviously BLEUwn proposed in this paper achieves

[1] http://wordnet.princeton.edu/

[2] http://www.sussex.ac.uk/Users/drh21/Java%20WordNet%20Similarity.beta.11.01.zip

significant improvement on correlation with both adequacy (Adq) and fluency (Flu), indicating that fuzzy matching does hold more useful information from translations to improve performance of automatic evaluation. Moreover, improvement is greater at document-level than segment-level and on adequacy than fluency.

Table 1. Pearson correlation coefficient between automatic manual evaluation scores

Metrics	Sentence-Level		Document-Level	
	Adq	Flu	Adq	Flu
BLEU	0.2379	0.2184	0.1148	0.1264
BLEUwn	0.3353	0.2745	0.2812	0.2770

Since only three systems have manual evaluation, this paper does not give the system-level coefficient. However, as shown in Table 2 which gives scores of automatic and human evaluation on each system (E05 and E09 are public web-based translation systems, E14 is a research system), both BLEU and BLEUwn do not correctly evaluate E14 system. We guess that the inconsistency of scores on E14 may result from its system model just like cases in [10] where BLEU is ineffective for evaluation on rule-based system. After all, BLEUwn does not have radical difference with BLEU and thus some flaws in BLEU may still retain.

Table 2. System score

System	Manual	BLEU	BLEUwn
E05	0.5085	0.1141	0.5364
E09	0.5123	01289	0.5670
E14	0.4685	0.1813	0.5739

4.2 Evaluation with Different N

Table 3 shows correlation coefficient between automatic and manual evaluation scores when N in (1) is set to different values. The experiment is mainly used to view change details brought by fuzzy matching.

It can be seen from Table 3 that compared to BLEU, the performance of BLEUwn when evaluating segment and considering only unigram, declines. A possible reason is that without high-order n-grams which can handle local context and with a relativ short length of segment, the ambiguity of words and fewer synonyms bring much "noise" into evaluation. However, as N increases, evaluation performance of BLEUwn is improved gradually while BLEU develops in the opposite direction. This indicates that the method proposed in this paper is more suitable for high-order n-grams. This may be because high-order n-grams can better regulate WordNet-based lexical similarities with the aid of close words. Furthermore, Table 3 shows that the increment is becoming less and less as N goes greater. So for a larger N value, BLEUwn may reach a peak.

Table 3. Pearson correlation coefficient when N is set to different value

N	Metrics	Sentence-Level		Document-level	
		Adq	Flu	Adq	Flu
N=1	BLEU	0.3201	0.2548	0.2434	0.2469
	BLEUwn	0.3067	0.2451	0.2510	0.2512
N=2	BLEU	0.3059	0.2592	0.2141	0.2238
	BLEUwn	0.3277	0.2649	0.2699	0.2669
N=3	BLEU	0.2727	0.2399	0.1732	0.1787
	BLEUwn	0.3349	0.2726	0.2777	0.2734
N=4	BLEU	0.2379	0.2184	0.1148	0.1264
	BLEUwn	0.3353	0.2745	0.2812	0.2770

4.3 Effectiveness of WordNet in Fuzzy Matching

During fuzzy matching, similarity of two words is based on two options: exact and WordNet. So this paper conducts an experiment which examines the role of WordNet in fuzzy matching. The experimental result is shown in Table 4. Compared with BLEUwn, BLEUexact ignores the calculation of WordNet-based lexical similarity in (6), that is, only uses exact word matching during fuzzy matching, so there is no difference between BLEUexact and BLEU when only considering unigram. It can be seen from Table 4 that compared to document-level evaluation the advantage of WordNet at the segment-level is not significant. This indicates WordNet plays a more important role in evaluation of document.

Table 4. Pearson correlation coefficient whether WordNet is used in fuzzy matching or not

Metrics	Sentence-Level		Document-level	
	Adq	Flu	Adq	Flu
BLEUexact	0.3313	0.2671	0.2524	0.2560
BLEUwn	0.3353	0.2745	0.2812	0.2770

In addition, since results in Table 4 shows WordNet alone does not seem to be promising, we are curious about whether WordNet acts as what we think, that is, it benefits high-order n-grams more. So we have additional experimental result shown in Table 5 which gives performance of BLEUexact with different N, just as what we do in subsection 4.2. For convenience and comparison, we also copy results of BLEUwn from Table 3 to Table 5.

Taking a close look at Table 5, we find that BLEUexact does perform well enough when comparing with results of BLEU in Table 3. However, when N is set to 4, its performance starts to decline at both segment-level and document-level evaluation. By contrast, BLEUwn keeps increasing all the time in the range. Another spot worthy of notice is that BLEUwn exceeds BLEUexact at segment-level when N is greater than 2 and at document-level all the time. Therefore, Table 5 verifies that WordNet is more suitable for document-level evaluation and high-order n-grams.

Table 5. Pearson correlation coefficient for BLEUexact when N is set to different value

N	Metrics	Sentence-Level		Document-level	
		Adq	Flu	Adq	Flu
N=1	BLEUexact	0.3201	0.2548	0.2434	0.2469
	BLEUwn	0.3067	0.2451	0.2510	0.2512
N=2	BLEUexact	0.3328	0.2651	0.2547	0.2559
	BLEUwn	0.3277	0.2649	0.2699	0.2669
N=3	BLEUexact	0.3338	0.2678	0.2545	0.2567
	BLEUwn	0.3349	0.2726	0.2777	0.2734
N=4	BLEUexact	0.3313	0.2671	0.2524	0.2560
	BLEUwn	0.3353	0.2745	0.2812	0.2770

4.4 Evaluation on Different Number of References

Typically, multiple references when available serve to reflect legitimate variance of translations. Their function overlaps with fuzzy matching in terms of synonym. So we investigate whether fuzzy matching can be an alternative when only one reference is given. Experimental results are shown in Table 6.

Table 6. Pearson correlation coefficient evaluation with different number of references

Num. of Ref.	Metrics	Sentence-Level		Document-level	
		Adq	Flu	Adq	Flu
1	BLEU	0.2023	0.1956	0.0974	0.1152
	BLEUwn	0.2465	0.1815	0.2550	0.2283
2	BLEU	0.2203	0.2074	0.0974	0.1146
	BLEUwn	0.2982	0.2337	0.2749	0.2630
3	BLEU	0.2301	0.2135	0.1046	0.1189
	BLEUwn	0.3214	0.2582	0.2780	0.2698
4	BLEU	0.2379	0.2184	0.1148	0.1264
	BLEUwn	0.3353	0.2745	0.2812	0.2770

We find that according to Table 6, BLEUwn on one reference is better than BLEU on all four references at both document-level and segment-level, with an exception at segment-level when coefficient is based on human fluency score. Moreover, BLEUwn performs better at document-level evaluation and has a greater correlation with human adequacy score.

5 Conclusion

This paper proposes a way of fuzzy matching, that is, calculating a similarity score between n-grams from candidates and references respectively, and adds this method to BLEU which originally uses only exact matching for n-grams. N-gram similarity is simply derived from lexical similarities of corresponding words. There are two options for calculating lexical similarity in this paper: “exact” and “wordnet”. When “exact” is applied, the lexical similarity is one; when using WordNet, the similarity is

based on synsets of the two words. Experimental results show that our method can significantly improve the correlation coefficient between automatic and manual evaluation scores, and fuzzy matching is more suitable for high-order n-grams and document-level evaluation, and WordNet also performs more effectively at document-level evaluation. Moreover, experimental results show fuzzy matching can also achieve promising performance when multiple references are not available.

Of course, our method does not distinguish different semantic roles of vocabulary or phrase. This could be our future work. For example, we may use semantic role labeling techniques, such as one proposed by [11], to allocate components different weights, or extract semantic relations using method of [12] to measure the completeness and correctness of translation. Also word sense disambiguation can also be used to capture the right meaning of ambiguous words.

References

1. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318 (2002)
2. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231 (2006)
3. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
4. Chan, Y.S., Ng, H.T.: MAXSIM: A maximum similarity metric for machine translation evaluation. In: Proceedings of ACL 2008: HLT, pp. 55–62 (2008)
5. Liu, D., Gildea, D.: Syntactic features for evaluation of machine translation. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 25–32 (2005)
6. Lo, C.K., Wu, D.: MEANT: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 220–229 (2011)
7. Liu, Y., Liu, Q., Lin, S.: Fuzzy matching in machine translation evaluation. Journal of Chinese Information Processing, 45–53 (2005)
8. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304 (1998)
9. Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., Ueffing, N.: Confidence estimation for machine translation. In: Proceedings of the 20th International Conference on Computational Linguistics (2004)
10. Callison-burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of bleu in machine translation research. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 249–256 (2006)
11. Zhou, G., Li, J., Fan, J., Zhu, Q.: Tree kernel-based semantic role labeling with enriched parse tree structure. Inf. Process. Manage. 47, 349–362 (2011)
12. Zhou, G., Qian, L., Fan, J.: Tree kernel-based semantic relation extraction with rich syntactic and semantic information. Inf. Sci. 180, 1313–1325 (2010)

Active Learning on Sentiment Classification by Selecting Both Words and Documents

Shengfeng Ju and Shoushan Li

Natural Language Processing Lab, Soochow University,1 Shizi Street, Suzhou, China 215006
{shengfeng.ju,shoushan.li}@gmail.com

Abstract. Currently, sentiment analysis has become a hot research topic in the natural language processing (NLP) field as it is highly valuable for many real applications.. One basic task in sentiment analysis is sentiment classification which aims to predict the sentiment orientation (positive or negative) of a document. Current approaches to this problem are mainly based on supervised machine learning technologies. The main drawback of such approaches lies in their needs of large amounts of labeled data. How to reduce the annotation cost has become an important issue in sentiment classification. In this study, we propose a novel active learning approach to select both "informative" word and document samples for annotation. Experimental results show that our approach apparently outperforms random selection or uncertainty sampling on documents.

Keywords: Chinese information processing, sentiment analysis, active learning, dual supervision.

1 Introduction

With the rapid development of Internet, the information on the Internet is more and more abundant. The various comments are valuable information to both customers and producers, which can be used for learning the satisfaction degree of the product or service. In order to acquire and analyze this kind of subjective information automatically, text sentiment analysis has got a rapid development which has aroused close attention from both academic and business research groups [1]. Sentiment Classification is a basic task of sentiment analysis, which is focused on the classification of semantic orientation, in other words, to classify the sentiment orientation as sentimental categories, such as positive, neutral and negative.

The research on sentiment classification has been carried out for many years, and currently the main methods for this task are generally based on supervised learning [1-2]. However, a significant disadvantage of supervised learning is that it requires a large amount of labeled samples during its training process while obtaining large amount of labeled samples is a very time-consuming and laborious work. Therefore, how to reduce the sample scale to be labeled and maintain a desired classification performance is an important issue which is really worth deep research. To achieve this, active learning is a method which can reduce the scale of

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 49–57, 2013.

labeling samples by choosing some “high-quality” samples actively for manual annotation, and it results in using the minimum number of labeled samples while keeps the classification performance at a high level.

The traditional active learning approaches focus on how to select the documents which could contribute most to the classification work, and select the most uncertain documents for the classifier usually before. Meanwhile the latest research shows that adding information of the words during active learning process helps improve the quality of the final classification [3], that’s to say the words with additional information (usually emotional words) make great contribution to the classification results. For instance, the word “comfortable” can be labeled as a positive word while “bad” can be labeled as a negative word in the hotel fields, therefore we can improve the classification performance with annotation of such words which contain additional information. In other words, we can have better classification performance when have both "informative" word and document samples for annotation. However, the cost of word and document annotation is different: the complexity of document makes the annotation of it much more costly of time and labor than that of word. In order to save the annotation cost, it is possibly preferable to choose the accurate sentiment words for effective annotation. In this study, based on the unit annotation scale, we calculate the information of every word and document respectively when the same annotation scale is given, then select the most helpful documents and words for manual annotation.

The remaining part of this paper is organized as follows: In the second section, we will introduce the previous research of active learning in sentiment classification; In the third section, we will introduce the classification approaches based on the coordination of word and document; The fourth section describes the active learning approaches based on collaborative selection of word and document; The fifth section i the results and analysis of the experiments and the last section is the conclusion of this paper.

2 Related Work

Recently, sentiment classification has gradually become a hot research topic in natural language processing. [3] employ the machine learning approaches based on supervised learning to conduct sentiment classification for the first time. The following studies aim at improving the performance of the supervised learning with various methods, such as extracting the subjective sentences [4], looking for the higher level classification character [5] and taking advantage of the theme related information [6]. Generally, sentiment classification research has been carried out on different text particle sizes, for example, document level [7-11] and word level [12-13]. This paper mainly focuses on document level, but also uses sentimental information for help in word level.

Active learning is an important research branch in machine learning for a long history[14-15]. The active learning algorithms can be roughly divided into three categories: The first category is to select the text which can reduce the inaccuracy

of the classifier mostly for manual annotation, like error reduction sampling approach [16]; The second category is to select the most uncertain texts of the classification results from the classifier for manual annotation, like uncertainty sampling approach [17]; The third category is based on the differences of predict result from multiple classifiers, like Query By Committee approach (QBC) [18].

In the research of sentiment classification, the relative achievements of active learning on both word and document are rare so far, compared to that of the traditional active learning on only document. Melville and Sindhwani (2009)[19-21] compare the performance of annotating both word and document with document only, and found out that the former one had better classification performance. However, they do not fully consider the different annotation cost of word and document during the selection procedure of word and document. Furthermore, the word polarity categories are not true label but a kind of simulation approach by feature selection method where the polarity categories of some words are not correct. Relatively speaking, we have fully considered the annotation cost of word and document in our approach and used the true polarity category labels in the experiments.

3 Sentiment Classification Method of Learning from Both Words and Documents

Most of the existing classification approaches are performed by training annotation of document. To incorporate the classification knowledge in labeled words, we adopt the classification method proposed by [19] that focus on the combination of document and dictionary when annotate both word and document. This method is based on Bayes classification method [22] and it degenerates to the ordinary simple Bayes when the dictionary is empty.

Simple Bayes classifier is a classification approach based on Bayes' principle and it has simple model and high operating speed as one of the most popular machine learning approaches. It has an assumption as prerequisite that the document features are independent of each other in a given document. It calculates maximum likelihood estimates to get the category of the document. The calculation formula is as below:

$$\arg\max_{c_i} P(c_i)\prod_k P(w_k \mid c_i) \tag{1}$$

Where $P_e(w_k \mid c_i)$ means conditional probability of the words w_k from training corpus to compute the document belongs to the sentiment category, $P_f(w_k \mid c_i)$ is the posterior probability of computing document belongs to the sentiment category by sentiment dictionary. α is weight ratio for both sides. We set α as 0.5 and $P(c_i)$ (prior probability of positive and negative category) as 0.5.

When learning the weight of $P_e(w_k \mid c_i)$, we calculate the estimate of the Laplasse of conditional probability word w_k in category c_i of document by ordinary simple Bayes approach. While the detailed derivation process of $P_f(w_k \mid c_i)$ has been completed by Melville etc. (2009)[12], so we just describe the calculation method of $P_f(w_k \mid c_i)$, which requires the parameters as below:

V : Collection of all words
P : Collection of positive words
N : Collection of negative words
U : Collection of neutral words , that's to say , $(V-(P+N))$
m : Number of words in collection V, that's to say , $|V|$
p : Number of words in collection P, that's to say , $|P|$
n : Number of words in collection N, that's to say , $|N|$

All the words can be divided into three categories: positive, negative and neutral. The neutral words contain two kinds of words: one are those included in collection V but haven't been annotated and the other are those have been annotated annually but can't be determined whether positive or negative. We denote the probability of positive word w_+ in positive document as $P_f(w_+ \mid +)$ and the probability of negative word w_- in negative document as $P_f(w_+ \mid +)$. Similarly, we denote the probability of neutral words w_u in positive document and negative document as $P_f(w_u \mid +)$ and $P_f(w_u \mid -)$ respectively. The detailed calculation formulas are listed as below:

$$P_f(w_+ \mid +) = P_f(w_- \mid -) = \frac{1}{p+n} \tag{2}$$

$$P_f(w_+ \mid -) = P_f(w_- \mid +) = \frac{1}{p+n} \times \frac{1}{r} \tag{3}$$

$$P_f(w_u \mid +) = \frac{n(1-1/r)}{(p+n)(m-p-n)} \tag{4}$$

$$P_f(w_u \mid -) = \frac{p(1-1/r)}{(p+n)(m-p-n)} \tag{5}$$

The parameter r means the proportion of the probability of positive words in positive documents and negative documents. In this study, we set the value to 100, i.e., r=100, as suggested by [19].

4 Active Learning with Collaborative Selection on Both Words and Documents

4.1 Annotation Costs

Previous active learning methods mainly focus on the selection of “informative" document, while the active learning approach in this paper annotate words and documents simultaneously. However, the annotation cost of word and document is different. Therefore, it is necessary to obtain the detailed annotation cost of each word and document.

In order to compare the cost of time between word and document, we propose the concept of unit annotation time and unit annotation scale:

Unit annotation time: the average annotation time of a document.

Unit annotation scale: the number of annotation words in unit annotation time.

So, we can annotate multiple words during the unit annotation time. We random sorted all the documents relating to package and hotel field together with all the words relating to these two fields, and assigned the sorted documents to two students (A and B) with Bachelor degree for manual annotation, then recorded the number of annotation words or documents in 15 minutes. Table 1 shows the final results as below:

Table 1. Annotation ratio

Annotation student	A	B
Number of document annotated	116	102
Number of word annotated	1848	1748

According to the table above, we can get the annotation overhead ratio of word and document: (1848+1748) / (116+102) =16.5. So we set the number of words can be annotated in unit annotation time as 16.5.

4.2 Selection of Sentiment Words

Sentiment words usually play a key role in the classification process of sentiment classification, so we should try to select sentiment words for annotation. Sentiment words can be divided into two categories: positive words and negative words. However, there are a large number of words are neutral words which cost much more annotation effort compared with sentiment words. Therefore, the effective selection of sentiment words for manual annotation can help save annotation cost very well.

Generally speaking, whether the word is a sentiment word is closely related to the part of speech. For instance, the probability of the adjective is sentiment word is apparently higher than that of the noun. Therefore, more attention should be paid to the adjective rather than noun while searching for sentiment words.

In order to calculate the information of the part of speech to approve that the adjective has the biggest chance to be a sentiment word, we had an experiment that

manually annotated 200 words randomly from all categories of part of speech respectively, and recorded the probability of the word is sentiment as weight. Table 2 shows the final results as below:

In addition, the frequency of a word in a document is an important basis of the importance of the word, in other words, the more a word appears in the document, the more it influences the document and the more important the word is.

$$V\left(POS\right)=\frac{\text{number of the words which are both emotional words and POS}}{\text{number of the all words which part of speech are POS}} \tag{6}$$

$$Weight(w)=V\left(POS(w)\right)\times\log(F(w)) \tag{7}$$

Note: $POS(w)$ is part of speech, $V(POS(w))$ is the weight of the part of speech which can be referred from Table 2. For instance, if the word is adjective (adj.), then $V(adj.)=0.77$. $F(w)$ means the total number of documents where the word appears.

Table 2. Statistical information of part of speech

part of speech	adjective	verb	noun	others
Weight of sentiment	0.77	0.36	0.11	0.08

4.3 Sort of Word and Document Based on Weight

We can obtain the weight of every word according to formula (7) when selecting word and document for manual annotation. In addition, we can calculate the weight of every word in each document, and it will help calculate the weight of each document.

$$Weight(d)=\frac{\sum_{w\in d}Weight(w)}{k\log\left(L(d)\right)} \tag{6}$$

$Weight(d)$ is the weight of document, $\sum_{w\in d}Weight(w)$ is the sum weight of the words in document, k is a constant which means the annotation scale of words in unit annotation time, and we set 16.5 in this paper. L is the length the document. We prefer the document with shorter length while with the same weight, because the longer the document is, the higher probability of containing words without classification information it has. These words which contain no classification information will cause noise, so we set the weight of document divided by the length of the log.

We can sort the words and documents by calculation of weight of them with formula (7) and formula (8), then select the words and documents with maximum weight to annotate.

In general, our procedure is as follows. 1) Input large number of unlabeled documents, use segmentation and part of speech annotation tool to divide words and annotate them. 2) Obtain the part of speech and occurrence frequency of each word.

3) Calculate the weight of word and document with formula (7&8). 4) Select the words and documents with maximum weight for manual annotation. 5) Train the model using the labeled word and document, then classify the test samples by Pooling Multinomials classifier.

5 Experiments

We collected the Chinese sentiment text from package and hotel field, and the corpus was from the comments on amazon website. Each field contained 2400 samples, in which 1200 positive and 1200 negative. We selected 400 positive and 400 negative samples as test samples, while all the rest as training samples. We used software called ICTCLAS by Chinese Academy of Sciences Institute for segmentation and part of speech annotation firstly. The evaluating indicator is standard accuracy in the experiment.

The comparison of different approaches in this paper:

Random Sampling (RAND): select the document randomly for manual annotation;

Uncertainty Sampling (UNCE): select the document with most uncertainty for manual annotation, it's based on the result by Bayes classifier;

Document-word co-selecting (DW): select the document and word with highest weight for manual annotation, it's based on the result of formula (8).

Since the first approach of selection is random, we performed five times and took the average results to make the final statistics stable.

Fig 1 and Fig 2 shows the performance of the three different selection categories in package and hotel field, and we can find that the performance of our DW approach is apparently better than the other two approaches when the number of text samples is small. For instance, our active learning approach based on co-selecting on words and documents performed much better than the other two approaches when only 20 or 40 unit annotation time cost. This improvement was six percentages in package field while even reached to more than 10 percentages in hotel field.

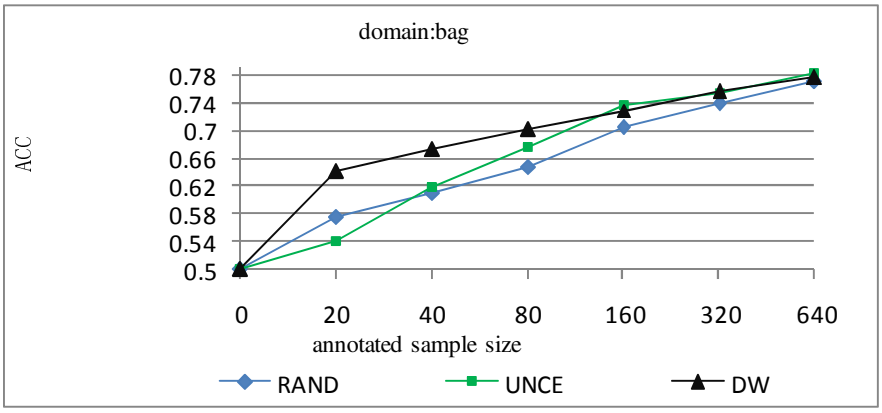


Fig. 1. Classification performance of the three approaches in each field respectively

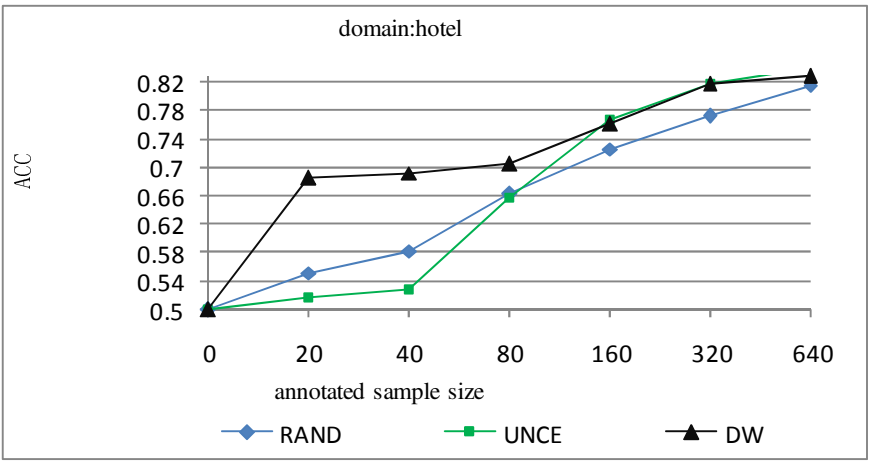


Fig. 2. Classification performance of the three approaches in each field respectively

After investigation of selected samples, we can find that the samples contained large amount of sentiment words at the beginning stage. This result shows the importance of annotating some sentiment words as classification resource at the beginning stage. As the sentiment scale increased to some extent, our approach performed similarly with UNCE, but still had obvious advantage on RAND.

6 Conclusion

This paper proposes a novel active learning approach for sentiment classification, where both "informative" words and documents are actively selected for training the classifier. To enable selecting words and documents simultaneously, we evaluate their annotation costs and informativeness values. The experimental results demonstrate that the proposed active learning approach greatly reduces the annotation cost and significantly outperforms random sampling and uncertainty sampling.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002, pp. 79–86 (2002)
2. Li, S., Zong, C.: Multi-domain Sentiment Classification (short paper). In: Proceedings of ACL 2008, pp. 257–260 (2008)
3. Melville, P., Gryc, W., Lawrence, R.: Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In: Proceedings of KDD 2009, pp. 1275–1284 (2009)
4. Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In: Proceedings of ACL 2004, pp. 271–278 (2004)
5. Riloff, E., Patwardhan, S., Wiebe, J.: Feature Subsumption for Opinion Analysis. In: Proceedings of EMNLP 2006, pp. 440–448 (2006)

6. McDonald, R., Hannan, K., Neylon, T., Wells, M., Reynar, J.: Structured Models for Fine-to-coarse Sentiment Analysis. In: Proceedings of ACL 2007, pp. 432–439 (2007)
7. Cui, H., Mittal, V., Datar, M.: Comparative Experiments on Sentiment Classification for Online Product Reviews. In: Proceedings of AAAI 2006, pp. 1265–1270 (2006)
8. Li, S., Huang, C., Zong, C.: Multi-domain Sentiment Classification with Classifier Combination. Journal of Computer Science and Technology (JCST) 26(1), 25–33 (2011)
9. Li, S., Lee, S., Chen, Y., Huang, C., Zhou, G.: Sentiment Classification and Polarity Shifting. In: Proceeding of COLING 2010, pp. 635–643 (2010b)
10. Li, S., Huang, C., Zhou, G., Lee, S.: Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In: Proceedings of ACL 2010, pp. 414–423 (2010a)
11. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis: Foundations and Trends. Information Retrieval 2(12), 1–135 (2008)
12. Hatzivassiloglou, V., McKeown, K.: Predicting the Semantic Orientation of Adjectives. In: Proceedings of ACL 1997, pp. 174–181 (1997)
13. Wiebe, J.: Learning Subjective Adjectives from Corpora. In: Proceedings of AAAI 2000 (2000)
14. McCallum, A., Nigam, K.: Employing EM in pool-based active learning for text classification. In: Proceedings of ICML 1998, pp. 350–358 (1998)
15. Long, J., Yin, J., Zhu, E., Zhao, W.: Active learning research. Research and Development of Computer 45, 300–304 (2008)
16. Roy, N., McCallum, A.: Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In: Proceedings of ICML 2001, pp. 441–448 (2001)
17. Lewis, D., Gale, W.: Training Text Classifiers by Uncertainty Sampling. In: Proceedings of SIGIR 1994, pp. 3–12 (1994)
18. Argamon-Engleson, S., Dagan, I.: Committee-Based Sample Selection For Probabilistic Classifiers. Journal of Artificial Intelligence Research, 335–360 (1999)
19. Melville, P., Sindhwani, V.: Active Dual Supervision: Reducing the Cost of Annotating Examples and Features. In: Proceedings of NAACL 2009, pp. 49–57 (2009)
20. Sindhwani, V., Melville, P.: Document-Word Co-Regularization for Semi-supervised Sentiment Analysis. In: Proceedings of ICDM 2008, pp. 1025–1030 (2008)
21. Sindhwani, V., Hu, J., Mojsilovic, A.: Regularized co-clustering with dual supervision. In: NIPS, pp. 1505–1512 (2008)
22. Zong, C.: Statistical natural language processing. Tsinghua University Publishing (2008)

Research on Intrinsic Plagiarism Detection Resolution: A Supervised Learning Approach

Xiuli Hua, Shoushan Li, Peifeng Li, and Qiaoming Zhu

Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu, 215006
School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, 215006
{20104227033,lishoushan,pfli,qmzhu}@suda.edu.cn

Abstract. Existing researches on text plagiarism detection mainly focus on external plagiarism detection which assumes a reference collection is given and the plagiarism detection task aims to compare suspicious documents against this collection to find the plagiarism articles with high similarity. The results of existing studies have performed well in identifying external plagiarized sections. However, in the real world, the reference collection is impossible to get. This paper focuses on this case and proposes an intrinsic plagiarism detection framework with supervised machine learning approach. The instance creation and the feature selection method are presented in detail. The experimental results on PAN'09 corpus demonstrate the effectiveness of our approach to intrinsic plagiarism.

Keywords: text plagiarism, intrinsic plagiarism detection, supervised learning, feature selection.

1 Introduction

The development of the Internet makes information sharing more and more rifeness, which helps people obtain information conveniently and provides an opportunity to plagiarize. Existing researches on text plagiarism detection always aim to compare suspicious documents against the paper collection to find the plagiarism article with high similarity. For example, the detection method for approximate copy has been studied intensively by Hoad [1]. Finkel [2] proposed an adjustable indicator of approximate plagiarism detection. Different from above, we assume that copy detection for the text in the premise of the paper collection can't get in advance and call this copy detection as the intrinsic plagiarism detection.

This paper's contribution is that it proposed an intrinsic plagiarism detection framework with supervised learning approach. In this framework, firstly we use the inconsistency between intrinsic plagiarism segments and whole article, then we learn the style model of article segment effectively to construct a classification model, finally we identify plagiarism segment.

2 Related Work

Some researchers have put forward intrinsic text plagiarism detection methods and achieved good achievements in recent years. But most researchers used the unsupervised

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 58–63, 2013.

method, namely that looking up the different features between intrinsic segment and the rest by using pre-defined rules, to identify abnormal segments. For example, Stamatatos [1] mined the string features in text and then detected plagiarism. And Essen [3] proposed the average word frequency and compared it with a text lexicalized computing method which proposed by Honore [4] and Yule [5]. Oberreuter [6] proposed a method to judge whether it was plagiarism by using words frequency features.

3 Intrinsic Plagiarism Detection Framework

3.1 System Framework

The system framework of intrinsic plagiarism detection based on supervised learning is shown in Fig. 1. This framework includes four stages: chunk, feature extraction, quantitative feature extraction and plagiarism recognization. It pre-processes the data set firstly, and then extracts features from pre-processed data and quantizes them, finally a plagiarism detection model is applied to recognize plagiaried segments.

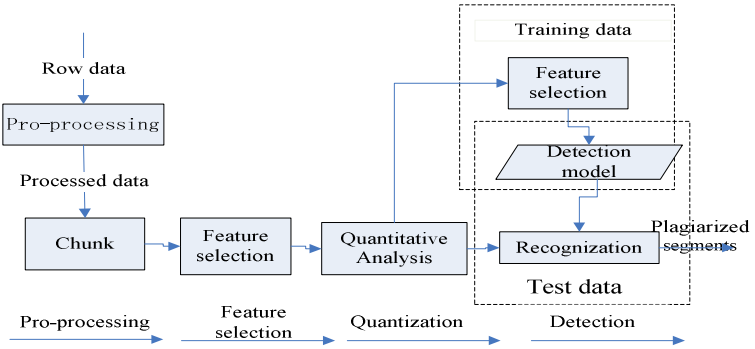


Fig. 1. The system framework of intrinsic plagiarism detection based on supervised learning

3.2 Feature Selection

We select three effective feature sets to detect intrinsic plagiarism, those specific feature sets showed in Table 1.

Table 1. Three effective feature sets

Feature sets	Feature name	Description
Character	NF	N-gram frequency of Character, n=3
Lexicon	WF	Word frequency
	AWF	Average word frequency
	H-measure	Honore's R measure
	Y-measure	Yule's K measure
	FFW	Frequency of function words
Part-of-speech	POS	Part-of-speech tag

Character feature describes the writing style feature defined from the point of view of characters' definition. Stamatatos [7] indicates that n-gram feature is very effective in intrinsic plagiarism detection. Lexicon features are extracted as a supplement in lacking of character feature. Essen [3] proposed a series of indicators to compute word frequency in the text. Many researches indicate function words such as prepositions without syntax information can recognize the writing style efficiently. Part-of-speech feature reflects writing habits. Part-of-speech feature is more reliable than lexicon feature because it reflects the writing habits.

3. 3 Quantizing Feature

The quantization method is that it compares $\sigma_j(s_i)$ with $\sigma_j(d)$,where $\sigma_j(s_i)$ is the value of j-th style feature in the segment s_i and $\sigma_j(d)$ is the average value of j-th style feature in the test document d where $\sigma_j(d) = \frac{\sum_{i=1}^{m} \sigma_j(s_i)}{\mathrm{n}}$. If$|\sigma_n(s_i) - \sigma_n(d)| > \delta$, segment $s_i$is plagiarized where δ is a threshold.

4 Experimental Results

4.1 Experiment Setting

Our experiments are based on PAN corpus and we divide it into four subsets showed in table 2.

Table 2. Four subsets pf PAN corpus

Subsets	#Files		#Total segments		#Average segments		Degree of plagiarism
	#P	#NP	#P	#NP	#P	#NP	
1	200	200	3762	2,5078	10.5	62	0.15
2	200	200	1407	9,378	4.0	26.3	0.15
3	200	200	7,116	1,9524	19.8	55	0.40
4	200	200	2,807	6,978	8.0	18	0.43

Note: set 1 & 3: long texts, set 2 & 4: short texts, set 1 & 2:minor plagiarism, set 3 & 4: heavy plagiarism. P-plagiarized files/segments, NP-non-plagiarized files/segments.

4.2 Evaluation Method

This paper evaluates the experimental result through the evaluate method used in PAN international plagiarism detection competition, including recall (R), precision (P) and F-measure (F).

Supposed d_{plg} represents a plagiarism text, then the plagiarism documentg can be denoted as $s =< s_{plag}, d_{plg}, s_{src}, d_{src} >$where the segment s_{plag} in the document d_{plg} is plagiarized from the segment s_{src} in the documentd_{src}. The case to detect plagiarism denoted as $r =< r_{plag}, d_{plg}, r_{src}, d'_{src} >$ where the segment r_{plag} plagiarized from the segment r_{src} in the document d'_{src}. If $r_{plag} \cap s_{plag} \neq \emptyset$,

$r_{src} \cap s_{src} \neq \emptyset$ and $d_{src} = d'_{src}$, the plagiarized segment is detected. We assume the set of all exiting real plagiarisms in the text denoted as S and the set of the plagiarism segments which have been detected denoted as R.

Thus P, R, F are given by

$$\mathrm{P}(S,R) = \frac{1}{|R|}\sum_{r\in R}\frac{|\bigcup_{s\in S} s \cap r|}{|r|}; \mathrm{R}(S,R) = \frac{1}{|s|}\sum_{s\in S}\frac{|\bigcup_{r\in R} s \cap r|}{|s|}; \mathrm{F}(S,R) = \frac{2\times P\times R}{P+R}$$

4.3 Results and Analyses

We introduce the SVMLight as the classifier and verify the result using 10-fold cross-validation. Table 3 shows the results.

Table 3. Experimental result on intrinsic plagiarism detection

Corpus	Non-plagiarized files			Plagiarized files		
	P	R	F	P	R	F
1	0.89	0.87	0.88	0.53	0.62	0.57
2	0.77	0.84	0.80	0.48	0.41	0.44
3	0.73	0.80	0.76	0.72	0.76	0.74
4	0.70	0.69	0.69	0.66	0.68	0.67

From table 3, we can draw a conclusion that long texts are easier to be detected than that of short texts, and plagiarism rate is the proportion to detection performance.

To explore the contribution of single feature on classification performance, we do the corresponding experiment, and table 4 shows the result. The results show that:

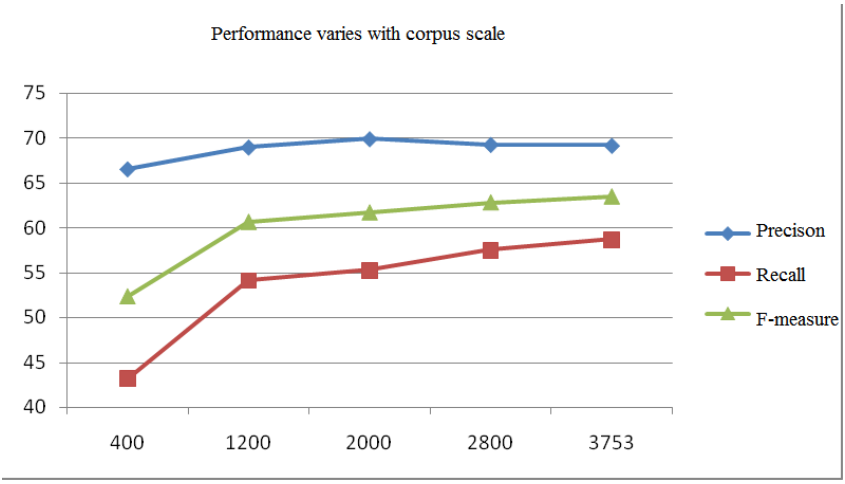
1. The Lexical features can get the optimal recall and precision according to the results which only use one of the three feature sets independently.
2. The combination of chararcter and POS feature can upgrade recall, so the overall performance is increased by 2%.

Table 4. Results based on single feature set

Feature	P	R	F
Cha	0.23	0.46	0.30
Lex	0.36	0.51	0.42
Pos	0.26	0.34	0.29
Cha+Lex	0.43	0.42	0.42
Lex+Pos	0.45	0.54	0.49
Cha+Pos	0.25	0.48	0.32
ALL	0.53	0.62	0.57

We also observe the effect to the performance through increasing the scale of data gradually, and get the result as shown in table 5.

Table 5. Performance varies with corpus scale



From the analysis, we can see that the precision is increased firstly and then has a small amount of volatility to be reduced on the increased size of training set, but the recall is increased sharply. The performance of the system is still improved in spite of the range is slowed down gradually. So the growth of training data still is helpful to improve the performance of the system.

5 Conclusions and Future Work

This paper proposes an intrinsic plagiarism detection resolution based on supervised learning approach. The result of the experiment shows that this resolution is very efficient. In the future work, more efficient structure features will be taken into consideration to upgrade the performance of intrinsic plagiarism detection.

References

1. Hoad, T.C., Zobel, J.: Methods for identifying versioned and plagiarized documents. American Society for Information Science and Technology, pp. 203–215 (2003)
2. Finkel, R.A., Zaslavsky, A.: Signature extraction for overlapdetection in documents. In: Proceedings of the 25th Australian Conference on Computer Science, pp. 59–64 (2002)
3. Meyer, E.S., Stein, B.: Intrinsic Plagiarism Detection. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 565–569. Springer, Heidelberg (2006)
4. Honore, A.: Some simple measures of richness of vocabulary. Association for Literary and Linguistic Computing Bulletin, pp. 172–177 (1979)
5. Yule, G.: The statistical study of literary vocabulary. Journal of the American Society for Information Science and Technology, 378–393 (1944)

6. Gabriel, O., Gaston, H.: Approaches for Intrinsic and External Plagiarism Detection: Notebook for PAN at CLEF, pp. 19–22 (2011)
7. Grieve, J.: Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 251–270 (2007)
8. Stamatatos, E.: Intrinsic plagiarism detection using character n-gram profiles. In: Notebook Papers of CLEF 2009 Labs and Workshops, pp. 16–17 (2009)

Employing Emotion Keywords to Improve Cross-Domain Sentiment Classification

Zhu Zhu, Daming Dai, Yaxing Ding, Jianbin Qian, and Shoushan Li

School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, 215006
{zhuzhu0020,mingroady,yxding1990,qianjianbin1990,
shoushan.li}@gmail.com

Abstract. Cross-domain classification is a challenging problem in the research of sentiment classification. In this study, we propose a novel approach to cross-domain sentiment classification by exploiting the classification knowledge from some emotion keywords. First, our approach uses some emotion keywords to extract the automatically-labeled samples with a high precision from the target area. Then, both the automatically-labeled samples from the target domain and the real labeled samples from the source domain are combined to be a new labeled data set. Third, all the labeled data and the unlabeled data in the target domain are used to perform cross-domain sentiment classification with a standard label-propagation algorithm. The empirical results demonstrate the effectiveness of our approach.

Keywords: Emotion keywords, Bipartite graph, Label propagation algorithm, Cross-domain sentiment classification.

1 Introduction

Sentiment classification aims to determine the sentimental polarity categories, e.g., positive or negative, of a given classification in the text (Pang et al., 2002) and this task has become a hot research in the natural language processing community. Nowadays, the main approach to sentiment classification is based on supervised learning approaches where a standard statistic classifier is trained on the labeled data to do the classification task. However, the main shortcoming of these approaches is that they need to annotate a great deal of labeled data manually when a new domain comes. In real-life applications, various of domains are often involved, which makes the annotation work too time-consuming and expensive. To overcome this problem, the issue of domain adaptation for sentiment classification (also called cross-domain sentiment classification) has recently been studied by some studies. In cross-domain sentiment classification, those domains which have abundant labeled samples are called source domains, while those have no labeled data but only unlabeled samples are called target domains.

In the existing approaches for cross-domain sentiment classification, both the labeled corpus of source domains and unlabeled corpus of target domains are employed to construct a model of sentiment classification in the target domain (Blitzer et al.,

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 64–71, 2013.

2007). However, the model may have a bad classification effect when the source and target domains differ a lot. It is generally believed that data among the same or similar domains have better learning ability. Therefore, we expect to get texts that are same as or similar to the target domain as far as possible. Based on this idea, we propose a novel domain adaptation approach where emotion keywords are used to extract some automatically labeled samples from the target domain to improve the adaptation performance. The novel approach is possible effective because some emotion keywords express a very clear meaning of semantic orientations (sentimental categories). Using these emotion keywords could extract a very high precision samples in the target domain. Given the classification knowledge of these newly '*labeled*' samples, domain adaption for sentiment classification achieves better performance than before.

2 Related Work

In the following we will discuss the related work regarding domain adaptation approach to sentiment classification. Recently, the study on cross-domain sentiment classification has been becoming the research focus in the area of natural language processing[2]. Blitzer et al. (2007)[4] exploits SCL (Structural Correspondence Learning) method for cross-domain sentiment classification where some pivot features are utilized to bridge the source and target domains, and improves the ability of domain adaptation to sentiment classification; Wu Q., etc. (2010)[5] combine the sentimental tendency of texts with graph ranking algorithm to select training samples for sentiment analysis cross domains, which also improves the domain applicability to sentiment classification.

Different from the previous methods, the domain adaptation approach we proposed employs some emotion keywords to automatically label samples in target domain, and then improve the ability of domain adaptation with the aid of unlabelled datasets in target domain. To the best of our knowledge, this study is the first attempt to utilize emotional knowledge to improve cross-domain sentiment classification.

3 Our Approach to Cross-Domain Sentiment Classification

3.1 Framework Overview

Emotion usually refers to human's inner reactions and feelings, such as happiness, anger, sorrow, enjoyment and so on. Generally, people express their opinions and attitudes, often with the expression of emotion. Therefore, we can consider emotion to infer people's sentimental tendencies of things. In our approach, we attempt to set the connection between sentimental tendency and emotion as followed: when people express a positive evaluation of something, it will show positive emotion and vice versa. Emotion words are the words that express emotions and that are the most obvious characteristic of emotion expressions, such as "disappointment" and "calm". In general, there are only a small number of emotion words, most of which are domain-independent, namely their emotional tendencies won't change while domain changes. Our approach proceeds in two phases:(1)First, we employ some emotion keywords to

extract the automatically-labeled samples with high precision from the target domain, and then add the automatically-labeled samples to the existing training data sets (labeled samples in other domains); (2) Second, the unlabeled data in the target domain are used to perform semi-supervised learning sentiment classification used the training data sets getting on the last step with label-propagation algorithm based on bipartite graph. Figure 1 depicts the framework of the proposed approach in the paper.

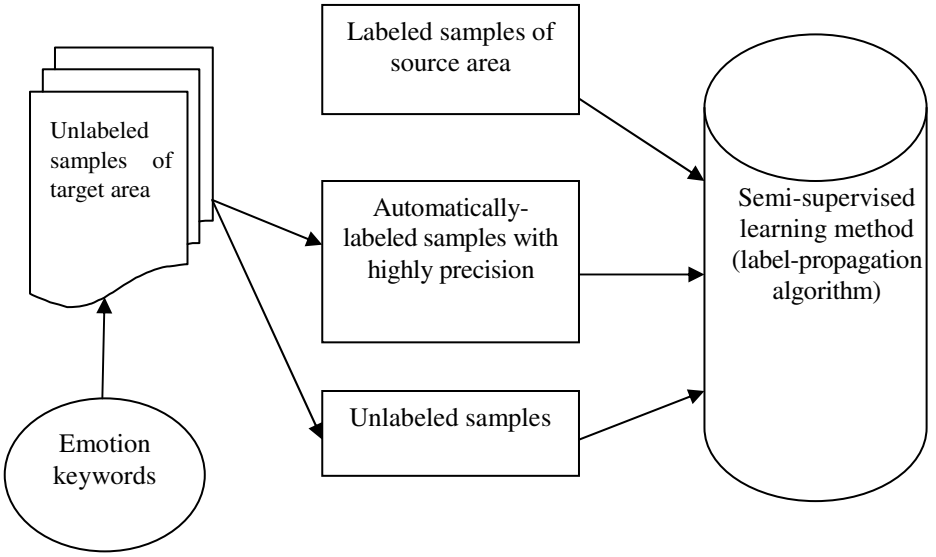


Fig. 1. The framework of cross-domain sentiment classification based on emotion keywords

3.2 Gathering and Labeling the Emotion Keywords

The English emotion words come from Plutchik_Turner[6] emotion classification system. We update the list of emotion words from Plutchik_Turner system according to the following manners: (1) add affixes (POS) information, such as sadden, saddens, saddened, sadness and so on; (2) remove the words that have semantic ambiguity. Table 1 shows the statistics of the emotion words used in this paper.

Table 1. The statistics of English emotion words

	Positive	Negative	Examples
Original emotion words	55	165	enjoy, happy, dissatisfy, hate, …
Ambiguous emotion words	49	62	obliging, surprise, shy, …
Emotion words after disambiguation and POS	112	301	enjoy, enjoying, enjoyed, happy, happily, happiness,...

3.3 Automatically Labeling Samples in Target Domain

The basic idea to extract automatically-labeled samples in target areas is to extract the samples concluded single or multiple emotion words that have a consistent emotional tendency (positive or negative).As linguistic phenomenon existed in texts may affect the true judgment of sentimental tendency, we do some filtration treatment with the following rules: :

Negation. Negation is a main reason for the difference of emotional tendency between emotion words and the whole sentence or document. For example,

*I have never been **satisfied** with the durability or performance.*

Our strategy to deal with negation is to decide the emotional tendency of sentences or documents through reversely annotating the emotion words' emotional tendencies, based on two or three negative words (not, never, hardly, etc.).

Uncertainty. Modals and if statements refer to the future emotional tendencies may or may not happen, namely uncertainty, such as *would, if, perhaps* and so on. As a result, samples containing such uncertain composition should not be regarded as having emotional tendencies. For example,

***If** I could replace just the mugs, I'd be a happy camper.*

Irrelevance. If emotion words are not the evaluation expected within the scope of a product or thing, they won't be considered. That is to say samples that only contain such emotion words will be removed. An example about expressing emotion in a movie scene or caption is as following,

*I think David Cross is a brilliant comic - even more amazing is that each show (shut up you..., **pride** is back, etc.) is different material, not the same jokes again and again.*

Our treatment to irrelevant texts is based on the operations of prompts such as "", '', ( ).

3.4 Label-Propagation Algorithm Based on Bipartite Graph (LP)

We use bipartite graph based on document-word to present the relationship between document and word [8]. If a document d_i conclude the word w_k , and the weight x_{ik} corresponds to the weight of w_k , we set an undirected edge (d_i, w_k). And the transition probability of document d_i to word w_k is $\frac{x_{ik}}{\sum_k x_{ik}}$. Similarly, the transition probability of word w_k to document d_j is $\frac{x_{jk}}{\sum_k x_{jk}}$. Hence, the transition

probability of d_i to d_j is worked out as $t_{ij} = \sum_k \frac{x_{ik}}{\sum_k x_{ik}} \cdot \frac{x_{jk}}{\sum_j x_{jk}}$. Figure 2 describes LP algorithm used to automatically label the unlabeled samples.

Input:

P: $n \times r$ mark matrix, and P_{ij} identifies the probability of document $i(i = 0...n)$ belonging to category $j(j = 1...r)$

P_L: m labeled examples corresponded to the first m lines of P^0

P_U: n-m unlabeled examples corresponded to the after n-m lines of P^0

$\bar{T}$: $n \times n$ matrix, $\bar{t}_{ij}$ is the transition probability of vector i to j

Procedure:

1) Initialization:
   a) Set iterative mark as $t = 0$, and set the value of P_L^0 according to labeled samples;
   b) Initialize P_U^0;
2) Iterating N times until convergence;
   a) Spread markup information of an example to the adjacent example according to the formula $P^{t+1} = \bar{T} P^t$;
   b) Restore markup information of mark examples, namely to replace P_L^{t+1} with P_L^0;
3) For every unlabeled mark example, to get its positive and negative labels according to $\arg\max_i P_{ij}(j = 0..r)$.

Fig. 2. Label-propagation algorithm based on bipartite graph

Input:

S_E : emotion word set;

L : labeled sample set of source domain;

U : unlabeled sample set of target domain

Procedure:

1) Extract automatically-labeled sample set L_1 from unlabeled data set U through emotion word set, and then add it to L, namely $L = L + L_1$;
2) Automatically label the rest unlabeled samples $U - L_1$ through LP algorithm, and then add it to L as well;
3) Use the learning classifier based on the latest training set L in the classification of testing samples of target domain.

Fig. 3. Implementation procedure of cross-domain sentiment classification based on emotion keywords

3.5 System Implementation Based on the Domain Adaptation Approach to Sentiment Classification with Emotion Keywords

In the framework of label-propagation algorithm based on bipartite graph, we have the automatically-labeled samples, the labeled samples in other domains and the unlabeled samples in target area semi-supervised sentiment classification. Figure 3 describes the procedure of the proposed approach.

4 Experiments

4.1 Experimental Setting

The experiment corpus come from multi-domain emotional comments corpus[1] which contains four domains: Book、DVD、Electronic and Kitchen. We select 800 positive and negative documents as unlabeled samples and 200 documents as testing data in per domain. In the experiment, we set the iterations as convergence so far. The classification method we employ is the maximum entropy classification model based on Mallet4[2] toolkit. Features used in classification are the combination of word unigrams and word bigrams. For sentiment classification, we use the standard accuracy as the evaluation measurement.

4.2 Classification Result of Automatically-Labeled Samples in Target Domain

In the conditions of all the 3 strategies (negation, uncertainty and irrelevance), Book 、 Kitchen 、 Electronic 、 DVD respectively achieves a precision of 81.17%、85.80%、88.14%、84.40% in automatically-labeled samples. Though we have achieved a high precision, there are also some errors. Through the error analysis, we find that the reasons led to such errors mainly due to three reasons: (1) There are both positive and negative emotion keywords in the reviews which makes it rather confusing. (2) Emotion words have nothing to do with the opinion target, namely the opinion target for the emotion expressed has changed. (3) Besides negation, there are some other kinds of polarity shift phenomenon, such as contrast transition with the trigger words ‘but’ or ‘however’.

4.3 Results of Our Approach to Cross-Domain Sentiment Classification

Our experiment contrasts based on three types of resources: (1) labeled samples of source area (S); (2) S and unlabeled samples of target area (S_D); (3) S_D and emotion word set (S_D_E). For clarity, Y->X means that Y is the source domain, and X is the target domain.

Figure 4 shows the comparisons of sentiment classification effect based on different resources. From the diagrams, we can find that the S_D_E approach increases

[1] http://www.seas.upenn.edu/~mdredze/datasets/sentiment/
[2] http://mallet.cs.umass.edu/

4.37% point on average than S and meanwhile get a better classification effect than S_D, which increases 1.55% point on average in the four domains. However, in the experiment of K->E、E->K、K->D, S_D_E is less effective than S_D. The main reason leading to the phenomenon is that the classification result should access to the result of using labeled samples in target area while the error existed in the automatically-labeled samples extracted by emotion words can affect the classification results. Hence, through the experiment, our approach has clearly shown the effectiveness used in two very different domains including Book and Electronic. Averagely, our approach (S_D_E) is more effective than the other approaches (S and S_D).

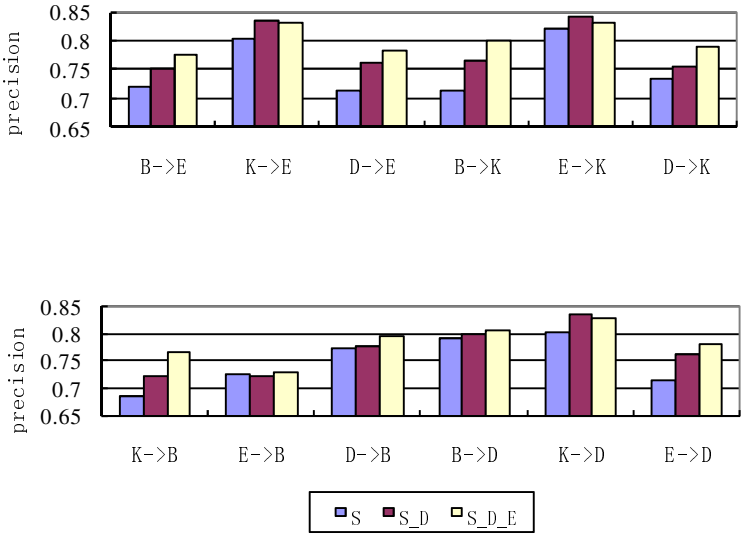


Fig. 4. Comparisons of S、S_D、S_D_E

5 Conclusions

In the paper, we propose a novel domain adaptation approach to sentiment classification with emotion keywords, which improve the ability of domain adaptation to sentiment classification with the aid of a few emotion word sets. First, we employ some emotion keywords to extract the automatically-labeled samples with high precision from the target domain. Then, we perform semi-supervised learning sentiment classification with label-propagation algorithm. The results of our experiment have shown that our approach can dramatically improve the ability of domain adaptation in sentiment classification. Especially in the case of the target area greatly different from source domain, our approach is obviously superior to the methods merely using unlabeled samples in the target domain.

Acknowledgments. The research work described in the paper has been partially supported by two grants, No. 111028524 and No. 61003155, one National Undergraduates Innovating Experimentation Project and National High-tech Research, and by Open Projects Program of National Laboratory of Pattern Recognition. We also thank the anonymous reviewers for their helpful comments.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002, pp. 79–86 (2002)
2. Li, S., Zong, C.: Multi-domain Sentiment Classification (short paper). In: Proceedings of ACL 2008: HLT, short paper, pp. 257–260 (2008)
3. Li, S., Huang, C., Zong, C.: Multi-domain Sentiment Classification with Classifier Combination. Journal of Computer Science and Technology (JCST) 26(1), 25–33 (2011)
4. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In: Proceedings of ACL 2007, pp. 440–447 (2007)
5. Wu, Q., Tan, S., Zhang, G.: Research on Cross-Domain Opinion Analysis. Journal of Chinese Information Processing 24(1), 77–83 (2010)
6. Plutchik, R.: Emotions: A Psychoevolutionary Synthesis. Harper & Row, New York (1980)
7. Turner, H.: On the Origins of Human Emotions: A Sociological Inquiry into the Evolution of Human Affect. Stanford University Press, CA (2000)
8. Sindhwani, V., Melville, P.: Document-Word Co-Regularization for Semi-supervised Sentiment Analysis. In. In: Proceedings of ICDM 2008, pp. 1025–1030 (2008)

Extracting Chinese Product Features: Representing a Sequence by a Set of Skip-Bigrams

Ge Xu[1,2,3], Chu-Ren Huang[1], and Houfeng Wang[2]

[1] Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong
churenhuang@gmail.com
[2] Institute of Computational Linguistics, Peking University, Beijing, 100871
wanghf@pku.edu.cn
[3] Department of Computer Science, MinJiang University, Fuzhou, 350108
xuge@pku.edu.cn

Abstract. A skip-bigram is a bigram that allows skips between words. In this paper, we use a set of skip bigrams (a SBGSet) to represent a short word sequence, which is the typical form of a product feature. The advantage of SBGSet representation for word sequences is that we can convert between a sequence and a set. Under the SBGSet representation we can employ association rule mining to find frequent itemsets from which frequent product features can be extracted.For infrequent product features, we use a pattern-based method to extract them. A pattern is also represented by a SBGSet, and contains a variable that can be instantiated to a product feature.We use two data sets to evaluate our method. The experimental result shows that our method is suitable for extracting Chinese product features, and the pattern-based method to extract infrequent product features is effective.

Keywords: sentiment analysis, product feature, word sequence, skip-bigram.

1 Introduction

In sentiment analysis, an important task is to extract people's opinions expressed on features of a product. Such information is valuable for both consumers and product manufacturers. For this task, the first step is to provide a set of product features. In this paper, we mainly focus on extracting product features from online Chinese product reviews.

In [1,2], the authors use association rule mining to obtain frequent itemsets, which are then pruned to find product features. However, association rule mining is unable to consider the order of words, which is very important in natural language texts. In our paper, we want to keep the order of words when using association rule mining, thus can obtain sequences of words of variable lengths, which are the candidates of product features.

For extracting Chinese product features, some problems can affect the performance.

First, although word segmentation and POS tagging in Chinese can reach over 90% accuracy, for domain-specified terms, the performance is less reliable. In Table 1, we offer some terms of digital cameras suffering from wrong word segmentation and POS

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 72–83, 2013.

Table 1. Examples of wrong word segmentation and POS tagging on digital cameras

w_1	w_2	meaning
热/a	靴/ng	hot shoe
微/ag	距/vg	macro
微/dg	距/v	macro
对/p	焦/j	focusing
滤/v	镜/ng	filter
像/v	素/dg	pixel
防/v	抖/v	anti-shake

tagging. Especially, it is noted that 微距 (macro) has two different results of POS tagging, which shows the difficulty to identify unknown domain-specific terms.

Second, in current Chinese product feature extraction, *noun+* (a noun or a sequence of nouns) is normally used to extract product feature candidates. Although *noun+* facilitates the following processing, it excludes many product features that include verbs or other POSs (mainly due to error POS tagging).

These two problems are related. Since we do not have satisfying word segmentation and POS tagging for Chinese domain-specific terms, we hesitate to use Chinese parsers that are based on word segmentation and POS tagging. So we choose to extract product features by *noun+* which require no parsing analysis.

To handle the above problems, we propose to use a set of skip bigrams (SBGSet) to represent a word sequence. A skip bigram (SBG) is a bigram that allow skips between the two words in the bigram. A word sequence and its corresponding SBGSet can be converted from each other. When combined with Apriori algorithm, it is convenient to extract frequent product features. Under this framework, wrong segmentation is corrected in the same way as multi-word phrases are identified, and filtering rules can be designed flexibly. Furthermore, we use a pattern-based method to extract infrequent product features, and a pattern is also represented as a SBGSet.

We will the give detailed introduction to our approach in the following sections.

2 Related Work

In [1,2], from a large number of reviews, the authors at first identify nouns and noun phrases using a parser, then their frequencies are counted and only the frequent ones are kept since important product features are more likely discussed by consumers. Furthermore, two pruning methods are used to improve the precision and recall. In [3], the authors use labeled data to mine Class Sequential Rules (CSR), which are then converted to language patterns for extracting product features.

Some attempts to improve the recall of product features are performed. In the paper of [4], the authors use nouns, noun phrases and verb phrases to extract feature candidates. The adding of verb phrases helps to identify more product features, and also incurs more noises. In [5], the authors claim that using noun phrases as product feature candidates limits the recall of product feature extraction, and propose a novel approach by generalizing syntactic structures of the product features.

In [6], the authors try to remove those noun phrases that do not have meronymy (part of) relationship with the target product. For example, the meronymy discriminators for the scanner class are, "of scanner", "scanner has", "scanner comes with", etc., which are used to find components or parts of scanners by searching on the Web. The algorithm also distinguishes parts from properties using WordNet's *is-a* hierarchy and morphological cues.

The double propagation method described in [7] can also be used to extract product features, which was highly dependent on a syntactic parser. The method started with only a set of seed opinion words (no seed features are required) and utilized the association or dependency relations between opinion words and features. The mutual reinforcement between feature candidates and opinion words was also used in a modified HITS algorithm [8].

In recent years, there has been an increasing interest in Chinese product feature extraction. The method in [9] starts from a small seed set of opinion words, and identifies product features by opinion words and vice verse through a bootstrapping iterative learning strategy; In [10], the authors propose a mutual reinforcement method to cluster both product feature candidates and opinion words. An association matrix between product features and opinion words was constructed. Together with traditional similarity approaches, the association matrix was used to calculate the similarity between homogeneous nodes. A similar iterative method also occurred in [11], which clustered product features and opinion words simultaneously by fusing both semantic information and co-occurrence information.

3 Representing a Word Sequence with a Set of Skip Bigrams

3.1 A Set of Skip Bigrams

Skip-grams are n-grams that allow words to be "skipped". In [12], the authors define k-skip-n-grams for a sentence $w_1, w_2, ...w_m$ to be the set $\{w_{i_1}, w_{i_2}, ...w_{i_n} | \sum_{j=1}^{n} i_j - i_{j-1} <= k+1 \; and \; i_j - i_{j-1} > 0\}$. For example, "4-skip-2-gram" results generalized bi-grams include 4 skips, 3 skips, 2 skips, 1 skip, and 0 skips (typical bi-grams formed from adjacent words).

In our paper, we only consider **skip-bigram** (SBG) because it is possible that we combine a set of skip-bigrams (SBGSet) to a sequence of any length. A SBG is denoted as **(**$word_1 \; word_2$**)**.

For example, for the following sequence of words,

"The price is high",

the SBGSet is {(The price), (The is), (The high), (price is), (price high), (is high)}. Furthermore, in most cases, we can recover the sequence using SBGSet. For example, given the SBGSet {(The price), (The is), (The high), (price is), (price high), (is high)}, "The price is high" can be recovered uniquely.

In addition, it is possible that the words in a sequence recovered from a SBGSet are not adjacent. For example, the SBGSet {(The price), (The high), (price high)} can be converted to "The price high", in which "price" and "high" are not adjacent in original sentence " The price is high".

3.2 The Conversion between a Sequence and a SBGSet

From a Sequence to a SBGSet. It is straight forward to convert a word sequence to a SBGSet. For our experiments on extracting Chinese product features, since product features are normally single words or short phrases, and also opinions occur with product features in local texts, we only consider skip-bigrams which has at most three skips between the two words.

For example, for the word sequence "abcdef" (a letter is seen as a word), the SBGSet generated is {(a b), (a c), (a d), (a e), (b c), (b d), (b e), (b f), (c d), (c e), (c f), (d e), (d f), (e f)}.

From a SBGSet to a Sequence. Definition 1: valid SBGSet
A valid SBGSet is the SBGSet that can be converted to a word sequence uniquely, and no abundant SBGs exist.

The necessary condition for a valid SBGSet is :

1. There exists an integer m, and the size of the SBGSet is $m * (m-1)/2$.
2. The number of terms[1] in the SBGSet is less than or equal to m. If there exists a word that occurs more than once in the original sequence, the number of terms is less than m, otherwise equal to m.

The above condition can not guarantee that a SBGSet is a word sequence. For example, {(a b), (b c), (c a)} meets the necessary condition, however, no word sequence can be recovered.

The necessary condition for a valid SBGSet is used to eliminate invalid SBGSets, which is not supposed to recover a word sequence. For a SBGSet that meets the necessary conditions, we simply consider all the permutation of the terms in the SBGSet, and keep the one that satisfies all the order relationship formed by SBGs in the SBGSet. Since a product feature is a short sequence of words, such processing is affordable.

However, it is possible that a valid SBGSet can be converted to multiple word sequences. For this we give the following theorem.

Theorem 1: If at most one word in a sequence occur more than once, the SBGSet generated by the sequence can be uniquely converted to the sequence.

The proof is not shown due to space limit.

When two or more words in a sequence occur more than once, it is possible that the generated SBGSet from the sequence can be converted to more than one sequence. For example, the generated SBGSet for "baab" is {(b a), (b a), (b b), (a a), (a b), (a b)}. However, this SBGSet can be recovered to two sequences: "baab" and "abba".

When we see a product feature as a word sequence, normally the length of the product feature is short, and the possibility that two or more words occur more than once is extremely low. So, under this observation, we can always get the original sequence of words for a product feature by the corresponding SBGSet.

Another problem may affect the conversion from SBGSet to a word sequence. It is possible that the SBGs in a valid SBGSet are far separated. For example, assume that we have a SBGSet {(a b), (b c), (a c)}, which can be recovered to the word sequence "abc".

[1] If a word occur more than once, it is seen as one term.

However, the three SBGs (a b), (b c), and (a c) can be far apart, such as in a sentence like "XXXXabXXXXbcXXXXac", so the possibility that "abc" is a semantic concept is low. Although the probability exists, no such cases are observed in our experiments, so we choose to ignore this.

4 Extracting Frequent Product Features

4.1 Creating SBGSet Corpus

For our experiments, each line of text is a sentence after preprocessing, which is seen as a word sequence, and then convert to a SBGSet. To processing single words simultaneously, we create a special type of SBG for any single word in the form of $(w = w)$, where w is a single word, see Table 2 for details.

Table 2. Examples of converting word sequences to SBGSets

sequence	quliaty of the battery
SBGSet	(quliaty=quliaty), (of=of), (the=the), (battery=battery), (quliaty of), (quliaty the), (quliaty battery), (of the), (of battery), (the battery)
sequence	quliaty of that bad battery
SBGSet	(quliaty=quliaty), (of=of), (that=that), (bad=bad), (battery=battery), (quliaty of), (quliaty that), (quliaty bad), (quliaty battery), (of that), (of bad), (of battery), (that bad), (that battery), (bad battery)
sequence	quliaty of battery is
SBGSet	(quliaty=quliaty), (of=of), (battery=battery), (is=is), (quliaty of), (quliaty battery), (quliaty is),(of battery),(of is), (battery is)

After the conversion, we have a SBGSet corpus, in which each line (a sentence) is represented a set, not a sequence, and can be seen as a transaction for Apriori algorithm. It should be mentioned that in the SBGSet corpus, a SBGSet (a line) normally is not a valid SBGSet on the whole (because we restrict skips between words), but contain many subsets which are valid SBGSet.

4.2 Extracting Frequent Candidate Product Features

In the SBGSet corpus, we see a SBG as an item, and a SBGSet generated from a line of text as a transaction. Thus, we can use the association rule mining to extract frequent itemsets (SBGSets), and then recover short word sequences that are possible product features.

In association rule mining, $I = i_1, i_2, ..., i_n$ is a set of items, and D is a set of transactions. Each transaction contains a subset of I. An association rule is an implication of the form $X \rightarrow Y$, where $X \subset I, Y \subset I, X \cap Y = \emptyset$.

As in [1], we use Apriori algorithm in [13] to find all frequent itemsets in the transaction set.

For the example in Table 2, after running Apriori algorithm with minimum support 3, we obtain 41 itemsets (SBGSets). Of them, many are invalid SBGSet and can not be converted to word sequences. After converting valid SBGSets to sequences, we have sequences of words in Table 3.

Table 3. Recovered sequences of words

sequences of words	support
quality	3
quality of battery	3
quality of	3
quality battery	3
of	3
of battery	3
battery	3

4.3 Filtering Product Features

After getting word sequences, we use some filtering rules to reduce the size of possible product features. For example, from Table 3, sequences such as "of", "of battery" are not product features and should be eliminated. The filtering rules are highly linguistic-dependent, for our experiments, we use the following filtering rules:

- A product feature contains 2∼8 Chinese characters.
- A product feature contains 1∼3 Chinese words.
- Remove the sequences that contain characters other than Chinese characters.
- POS filtering: We define a set of error POS tags (applied on all words) and a set of valid POS tags (precisely nouns and verbs, applied on the words contain at least two Chinese characters). In addition, if the last word in a product feature contain more than one Chinese character, the word must be a noun.
- Word filtering: Some words are not supposed to appear in product features, normally they are stop words such as 是 (be), 的 (of), 这 (this), 有 (have) etc.

Since a feature may be a part of another feature, we use p-support to recalculate the frequency of a feature. In [1], the authors defined that **p-support** of a feature *ftr* is the number of sentences that *ftr* appears, and these sentences must contain no features that contain *ftr*.

5 Extracting Infrequent Product Features by Patterns

We can extract more itemsets (possible product features) by decreasing the support for Apriori algorithm. However, when the support is very low, the size of extracted itemsets is quite large and makes the processing unaffordable.

In this paper, to extract infrequent product features, we use the existing product features to find the patterns, and then use these patterns to extract new product features. Since a pattern is also a sequence, we can also represent a pattern using a SBGSet.

5.1 Selecting Seed Product Features

To obtain patterns for extracting product features, we need some high-quality product features as seed product features. These seeds are used to obtain the contexts in which product features occur, then we extract product features using these contexts.

There are many ways to obtain high-quality product features. For example, we can rank the frequent product feature, and choose the ones with high scores as product features. In this paper, since the focus is not ranking product feature and the size of frequent product feature is small (normally several hundreds after filtering), we simply manually label the frequent product features to get seed product features.

5.2 Creating the SBGSet Corpus for Extracting Patterns

Having seed product features, we use the following steps to obtain patterns:

1. Create the SBGSet corpus by extracting sentences which contain seed product features. Note that a seed product feature in a sentence is replaced by a variable X.
2. Use Apriori algorithm to extract frequent SBGSets.
3. SBGSets are filtered and converted to patterns

Definition 2: pattern
A *pattern* is a sequence with a *variable* and at least one word.

Definition 3: variable
A *variable* can be instantiated as a word sequence.

We provide an example to illustrate the steps.

Suppose we already have seed product features, and one of the product features is "touch screen". Table 4 shows how we replace "touch screen" in the line with a variable X, and get a SBGSet containing a variable X.

Table 4. Create SBGSet corpus for extracting patterns

The touch screen is big
⇓
The X is big.
⇓
{(The X), (The is), (The big), (X is), (X big), (is big)}

To get the SBGSet corpus for extracting patterns, we go thorough the whole text file, and extract all the lines that contain any seed product feature; all the product features will be replaced by the variable X.

We also create two pseudo-words (START and END) for the beginning and ending of a sentence respectively, which is useful in extracting patterns for infrequent product features.

5.3 Extract Frequent Patterns

Then, each line in SBGSet corpus created in last section is again represented by SBGSet, and can be seen as a transaction for Apriori algorithm, and we use the algorithm to obtain all frequent itemsets (SBGSets). Finally, we convert SBGSets to sequences, and use definition of patterns to get frequent patterns.

For the case above, "The X is big", "The X is" or "X big" etc. are all patterns. If their frequencies in the SBGSet corpus for extracting patterns are over minimum support, they will be in the set of frequent patterns.

5.4 Selecting and Scoring Patterns

We can score patterns and assume that the word sequences extracted by good patterns are more likely to be product features.

However, although ranking is important for performance, it is not the focus in this paper. So we use simple scheme to select and score patterns. In our experiments, we only consider patterns with the form "$word_1$ X $word_2$", where X is a variable. The score of a pattern p is defined as: $\frac{the\ frequency\ that\ p\ extracts\ seed\ product\ features}{the\ frequency\ that\ p\ extract\ frequent\ product\ features}$.

5.5 Extracting Product Features with Patterns

Suppose that we have the pattern "The X is", two examples are shown in Table 5. It should be pointed out that each sentence is represented as a SBGSet during processing, but we use the original sentence in Table 5 for clarity.

Table 5. Two product features extracted by the pattern "The X is"

The keyboard is big. ⇓ X is instantiated as "keyboard"
The quality of signal is bad. ⇓ X is instantiated as "quality of signal"

In our experiments, we assume that a variable is composed of at most three words, which is reasonable since product features are single words or short phrases. After instantiating all the variables in patterns when going through the corpus, we use the same filtering rules in section 4.3 and exclude the frequent product features, finally we obtain the set of infrequent product features.

Since the number of infrequent product features is normally much larger than frequent ones, we use the simple scheme to rank them in order to extract the ones with high scores. We define the score of an infrequent product feature ip is:

$$the\ sum\ of\ scores\ of\ patterns\ that\ can\ extract\ ip.$$

6 Evaluation

6.1 Experiment Setting

In our experiment, we set minimum support to 0.1% (ratio) for Apriori algorithm in extracting frequent product feature for both corpora; in extracting patterns, we set minimum support to 10 (frequency) for the mobile phone corpus and 3 (frequency) the digital camera corpus.

For infrequent product features, if the number of extracted ones are larger than 500, we only keep the top 500 according to the scores defined in section 5.5.

6.2 Evaluation

Given a set of extracted product features E and a gold standard G (a set of product features manually labeled). We use recall (R) and precision (P) to evaluate the extracted product features, $R = \frac{\sum_{x \in E \cap G} Freq(x)}{\sum_{x \in G} Freq(x)}$ and $P = \frac{\sum_{x \in E \cap G} Freq(x)}{\sum_{x \in E} Freq(x)}$, where $Freq(x)$ is the number of the feature x in the corpus.

6.3 The Corpora

We download product reviews on mobile phones and digital cameras from *www.taobao.com*. We use ICTCLAS[2] package to perform word segmentation and POS tagging, during which Specification for Corpus Processing at Peking University in [14] is adopted. Some statistics of two corpora are shown in the Table 6.

Table 6. Statistics of two corpora

Product	Phone	Camera
Corpus size	2396KB	403KB
No. of annotated features	167	454

6.4 Test Sets

Since the corpora are large, it is costly to annotate them in the corpus directly. Instead, after word segmentation and POS tagging, we extract all the word sequences according to the filtering rules in section 4.3, and then perform the annotation. The size of two test sets (two sets of product features) is shown in the last row of Table 6.

6.5 Results of Experiments

The precision in our experiments is lower than some work in English [1,3,6], which is explained from two aspects:

1. In this paper, we pay more attention on the representation of short sequences which can be used to represent product features and patterns, and we do not use any

[2] http://www.ictclas.cn

ranking schemes (meronym relationship, feature-opinion association etc.) in our experiments. There ranking schemes[3] help to find a more compact set of product features or patterns.

2. Since we use no Chinese parser for our experiments, we can not obtain noun phrases from a parser. To cover more product features, we allow verbs and other POSs in product features, so the size of candidate product features is large and the precision is low.

Table 7. Results of extracting product features

Product	Mobile phone corpus		Digital camera corpus	
Size of annotated product features	167		454	
Extracted frequent product features	No. of correct	67	No. of correct	146
	No. of Extracted	284	No. of Extracted	630
	Precision	0.4547	Precision	0.4097
	Recall	0.9117	Recall	0.6845
Extracted infrequent product features	No. of correct	37	No. of correct	46
	No. of Extracted	500	No. of Extracted	500
	Precision	0.1249	Precision	0.1557
	Recall	0.0412	Recall	0.0307
All extracted product features	No. of correct	104	No. of correct	192
	No. of Extracted	784	No. of Extracted	1130
	Precision	0.4081	Precision	0.3828
	Recall	0.9529	Recall	0.7153

In Table 7, we can see that frequency is a good indicator for product features. For the extracted frequent product features, recall is 0.9117 for mobile reviews and 0.6845 for camera reviews, which show that frequent product features account for the majority of product features in the corpus.

It is encouraging to find that pattern-based method can find many infrequent product features. Although these product features occur less in the corpus and contribute less to recall, the number of them is not small.

Compared with the experiment on mobile reviews, there are more product features in the domain of digital cameras and the corpus is small, which explains the low performance on camera reviews because it is more difficult to extract reliable sequences (product features or patterns) when product features and the contexts for them are diversified and of low frequency.

7 Conclusion and Future Work

In this paper, aiming to extract product features in Chinese reviews on a product, we propose to use a SBGSet (a set of skip-bigrams) to represent a short word sequence,

[3] Of them, meronym (part of) relationship is most useful. Many high frequent words such as "girl", "old person" can be filtered out by this relationship.

which is of variable length and the common form of product features. The advantage of SBGSet representation is to convert a sequence to a set, thus we can use Apriori algorithm to find frequent itemsets, and then recover the word sequences from the itemsets. Furthermore, we propose a pattern-based method to extract infrequent product features in the corpus, and we also see a pattern as a short sequence that can be represented by a SBGSet.

By adopting SBGSet representation for short word sequences, we can devise filtering rules flexibly and such representation is helpful to wrong word segmentation in Chinese.

In our future work, we will cluster the product features and also make a deep analysis on the association between product features and opinions.

Acknowledgements. This work was finished when the first author worked in the group led by Professor Churen Huang in the Kong Kong Polytechnic University, supported by Joint Supervision Scheme(G-U966) and GRF project(544011,543810). The work is also supported by National High Technology Research and Development Program of China (863 Program) (No.2012AA011101),National Natural Science Foundation of China (No.60973053, No.90920011), and the Specialized Research Fund or the Doctoral Program of Higher Education of China (Grant No.20090001110047).

References

1. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of Nineteeth National Conference on Artificial Intellgience, AAAI 2004 (2004)
2. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD (2004)
3. Hu, M., Liu, B.: Opinion feature extraction using class sequential rules. In: AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, Palo Alto, USA (2006)
4. Fujii, A., Ishikawa, T.: A system for summarizing and visualizing arguments in subjective documents: Toward supporting decision making. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text, ACL 2006 (2006)
5. Zhao, Y., Qin, B., Hu, S., Liu, T.: Generalizing syntactic structures for product attribute candidate extraction. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (2007)
6. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, HLT/EMNLP (2005)
7. Qiu, G., Liu, B., Bu, J., Chen, C.: Expanding domain sentiment lexicon through double propagation. In: International Joint Conference on Artificial Intelligence, IJCAI 2009 (2009)
8. Zhang, L., Liu, B.: Extracting and ranking product features in opinion documents. In: Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010 (2010)
9. Wang, B., Wang, H.: Bootstrapping both product features and opinion words from chinese customer reviews with cross-inducing. In: Proceedings of IJCNLP 2008 (2008)
10. Su, Q., Xu, X., Guo, H., Wu, X., Zhang, X., Swen, B., Su, Z.: Hidden sentiment association in chinese web opinion mining. In: WWW 2008 (2008)
11. Du, W.F., Tan, S.B.: An iterative reinforcement approach for fine-grained opinion mining. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (2009)

12. Guthrie, D., Allison, B., Liu, W., Guthrie, L., Wilks, Y.: A closer look at skip-gram modelling. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006 (2006)
13. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules. In: VLDB 1994 (1994)
14. Yu, S., Duan, H., Swen, B., Chang, B.: Specification for corpus processing at peking university: Word segmentation, pos tagging and phonetic notation. Journal of Chinese Language and Computing 13 (2003) (in Chinese)

Ensemble Learning for Sentiment Classification

Ying Su[1], Yong Zhang[2,3], Donghong Ji[2], Yibing Wang[4], and Hongmiao Wu[5]

[1] Department of Computer and Electronic, Huazhong University of Science and Technology Wuchang Branch, Wuhan, P.R. China
[2] Computer School, Wuhan University, Wuhan, P.R. China
[3] Department of Computer Science, Huazhong Normal University, Wuhan, P.R. China
[4] Third Faculty, Second Artillery Command College, P.R. China
[5] School of Foreign Languages and Literature, Wuhan University, P.R. China
ychang.cn@gmail.com, donghong_ji2000@yahoo.com.cn,
{suying929,cjt422,hongmiao23}@163.com

Abstract. This paper presents an ensemble learning method for sentiment classification of reviews. The diversity among the machine learning algorithms for sentiment classification with different settings, which includes different features, different weight measures and the modeling of negation, is investigated in three domains, which gives a space for improving the performance. Then the ensemble learning framework, stacking generalization is introduced based on different algorithms with different settings, and compared with the majority voting. According to the characteristic of reviews, the opinion summary of review is proposed in this paper, which is composed of the first two and last two sentences of review. Results show that stacking has been proven to be consistently effective over all domains, working better than majority voting, and that using the opinion summary can improve the performance further.

Keywords: sentiment classification, sentiment analysis, stacked generalization, diversity measure.

1 Introduction

Today, there are more and more customers' reviews about products or services on the Web, which are shown the authors' overall opinion towards the subject matter. Sentiment classification of these reviews would be helpful in business intelligence applications and recommender systems, where the reviews could be quickly summarized. But compared to topical text classification, the sentiments of reviews are more difficult to analyze.

Although a lot of studies have been done on the sentiment analysis in recent years [1-5], especially for online reviews [6], [7], [8], sometimes we need classified documents with high accuracy for other researches, such as building of sentiment lexicon [9], finer-grained sentiment analysis and so on. Many machine learning techniques are applied to the sentiment classification problem and have shown the effectiveness. At the same time, we found the performances of different methods vary so much when using different features and weight measures [3]. For example, the

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 84–93, 2013.

maximum and minimum accuracies in the study [3] are 82.9% and 72.8% respectively for English. For reviews, the gaps between the maximum and minimum accuracies are also more than 10% [5]. One of the most interesting topics exploiting the diversity is studying how to combine the individual predictions of multiple classifiers. The motivation derives from the opportunity of obtaining higher prediction accuracy at meta-level, while treating classifiers as black boxes, i.e., using only their output, without considering the details of their implementation.

Stacked generalization [10], or stacking, is a common ensemble learning method for constructing classifier ensembles. A classifier ensemble, or committee, is a set of classifiers whose individual decisions are combined in some way to classify new instances. Stacking combines multiple classifiers to induce a higher-level (meta-level) classifier with improved performance. The latter can be thought of as the president of a committee with the base-level classifiers as members. Each unclassified review is first given to the members; the president then decides on the polarity of the review by considering the opinions of the members and the review itself. Base-level classifiers often make different classification errors. Hence, a president that has successfully learned when to trust each of the members can improve overall performance. In contrast, no learning takes place when voting on the predictions of multiple classifiers. Voting is typically used as a baseline against which the performance of stacking is compared.

In this paper, a stacking framework is then introduced that combines a wide range of base-level sentiment classifiers at the meta-level. In the framework, only the meta-level data set consists of feature vectors that are constructed by the predictions of the base-level classifiers.

The remainder of this article is structured as follows: Section 2 introduces the stacking framework for sentiment classification and the opinion summary of a review. Section 3 describes the experimental design and the results obtained by stacking, and compares all classifiers both base-level and meta-level. Section 4 presents our conclusions, and discusses the results obtained in the experiments and potential extensions.

2 Methodology

2.1 Motivation

In many reviews we observe that there are some summary sentences expressing the author's overall opinion in the beginning or the end. For example, in a review of movie review dataset [3], the last two sentence is "*" lumumba " is a solid , interesting , educational and honest docudrama that should appeal to film buffs and politicos , both . it has more intelligence in its telling than anything i've seen out of hollywood for months and i give it a b+ .*". The two sentences conclusively tell us that the author's attitude is positive. And most of the other 32 sentences describes the story of the movie, and are objectivity expression. We called the last two sentences "opinion summary" of the review. Comparatively, the opinion summary sentences have more strong sentiments and are more easily classified. Intuitionally, the characteristic could be utilized to improve the performance of sentiment classification.

2.2 Overview

We have experimented with five base-level algorithms, which results on a public corpus. The five algorithms are Naive Bayes (NB for short) [11], [12], centroid-based classification (CB for short) [13], k-nearest neighbors algorithms (KNN for short) [14], maximum entropy model (ME for short) and support vector machines (SVM for short). At the same time, we investigated the performances of six algorithms as meta-level classifiers respectively. For our experiments, we used a publicly available implementation[1] of maximum entropy model with all parameters set to their default values except the iteration number, which we set to 10. For the algorithm of SVM, we adopted SVM-light[2] package for training and testing with all parameters set to their default values. As to KNN, we set k to 1 for simplicity.

Overall, the results suggest that stacking can improve the performance of the base-level classifiers, and that using the summary can improve the performance further.

As stated in the title, we constructed the meta-level classifier by the combined five simple base-level algorithms. By stacking we generated a metal-level classifier by making simple changes in the value of their output. Figure 1 presents the flowchart for the methodology.

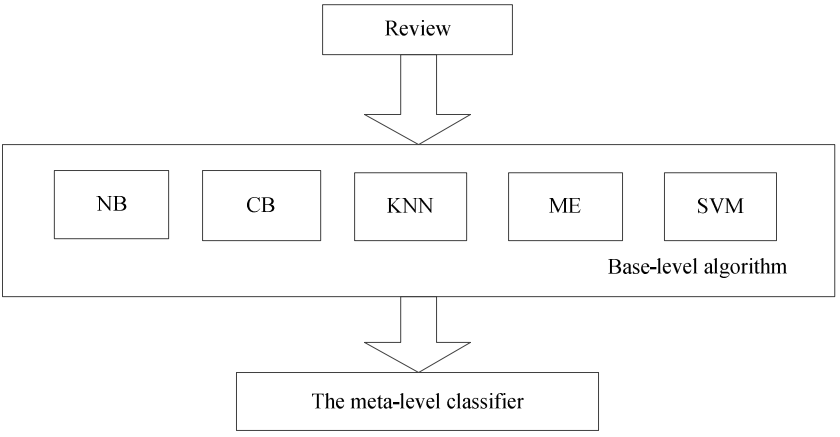


Fig. 1. The schematic of the algorithms and system design

2.3 Base-Level Classification Algorithms

Each of our five classification algorithms relies on a different approach to review interpretation. Each approach is intuitive. They are all analytical, and do not require any lengthy optimization or convergence issues.

[1] `http://homepages.inf.ed.ac.uk/lzhang10/maxent.html`
[2] version 6.0.1 `http://svmlight.joachims.org/`

In the experiment, we focused on three types of features based on unigrams[3], bigrams[4] and words. For Chinese segmentation, we adopted the "smartcn" analyzer for Chinese in the lucene packages[5]. Besides, we used the all of these three sets of features as a new set of features. Therefore, in the experiment four types of features were used.

As the weight of features has the most significant impact on performance of classification, we used three kinds of weight measures, listed below.

- PRESENCE, the feature value is set to 1 if the feature is contained, otherwise to 0.
- TF, the feature value is set to the frequency of the feature.
- TF-IDF, the famous weight formula in information retrieval.

For the particularity of the algorithm of naive bayes, the weight TF-IDF can not be fit for the NB classifier, but the performances of experiments using the TF-IDF were comparable to the other weight measures dramatically. So we employed it at base-level.

To model the potentially important contextual effect of negation, we added the tag n_ to every word in the sentence which has odd negation words. This approach to negation is used by other researchers [3], [4], [15]. For this, we developed a little utility tool to detect the number of negative words in the Chinese sentence. The experiments indicate that this approach to negation is slightly effective on performance for Chinese reviews.

As mentioned above, we used four kinds of features, three types of weight measures and two strategies for negation. So for each algorithm of classification we constructed 24 classifiers. In a word, we could train 120 classifiers for five types of classification algorithms.

2.4 Stacking

Algorithm at Meta-level

At meta-level, we employed the six algorithms, implemented in the WEKA data mining platform: J48, NaiveBayes, IB1, SMO and LogitBoost [16], [17]. Most of these algorithms have already been evaluated as meta-level classifiers in recent studies for stacking [18].

To construct the meta-leval classifier, we used a 4-fold cross-validation. We split the data set into four sets. Label such a set of partitions as L_i (i = 1, 2, 3, 4). Table 1 presents an algorithmic description of the stacking framework with the use of pseudo code.

[3] ChineseAnalyzer used in lucene 3.0.1.
[4] CJKAnalyzer used in lucene 3.0.1.
[5] `http://lucene.apache.org/` version 3.0.1.

Table 1. The stacking procedure used 4-fold cross-validation

```
For each set L_i
Reserve L_i for test of the base-level classifiers
  For the left three sets
    Reserve set L_j for the training of the meta-level classifier
    Train the base-level classifiers on the left two sets
    Test the base-level classifiers on set L_j
    Train the meta-level classifier on the predictions of the base-level classifiers on set L_j
  End for
  Retrain the base-level classifiers on the data sets except L_i
  Test the base-level classifiers on set L_i
  Test the meta-level classifier on the predictions of the base-level classifiers on set L_i
End for
```

The key difference between the stacking and common stacking is that only the predictions of the base-level classifiers are used at the meta-level. Thus any other classifier of sentiment can be integrated into the base-level easily.

Features at Meta-level

We generated the meta-level classifier by making simple changes in the value of the base-level classifiers' output. We added two features into vector at meta-level for the output of each classifier at base-level, which stood for positive confidence score and negative confidence score. Thus each vector at meta-level has 240 features since there are 120 classifiers at base-level.

For SVM and KNN, their classifiers output only one value for a review. When the value is greater than zero, we set positive confidence score to the value, and set negative confidence score to zero, and vice versa. For ME, CB and NB, their classifiers can output the probabilities, in which a review is positive or negative.

Opinion Summary of a Review

In general, an overall opinion is often expressed in the first or the last two sentences in the reviews. These sentences often have strong sentiments and are classified easily, and the other sentences often describe the detail about the attitude. Intuitionally, the characteristic could be utilized. So we extracted the first and the last two sentences at most in a review. Thus the no more than four sentences formed a new "document", which we called opinion summary of the review. Comparatively, the opinion

summary is shorter than the review, but has stronger sentiment and is more easily classified relatively. Note that these sentences can not be guaranteed to be complete. Our sentence splitter tool is simple, which just splits the text by some punctuation marks.

To evaluate the effect of opinion summary we proposed, the predictions of the base-level classifiers on these opinion summaries were used to extend the vectors at meta-level.

3 Experiments

3.1 Corpus

For our experiments, we chose three domains of Chinese reviews, which are about book, hotel and notebook computer[6]. They were all downloaded from web. Each domain has 2000 positive sentiment and 2000 negative sentiment documents. Then we divided every domain documents into four equal-sized folds, maintaining balanced class distributions in each fold. All results reported below are the average four-fold cross-validation results on each category.

Most of reviews are very short. Even some reviews just have only one sentence. As shown in Table 2, every review contains about 200 characters on average, which makes it more difficult for sentiment classification.

Table 2. The length of reviews, in character

Domains	Max	Min	Average
Book	118868	26	191
Hotel	122890	16	211
Notebook	113712	39	129

3.2 Diversity among the Algorithms at Base-Level

Since there are 24 classifiers for each algorithm, Table 2 shows the maximum and the minimum accuracies obtained by the five algorithms in three domains. Most of the five algorithms have already been evaluated in recent studies [3], [4]. But in these experiments, the same algorithms with different features and different weight measures in different domains have performed very differently as shown in Table 3.

Generally, the ME and SVM algorithms had better performances. The simple choice is to select the best base-level classifier for each domain. On the other hand, a more desirable approach is to try to exploit the diversity in the outputs of all classifiers, hoping to improve the results.

[6] The corpus are provided by professor songbo Tan in the Institute of Computing Technology, Chinese Academy of Sciences.

Table 3. The max and min average 4-fold cross-validation accuracies of different algorithms in three domains, in percent

	Book		Hotel		Notebook	
	max	min	max	min	max	min
NB	**96.00**	84.00	86.8	79.8	91.30	83.60
CB	90.40	70.70	85.10	76.50	90.20	80.80
KNN	90.90	79.30	75.60	65.40	83.70	74.40
ME	95.30	82.60	**87.80**	**83.40**	**93.00**	86.40
SVM	93.80	**89.90**	87.70	82.90	92.10	**89.20**

3.3 The Effect of Negation

According to the experiments, we found the effect of negation were not stable. Usually, adding the negation tag could improve the accuracies slightly, but not always. Table 4 show the performances of ME model on the domain of hotel. Even in some cases, the negation was bad for classification, such as the accuracies in bold as shown Table 4, because adding negation tag caused more features or more noise data. However the results may be due to that our treatment for negation is too simple, and we should employ more techniques to detect the scope of negation [19].

Table 4. The accuracies of ME in the domain of hotel with different features, weight measures and the modeling of negation, in percent

	Features	PRESENCE	TF	TF-IDF
ME	unigrams	84.4	83.4	84.6
ME+negation	unigrams	85.6	85.6	85.1
ME	bigrams	86.4	86.3	87.2
ME+negation	bigrams	86.4	86.3	**87.0**
ME	words	85.9	86.3	86.6
ME+negation	words	**85.1**	86.8	**86.3**
ME	all	86.7	86.5	87.8
ME+negation	all	87.3	87.3	87.6

3.4 Stacking vs Voting

Besides stacking, additional classifier combinations are performed through a simple majority voting, where we count the predicted class by the base-classifiers and select the class with the highest count. In the case of a tie, a random selection is typically performed.

Table 5 shows the performances of voting, which are very close to the best results obtained among the base-level classifiers. These results suggest that voting is an effective combination of classifiers for stability.

By comparing voting against stacking, we observe that stacking outperforms voting in all three domains, although the difference is statistically significant only in the domains of hotel and notebook as shown Table 5.

Table 5. Accuracies obtained by stacking and voting, and the best accuracies obtained at base-level in three domains, in percent

	Book	Hotel	Notebook
Best at base-level	96.0	87.8	93.0
Voting	95.9	87.3	92.6
NaïveBayes	96.1	87.4	93.0
MLR	96.2	88.2	93.8
IB1	96.1	87.5	94.1
SMO	96.3	89.4	94.7
j48	96.8	89.5	94.6
LogitBoost	96.9	90.3	95.0

On the other hand, LogitBoost was particularly effective at meta-level. However the difference among these algorithms was measured as statistically insignificant.

3.5 The Effect of Opinion Summary

As already mentioned in section 2.3, we investigated the effect of opinion summary, which is composed of the first two and last two sentences for each review.

Table 6. Accuracies obtained by stacking with opinion summary in three domains using six algorithms, in percent

	Book	Hotel	Notebook
NaïveBayes	**95.3**	87.8	93.5
MLR	96.5	88.9	94.4
IB1	96.1	87.8	94.2
SMO	97.1	89.7	94.8
j48	97.2	89.8	94.9
LogitBoost	97.1	91.2	95.4

Table 6 shows the accuracies of the six kinds of meta-level algorithms with opinion summaries. By comparing, the classifiers with opinion summaries perform better than the standard ones except NaïveBayes in domain of book. These experiments indicate the effectiveness of opinion summary we proposed.

3.6 Diversity Analysis

Since the success of stacking relies on the disagreement in the output of the base-level classifiers, we were particularly interested in how stacking behaves with respect to the

diversity in the output of the base-level classifiers. Ten diversity measures are surveyed in the literature [20], and there is no generally accepted measure for diversity. The entropy measure E, one of the ten measures, is non-pairwise and varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity. Table 7 shows the average diversities evaluated by E in three domains. Generally more diversity gives more space for improvement, but not absolutely. As shown in Table 7, the opinion summary can increase the diversity slightly.

Table 7. Average diversities evaluated by the entropy measure

entropy measure	Book	Hotel	Notebook
normal	0.190	0.249	0.175
+negation	0.197	**0.246**	0.181
+summary	0.198	0.254	0.186
+both	0.201	0.253	0.191

4 Conclusion

This article employed one of the ensemble learning methods, stacking generalization to sentiment classification, and demonstrated the effectiveness using a variety of different algorithms and domains. The disagreement in the output of the sentiment classifiers that were employed at base-level has been successfully exploited by stacking, leading to higher performance at meta-level. At the same time, the method gave a way to obtain high performance classified corpus automatically for other purpose, such as building of sentiment lexicon. However any other classifier of sentiment can be integrated into the framework easily.

We investigated the performances of five machine learning techniques with different features in three domains. In terms of performance, these algorithms with different features and different weight measures in different domains vary greatly. The diversity gives a space for improving sentiment classification.

The effect of negation on sentiment classification was unstable according to our experiments. As we observed, when the accuracy of the classifier was less then 80%, adding negation tag could often improve the performance, otherwise hardly. This characteristic of negation increases the diversity of the base-level classifiers, which gives more probability for improving performance by stacking.

Finally, we proposed a method to exploit the opinion summary of review, which is composed of the first two and the last two sentences. Usually the opinion summary has more strong sentiment and is more easily classified. The experiments have shown the effectiveness of opinion summary for the sentiment classification.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Nos. 61133012, 61173062, 61202193 and 61070082) and the Major Project of Invitation for Bid of National Social Science Foundation (No. 11&ZD189). We are grateful to Professor Songbo Tan for offering the labeled corpus. Yong Zhang is the corresponding author.

References

1. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity an exploration of features for phrase-level sentiment analysis. Computational Linguistics 35, 399–433 (2009)
2. Dasgupta, S., Ng, V.: Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification. In: Preceeding of ACL 2009, pp. 701–709 (2009)
3. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: Proceeding of EMNLP (2002)
4. Tang, H.F., Tan, S.B., Cheng, X.Q.: A survey on sentiment detection of reviews. Expert Syst. Appl. 36(7), 10760–10773 (2009)
5. Xu, J., Ding, Y.X., Wang, X.L.: Sentiment Classification for Chinese News Using Machine Learning Methods. Journal of Chinese Information Processing 21(6) (2007)
6. Zhang, Y., Ji, D.-H., Su, Y., Sun, C.: Sentiment Analysis for Online Reviews Using an Author-Review-Object Model. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (eds.) AIRS 2011. LNCS, vol. 7097, pp. 362–371. Springer, Heidelberg (2011)
7. Täckström, O., McDonald, R.: Semi-supervised Latent Variable Models for Sentence-level Sentiment Analysis. In: Proceeding of Association for Computational Linguistics, ACL (2011)
8. Mukherjee, A., Liu, B.: Modeling Review Comments. In: Proceedings of ACL 2012, Jeju, Republic of Korea, July 8-14 (2012)
9. Du, W.F., Tan, S.B., Cheng, X.Q., Yun, X.C.: Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In: Proceeding of WSDM 2010, pp. 111–120 (2010)
10. Wolpert, David, H.: Stacked Generalization. Neural Networks 5(2), 241–260 (1992)
11. Lewis, David, D.: Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, Springer, Heidelberg (1998)
12. Domingos, P., Pazzani, M.J.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29(2-3), 103–130 (1997)
13. Han, E.H., Karypis, G.: Principles of Data Mining and Knowledge Discovery. Springer (2000)
14. Pan, J.S., Qiao, Y.L., Sun, S.H.: A fast K nearest neighbors classification algorithm. J. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. E87-A(4), 961–963 (2004)
15. Das, S., Chen, M.: Yahoo! for Amazon: Extracting market sentiment from stock message boards. In: Proceeding of the 8th Asia Pacific Finance Association Annual Conference (2001)
16. Sigletos, G., Paliouras, G., Spyropoulos, C.D., Hatzopoulos, M.: Combining Information Extraction Systems Using Voting and Stacked Generalization. Journal of Machine Learning Research 6, 1751–1782 (2005)
17. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (2000)
18. Ting, K., Witten, M.: Issues in stacked generalization. Journal of Artificial Intelligence Research (JAIR) 10, 271–289 (1999)
19. Jia, L.F., Yu, C., Meng, W.Y.: The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management, pp. 1827–1830 (2009)
20. Kuncheva, L.I., Whitaker, C.J.: Measures of Diversity in Classifier Ensembles and their Relationship with the Ensemble Accuracy. Machine Learning 51, 181–207 (2003)

Social Relation Extraction Based on Chinese Wikipedia Articles

Maofu Liu, Yu Xiao, Chunwei Lei, and Xin Zhou

College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
e_mfliu@163.com, fish1953@yeah.net

Abstract. Our work in this paper pays more attention to information extraction about social relations from Chinese Wikipedia articles and construction of social relation network. After obtaining the Chinese Wikipedia articles according to the provided person name, locating the relationship description sentences in the Chinese Wikipedia articles and extracting the social relation information based on the sentence semantic parser, we can construct the social network centered with the provided person name, using the social relation information. The relation set also can be iteratively expanded based on the person names associated with the provided person name in the related Chinese Wikipedia articles.

Keywords: Social Relation Extraction, Chinese Wikipedia Article, Social Relation Network.

1 Introduction

With the rapid development of Internet and its extensive application, the scope of Web is enlarging at an exponential rate. The vast amounts of Web data imply an abundance of information. Many social networks and informative sites are focusing on establishing social relationships so as to take in huge user groups, which contains very important information about social relationships, but it has not been used effectively. We can construct a small social relation network on the base of these Web documents, which helps us to extract the implicit information of social relations within interested domain. The network illustrates relationships among people within the specific domain and their real social relations, and with a further analysis we can easily figure out some deep information, covering family background, living surroundings, vocation, life experience, identity and status, etc. The direct usage of this kind of basic information about one person can help us to obtain a quick outline on him or her from all aspects. At the same time, the implicit value is that it can act as a basic tool for the research on discovering some obvious or less obvious statistical laws, in hope to lead a better life.

We can probably find some unknown social laws or rediscover the well-known laws of society when analyzing these socal data. In the establishment of social networks, the most important process should be the information extraction about social relations, including filteration and purification. The existing social relation

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 94–101, 2013.

extraction systems based on Web documents is mostly dependent on manually screening Web documents, but the normal Web documents have a lot of limitations in quantity and accuracy. Taking the efficiency and accuracy into consideration, this paper designs and implements a social relation extraction system based on the Chinese Wikipedia articles.

Culotta et al [1] applied the probability extraction model to both top-down relational pattern discovery and bottom-up relation extraction in the Wikipedia biography articles. Tnguyen et al [2] extracted relations among entities from Wikipedia's English articles based on the sub-trees mining from the syntactic structure of text. Vo and Ock [3] proposed a feature vector to extract semantic relations using dependency tree and parse tree from infoboxes on Wikipedia documents. Yang et al [4] built and analyzed social networks of person entities on Wikipedia after employing the systematic similarity measure theory to compute the person entity similarity. Massa [5] studied the public conversations in Venetian Wikipedia from the social network analysis perspective in order to highlight the structure of the user talk network. Xu et al [6] focused on the networks of links between these biographical articles starting with the set of biographies in the English Wikipedia and pointed out the most central and culture-related peculiarities.

This paper is intended to use the Chinese Wikipedia articles as the information corpus and study a solution on relation pattern matching and relation extraction based on Chinese Wikipedia articles.

2 System Description

The system basically consist of two modules, the Wikipedia articles collection and social relation extraction. The former is responsible for obtaining articles related to the provided person name from the Chinese Wikipedia, while the latter takes charge of extracting social relation from these articles. The system architecture is shown in Fig. 1.

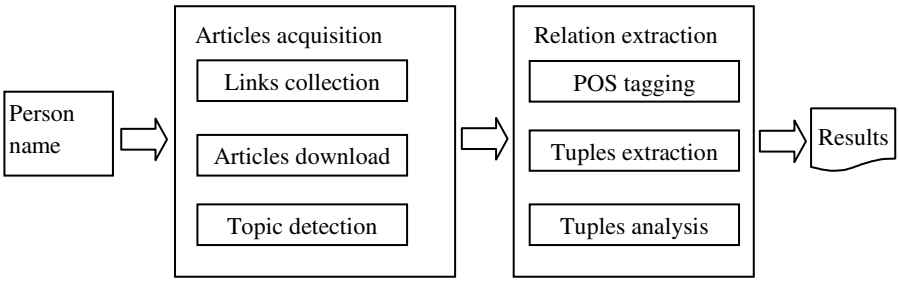


Fig. 1. System architecture

Given the name of a person as starting point, the articles collection module can collect the Chinese Wikipedia articles related to the provided person name. Through the API provided by the Wikipedia, attaching the person name as a keyword and

concatenating a specific query string, we can query a list of the figures related to the provided person name. And then with a regular expression we can get all the related links, through which we access to the targeted articles.

After downloading these articles, we need to extract the text content and filter out the tags in the source file and some other parts that can not make up a sentence, such as lists, tables, and so like. And then using special punctuation, we split the text into sentences on the condition that the contents of each line are of proper length and a simple context, as a result, the subject text per line can be able to express certain information.

As for the social relation extraction module, the segmentation tool ICTCLAS[1], provided by Chinese Academy of Sciences, performs word segmentation and POS tagging, name and relation words recognition on the topic text. Finally, we get the complete, correct and minimized information of social relations.

3 Social Relation Extraction

3.1 Relation Words Extraction

The filtering rules can be used to extract topic text from the related Chinese Wikipedia articles and the main filtering rules expressed by regular expressions are listed in Table 1.

Table 1. Filter rules (partial) on Chinese Wikipedia articles

regex	Functional Description
</(?:li|p|div)>	Each item of the list stands a separation line, as well as the logical block (illustration,paragraph,etc.) to avoid issues such as gathering of person names.
<(\n|.)*?>	Filter tags (<td>/ reserved)
<\s{2,}>	Combine continuous space (mostly generated by the filter)
<(td|li)[^>]*>.{0,15}</\1>	Filter off short content in table and list elements. If the text contains person names, the follow-up filter will cause these names gathered to pose serious problems to the recognization of relationship description sentence.
<(?[| .\] ?.+){2,}|<.+>)>	Filter off other lists without tags （usually with the "|•." seperater） and remaining tags.
<[; 。 ： ？ ！,…]>	special punctuation which splits the text into sentences
\s(?:[\d\-－\?？\s]+[年月 日]){1,6}[\s,]	Filter off date information, which will probably disturb the later process.
(?m)(^.{0,10}[, ：])\n	Replaced with the first sub match, so as to splice parts of one sentence that is cut off by line break.

[1] http://ictclas.org/

In order to collect the relative words, some typical words, such as"爸爸(fatler)"," 老师(teacher)", "妻子(wife)" and "朋友(friend)", are added manually as sample seeds, and the sample seeds cover all categories of relationship, with the most common description word. Therefore, for each sample word in the sample seeds, we need to look into the Tongyici Cilin[7] to find out all words having the similar meaning with the sample word, and apply a word semantic similarity algorithm [8] designed for the Tongyici Cilin. We collect the words or phrases which semantic similarity with current sample word is greater than a thresholding value. The collected words are incorporated into the sample seeds as the expansion. So, with a reasonable similarity threshold, we can get a relatively comprehensive sample seed set.

3.2 Social Relation Information Extraction

The relation sample seed set is added into a user dictionary file as an expansion, and all these words are marked as 'xx@rel', in which 'xx' represents the word and 'rel' is a customized POS tagging. The extraction of linguistic information is always inseparable from the language itself. In order to find out whether a sentence contains relationship description information or not, this paper proposes a solution on the fact that a POS tagging sequence carries sentence patterns information, providing a lot helps to syntactic and semantic analysis.

POS tagging sequence is generated by removing all the Chinese words and punctuations, so that the coding sequence represents a specific sentence pattern, hereinafter uniformly referred to as the "context code". From Example 1, we can get a context code like "/nr /rel /nr /cc /ren /nr /w /rel /w /t /rel /nr ".

Example 1: 三毛/nr 姐姐/rel 陈田心/nr 及/cc 弟弟/rel 陈圣/nr 、/w 陈杰/rel 、/w 生前/t 好友/rel 丁松青/nr

In order to confirm the relationship description information, judging by specific words with special characteristics of POS, we can take '/cc' for the POS tagging in Example 1, which represents a parallel relationship, and apparently the word '及(Ji)' is the same part of speech of the conjunction '和(He)' and '与(Yu)', so these words can mostly express the same meaning. The relative position of conjunctions, prepositions and auxiliary, can be treated with special concern. By comparing the context code of the target sentence with each of those sample codes gathered by manual or automatic process, the massive corpus of sentences will be classified reasonably.

The relation information extraction process is a conversion from the Chinese language to expected context code. And then we need to build up a system that can automatically forecast the relationship matching regex, on the basis of both the sentence and corresponding context code. One context code is logically treated as a specific syntax, and thus the massive corpus is easily classified into diffrent groups. The matching rules and prediction method are continuously improved by iterative feedback, until a complete relation matching repository is acquired ultimately. The specific process is shown in Fig. 2.

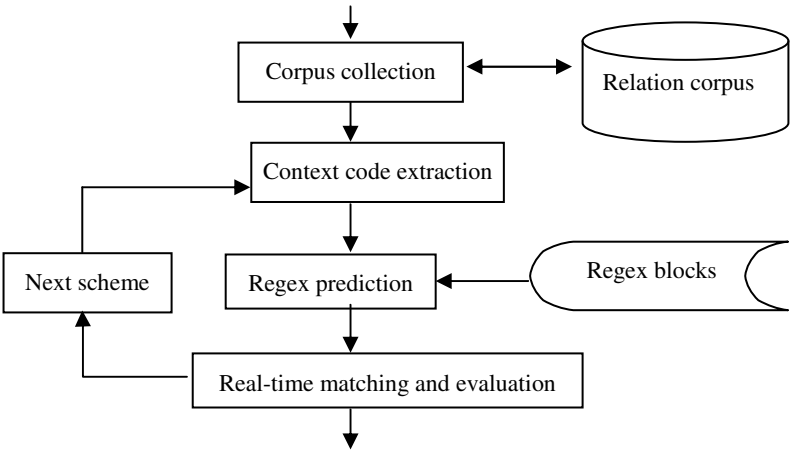


Fig. 2. The relationship matching regex forecast system

The regex template is constructed by sub-block splicing, according to the characteristics of context code. And in turn, to constitute a comprehensive regex, we conduct a matching test upon the relation description corpus, with immediate matching results to evaluate its matching ability. The regex template can be improved if necessary and is shown in Example 2.

Example 2:
```
\[([•\u4e00-\u9fa5]*)/nr\] (?:/p/m/q)
\[([\u4e00-\u9fa5]{1,5})\]
```

In Example 2, the regex consists of four parts, i.e. person matching, feature sequence, relationship word matching and fragment filling. The feature sequence will match a fixed part of syntax, while the fragment filling is a definite word that matches the word itself in a sentence. Both of the two parts mentioned above aim at describing a unique syntax. According to the characteristics of the context code, the system tries out by combining these regular blocks in diffrent ways and showing matching results instantly, in a predefined order.

4 Experiment Results and Analysis

Our system obtains a total of 118 relationship description sentences from the Wikipedia articles introducing "Chen duxiu" and "Cao cao" to Chinese Wikipedia. The mannual evaluation picks up 31 of them, and the accuracy rate is up to 73.7%. During the mannual evaluation, we find that the process to split text into sentences has a certain influence on the number of relationship clause. Many sentences carry more than one relationship description fragment and the real natural language text of the subject from the different articles may also contain partial repetition of the description, so the number of the relationships is far more than the number of sentences in the evaluation. One screenshot of the provided person names, "Chen duxiu" and "Cao cao", is shown in Fig. 3.

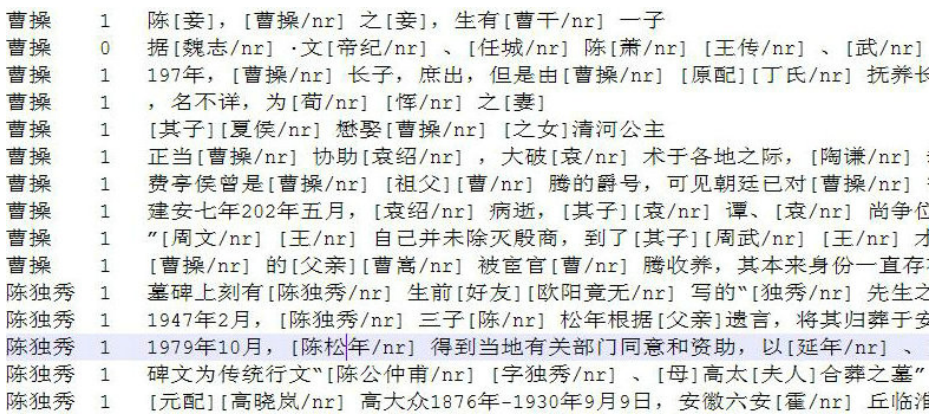

曹操 1 陈[妾]，[曹操/nr] 之[妾]，生有[曹干/nr] 一子
曹操 0 据[魏志/nr] ·文[帝纪/nr] 、[任城/nr] 陈[萧/nr] [王传/nr] 、[武/nr]
曹操 1 197年，[曹操/nr] 长子，庶出，但是由[曹操/nr] [原配][丁氏/nr] 抚养长
曹操 1 ，名不详，为[荀/nr] [恽/nr] 之[妻]
曹操 1 [其子][夏侯/nr] 懋娶[曹操/nr] [之女]清河公主
曹操 1 正当[曹操/nr] 协助[袁绍/nr] ，大破[袁/nr] 术于各地之际，[陶谦/nr]
曹操 1 费亭侯曾是[曹操/nr] [祖父][曹/nr] 腾的爵号，可见朝廷已对[曹操/nr]
曹操 1 建安七年202年五月，[袁绍/nr] 病逝，[其子][袁/nr] 谭、[袁/nr] 尚争位
曹操 1 "[周文/nr] [王/nr] 自己并未除灭殷商，到了[其子][周武/nr] [王/nr] 才
曹操 1 [曹操/nr] 的[父亲][曹嵩/nr] 被宦官[曹/nr] 腾收养，其本来身份一直存
陈独秀 1 墓碑上刻有[陈独秀/nr] 生前[好友][欧阳竟无/nr] 写的"[独秀/nr] 先生之
陈独秀 1 1947年2月，[陈独秀/nr] 三子[陈/nr] 松年根据[父亲]遗言，将其归葬于安
陈独秀 1 1979年10月，[陈松年/nr] 得到当地有关部门同意和资助，以[延年/nr] 、
陈独秀 1 碑文为传统行文"[陈公仲甫/nr] [字独秀/nr] 、[母]高太[夫人]合葬之墓"
陈独秀 1 [元配][高晓岚/nr] 高大众1876年-1930年9月9日，安徽六安[霍/nr] 丘临淮

Fig. 3. The accuracy assessment of relationship description sentence

In experimental evaluation, we select more than 200 singers, poets and historical figures from the list of Baidu MP3 names, ancient poetry and Chinese Wikipedia. And then more than 700 person names have been chosen as samples for content retrieval and filtering, and ultimately we find 2640 relationship description sentences, but experiments have shown that these sentences are related to only nearly 200 of the names, with about 400 articles involved. The availability of topic text about the given person name is dependent on Wikipedia articles, and it has a great relevance with the visibility of the name itself.

Experiments show that the matching count does not rise linearly with the number increase of regex due to the fact that the amount of information about relationship is limited in many articles. But the increase will become visibly only when the corpus is abundant enough, as we will see in later experiments. In this case, it is better to reflect the matching results with a separating investigation of a few regex. The experiments, using three regular expressions for the 2640 sentences, are shown as follows.

Regex 1：`\[([•\u4e00-\u9fa5]*)/nr\][^\[]*的\[([\u4e00-\u9fa5]{1,5})\]\[([•\u4e00-\u9fa5]*)/nr\]`

Regex 2：`\[([•\u4e00-\u9fa5]*)\/nr\]之\[([\u4e00-\u9fa5]{1,5})\]\[([•\u4e00-\u9fa5]*)/nr\]`

Regex 3：`\[([•\u4e00-\u9fa5]*)\/nr\][^\[]*/t\[([\u4e00-\u9fa5]{1,5})\]\[([•\u4e00-\u9fa5]*)\/nr\]`

Owing to limited regex templates and a lack of co-reference resolution and extra consideration of the particular syntax, 73 pairs of relation components have been discovered.

We can see the specific type of relationship directly from matching results, but issues, such as the direction of relationship and the reference of personal pronoun, still exist. The first issue is easy to solve by just adding extra column in the data table because the direction of the relationship in a specific syntax is basically fixed. The second issue is relatively harder, but it works out well by searching named entity in

the preceding context. Generally, additional information must be transferred to the client browser so as to complete the visualization of the network of relationships.

By manual review, 13 of the 73 pairs do not express the same information as the corresponding sentence due to a name recognition error, but all 73 pairs of relationship recognition are correct, and the accuracy of relationship recognition rate is up to 100%, while the accuracy of relationship extraction rate is up to 82.2%.

5 Conclusions

This paper put forward an approach to designing and building a social relationship extraction system based on Chinese Wikipedia articles. The system conducted word segmentation and POS tagging using the ICTCLAS tool to the Chinese Wikipedia articles, and then acquired regex templates out of a sub-system by discriminating and analyzing the statistics of context code. Ultimately, the system produced a tuple in the form of vector for each of the relationship description sentence in the massive corpus.

Experiments have obtained valuable templates for relationship matching, and the follow-up work will focus on the improvement of relation extraction accuracy and the enhancement of abilities of long sentence processing. One of the vital point is to improve the accuracy of word segmentation and POS tagging. At the same time, we will try to visualize the relationship by generating graph with Canvas or in a SVG format.

Acknowledgements. The work presented in this paper is supported by the National Natural Science Foundation of China (No. 61100133), the Major Projects of National Social Science Foundation of China (No. 11&ZD189), the Major Projects of Social Science Foundation of Department of Education of Hubei Province of China (No. 2011jyte126) and the University Students Science and Technology Innovation Foundation of Wuhan University of Science and Technology (No. 11ZRA101).

References

1. Culotta, A., McCallum, A., Betz, J.: Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text. In: Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, USA, pp. 296–303 (2006)
2. Nguyen, D.P.T., Matsuo, Y., Ishizuka, M.: Relation Extraction from Wikipedia Using Subtree Mining. In: 22nd Conference on the Advancement of Artificial Intelligence, Vancouver, B.C., Canada, pp. 1414–1420 (2007)
3. Vo, D.-T., Ock, C.-Y.: Extraction of Semantic Relation Based on Feature Vector from Wikipedia. In: Anthony, P., Ishizuka, M., Lukose, D. (eds.) PRICAI 2012. LNCS, vol. 7458, pp. 814–819. Springer, Heidelberg (2012)
4. Yang, F.F., Xu, Z.M., Li, S., Xu, Z.K.: Social Network Mining Based on Wikipedia. In: International Conference on Asian Language Processing, Harbin, China, pp. 223–226 (2010)

5. Massa, P.: Social Networks of Wikipedia. In: 22nd ACM Conference on Hypertext and Hypermedia, New York, USA, pp. 221–230 (2011)
6. Aragon, P., Kaltenbrunner, A., Laniado, D., Volkovich, Y.: Biographical Social Networks on Wikipedia. In: 8th International Symposium on Wikis and Open Collaboration, Linz, Austria, pp. 1–4 (2012)
7. Mei, J.J., Zhu, Y.M., Gao, Y.Q.: Tongyici Cilin. Shanghai Lexicographical Publishing House, Shanghai (1983)
8. Tian, J.L., Zhao, W.: Synonym for the Word Forest-Based Word Similarity Calculation Method. Journal of Jilin University (Information Science) 28(6), 602–608 (2010)

Event Recognition Based on Co-occurrence Concept Analysis

Yi Zheng[1], Shi Ying[2], and Yibing Wang[3]

[1] School of Information Management, Wuhan University, Wuhan, 430072, China
zhengyijenny@gmail.com
[2] The State Key Laboratory of Software Engineering, Wuhan University, Wuhan, 430072, China
yingshi@whu.edu.cn
[3] Third Faculty, Second Artillery Command College, Wuhan, 430012, China
cjt422@163.com

Abstract. This paper proposes a kind of event recognition technique on news events which analyzes the words in the news documents using co-occurrence analysis for mining the key meta-event, and realizes more pervasive event recognition by leveraging on Markov chain. This approach overcomes the over-sensitivity problems in the traditional event recognition domain based on the words, and it can adapt to intelligent recognition of news event in different domains.

Keywords: Meta-event, Co-occurrence analysis, News event recognition, Markov chain.

1 Introduction

On February 2011, by analyzing the data from 1986 to 2007, the University of Southern California in America announced a research result: till the end of 2007, the amount of information stored, communicated, and computed by human society is about 295EB; In the same month, the American Journal "Science" reported that the existing amount of information in the world by all various of channels has increased by 4 times than the figure in 2007, almost 1300EB, the polymorphism information content has reached the 2 times of the storage capacity and the total information redundancy has also approached 75%. In this era of information explosion, traditional techniques of information retrieval based on the key words have been more and more difficult to meet the requirements of the existing information positioning. Although the information space has been making rapid explosion, the human being has transferred their information requirements from single messages to those semantic information related to their interests, jobs, and majors. In this situation, the mass of information needs to be filtered and disposed effectively in order to achieve users' satisfaction. The semantic knowledge mainly includes "object", "scene", "event" and other key elements, in which the most important element is the event semantics. Especially in the news domain, the event semantics is the core of news reporting. In

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 102–109, 2013.

this paper, we will put forward a kind of technique about event recognition and especially apply it to the news domain.

There is no unified concept for the "event", for the different application domain has different cognitions, but we can say that the way of regarding the "event" as the basic unit of knowledge representation and the means of information organization has become more and more popular and valued. The "event" could describe the knowledge that is bigger, more dynamic, more structured, and having more integrated meanings than the granularity of "concept", and so in the news event processing, researchers try to transform non-structured news data into structured or semi-structured data from various sources so as to facilitate the applications on similar news filtration, news-levels classification, and propelling movement of the intelligent news, and gain news knowledge more quickly and orderly for users.

At present, the news information processing based on "event" includes three aspects: event-oriented corpus construction; event recognition and event relationship mining; and event application and so on. If we regard "event" as a basic text unit containing news elements, a news article is normally constituted by more than one event, namely Latent Ingredients. The event recognition aims to mine out these Latent Ingredients [1].

In this paper, we mainly investigate the news event recognition issue, and propose a method on news Event Recognition based on Co-occurrence Analysis (ERCA) which forms meta-events by extracting the core concept in news reporting, mine out the sequence relationships between meta-events, and compute the most stable distribution by Markov chain, and finally infer the probability of specific news events.

2 Related Work

Two novel event recognition models based on Latent Ingredients extraction are presented in [1] and it exploits a set of useful features, either semantic or syntactical, consisting of context similarity, distance restriction, entity influence from thesaurus and temporal proximity. Temporal information is proved to be quite essential in Latent Ingredients extraction task and generally the mixture of all four features brings the best performance in balance. But this work did not investigate further on temporal information analysis.

[2] proposes an integrated technical framework on emergency-event recognition and presents the key problems and solving strategies of different constituent parts of system; Meanwhile, it combines the text contents and structure characteristics of news reporting and the distribution characteristics of reporting sources, and proposes an improved method for text-cutting and a model for feature weight computation.

Based on the idea of the distance between the words, a model for emergency-event recognition in [3] is constructed based on Internet news reporting. This model is mainly composed of two parts: the finding of hot lemma and detection of new words, and namely by improving TF-IDF algorithm to catch the current lemma to be concerned for forming the hot lemma, and by using the distance between the words to search objective distribution state between hot lemma. Thereby it realizes emergency-event recognition

based on the combination of relative stability between hot lemma. However, the shortage of this model is that it is quite sensitive to time property.

[4] indicates that the amount variation of emergency-event news reporting reflects the emergency-event trend and the reflection change of the media and public to the events, and it is also the important source of emergency management decision-making information. This work constructs a model to the explosion feature of emergency-event news reporting based on HMM (Hidden Markov Model) in order to reflect the trend of the amount of emergency-event news reporting. In addition, it also proposes a method of time-sequence aggregation algorithm to identify the evolution pattern of the amount of emergency-event news reporting.

An HMM-based method for passage extraction which can naturally exploit the coherence in the text to accurately identify coherent relevant passages of variable lengths is put forward in [5]. This result will be valuable to event recognition but this paper does not make further research for this objective.

Some work addresses the news event recognition by extracting the topic sentence for critical event information in [6]. This method is to extract topic sentence in single text for gaining key events information to solve the problem of news event recognition. According to the characteristics of news, it analyses the relationship between news reporting and events and the features of news titles by their contents, forms and languages, and proposes the method by using the information of title tips to retrieve topic sentence for describing news key events.

[7] describes a system which recognizes events in news stories and extracts knowledge using an ontology, which classifies stories and populates a hand-crafted ontology with new instances of classes defined in it. In each case, the system provides a confidence value associated to the suggested classification by using information extraction and machine learning technologies.

Furthermore, in algorithm, the [8] presents a partitioning algorithm for recursively computing the steady state probability for a finite, irreducible Markov chain or a Markov process. The algorithm contains a matrix reduction routine, followed by a vector enlargement routine which computes the components of the steady state probability vector by starting with the smallest reduced matrix and working sequentially toward the original transition matrix.

In summary, we can make a conclusion on the main problems of existing research on news event recognition in the following aspects:

(1) The method based on the association analysis of the words can be applied pervasively, but this method has a low accuracy rate;

(2) Comparatively, although the method based on events model has a high accuracy rate, it cannot accommodate the situation of crossing domains;

(3) When dealing with a large amount of news texts, the existing methods have a low processing efficiency.

To address above problems, this paper proposes a kind of news event recognition method, by co-occurrence analysis which introduces the concept of meta-event and further constructs the meta-event cluster by co-occurrence probability model which is appropriate for various domains, and then realizes more pervasive event recognition by Markov chain.

3 Creating Meta-event by Co-occurrence Analysis

The meaning of a word is the core of a word which reflects the understanding of people to the outside world. At the first time, a word only represents a concept. With the understanding development of people to the objective world, a single word contains more concepts than only one, namely now a single word usually has the ambiguity in understanding.

The ambiguity of word makes the precision rate low if mining and identifying news event only by the space-time relationship. To solve this issue, we introduce the concept of "meta-event" in this work. In detail, the "meta-event" is the basic unit composing a news event that is made up of several co-related words. When we put these words combined together, it will make each word have a certain context. As a result, this method can avoid the disturbance in news event recognition due to the words ambiguity. Let's set a sample:

The word "Apple" simultaneously appears in two different news texts, and we could not know the specific meaning only by this single word, for it could be interpreted as a kind of fruit or mobile telephone or computer with "Apple"-brand and etc... However, if in a certain news text, the word "Apple" appears simultaneously with the other words such as "Zhongguancun" and "Mouse", and then we can confirm that the word "Apple" means a kind of electronic products other than a kind of fruit. In other words, by the context the concepts of "Apple"、"Zhongguancun" and "Mouse" restrict the ambiguity of the word "Apple" and constitute the concept muster which has clear semantic, that is "meta-event", which avoids the confusion and interference of news event recognition caused by words ambiguity.

As to above-mentioned the idea, in this paper we propose the extraction and generation methods of meta-event based on co-occurrence analysis. The co-occurrence analysis is a kind of analysis method qualifying the co-occurrence information in all kinds of information carrier, which are the important means and tools supporting knowledge mining and knowledge service.

Different from the classical TF-IDF method which assigns the key words different weight, the proposed method treats each key word equally. In a news text, all key words construct an undirected complete graph, and the co-occurrence times between each two words are noted as "1". After analyzing all the texts in the documents collection, the key words that are closely related will have larger co-occurrence times, namely the generated meta-event. The calculation procedure is as follows:

(1) Scan sequentially each document "d_i" in news text collection, and find out each key word which presents in keywords collection;

(2) Maintain dynamically an inverted file for keywords which has a unified number. If the inverted list of keyword k_j does not exist in the inverted list, a corresponding structure $list_j$ is created, and d_i is added into the corresponding list for each keyword;

(3) Exit after finishing the scanning of file set.

After processing all the documents, we could employ the following methods to calculate the co-occurrence times between any two words.

$$CO_{ij} = |list_i \bigcap list_j| \tag{1}$$

Due to the fact that only text lists of two words are needed in calculation, getting the intersection between them can result in co-occurrence times, which can complete co-occurrence matrix analysis of random key words collection by setting up inverted index.

The co-occurrence times between the words represent strong or weak relationship between the words to a certain extent, and this paper regards the words collection whose relationship intensity is greater than the features threshold as a meta-event. However, the greater co-occurrence times does not mean the stronger incidence relation, so a formula of balancing the similarity between words is needed to optimize co-occurrence network. In this work, we employ a words similarity measurement method combing the important information of nodes and geometry similarity. The formula is presented as follows:

$$\begin{aligned} Sim(A,B) &= \sqrt{\max(D(A),D(B))} * \\ &\left|list_A \bigcap list_B\right| * \left(\frac{1}{|list_A|} + \frac{1}{|list_B|}\right) \end{aligned} \tag{2}$$

In this formula, D(A) and D(B) denote the degree of a vertex of panel points A and B in a co-words network; $list_A$ and $list_B$ denote the corresponding text list of A and B. The three parts of the above formula from the left to the right are described as follows:

(1) By considering the information of node degree, making it as one of the factors that balancing nodes importance, the bigger the vertex degrees are, and the bigger the weights of the mid-side nodes are;

(2) The size of two nodes texts list intersection is obviously an important element to balance the edge importance degree between the nodes;

(3) Considering the size of nodes texts list, the calculation of reciprocal sum can avoid the less-similarity weakness when computing big collection and small collection using traditional method of "Jaccard".

4 News Event Recognition Based on Markov Chain

"Meta-event" is the basic unit composing a news event that is made up of several co-related words, whose feature is to hide the ambiguity of each word using the context relationship of words so as to get a higher semantic precision than the traditional methods that only consider the words. This paper argues that a news event is composed of several meta-events in the sequence of a certain timing relationship. If we regard each meta-event as a kind of state, the news event will be evolved into a series of state-transferring results.

This paper realizes the modeling and recognition of news event semantics by introducing the idea of Markov chain. Let the occurrence probability of meta-event E^t is $\pi_t(E^t)$ in the time point of t, thereby the occurrence probability of meta-event E^{t+1} is $\pi_{t+1}(E^{t+1})$ in the time point of t+1. If given the $\pi_t(E^t)$, compute $\pi_{t+1}(E^{t+1})$ by product

summation between the occurrence probability of current meta-event and the probability $p(E^t \to E^{t+1})$ of the current meta-event E^t transferring to the next event unit E^{t+1}:

$$\pi_{t+1}(E^{t+1}) = \sum_{E^t} \pi_t(E^t) p(E^t \to E^{t+1}) \tag{3}$$

When $\pi_t=\pi_{t+1}$, we can tell that the Markov chain satisfies the stable-state distribution, namely at this time point, news event semantics has been composed by the most stable meta-event in the sequence of a certain timing relationship, and consequently the recognition problem of news event semantics is transformed into the stable-state distribution problem using Markov chain which is described in Formula (3). By introducing the algorithm "Sheskin" in [8], we repeatedly partition the corresponding state space of meta-event E= $\{E_1, E_2,\ldots, E_N\}$, namely computing the probability of stable-state distribution by dimension-reduction analogy method.

Let P= $[p_{i,j}](i,j\in E)$ to be the most simple matrix of the state transferring probability in the state-space E of Markov chain. Firstly, the state-space is decomposed into E= $\{E_1, E_2,\ldots, E_{N-1}\}\cup\{E_N\}$, thus matrix P is decomposed accordingly as presented in Formula (4):

$$P = \begin{pmatrix} T & W \\ R & Q \end{pmatrix} \tag{4}$$

In this formula, T denotes the dimensional matrix of (N-1) × (N-1), W denotes the vertical vector of N-1 dimension, R denotes the horizontal vector of N-1 dimension, and Q represents the invariant $p_{n,n}$. Let π to be the stable-state distribution vector of Markov chain, and then π = πP, thus the random simplest matrix P' of N-1dimension can be defined as:

$$P' = T + W(1-Q)^{-1}R \tag{5}$$

We use π' to denote the stable-state distribution, meanwhile π'=π'P'. If we divide π into two parts by (x, π_N), in these two parameters, x is the horizontal vector of N-1 dimension, thus

$$\pi_N = xW + \pi_N Q \tag{6}$$

The above formula can be changed into the Formula (7):

$$\pi_N = xW(1-Q)^{-1} \tag{7}$$

So π' is proportional to x, x = cπ', among them c denotes the scale factor. Thus we can get the conclusion using Formula (8):

$$c= 1-\pi_N,\ \pi_N = \frac{\pi' W (1-Q)^{-1}}{1+\pi' W (1-Q)^{-1}} \tag{8}$$

Using π', we can calculate the value of x. Once getting P', the velocity R is not needed any more. Starting from Matrix P', with N-2 times of same segmentation

procedure, we can calculate all the stable-state distribution velocity in the original Markov chain; thereby confirm various parameters in news event semantics.

The process above can be called news event semantics modeling process. Using specific news database as the training set, we can get various related parameters of news event semantics to form semantic knowledge, and the news event recognition is the inverse-processing of above procedure.

5 Experiment

We implement the above event recognition algorithm in Java. In detail, we design and conduct an experiment to evaluate the accuracy of the proposed approach and the experiment is conducted on a PC with 2.8 GHz CPU and 3 GB memory, running the Windows 7 operating system.

The experience dataset adapted in this paper is the News in [11], including 2 million news documents and the theme news dataset is composed by theme news gathered from the Internet certainly containing some world news network, such as Natural Gas Explosion、Armour Flow Epidemic and so on. By Word Segmentation, more than 34,000 words can be gained.

Figure 1 shows the precision comparison of news event recognition between the ERCA algorithm presented in this paper and the CRP method in [5].

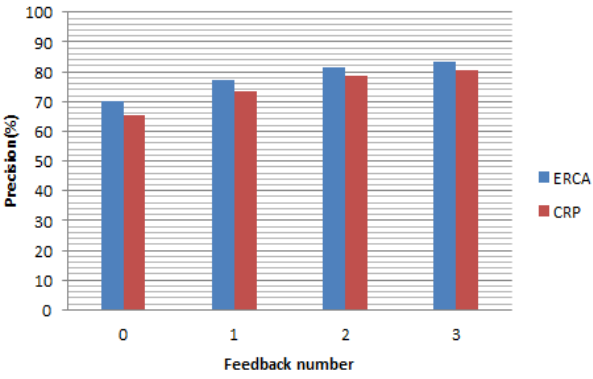


Fig. 1. Precision comparision between MMC and CRP

We can find from the above figure that with the increasing of the feedback time, the accuracy rate of these two kinds of methods are all improved, and even exceeds the accuracy of 80%. Furthermore, the event recognition precision probability is higher than that of the MMC algorithm proposed in this paper in the feedback processing of 0~3 times, and the reason is that we introduce the concept of meta-event by co-occurrence analysis different from the traditional concepts so as to avoid of the errors caused by the ambiguity of concept.

6 Conclusion

The news event recognition is one of hot research topics in recent years, however, the smallest unit of traditional news event recognition processing is words, and the ambiguity of words affects negatively the precision of events recognition. This paper proposes the concept of "meta-event" to mine out stronger-relevancy words collection by words co-occurrence analysis in order to form the meta-event, and finally realizes the modeling and identification of news event semantics by computing the most stable-state distribution of the Markov chain of several meta-events. Compared with traditional methods, the proposed method in this paper has the best performance especially in the pervasive aspect of cross-domain.

Acknowledgement. This paper is supported by Overseas, Hong Kong & Macao Scholars Collaborated Researching Fund (61028003).

References

1. Yan, R., Li, Y., Zhang, Y., Li, X.Y.: Event Recognition from News Webpages through Latent Ingredients Extraction. In: Cheng, P.-J., Kan, M.-Y., Lam, W., Nakov, P. (eds.) AIRS 2010. LNCS, vol. 6458, pp. 490–501. Springer, Heidelberg (2010)
2. Chen, L.P., Du, J.P.: Hot Topics Identification System and the Key Problem Research on Emergency-event. Journal of Computer Engineering and Application 47(32) (2011)
3. Yao, Z.L., Xu, X.: Emergency-event Recognition Research in Internet News Reporting. Journal of Intelligence Analysis and Research 4, 52–57 (2011)
4. Miao, R., Liu, L., Liu, Z.M.: The Explosive Analysis of Emergency-event News Reporting based on Hidden Markov Models. Journal of System Engineering 8, 89–95 (2010)
5. Jiang, J., Zhai, C.: Extraction of Coherent Relevant Passages Using Hidden Markov Models. ACM Transactions on Information Systems 24, 295–319 (2006)
6. Wang, W., Zhao, D.Y., Zhao, W.: Topic Sentence Recognition of Chinese News Key Events. Journal of Peking University (Natural science edition) 47(5), 789–796 (2011)
7. Vargas-Vera, M., Celjuska, D.: Event Recognition on News Stories and Semi-automatic Population of an Ontology. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), Beijing, China, September 20-24. For guidance on citations see FAQs (2004)
8. Sheskin, T.J.: A Markov Chain Partitioning Algorithm for Computing Steady State Probabilities. Journal of Operations Research 33(1), 228–235 (1985)
9. Fu, J.F.: Even-oriented Knowledge Treatment Research. Shanghai University (2010)
10. Sun, H.: The Research and Development on Event Timing Relationships Recognition. Harbin Industrial University (2010)
11. `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`

Atomic Event Semantic Roles and Chinese Instances Analysis

Maofu Liu[1], Yan Li[1], Donghong Ji[2], and Yi Zheng[3]

[1] College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
[2] School of Computer, Wuhan University, Wuhan 430072, China
[3] School of Information Management, Wuhan University, Wuhan 430072, China
e_mfliu@163.com, liyan880923@sina.com,
donghong_ji2000@yahoo.com.cn, zhengyijenny@gmail.com

Abstract. The research on event has been becoming more and more popular in natural language processing and text analysis. The event semantic roles for natural language text corpus should be paid more attention in order to make great progress in the event semantic analysis and computation. Compared with the sentence-level semantic roles, this paper discusses some important issues from the view of atomic event. Besides the subjective and objective semantic roles, this paper also puts emphasis on the space and time event semantic roles in accordance with the nature of event. In order to deal with the embedded event, this paper puts forward the concept of recursive event. After labeling and analyzing some Chinese instances for the natural language text corpus collected from Web, the results show that it is necessary to discuss the problems and their solutions about atomic event semantic roles.

Keywords: Atomic Event, Event Semantic Roles, Recursive Event.

1 Introduction

With the development of the Internet, the Web natural language text data have been rapidly increasing. If the machine can communicate with human beings or other machines, the computer should firstly understand the meaning of natural language text, and therefore natural language semantic analysis and computation is necessary. In recent years, the event study has become a hot spot in natural language processing and applied semantics, and has been playing very important role in a lot of applications, such as automatic summarization and question answering [1-6]. Event is worth exploring to understand the meaning of natural language text based on event semantics.

At present, information processing has been shifting from the fine-grained static units of language, such as word, phrase and so on, to the high-level dynamic units of language, including sentence, paragraph, text and so on. From the view of top-down macro-logical structure of the text, one text is comprised by one or several coherent paragraphs and one paragraph contains one or several coherent sentences. Lv considered that one sentence includes one or several sub-sentences and the sub-sentence is

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 110–121, 2013.

the basic unit of language dynamic semantics [7], but from the computational perspective, atomic event is the most basic information processing unit and starting point of the event semantics.

Table 1. The sub-sentence and atomic event

Sentences
In Simplified Chinese (CS)
S1: 朴文垚首夺世界冠军，荣升中国第30位九段围棋手。
S2: 1814年至1816年英国人从北印度入侵，后签订条约使尼泊尔丧失部分领土。
In English (EN)
S1': Pu Wenyao won the world champion for the first time, promoted to the 30th Kudan go hand in China.
S2': 1814 to 1816 the British invaded from the North Indian, then signed the treaty, Nepal lost part of its territory.
Sub-Sentences
In Simplified Chinese (CS)
C1_1: 朴文垚首夺世界冠军
C1_2: 荣升中国第30位九段围棋手
C2_1: 1814年至1816年英国人从北印度入侵
C2_2: 后签订条约使尼泊尔丧失部分领土
In English (EN)
C1_1': Pu Wenyao won the world champion for the first time
C1_2': Promoted to the 30th Kudan go hand in China
C2_1': 1814 to 1816 the British invaded from the North Indian
C2_2': Then signed the treaty, Nepal lost part of its territory
Atomic Events
In Simplified Chinese (CS)
E1_1_1: 朴文垚首夺世界冠军
E1_2_1: 荣升中国第30位九段围棋手
E2_1_1: 1814年至1816年英国人从北印度入侵
E2_2_1: 后签订条约
E2_2_2: 使尼泊尔丧失部分领土
In English (EN)
E1_1_1': Pu Wenyao won the world champion for the first time
E1_2_1': Promoted to the 30th Kudan go hand in China
E2_1_1': 1814 to 1816 the British invaded from the North Indian
E2_2_1': Then signed the treaty
E2_2_2': Nepal lost part of its territory

In Table 1, the sentence S1 includes two sub-sentences, i.e. C1_1 and C1_2, and two corresponding atomic events, i.e. E1_1_1 and E1_2_1. The relationship type between the sub-sentence and atomic event is one-to-one. However, sentence S2 also contains two sub-sentences, i.e. C2_1 and C2_2, but there are two atomic events, i.e. E2_2_1 and E2_2_2, in the sub-sentence C2_2. The sub-sentence can be divided into one or several atomic events and the relationship type between sub-sentence and atomic event is one-to-many. Therefore, we can conclude that sub-sentence is not the basic unit of

the event semantics, but the atomic event is the basic unit of the dynamic semantic analysis and computation.

For the current semantic labeling, Yuan [8, 9] studied the Chinese argument structure, described the framework from a cognitive perspective, and carried out semantic labeling practices of natural language text. But his study involved too many basic and additional semantic roles and the labeling practices were very complex and difficult for natural language semantic analysis and computation. The number of semantic roles should be condensed from the cognitive event semantics perspective.

This paper aims to determine the number of event-based semantic roles and discusses the two most important types of semantic roles in event semantics, i.e. space and time. Moreover, according to the embedded event, this paper proposes the concept of recursive event. After labeling some instances for the natural language texts, we analyze and discuss the necessity and feasibility to consider these issues in the event semantics.

2 Event-Based Semantic Roles

The sentence-level semantic roles are generally divided into three categories, i.e. subjective, objective and auxiliary semantic roles. The subjective semantic roles mainly include agent, sentient, causer, theme and experiencer, and the objective semantic roles generally contain patient, dative, result, target and relevant. The auxiliary semantic roles are usually comprised by instrument, material, manner, location, source, goal, range, time, reason, aim, logical operators and so on.

At present, sentence-level semantic roles are defined from the perspective of applied linguistics and in the finer level. If the sentence-level semantic roles are directly used to label the event-based semantics, the number of semantic roles is too large. Therefore, we should redefine and analyze the event-based semantic roles from the view of the event semantic analysis and computation based on the sentence-level semantic roles according to the nature of event.

According to the labeling practices of natural language texts, we define event-based semantic roles and the semantic roles for event semantics generally include three categories, i.e. core semantic roles, additional semantic roles and logical operators. Moreover, the core semantic roles can also be divided into three sub-categories, the subjective, objective and space-time semantic roles. The structure of the semantic roles for event semantics is shown in Fig. 1.

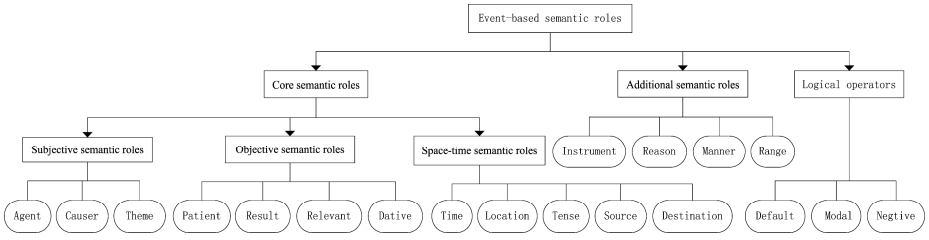


Fig. 1. The structure of semantic roles for event semantics

2.1 Core Semantic Roles

The core semantic roles are the most important parts of the event-based semantic roles and occur frequently in the natural language texts.

Subjective and Objective Semantic Roles. The sentence-level subjective semantic roles generally include agent, sentient, causer, theme and experiencer. The sentient and experiencer are deleted in the event-based subjective semantic roles, so the event-based subjective semantic roles only include agent, causer and theme.

Sentient is the non-autonomous subject of the psychological perception. Generally speaking, psychological verbs and adjectives, which express the sense, dominate the sentient. If we ignore the difference between the agent and the sentient on causative usage, the functions of the agent and the sentient will overlap. Therefore, we delete the sentient, and use the agent to substitute the sentient when the event predicate (EP) is psychological verb, and theme to replace the sentient when the EP is adjective. The following instances in Table 2 are used to illustrate the replacement of the sentient semantic role.

Table 2. Instances of replacement of the sentient

Psychological Verbs
In Simplified Chinese (CS)
E1: [纽约双雄 A][希望 EP][林书豪归来 Re]
In English (EN)
E1': [New York duo A] [hoped EP] [Jeremy Lin returned Re]
Adjectives
In Simplified Chinese (CS)
E2: [我 Th]实在太[困 EP]了
In English (EN)
E2': [I Th] am too [sleepy EP]

The "希望", the EP of event E1 in Table 2, is a psychological verb, so the sentient "纽约双雄" is labeled as the agent. And in event E2 of Table 2, the EP "困" is the adjectives that express the sense, so the sentient "我" is labeled as the theme.

Experiencer is the specific perception subject that shows some changes of the state. The EPs that describe job change are finite and divided into six sub-categories, i.e. appoint, hold, remove, resign, dispatch and transfer. The subjects of the event, which use hold, resign and transfer EPs, can be directly labeled as the agents, and the subjects of the event, which use appoint and remove EPs, can be labeled as the datives. The subjects of the event, which use dispatch EP, can be divided into two cases, the agents or the datives. The following instances in Table 3 are used to illustrate the replacement of the experiencer semantic role.

Table 3. Instances of replacement of the experiencer

Experiencer as the Agent
In Simplified Chinese (CS)
E3: [景俊海 A][辞去 EP][陕西省副省长职务 Re]
In English (EN)
E3': [King Junhai A] [resigned as EP] [vice governor of Shaanxi Province duties Re]
Experiencer as the Dative
In Simplified Chinese (CS)
E4: [撤消 EP][王濛 D][队长职务 Re]
In English (EN)
E4': [Undo EP] [Wang Meng D] [captain duties Re]
Experiencer as Recursive Event
In Simplified Chinese (CS)
E5: [AOL A][任命 EP][Jolie Hunt为首席营销官兼公关官 Re]
E6: [Jolie Hunt Th][为 EP][首席营销官兼公关官 Re]
In English (EN)
E5': [AOL A] [appointment of EP] [Jolie Hunt as chief marketing officer, chief public relations officer Re]
E6': [Jolie Hunt Th] [as EP] [chief marketing officer, chief public relations officer Re]

In event E3 of Table 3, the EP "辞去" is the resign verb, so experiencer "景俊海" should be labeled as the agent. And in event E4, the EP "撤消" is the remove verb, so experiencer "王濛" should be labeled as the dative. The event E5 is the special situation, and there is an EP "为" after the experiencer "Jolie Hunt", but in an event, only one EP can be labeled, so the event should be labeled as the recursive event.

The sentence-level objective semantic roles include patient, dative, result, target and relevant. The target is deleted from the perspective of event, so the event-based objective semantic roles only include patient, dative, result and relevant.

Target semantic role is the target of perceived behavior and often occurs with the sentient semantic role, so it can be labeled as the patient or relevant. In event E1 of Table 2, the objective semantic role "林书豪归来" corresponds to the sentient "纽约双雄", which uses psychological verb as the EP, and the "林书豪归来" is obviously another atomic event because of the EP "归来", so we label it as the relevant. There is a situation that the objective semantic role must be labeled as the patient when the event uses psychological verb as the EP, e.g. "我认识你" (I know you). In this case, the objective semantic role "你" can be labeled as the patient. So the target is often labeled as the patient when the objective semantic role is not recursive event and the event uses psychological verb as the EP. The event E2 of Table 2 uses the adjective as the EP and no objective semantic role in it, so there is no any influence to this kind of the events when we delete the target semantic role.

Experiencer semantic role is deleted with no any affect to the objective semantic role. In event E3 of Table 3, the objective semantic role "陕西省副省长职务" is labeled as the relevant no matter what "景俊海" is labeled as the agent or experiencer. In

event E4, the "队长职务" is labeled as relevant no matter what "王濛" is labeled as the dative or experiencer.

The event-based subjective and objective semantic roles are listed in Table 4.

Table 4. The event-based subjective and objective semantic roles

Subjective	Label	Objective	Label
Agent	A	Patient	P
Causer	Cau	Dative	D
Theme	Th	Result	R
		Relevant	Re

Space-Time Semantic Roles. The event model is accordance with the model of human cognition, in which all things are constantly moving and changing within a specific time and space [14]. Therefore, the space-time semantic roles are very important for event-based analysis and computation. Time semantic roles include the roles which show the clear time information of the event and the ones which show tense. The specific time semantic roles are shown in Table 5.

Table 5. The time semantic roles

Semantic roles	Label	Clue words
Time	T	今天 today, 11月30日 November 30th
Future tense	fut	将 will, 即将 on the horizon, 再 again
Past tense	past	刚刚 just now , 已经 already, 曾经 once
Past perfect tense	past-perf	过 over
Continuous tense	prog	着 zhe, 正 now, 在 at
Perfect tense	perf	了 have
Present perfect tense	pres-perf	来着 laizhe, 来的 laide

The space semantic roles include location, source and destination, which are assigned labels as "L", "Ls" and "Ld" respectively. The instances of space-time semantic roles are listed in table 6.

Table 6. The instances of space-time semantic roles

Time Semantic Roles
In Simplified Chinese (CS)
E7: [黄金周期间 T], [国美、苏宁等家电连锁企业 A][创造 EP][了 perf][较好的销售业绩 P]
In English (EN)
E7': [In Golden Week T], [Gome and Suning appliance chain enterprises A] [have perf] [created EP][a better sales P]

Table 6. *(continued)*

Space Semantic Roles
In Simplified Chinese (CS)
E8: [战俘移交仪式 P][4月19日 T]在[阿富汗最高法院 L][举行 EP]
In English (EN)
E8': [POW handover ceremony P] [held EP] at [Supreme Court of Afghanistan L] on [April 19th T]

In Table 6, the "黄金周期间" is the time when the event E7 happens, and the "了" expresses the meaning that the event E7 has happened, so the "了" is labeled as the perfect tense. In event E8 of Table 6, the "阿富汗最高法院" is the location where the event E8 happens.

2.2 Additional Semantic Roles

The additional semantic roles are the complement to the core semantic roles and listed in Table 7. The instances of additional semantic roles are shown in Table 8.

Table 7. Additional semantic roles

Additional semantic roles	Label	Additional semantic roles	Label
Instrument	I	Manner	M
Reason	Rn	Range	Ra

Table 8. The instances of additional semantic roles

Instrument
In Simplified Chinese (CS)
E9: 以[美国为首的北约 A]使用[五枚导弹 I][袭击 EP][了 perf][中国驻南斯拉夫大使馆 P]
In English (EN)
E9': [The US-led NATO A] using [five missiles I] [struck EP] [the Chinese Embassy in Yugoslavia P]
Reason
In Simplified Chinese (CS)
E10: [陈省身先生 A]因[病 Rn]于[2004年12月3日19时14分 T]在[天津 L][逝世 EP]
In English (EN)
E10': [On 19:14 December 3, 2004 T], [Chen Xingshen A] [died EP] in [Tianjin L] because of [illness Rn]
Manner
In Simplified Chinese (CS)
E11: [恐怖分子或"流氓国家" A][可能 mod]以[购买或偷窃的方法 M][取得 EP][俄罗斯的武器级钚 P]

Table 8. *(continued)*

Manner
In English (EN)
E11': [Terrorists or "rogue states" A] [may mod] use [the methods of buying or stealing M] to [get EP] [the weapons-grade plutonium of Russia P]
Range
In Simplified Chinese (CS)
E12: [李宁 A]一共[获得 EP][了 perf][14个 Ra][世界冠军 P]
In English (EN)
E12': [Li Ning A] [won EP] a total of [14 Ra] [world champions P]

In Table 8, the "五枚导弹" is the instrument of the event E9. In event E10, the "病" is the reason that "Chen Xingshen" died. In event E11, the "购买或偷窃的方法" is the means to get the weapons-grade plutonium of Russia. The range semantic role is the number, frequency or range that occurs in the event, just like "14个" in event E12.

2.3 Logical Operators

Logical operators can change the event meaning or reflect the relationship among the semantic roles within the event. The logical operators are shown in Table 9.

Table 9. Logical operators

Logical operators	Label	Clue words
Negative	neg	不 not, 没 not, 没有 not
Modal	mod	能 can, 能够 can, 可以 can, 会 can, 可能 can

In Table 9, there are two logical operators, negative and modal. The instances of logical operators are listed in Table 10.

Table 10. The instances of logical operators

Negative Operator
In Simplified Chinese (CS)
E13: [小泉纯一郎 A][2001年 T][未 neg][赢得 EP][自民党总裁选战 P]
In English (EN)
E13': [Junichiro Koizumi A] [did not neg] [win EP] [the election campaign of president of the LDP P] [in 2001 T].
Modal Operator
In Simplified Chinese (CS)
E14: [桥本龙太郎 A][应该 mod][辞职 EP]
In English (EN)
E14': [Hashimoto Ryutaro A] [should mod] [resign EP]

In event E13 of Table 10, the "未" expresses the negative meaning and is a negative operator. In event E14, the "应该" is the auxiliary verb that expresses modality, so it is labeled as the modal operator.

3 Recursive Event

The recursive event is the situation that the semantic role in an event acts as another atomic event. In event E5 of Table 3, the whole event is an atomic event, and the "Jolie Hunt为首席营销官兼公关官" is the relevant of the EP "任命". However, the semantic role relevant of E5 is another atomic event, because of the EP "为" in it, and the semantic role relevant is labeled as another atomic event E6. The recursive event makes the structure of the events in the sentence clearly, and the semantic analysis of the sentence becomes easily. So we use the recursive method to label this kind of event.

From our labeling practices, there are four subtypes of recursive event on the natural language text. The first subtype is the whole recursive event acts as one atomic event, and this subtype of atomic event does not have the further recursive event, just like the event E5 of Table 3. The second subtype is the whole recursive event can be divided into several parallel events, and each parallel event does not have the further recursive event. The third subtype is that there is the further recursive event in the recursive event. The last subtype is the hybrid of the second and third subtypes. The instances of the recursive event are listed in Table 11.

Table 11. The instances of the recursive event

The Second Subtype
In Simplified Chinese (CS)
E15: [一九七一年 T][印巴战事 Th][再起 EP]，因[印度帮助东巴基斯坦独立成为孟加拉 Rn]
E16: [印度 A][帮助 EP][东巴基斯坦 P]
E17: [东巴基斯坦 A][独立 EP]
E18: [东巴基斯坦 Th][成为 EP][孟加拉 Re]
In English (EN)
E15': [1971 T] [India-Pakistan war Th] [breaks out EP], because [India helped East Pakistan be independent and become Bangladesh Rn]
E16': [India A] [helped EP][East Pakistan P]
E17': [East Pakistan A] was [independent EP]
E18': [East Pakistan Th] [become EP] [Bangladesh Re]
The Third Subtype
In Simplified Chinese (CS)
E19: [印尼政府 A]对于[为躲避澳洲暴民的暴力行为而逃抵此地的华裔 D]而[可能 mod][放宽 EP][移民法规 P]
E20: 为[躲避澳洲暴民的暴力行为 Rn]而[逃抵 EP][此地 P]的[华裔 A]
E21: [躲避 EP][澳洲暴民 D]的[暴力行为 Re]

Table 11. *(continued)*

The Third Subtype
In English (EN)
E19': [The Government of Indonesia A] [is possible to mod] [relax EP] [the immigration rules P] [for the Chinese who escaped here in order to avoid the violence of Australian D].
E20': for the [Chinese A] who [escaped EP] [here P] in order to [avoid the violence of Australian Rn]
E21': [avoid EP] the [violence Re] of [Australian D]
The Fourth Subtype
In Simplified Chinese (CS)
E22: [哈塔米 Th][被看作为 EP][伊朗 L][第一位 Ra][改良主义总统 Re]，因为[他在选举中重视法治和民主，主张所有伊朗人参与政治决策的进程 Rn]。
E23: [他 A]在[选举 L]中[重视 EP][法治和民主 P]
E24: [主张 EP][所有伊朗人参与政治决策的进程 Re]
E25: [所有 Ra][伊朗人 A][参与 EP][政治决策的进程 P]
In English (EN)
E22': [Khatami Th] [was to see as EP] [Iran's L] [first Ra] [reformist president Re], because of [his emphasis on the rule of law and democracy in the election, the proposition that all Iranians to participate in political decision-making process Rn]
E23': [his A] [emphasis on EP] [the rule of law and democracy P] [in the election L]
E24': [the proposition EP] that [all Iranians to participate in political decision-making process Re]
E25': [all Ra] [Iranians A] to [participate in EP] [political decision-making process P]

In event E15 of Table 11, the reason "印度帮助东巴基斯坦独立成为孟加拉" is a recursive event, and it can be divided into three parallel atomic events, i.e. E16, E17 and E18. In event E19, the dative "为躲避澳洲暴民的暴力行为而逃抵此地的华裔" is a recursive event E20 and the reason in E20 "躲避澳洲暴民的暴力行为" is also a recursive event E21. In event E22, the reason "他在选举中重视法治和民主，主张所有伊朗人参与政治决策的进程" is the recursive event, and this recursive event can be divided into two parallel events, i.e. E23 and E24. In event E24, the relevant "所有伊朗人参与政治决策的进程" is also a recursive event E25.

The main reason of the recursive event is that the long and complex sentences are common in Chinese.

4 Conclusions

The event has been becoming the hot topic in natural language processing and text analysis. In order to further study of the event-based semantic analysis and computation, it is necessary to label the event semantics for raw natural language text corpus. Compared with sentence-level semantic roles, this paper discussed some problems about event-based semantic roles. According to labeling practices of the natural

language text, we condensed the event-based subjective semantic roles into agent, causer and theme and held patient, dative, result and relevant as the event-based objective semantic roles. The space-time semantic roles which closely related to the event semantics include time, tense, location, source and destination. The additional semantic roles mainly contain instrument, reason, manner and range. The negative, modal and default are taken into the logical operators. In order to deal with the problem of embedded event, this paper put forward the concept of recursive event. Our labeling practice in the natural language texts show that it is necessary to discuss the issues about the semantic roles in the event semantics.

The discussions of the event-based semantic roles in this paper are not enough, and we will study event-based semantic roles deeply through more natural language text labeling practices and analysis. In addition, we can also explore the dynamic event semantics, and then deepen the text analysis and understanding. Therefore, we need to strengthen the study on the event-based semantic relations and computation in the future.

Acknowledgements. The work presented in this paper is supported by the National Natural Science Foundation of China (No. 61100133 and 61173062), the Major Projects of National Social Science Foundation of China (No. 11&ZD189), the Major Projects of Social Science Foundation of Department of Education of Hubei Province of China (No. 2011jyte126) and the University Students Science and Technology Innovation Foundation of Wuhan University of Science and Technology (No. 11ZRA105).

References

1. Wu, P.B., Chen, Q.X., Ma, L.: Study on Intelligent Retrieval of Event Relevant Documents Based on Event Frame. Journal of Chinese Information Processing 17(6), 25–31 (2003)
2. Liang, H., Chen, Q.X., Wu, P.B.: Information Extraction System Based on Event Frame. Journal of Chinese Information Processing 20(2), 40–46 (2006)
3. Wu, P.B., Chen, Q.X., Ma, L.: Research on Extraction and Integration of Developing Event Based on Analysis of Space-Time Information. Journal of Chinese Information Processing 20(1), 21–28 (2006)
4. Liu, M.F., Jin, K.J., Ji, D.H., Zhang, X.L.: Application of Anaphora Resolution Based on Statistics and Rules in Event-based Automatic Summarization. In: 10th Chinese National Conference on Computational Linguistics, pp. 534–539. Tsinghua University Press, Beijing (2009)
5. Liu, M.F., Li, W.J., Wu, M.L., Lu, Q.: Extractive Summarization Based on Event Term Clustering. In: 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), pp. 185–188. Association for Computational Linguistics, Stroudsburg (2007)
6. Liu, M.F., Li, W.J., Hu, H.J.: Extractive Summarization Based on Event Term Temporal Relation Graph and Critical Chain. In: Lee, G.G., Song, D., Lin, C.-Y., Aizawa, A., Kuriyama, K., Yoshioka, M., Sakai, T. (eds.) AIRS 2009. LNCS, vol. 5839, pp. 87–99. Springer, Heidelberg (2009)
7. Lv, S.X.: Hanyu Yufa Fenxi Wenti. The Commercial Press, Beijing (2010)

8. Yuan, Y.L.: Jiyu Renzhi De Hanyu Jisuan Yuyanxue Yanjiu. Peking University Press, Beijing (2008)
9. Yuan, Y.L.: Matching Even-template with Argument Structure of Verbs: Towards a Verb-driven Approach of Information Extraction. Journal of Chinese Information Processing 19(5), 37–43 (2005)
10. Wu, P.: Hanyu Teshu Jushi De Shijian Yuyi Fenxi Yu Jisuan. China Social Sciences Press, Beijing (2009)
11. Wu, P.: Logic Semantic Analysis and Computation of Argument Controlling Predicates and Non-argument Controlling Predicates. Foreign Languages and Their Teaching 204, 5–10 (2006)
12. Wu, P.: A Semantic Analysis of Event Structure of Shi-Construction. Journal of Zhejiang University (Humanities and Social Sciences) 39(3), 157–164 (2009)
13. Yu, J.D., Fan, X.Z., Pang, W.B.: Research on Semantic Role Labeling for Event Information Extraction. Computer Science 35(3), 155–157 (2008)
14. Dong, Z.D., Dong, Q.: HowNet, http://www.keenage.com/zhiwang/c_zhiwang.html

Construction and Application of Chinese Emotional Corpus

Liang Yang and Hongfei Lin

School of Computer Science and Technology, Dalian University of Technology, Dalian
yangliang@mail.dlut.edu.cn, hflin@dlut.edu.cn

Abstract. Text affective computing or sentiment analysis is an important research domain for natural language process, and it requires a large-scale emotion corpus, which can significantly support emotion classification and recognition. In this paper, some experiences and basic rules about constructing a Chinese emotional corpus are discussed, which include data collecting coverage, annotation system and quality criterion. Now there are 163,813 sentences labeled in our Chinese emotional corpus, and works like automatic emotion ontology learning, emotion transformation, gender analysis and emotion schema mining are available based on its statistic and labeled data. This effectively proves the value of our Chinese emotional corpus.

Keywords: Affective computing, sentiment analysis, Chinese Emotional Corpus.

1 Introduction

At present, affective computing or sentiment analysis has become the focus of artificial intelligence and natural language process, which aims to assist computers to recognize the human emotion. To achieve this target, available recognition and identification models are requisites. In order to train the models with higher precision and fault tolerant capability, a large-scale Chinese emotional corpus is prerequisite. Considerable research works have been done in abroad by far and a series of corpora have been constructed, such as Brown corpus, Lob corpus, and Longman corpus. In domestic, the construction of Chinese corpus started at 1980s, which include People Daily corpus, Modern National Chinese corpus [1], and Academia Sinica balanced corpus [2]. These corpora mentioned above have accumulated a great deal of valuable experience on corpus design, data collecting, annotation standard, and quality control. Emotional corpus is a specialized corpus, which can significantly improve the accuracy of opinion mining and sentiment analysis. Currently, the existing text emotional corpora are Pang corpus [3], Whissell corpus [4], Berardinelli Movie Review corpus and Products Review corpus [5]. However, none of them has a wide universality for affective computing or opinion mining, and only can be applied in a specific domain. For example, Pang corpus [3] is apparently available for movie review domain, but not available for other domains. Therefore, emotional corpus has plenty room to improve. So our work tends to construct a Chinese emotional corpus,

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 122–133, 2013.

which provides a formal standard for Chinese text emotional corpus construction, and contains a better universality for affective computing and sentiment analysis.

With the development of Web 2.0, the number of web users and applications is soaring. So the User Generation Contexts, such as Blog and Micro-Blog, are indispensable to corpus construction. Blog based Corpora have been done by Quan and Ren [6] and Yang et al. [7], while our corpus takes Micro-Blog into consideration, though the length of its context is shorter compared with Blog, the research value of Micro-Blog is the same to researchers of sentiment analysis.

Corpus construction consists of corpus design, data collecting and preprocesses annotation rule and regulation, quality control; later the paper will state them respectively. In short, affective computing and sentiment analysis just started in domestic, and requires an authority and large-scale Chinese emotional corpus as its foundation.

The rest of this paper is organized as follows. Section 2 describes the design criterion of the Chinese emotional corpus. Section 3 presents the collecting principle and the coverage of data. Section 4 discusses the annotation system. Section 5 overviews the corpus quality control. Section 6 illustrates the auto-extending method and the applications of our emotional corpus. Section 7 draws the conclusion and outlines the future work.

2 Corpus Design

The first step of constructing a corpus is design, and the aim of corpus design is to guarantee the constructed corpus in accordance with specific application. There are two principles to be followed, and they are versatility and descriptiveness [1].

2.1 Versatility

In order to achieve versatility, our emotional corpus firstly analyzes the particle size. If the particle size is too coarse, the description of complex language phenomena will not be comprehensive and detailed; on the other hand, if the particle size is too fine, there will be a great number of annotation information, which lowers annotation efficiency. After the above mentioned analysis, the sentence is selected to be the label unit. With the labeled emotion sentences, the corpus can be widely used to train the emotion identification or classification model. Different topics and domains are also included to meet specific domain sentiment analysis.

2.2 Descriptiveness

The application value of a corpus depends on whether the collecting data can objectively reflect the aspects of modern Chinese, such as words, phrases, syntax, semantics and pragmatics. Descriptiveness requires that the corpus data has a practical frequency of use, and makes the corpus universality and representativeness. In our Chinese emotional corpus, descriptiveness emphasizes that the corpus is faithful to original language facts.

3 Data Collecting

The second step of constructing a corpus is data collecting, and it is a preparation for annotation so as to provide high quality training sets, statistic regularity, relevant rules and commonsense knowledge. The target of data collecting is to collect proper data for emotional corpus and preprocess. The advantages of our corpus in above mentioned aspects are significant, and these can enhance the qualities and quantities of affective lexicon ontology [8] and affective commonsense knowledgebase [9].

The coverage of a corpus is decided by the selection strategy. Coverage means the distributions of a corpus in different domains. A standard corpus should cover different domains as many as possible. The domain generally contains four dimensions [10], which are time axis (time feature), space axis (the regional feature), subject axis (knowledge feature), and style axis (type of writing feature).

Primary school textbook, movie script, fairy tale, literature periodical, Blogs and Micro-Blogs are collected to be the materials of our emotional corpus. The reasons are as follows. From time axis to analyze, Fairy tale and primary school textbook are earlier classical articles, while literature periodicals and movie scripts are recent works, so there is a large time span, which is from 1939 to 2010 in our emotional corpus. Such as the document we find in the corpus below:

(1)Chinese:“骆 luo 驼 tuo 祥 xiang 子 zi，老 lao 舍 she，1939。”

English: “Rickshaw Boy, She Lao, 1939.”

(2)Chinese:“春 chun 天 tian 的 de 故 gu 事 shi，韩 han 寒 han，2010。”

English: “Story of spring, Han Han, 2010.”

Although materials mainly consist of Chinese articles, movie scripts are translated from foreign ones, therefore regional features are taken into consideration. The example is given below:

(3)Chinese: “狮 shi 子 zi 王 wang。”

English: “The Lion King.” (Translated)

(4)Chinese: “汽 qi 车 che 总 zong 动 dong 员 yuan。”

English: “Cars.” (Translated)

In subject aspect, Blogs downloaded contain IT, Finance, Sports and so on. These blogs belong to different subjects. The samples are given as follows:

(4)Chinese: “南 nan 非 fei 世 shi 界 jie 杯 bei，我 wo 支 zhi 持 chi 巴 ba西 xi 队dui。”

English: “South Africa World Cup, I support Brazil.” (Sports)

(5)Chinese: “股 gu 市 shi 今 jin 日 ri 下 xia 跌 die。”

English: “Stock market fell today.” (Finance)

In style aspect, primary school textbooks are more normative, but movie scripts and blogs are more oral. Consider the sentences we find below:

(6)Chinese: “刘 liu 四 si 恶 e 狠 hen 狠 hen地 di 看 kan着 zhe 祥 xiang 子 zi。”

English: “Liu Si stares at Xiang Zi fiercely.” (Normative in textbook)

(7)Chinese: “爷 ye 们 men 们 men，请 qing 你 ni 们 men 尽 jin 兴 xing。”

English: “Boys, enjoy yourselves.”(Oral in Blogs)

In brief, firstly our data collecting principles satisfy the basic corpora material demands; Secondly, selected ones are incline to have a strong literary flavor and abundant emotional expressions, which are significantly important to affective computing. Table 1 lists detailed information of our corpus.

Table 1. Information of Our Corpus

Source	Descriptions	Word	Phrase	Sen	Doc
Primary Textbook	People's Education Press	129,486	91,032	4,809	171
Movie Script	The Lion King, Cars and so on	84,118	54,092	5,911	237
Fairy Tale	Part of Green and Anderson Fairy Tale	5,4066	39,005	2,011	73
Literature Periodical	Young People Digest, Young Literature and Art, New Youth	6,308,526	4,375,396	237,290	3754
Sina Blog	Literature, Sports, IT and Finance	1,331,573	980,651	13,574	365
Sina M-Blog	Literature, Sports, IT and Finance	2,552,248	1,953,156	70,698	-
Total		10,460,017	7,493,332	334,293	4,600

4 Emotional Corpus Annotation System

In this section, the paper presents annotation system, which is based on TEI (Text Encoding Initiative) and emotion annotation system. Corpus Annotation System is to decide the processing degree of a corpus and particle size, and to change unstructured texts into knowledge or commonsense. Apparently, annotation system is a key factor to construct a high-quality and large-scale corpus.

4.1 Annotation Based on TEI

Annotations in our emotional corpus obey the basic rules proposed by the corpus linguistic expert Leech:

(1) The user should have a clear idea about the meaning of each tag;
(2) Any tag can be extracted stored, deleted;
(3) The annotation can be generally accepted by users, and used as a tool;
(4) There should be instruction files attached with the corpus.

The Text Encoding Initiative (TEI) is a text-centric community of practice in the academic field of digital humanities, operating continuously. They collectively define an XML format, are the defining output of the community of practice. The format differs from other well-known open formats for text (such as HTML) in that it is primarily semantic rather than presentational; the semantics and interpretation of very tag and attribute are specified. So in this paper, a passage in our emotional corpus consists of header and body, and different textual components and concepts. The part of header contains background information, such as author, title, topic, date. The body part is the context. Words, sentences and document all can be annotation units with a start tag and an end tag. This semi-structure way is similar to XML, and convenient to be applied to train models. An annotation example based on TEI is presented below in Figure 1:

4.2 Emotion Annotation System

The emotion annotation standards should be decided and stay unchangeable to guarantee the annotation consistency. Due to the diversity and complexity of the data, some standards need to be adjusted, and that will lower the quality. To avoid adjusting annotation standards, in this paper part of the corpus are labeled in advance, summarize problems, and take the feature of emotion annotation into consideration. The annotation system is as follows:

$$DocumentModel = (title,\ author,\ [gender],\ style,\ source,\ persons,\ sentences,\ [topic],\ keynote) \quad (1)$$

$$SentenceModel = (origin,\ sender,\ [accepter],\ [rhetoric],\ emotions,\ [keywords]) \quad (2)$$

Three particle sizes are defined: word, sentence, and document. Sentence size annotation is chosen as the main part, and the information of word and document size annotation assists it to analyze the emotion. The variable inside *[]* are optional, while others are necessary.

The meanings and extents of variables in two models are listed in Table 2.

Persons, sentences, emotions and keywords in Table 2 are all sets, which can be defined as a vector. For example, $persons = (persona_1, persona_2\ldots, persona_i\ldots)$. There are two special emotion senders in our corpus, which are aside and others. The sender aside emphasizes the sentence is narrated by the author, and no obvious emotion sender exists. Others are used to replay the unimportant, and this can effectively reduce the burden of annotator, and avoid the sparse of some emotion senders. *Topic* and *gender* variable are mainly used in Blog data to assist analysis based on topic.

```
<document>
  <header>
    <title>想八卦，被八卦，抱大腿</title>
    <author>
      <name>徐静蕾</name>
      <year>2009</year>
    </author>
    <style>网络</style>
    <source>新浪博客</source>
    <persons>
      <persona>旁白,我</persona>
    </persons>
    <topic>八卦，好奇，感恩</topic>
  </header>
  <body>
    <topic>八卦，好奇</topic>
    <emotion>q</emotion>
    <sentence summarize>2</sentence summarize>
    <sect>
<orgin>职场小说，是被我当作八卦来看的，十分好奇人家过的是什么样的生活!</orgin> <segment>职场/n 小说/ n,
/w是/v 被/v 我/r 当作/v 八卦/n 来看的/v，十分/ad 好奇/v 人家/r 过的/v 是/v 什么样的/a 生活/n!/w</segment>
      <sender>我</sender>
```

Fig. 1. The Example of Emotion Annotation

Table 2. Explanations of Two Models

Category	Variable	Explanations	Extension
Document Model	*Title*	Article Title	
	Author	Article Author	Name, Country, Times
	gender	Author Gender	
	Style	Writing Style	prose| poetry| friction| drama|Blog|Micro-Blog
	source	Source Docs'	Primary Textbook| Fairy Tale| Movie Script|
	topic	topic	Literature Periodical
	Persons	Emotional Subject	$Protagonist_1$|$Protagonist_2$…| $Protagonist_i$…
	Sentences	Sentences in Document	Details in SentenceModel
	Keynote	Emotion	o| h| e| i| m| f| d| s
Sentence Model	*origin*	Original Sentence	
	sender	Emotional Subject	$Protagonist_i$
	Accepter	Emotional Acceptor	$Protagonist_i$
	Rhetoric	Rhetoric category	Metaphor|Assimilate|Metonymy|Exaggerate| Parallelism| Photonic| Question| Repeat
	Emotions	Sentence Emotion	o| h| e| p| r| b| l| k| c| i| s| w| g| m| u| f| x| t| d| a| j| y| q
	Keywords	Emotion Words	$Word_1$| $Word_2$…| $Word_i$…

Experiment in [8] draws the conclusion that emotion lexicon feature has a better ability to automatically distinguish the emotion of a sentence. So *Keywords* are important to decide emotions of sentences, in this paper DUTIR emotion lexicon [11] are applied. Due to improve the annotation efficiency, this paper does not annotate negative words and adverb of degree, though emotion classification is influenced by them too. The extent of variable *Keynote* and *emotions* are listed in Figure 2.

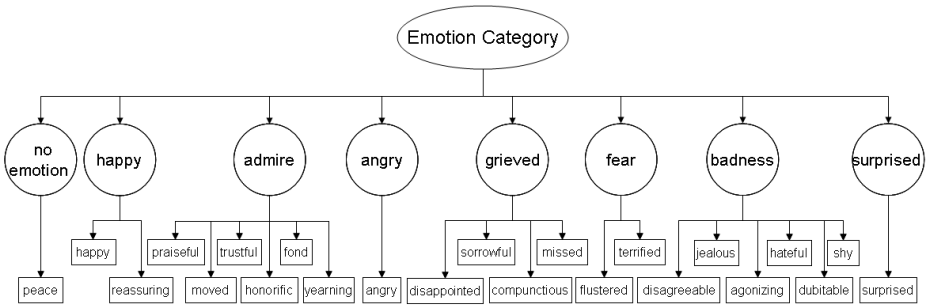


Fig. 2. Emotion Category in Corpus

5 Corpus Quality Control

In this section, corpus quality control is illustrated in 4 aspects: (1) annotation criterion and input system; (3) Artificial calibration; (3) mechanism for correcting errors; (4) automatic update system.

5.1 Annotation Criterion and Input System

The aim of annotation criterion and input system is to reduce the error rate of operation and to increase the consistency annotation speed effectively. With the uniform annotation criterion and standard, the differences among annotator can lessen. The criterions are dynamically updated, and part of them is listed as follows:

(1) Based on the statistic, if the next sentence has the same emotion subject with the former one, the emotion tends to be continual.

(2) The Keyword can be a word or a phrase except one sentence.

(3) Only topic and gender are optional in header part, while in body part, topic, accepter and rhetoric are optional.

(4) One sentence can have more than one emotion, but any emotion cannot co-exist with peace emotion in the sentence.

(5) Two special characters, aside and other, can be used to replay unimportant character or unclear protagonist.

An input system is provided so as to increase the efficiency and accuracy. Validity check and Heuristic algorithm [12] are applied to split sentence automatically and avoid wrong materials entering the corpus.

5.2 Artificial Calibration

The Chinese emotional corpus was annotated by eight annotators. Prior to formal annotation, they were given a basic instruction about the purpose of the annotation and annotation rules based on Chinese Opinion Analysis Evaluation (COAE). Firstly, two annotators are selected to do the sample annotation. They annotated three sample topics and held a meeting to discuss the discrepancies and the ambiguous statements in the instruction rules. Then all the annotators held another meeting to learn how to do the annotation work based on the finished sample. After that, two teams are set up, and the person who done the sample annotation are set as the leader. The annotators can ask any questions to the team leader about the specific sentences if they were unsure of the labeling. When more than one emotion is present in a sentence, the annotators will analyze the sentence together to confirm the emotion category.

5.3 Mechanism for Correcting Errors

Mechanism for correcting errors is to check the correctness and consistency uniformly after the corpus annotation. We adopt cross validation, which is widely used in large-scale corpora validation. It can uniform the annotation standards in most cases among annotator. In the validation of consistency, automatic machine check and human modification are combined. Considering the feature of emotion materials, the consistency are analyzed in two aspects: word and emotion continuity.

To make our work clear, the functions and variables are introduced in Table 3.

From the consideration on emotion keywords, emotion keywords are the clues to check consistency. Consistency is calculated in equation (3):

$$wordConsistency(i,j) = wordSame(S_i, S_j) \cap \overline{larSame(E_i, E_j)} \cap \overline{Neg(S_i)} \cap \overline{Neg(S_j)} \tag{3}$$

S_i, S_j are sentences in one document; E_i, E_j are the emotions of S_i and S_j; Validation check of consistency is necessary when the value of wordConsistency is one. Here is the case here:

Two sentences without negative words have the same emotion keywords, if they have different emotions, then the value is one.

From the consideration on emotion continuity, if the sentences before and after have different emotion, but they have the same emotion subject, there may be errors in this sentence. The consistency is calculated in equation (4):

$$contextConsistency(i) = personSame(S_i) \cap \overline{emotionSame(E_i)} \tag{4}$$

Table 3. Introductions of Functions and Variables

Function	Variable	Explanation	Value	Condition
Neg	S_i	Whether $sentence_i$ contains negative words	0 1	No YES
larSame	E_i, E_j	Whether $sentence_i$ and $sentence_j$ belong to the same branch note of emotion tree	0 1	Different Same
wordSame	S_i, S_j	Whether $sentence_i$ and $sentence_j$ contain the same emotion keywords	0 1	Different Same
personSame	S_i	Whether $sentence_{i-1}$ to $sentence_{i+1}$ have the same emotion subject	0 1	Different Same
emotionSame	E_i	Whether $sentence_{i-1}$, $sentence_i$, $sentence_{i+1}$ have the same emotion	0 1	Different Same

The above mentioned methods both check the consistency by machine automatically, and then human modification is adopted. Word continuity has a higher error rate and need to be analyzed further. So the context continuity check is provided to modify the occasion, which may be an error.

5.4 Automatic Update System

In Web 2.0, user generation content is a resource for updating the corpus. We develop an automatic update system, which contains a crawler and emotion classification module. The information corpus contained can easily get after analyzing the source code of Sina Blog. The example source code of the body part is illustrated in Figure 3.

And then the information is extracted from the original. The module of emotion classification splits the body content into sentences, and finishes the emotion classification job. After machine and human modification, the result will be stored in our emotional corpus finally.

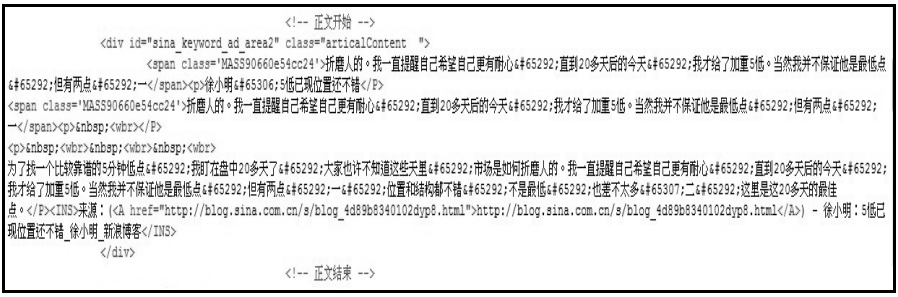


Fig. 3. The Source Code Example of Body

6 Statistic Data and Application of Emotion Corpus

Nowadays 3,594,887 words, 2,684,866 phrases and 163,813 sentences have been labeled. Among the labeled sentence, we found that, sentences with no emotion have the maximum portion, and then is happy, praiseful, agonizing and dubitable. This is the first phase of the project to complete emotional corpus of corpora. In the future, we expect to label ten million words for our Chinese emotion corpus.

The content and annotation of emotion corpus system decides its applied scope. Currently, the emotion corpus [13] mainly used in training emotion recognition model, automatic learning of emotion ontology, emotion transfer learning [14], and gender analysis [15], Question Answering [16] and emotion schema mining.

Sentence labeled corpus not only provides the emotion category, but also the keywords, rhetoric and emotion subject. All these are abundant and distinguished features for emotion recognition model, and they can significantly improve its accuracy. The keywords are the first hand materials for emotion ontology learning. The gender and writing styles are used to predict the author's gender of unknown document. Also, emotion schemas can be mined based on the context, and it is a perquisite for cognitive theory learning.

Emotion transfer learning and Emotion Schema have been done based on the statistic data. The figure 4 below is the transfer probabilities figure, and figure 5 is about the Emotion Schema.

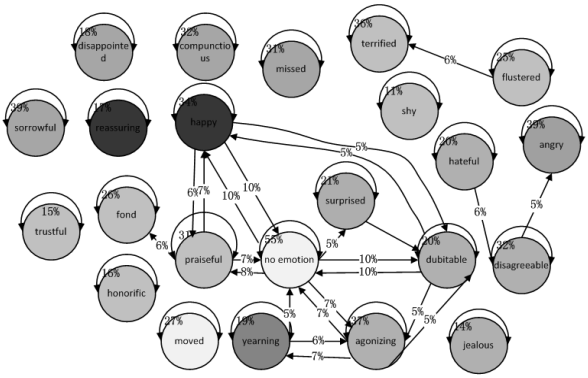


Fig. 4. Emotion Transfer Probabilities

From Figure 4, we can find that there is a better cohesion among the emotions belonging to the same branch node of emotion tree, especially the emotion "hate". And continuity exists in emotion transfer, which means that there is a higher probability transferring to the same emotion.

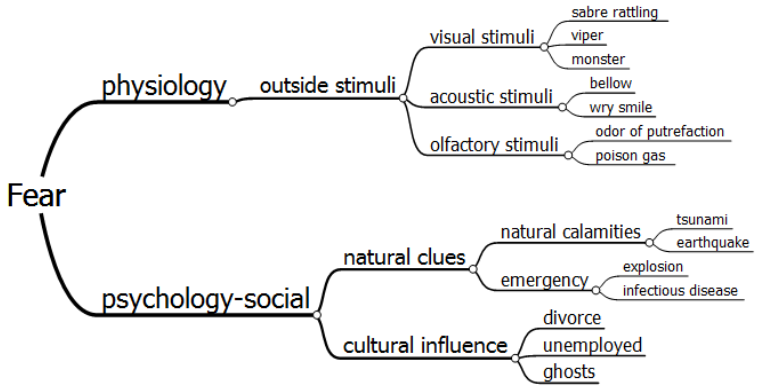


Fig. 5. Emotion Schema of Fear

The emotion schema consists of physiology and psychology factors. We extract the fear schema from our Chinese emotional corpus as an example, and details can be seen in figure 5. It can be applied as priori knowledge to assist sentiment analysis and opinion mining.

7 Conclusion and Future Works

In this paper, how to construct an emotion corpus is introduced, which consists of corpus design, data collecting, annotation system, and quality control. We provide some suggestion to improve the annotation speed and quality.

In the first phase, 3,594,887 words, 2,684,866 phrases and 163,813 sentences have been labeled for the corpus. After analyzing the annotation errors, we find that there are still some problems to be solved. Then we will summarize the experience, and ten million sentences will be labeled in the second phase. Since any corpus construction cannot be perfect at first, we will modify as many as possible errors in the future, such as the distributions of emotion are not even. In the future, we will make continuous improvement.

Acknowledgments. This research is supported by the National Natural Science Foundation of China (No: 60973068) and State Education Ministry Research Fund for the Doctoral Program of Higher Education (No.20090041110002).

References

1. Liu, L.Y.: Study of Corpus for Contemporary Chinese Language. Applied Linguistics 3, 114–133 (1996)
2. Chen, Y., Lee, Y.M., Li, S., Huang, C.-R.: Construction of Chinese Emotion Corpus with an Unsupervised Approach. In: 10th Chinese National Conference on Computational Linguistics, pp. 325–331 (2009)

3. Pang, B., Lee, L.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics, Stroudsburg (2002)
4. Athanaselis, T., Bakamidis, S., Dologlou, I.: Recognizing Verbal Content of Emotionally Colored Speech. In: 14th European Signal Processing Conference. European Association for Signal Processing (2006)
5. Massa, P., Avesani, P.: Trust-aware Bootstrapping of Recommender Systems. In: ECAI 2006 Workshop on Recommender Systems, pp. 29–33 (2006)
6. Quan, C.Q., Ren, F.J.: Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis. In: ACL 2009 Conference on Empirical Methods in Natural Language Processing, vol. 3, pp. 1446–1454. Association for Computational Linguistics, Stroudsburg (2009)
7. Yang, C.H., Lin, K.H.-Y., Chen, H.-H.: Building Emotion Lexicon from Weblog Corpora. In: 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 133–136. Association for Computational Linguistics, Stroudsburg (2007)
8. Xu, L.H., Lin, H.F.: Discourse Affective Computing Based on Semantic Features and Ontology. Journal of Computer Research and Development 44, 356–360 (2007)
9. Chen, J.M., Lin, H.F.: Constructing the Affective Commonsense Knowledgebase. Journal of the China Society for Scientific and Technical Information 28, 492–498 (2009)
10. Zhou, M.: Approach to the Chinese Dependency Formalism for the Tagging of Corpus. Journal of Chinese Information Processing 8, 35–52 (1994)
11. Xu, L.H., Lin, H.F., Pan, Y., Ren, H., Chen, J.M.: Constructing the Affective Lexicon Ontology. Journal of the China Society for Scientific and Technical Information 27, 180–185 (2008)
12. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing, 1st edn. MIT Press, Cambridge (1999)
13. Xu, L.H., Lin, H.F.: Ontology-Driven Affective Chinese Text Analysis and Evaluation Method. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 723–724. Springer, Heidelberg (2007)
14. Yang, L., Guo, W., Lin, H.F.: Emotion Transformation Analysis Based on Text. Computer Engineering and Science 39, 123–129 (2011)
15. Mukherjee, A., Liu, B.: Improving Gender Classification of Blog Authors. In: The 2010 Conference on Empirical Methods in Natural Language Processing, pp. 207–217. Association for Computational Linguistics, Stroudsburg (2007)
16. Tang, Q., Lin, H.F.: Research on Focal Figures-Oriented Sentimental Question Answering. In: 2008 Advanced Language Processing and Web Information Technology, pp. 168–174. IEEE Press, New York (2008)

Termhood-Based Comparability Metrics of Comparable Corpus in Special Domain

Sa Liu and Chengzhi Zhang

Department of Information Management, Nanjing University of Science and Technology, Nanjing, China
liusa321@163.com, zhangchz@istic.ac.cn

Abstract. Cross-Language Information Retrieval (CLIR) and machine translation (MT) resources, such as dictionaries and parallel corpora, are scarce and hard to come by for special domains. Besides, these resources are just limited to a few languages, such as English, French, and Spanish and so on. So, obtaining comparable corpora automatically for such domains could be an answer to this problem effectively. Comparable corpora, that the subcorpora are not translations of each other, can be easily obtained from web. Therefore, building and using comparable corpora is often a more feasible option in multilingual information processing.

Comparability metrics is one of key issues in the field of building and using comparable corpus. Currently, there is no widely accepted definition or metrics method of corpus comparability. In fact, Different definitions or metrics methods of comparability might be given to suit various tasks about natural language processing. A new comparability, namely, termhood-based metrics, oriented to the task of bilingual terminology extraction, is proposed in this paper. In this method, words are ranked by termhood not frequency, and then the cosine similarities, calculated based on the ranking lists of word termhood, is used as comparability. Experiments results show that termhood-based metrics performs better than traditional frequency-based metrics.

Keywords: Termhood-based Comparability, Comparable Corpus, Frequency-based Metrics, Terminology Extraction.

1 Introduction

Parallel corpus which contains source documents and their translations plays an important role in multilingual information service [1], such as Cross-Language Information Retrieval (CLIR), and machine translation (MT). However, parallel corpus is scarce resources and not easy to be obtained in some under-resourced languages or special domains. Due to these shortcomings, building and using comparable corpora is often a more feasible option. It is obviously easier to find document collections with similar topics in multiple languages than to find parallel corpus [2]. The Web, with its vast volumes of data, offers a natural source for this. For example, bilingual website, and online Wikipedia, are very good resources for collecting and obtaining

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 134–144, 2013.

comparable data. Meanwhile, comparable data extracted from these resources can update with the increasing of the source data, and then more new words can be extracted easily and accurately. Therefore, building and using comparable corpora is becoming more and more important and urgent.

Comparability is the key concept in the research of comparable corpus. However, so far there has been no widely accepted definition of comparability. Different definitions or metrics methods of comparability might be given to suit various NLP tasks. In the task of machine translation, comparability is focused on distribution and quality of translated knowledge [3]. In multilingual terminology extraction, comparability is focused on distribution and quality of the vocabulary of translated forms [4].

So far, method of based word frequency list has been always used to measure corpus homogeneity and similarity between corpora [5]. This method is useful for measure corpus similarity in the respect of comparing the different language styles, however, this method perform badly in the task of bilingual term extraction. In our previous experiments, we verify this point.

A new comparability, namely, termhood-based metrics, especially used for comparability metrics of comparable corpus in special domain, is proposed in this paper. Experiments results show that this method performs better than traditional frequency-based metrics in the task of bilingual term extraction. The remainder of this paper is organized as follows. In section 2, related works are introduced. Then the termhood-based metrics is described in Section 3. Section 4 presents the experiment results and the proposed method is evaluated according to the task of bilingual terminology extraction in section 5. The paper is concluded with a summary and directions for future works.

2 Related Works

In this section, we review related works relevant to our research, including a brief review of building comparable corpus and comparability metrics of comparable corpus.

2.1 Building Comparable Corpus

Generally, comparable corpus is generated from news agencies or by crawling certain sites [2]. Talvensaari et al. built comparable corpora based on focused crawling [6]. Leturia et al. (2009) used search engine queries for collecting comparable corpora from the Internet [7]. Otero and López exploited Wikipedia for collecting domain comparable corpora by using categories as topic restrictions [8].

In our previous experiments, we used three different Internet data source for collecting comparable corpus: querying bilingual domain keywords in search engine, exploiting the online encyclopedia-Wikipedia, and searching academic databases. At last, we choose data from academic databases as our experiment corpus for its suitable size and quality.

2.2 Comparability Metrics of Comparable Corpus

In order to evaluate the quality of comparable corpus, we need some way to measure the degree of comparability of the comparable corpus. The most used metrics of comparable corpus are Chi-square statistics and word similarity. Leturia et al. used these two methods to measure the comparability of domain-specific comparable corpora obtained from the Internet by using search engine [2]. One method is calculating the cosine value between the vectors containing all the keywords of each corpus; the other is calculating Chi-square statistics for the most N frequent keywords (Top-N keywords). The ACCURAT (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation) project used asymmetric Chi-square statistics to measure comparability [9]. The TTC (Terminology Extraction, Translation Tools and Comparable Corpora) project concentrate on two dimensions for comparability calculation: one is similarity between anchor texts in its own language; the other is dissimilarity between anchor points texts in two corpora [10].

The aforementioned works are based on word frequency lists. This kind of method is simple and effective for measure language style. However, this method performs poor between different domain corpora. In our previous investigation, we find this method can't distinguish different domain-specific corpus efficiently.

Li and Gaussier purposed a metrics of comparable corpus for bilingual lexicon extraction [11]. Given a comparable corpus P consisting of an English part P_e, and a French part P_f , the degree of comparability of P is defined as the expectation of finding the translation of any given source/target word in the target/source corpus vocabulary. This method needs a bilingual dictionary for mapping between two corpora. Thus the size and quality of dictionary may heavily affect the result of comparability measure.

In this paper, we propose termhood-based comparability metrics, according to bilingual terminology extraction task. It is noted that our method is oriented to bilingual terminology. It is in this point that our method is different to Li et al [11], which is oriented to bilingual lexicon. As for comparability of domain-specific comparable corpora, our method, based on termhood calculating, is more suitable to highlight terminology.

3 Termhood-Based Comparability Metrics of Comparable Corpus

As for terminology extraction based on comparable corpus, comparability should concern the distribution and quality of terminology. Termhood is defined as degree of terminolgy to be term in a specific field. Quality of term can be measured by termhood and distribution of words be measured by ranking list of word weighting. So, we proposed termhood-based comparability metrics. In this method, words are ranked by termhood not frequency, and then the cosine similarities are calculated based on the ranking lists of word termhood. The similarity obtained is used as comparability of comparable corpus.

3.1 The Basic Idea

For calculating termhood-based comparability, the whole process we used is described as follows.

(1) Chinese-English domain comparable corpus collecting: comparable corpus we exploit in the experiment from two online academic databases (Chinese corpus from CNKI [12], English corpus from EBSCO[13]);

(2) Preprocessing: Chinese corpus is preprocessing and word lists from keywords (come from the abovementioned academic databases) and full-text words (come from full-text of document) are both obtained;

(3) Translating and processing: English corpus is translated into Chinese through Google translate [14]. Then the translated corpus is segmentated by ICTCLAS[15], and finally word lists from keywords and full-text words are acquired;

(4) Termhood measure: Termhood of words is computed by using the method of corpus comparison after acquiring word frequency;

(5) Similarity calculating: Ranking the word list again based on termhood and calculating cosine similarity between vectors.

3.2 Key Technology in the Proposed Method

In this section, key technology used in our proposed method is described in detail, including termhood measure and comparability of termhood-based metrics.

(1) Termhood measure by corpus comparison

It is observed that a true term is more outstanding (or peculiar) to its own subject domain than to a general domain or another field. Kit and Liu proposed a measure for mono-word termhood in terminology of such peculiarity and quantify it in terms of a word's ranking difference in a domain and background corpus [16]. We use this method to measure the termhood of terminology. We use People's Daily corpus of 1998 from January to June as background corpus and Library and information (LIS) corpus as domain corpus. Given a domain corpus D (with a vocabulary VD) and a background corpus B (with a vocabulary VB), the termhood of a word w is defined as formula (1).

Where r (w) is the ranking number of word w in a corpus in question, |VD|,| VB | is the size of domain and background corpus respectively. A word rank is normalized by the vocabulary size of its corpus in order to make the word ranks in two corpora comparable.

$$\text{Termhood}(w) = \frac{r_D(w)}{|V_D|} - \frac{r_B(w)}{|V_B|} \tag{1}$$

(2) Comparability of termhood-based metrics in comparable corpus

According to the termhood, word lists are ranked in descending order. Then we calculate similarity of the new word list by vector space model [17]. The similarity obtained is used to be comparability of comparable corpus.

4 Experiments and Result Analysis

In this section, we first introduce experiments data used for comparability metrics. Then two different way of data processing are described in detail. At last, the experiment results are given and analyzed.

4.1 Data

The maximum and minimum of comparability is comparability of parallel corpus and comparability of non-comparable corpus respectively. In the experiments, we select three kinds of comparable corpus with different comparability, i.e. parallel corpus, comparable corpus, and non-comparable corpus. It is assumed that comparability of comparable corpus lies between parallel and non-comparable corpus. In our experiments, the domain of parallel corpus is Library and Information Science (LIS), obtained from abstracts of records from CNKI database; comparable corpus is also LIS domain, collected from two aforementioned academic databases. We build non-comparable corpus through combining Chinese corpus in domain of law and English corpus in domain of LIS. Table 1 describes basic information of corpus.

Table 1. Description of Experiment Corpus

Experiment Corpus	Domain	number of tokens	
		Chinese	English
parallel corpus	LIS	81981	60841
comparable corpus	LIS	82024	60928
non-comparable	Law & LIS	29350	60928

Note: LIS denotes Library and Information Science.

4.2 Data Processing

In order to verify the effectiveness of the proposed method, we compute the comparability of three kinds of comparable corpus, i.e. parallel corpus, comparable corpus and non-comparable corpus respectively. In the experiments, the comparability of each kind of comparable corpus is computed based on termhood-based and traditional frequency-based metrics respectively.

(1) Frequency-based metrics

Word was ranked based on their frequency in the corpus, and then comparability was calculated by cosine value between two vectors represented by words and their frequency. In the experiment, word frequency is normalized because there is some difference between size of Chinese corpus and English corpus. Experiments were carried out for six different corpus sizes, i.e. Top100, Top200, Top500, Top1000, Top2000 and Top5000 respectively.

(2) Termhood-based metrics

The comparability metrics based on termhood computes word termhood by corpus comparison method after word frequency statistic. Then words vectors are generated based on their termhood. Finally comparability was calculated by cosine value between two vectors represented by words and their termhood. In this method we also compute comparability metrics in six different corpus sizes, i.e. Top100, Top200, Top500, Top1000, Top2000, and Top5000.

Besides, keywords and all words were both used in our experiments for comparison. It should be noted that keywords come from the abovementioned academic database and all of words come from full-text of document after segment.

4.3 Experiments and Results

Fig.1 and Fig.2 are results of comparability based on keywords according to the two measurement methods.

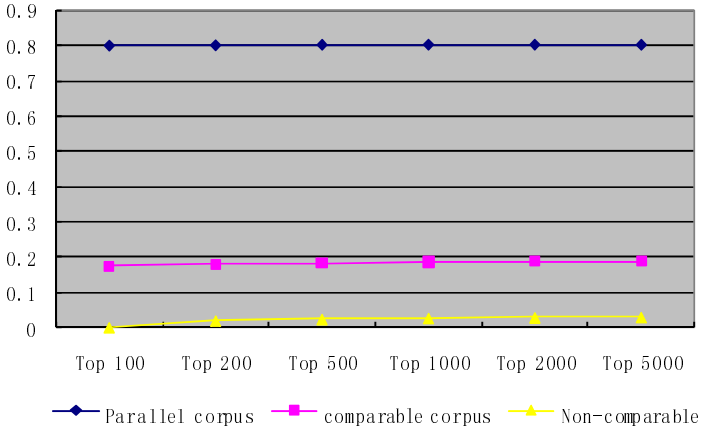


Fig. 1. Frequency-based metrics using keywords

Notice that both methods reflect the fact that comparability of parallel corpus > comparability of comparable corpus> comparability of non-comparable corpus. However, it is assumed that comparability of comparable corpus should be in the middle of the two; parallel, comparable, non-comparable, comparability of three kinds of corpus should present an even decreasing trends. In comparison, termhood-based approach reflects this point more obviously.

The performance of frequency-based method for all word is presented in Fig.3. As a whole, it reflects the fact that comparability of parallel corpus > comparability of non-comparable corpus> comparability of comparable corpus. This is against with our hypothesis. Moreover, this result is inconsistent with frequency-based method using keywords. Furthermore, we can also find that comparability of these kinds of corpus is very close to each other, all above 0.9. It is likely that there are so many noisy words that results are affected, further cause incorrect or incredible results.

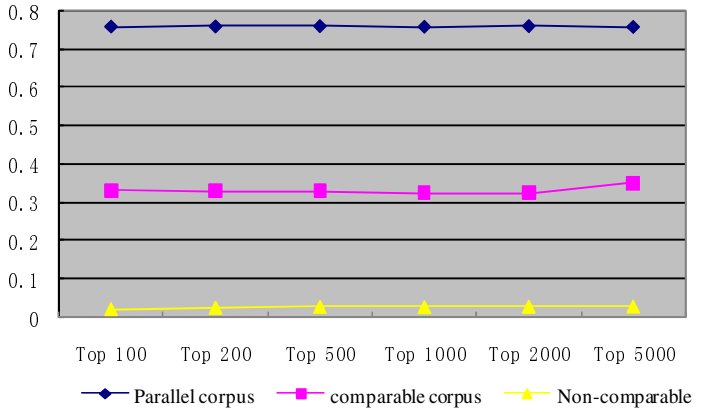


Fig. 2. Termhood-based metrics using keywords

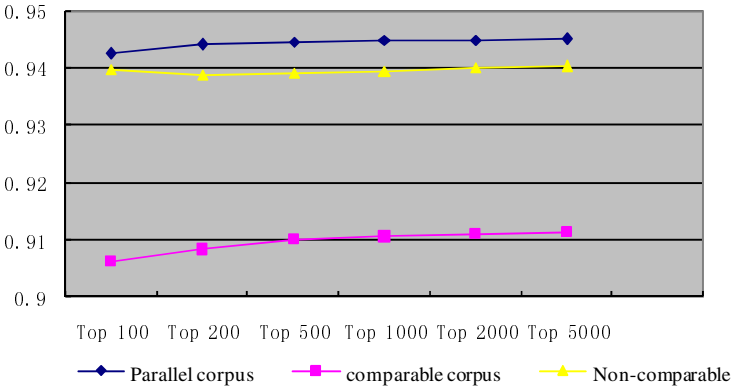


Fig. 3. Frequency-based metrics using all words

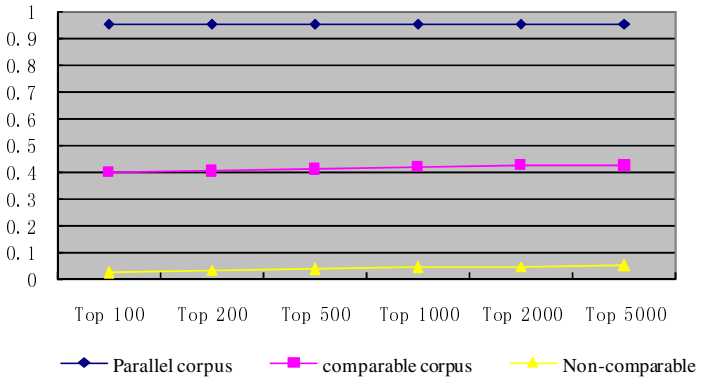


Fig. 4. Termhood-based metrics using all words

The performance of termhood-based method for all word is presented in Fig.4. Notice that comparability of parallel corpus > comparability of comparable corpus> comparability of non-comparable corpus; comparability of three kinds of corpus present an even decreasing trends. Furthermore, these results using all words are consistent with using only keywords.

According to the above analysis, we can conclude that: according to reflecting the real comparable degree between corpora, termhood-based method is better than frequency-based method. Meanwhile, the results of termhood-based method using keywords are consistent with the situation of using full-text words. It also shows that termhood-based method is more stable and reliable than frequency-based method. Therefore, we can conclude that the performance of termhood-based metric method is better than frequency-based approach.

5 Evaluation

Furthermore, we verify the validity of the comparability measure of termhood-based metrics by the task of bilingual term extraction.

In this section, we need to learn whether comparable corpus with high comparability can generate bilingual terminology with high quality. Therefore, our experiment is designed to extract bilingual terminology from three corpora with different comparability. These corpora are parallel corpus, comparable corpus and non-comparable corpus respectively. Again, it is assumed that comparability of comparable corpus lies between parallel and non-comparable corpus. We expect that the higher corpus comparability is, the higher quality bilingual terminology we can obtain.

5.1 Methods of Bilingual Term Extraction and Evaluation

The method of terminology extraction in our experiment is one of the most popular methods, namely, context vector-based method [18], which includes the following steps.

1) Preprocessing. For Chinese corpus: segmentation and part of speech. For English: stemming and part of speech;
2) Generating candidate monolingual terminology;
3) Creating context vector of monolingual terminology based on co-occurrence statistics;
4) Translating Chinese context vector to Eng-lish through bilingual dictionary from LDC [19];
5) Computing similarity of context vector in singular space of English language.
6) Extracting terminology pairs of which similarity larger than the given threshold.
7) Evaluation of terminology quality.

We use Top@N method to evaluate the result of bilingual terminology extraction. That is, for every Top@N English terminology together with N candidate Chinese terms, if there is one is the right translation, we consider the result is correct. Here we

take 10 for N. In the evaluation criteria of human judgments, if translation relation is completely correct, the score of matching will be 1; partly correct, score will be calculated by dice coefficient [20]; completely incorrect, score will be zero.

$$\text{Dice} = \frac{2 * overlaps}{\text{sum of number of tokens}} \tag{2}$$

5.2 Results and Analysis

We use two indicators for overall analysis, the overall similarity of terminology pairs and the overall degree of matching. Table 2 is the result of evaluation. Notice that the overall similarity of term pairs is increasing with the growth of comparability.

Table 2. Results of Evaluation

Experiment Corpus	comparability	similarity	degree of matching	
			Machine	Human
non-comparable	0.0478	0.2414	0.0480	0.0612
comparable	0.4226	0.3237	0.0395	0.0944
parallel	0.9527	0.3792	0.0566	0.0953

According to table 2, the overall degree of matching obtained by machine discrimination is not always increasing with the growth of comparability. In fact, machine discrimination is carried out with the help of dictionary of LDC. As far as we know, LDC's dictionary is a general dictionary which only includes about 80, 000 pairs bilingual lexicons, but the corpus in our experiments is corpus in special domain. So it is inevitable that there is some deviation because of limited size of bilingual dictionary. Therefore, human evaluation is very necessary. Notice that the overall degree of matching obtained by human judgments is increasing with the growth of comparability.

We can also find out that the overall degree of matching, no matter machine or human, is very low. This could well be due to the limited size and quality of bilingual dictionary, small size of our experimental corpus, or the bias caused by parameter settings in terminology extraction based on context vector method. At the same time, the overall similarity of terminology pairs is only related to terminology frequency. It is not affected by other factors. Therefore the result should be more reliable than the overall degree of matching.

In summary, we can conclude that the overall similarity of term pairs and the overall degree of matching is increasing with the growth of comparability. Accordingly, we can learn that high comparability of corpus generate high quality bilingual terminology. This also shows that our termhood-based method of comparability is effective in the application of bilingual terminology extraction.

6 Conclusions and Future Works

In this paper, we proposed termhood-based method to measure comparability of comparable corpus. Experiment results showed that this method performs better than traditional frequency-based method. It is likely to be that the candidate terms are ranked more reasonable because of constrain of termhood. It is possible that this method is less affected by common words, and it considers quality of term in special domain, so its performance is more stable and better in the task of terminology Extraction.

Experiments also show that results of comparability are more reliable when using keywords not full-text in the frequency-based method. This is because when using full-text data after preprocessing, there are so many common words that they influence the rank lists, further causing inaccurate results. Regardless of keywords or full-text data, the results are consistent in the termhood-based method. This again shows that termhood method has a better stability than frequency method.

Along the direction of our current work there are some directions for future works. One is to measure the effectiveness of termhood-base method in a more fine-grained comparable level. Another piece of work is to filter out the stop and common words after calculating word frequency, and then calculates the similarity of two words sequence; and finally we can make a comparison between this improved frequency-based method and termhood-based approach.

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grant No. 70903032.

References

1. McEnery, A.M., Xiao, R.Z.: Parallel and comparable corpora: What are they up to. In: Proceedings of Incorporating Corpora: Translation and the Linguist Translating Europe Multilingual Matters, Clevedon, UK (2007)
2. Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., Keskustalo, H.: Creating and Exploiting a Comparable Corpus in Cross - Language Information Retrieval. J. ACM Transactions on Information Systems (TOIS) 25(1), 322–334 (2007)
3. D1.3: Final report on criteria and metrics of comparability and parallelism and their suitability, http://www.accurat-project.eu/index.php?p=deliverables
4. Dejean, H., Gaussier, E., Sadat, F.: Bilingual terminology extraction: an approach based on multilingual thesaurus applicable to comparable corpora. In: Proceedings of COLING 2002, Tapei, Taiwan, pp. 218–224 (2002)
5. Kilgarriff, A.: Using word frequency lists to measure corpus homogeneity and similarity between corpora. In: Proceedings of the Fifth Workshop on Very Large Corpora, Hong Kong, China, pp. 231–245 (1997)
6. Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., Laurikkala, J.: Focused web crawling in the acquisition of comparable corpora. J. Information Retrieval 11(5), 427–445 (2008)
7. Leturia, I., Vicente, I.S., Saralegi, X.: Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In: Proceedings of the Fifth Web as Corpus Workshop (WAC5), Basque Country, Spain, pp. 53–61 (2009)

8. Otero, P.G., López, I.G.: Wikipedia as Multilingual Source of Comparable Corpora. In: Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010, Malta, pp. 21–25 (2010)
9. Vasiļjevs, A.: ACCURAT: Metrics for the evaluation of comparability of multilingual corpora. In: Proceedings of the Workshop on Methods for the Automatic Acquisition of Language Resources and their Evaluation Methods, LREC 2010, Malta (2010)
10. TTC Project, http://www.ttc-project.eu/releases
11. Li, B., Gaussier, E.: Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, pp. 644–652 (2010)
12. CNKI, http://www.cnki.net/
13. EBscohost, http://www.ebscohost.com/
14. Google Translate, http://translate.google.cn/
15. ICTCLAS, http://ictclas.org/
16. Kit, C.Y., Liu, X.Y.: Measuring mono-word termhood by rank difference via corpus comparison. J. Terminology 14(2), 204–229 (2008)
17. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM, Malta (1975)
18. Fung, P.: A Statistical view on Bilingual lexicon extraction: From Parallel Corpora to non-parallel corpora. In: Proceedings of Jean Veronis. Parallel Text Processing (2000)
19. LCD, http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2002L27
20. Kondrak, G., Marcu, D., Knight, K.: Cognates Can Improve Statistical Translation Models. In: Proceedings of HLT-NAACL 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (2003)

Corpus-Based Statistics of Pre-Qin Chinese

Bin Li[1,2], Ning Xi[2], Minxuan Feng[1], and Xiaohe Chen[1]

[1] Research Center of Language and Informatics,
Nanjing Normal University, 210097 Nanjing, China
fennel_2006@163.com, chenxiaohe5209@126.com
[2] State Key Lab for Novel Software Technology,
Nanjing University, 210023 Nanjing, China
{lib,xin}@nlp.nju.edu.cn

Abstract. The Pre-Qin Chinese plays a key role in the history of Chinese. However, for the lack of annotated corpus, the overview of Pre-Qin Chinese vocabulary is still not clear. This paper introduces the corpus of 25 Pre-Qin classical texts, which are under manual word segmentation and part-of-speech tagging. Then, the character and word frequencies are calculated based on the corpus. The character entropy, the syllables of words and the multiple part-of-speech words are also statistically analyzed.

Keywords: Chinese information processing, Pre-Qin Chinese, lexical statistics, multiple part-of-speech word.

1 Introduction

The Pre-Qin Chinese classical texts are of great importance in the history of Chinese language. There are about 60 texts before the Qin dynasty (221BC) inherited to today. But there is no free database which can be used to get the statistical data of the Pre-Qin Chinese. We select 25 most important literatures to construct a corpus of Pre-Qin Chinese. The 25 literatures have been manually segmented and annotated word part-of-speech (POS) in the past 4 years [1-2]. Based on the corpus, it is easy to get a detailed overview of the Pre-Qin Chinese. For example, we can see the basic characters and words of Pre-Qin Chinese, as well as the character entropy and POS distributions.

2 Related Work

There have been many researches done on Pre-Qin Chinese, which can be divided into 3 types. Most researchers focus on the pronunciation, shape and sense of characters and words in specific literatures like [3-5]. Several scholars have made brief descriptions of the vocabulary of Pre-Qin Chinese by summarizing these works [6], but still in lack of statistical data support. In recent years, the Academia Sinica has constructed a corpus of Pre-Qin Chinese containing 20 classical literatures [7]. However, this important resource only supplies online queries, and has not been used to get a

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 145–153, 2013.

statistical overview of the Pre-Qin vocabulary by the developers. In short, there is still not a deep statistical research on Pre-Qin Chinese, which can give us the clear sight of distributions on the characters, words, part-of-speeches, etc.

On the other hand, the basic words in Chinese are not clear yet. As [8-9] pointed out, the basic words in Chinese should be used widely along the long history(about 3000 years). So it is necessary to investigate the Chinese in different times. Now, by investigating the annotated Chinese texts in Pre-Qin, it is convenient to get the most frequent characters and words, to see which characters and words appear in every literature, and to know which words have more than one POS tags, etc.

Therefore, we get the opportunity to get an overview of the Pre-Qin Chinese, and to find the basic words of Pre-Qin Chinese.

3 Statistics of Characters in Pre-Qin Literatures

Before elaborate discussion, we first give the statistics of characters in 25 Pre-Qin literatures. There are totally 1,334,780 character tokens (7,049 character types) in these literatures, of which *Zuo Zhuan* containing 179,814 character tokens (3,312 character types) makes up the largest part.

Table 1. The 25 literatures and their character distributions

Literatures	Char tokens	Char types	Most frequent char "之zhi" as default
左传 *Zuo Zhuan*	179814	3312	7260
管子 *Guanzi*	127901	2876	5873
韩非子 *Hanfeizi*	105835	2713	5325
吕氏春秋 *Lv Shi Chun Qiu*	101452	3010	4838
礼记 *Liji*	97994	2999	4129
墨子 *Mohism*	79394	2458	4000
荀子 *Xunzi*	72584	2649	3780
国语 *Guo Yu*	72407	2586	3313
仪礼 *Yili*	71342	1507	于yu：1831
庄子 *Zhuangzi*	64744	2888	3090
周礼 *The Rites of Zhou*	49238	2167	2519
公羊传 *Gongyang Zhuan*	44366	1642	也ye：1496
谷梁传 *Guliang Zhuan*	40913	1593	也ye：2128
晏子春秋 *Yanzi Chun Qiu*	40836	2020	1889
孟子 *Mengzi*	35389	1897	1900
诗经 *Book of Poetry*	30954	2806	1176
尚书 *Shang Shu*	28146	1995	惟wei：680
周易 *Book of Changes*	21152	1363	也ye：961

Table 1. (*Continued*)

商君书 *Shang Jun Shu*	20035	1195	781
论语 *The Analects*	15935	1349	子zi：973
楚辞 *Chu Ci*	15262	2360	兮xi：1052
孙子兵法 *The Art of War*	6707	772	364
道德经 *Daoism*	5850	816	275
吴子 *Wuzi*	4729	851	187
孝经 *Xiao Jing*	1801	373	92
SUM	1334780	7049	之zhi：56862

3.1 High Frequency Characters and General Characters

In order to investigate how characters are frequently and widely occurred in the 25 pre-Qin literatures. We listed the most frequent characters and the characters which occur in all literatures (general characters) respectively in Table 2 and Table 3.

Table 2. The list of the most frequent 100 characters in 25 Pre-Qin literatures, sorted by the global counts of occurrence in the 25 literatures

Top-100 frequent characters
之、不、也、而、以、其、子、曰、人、者、有、為、則、於、公、君、無、大、故、王、天、所、于、夫、可、是、國、下、矣、民、何、與、乎、上、事、使、行、三、知、一、此、能、侯、言、必、如、得、若、謂、臣、主、道、二、焉、非、齊、十、將、吾、用、諸、然、我、月、在、至、士、禮、命、自、中、師、晉、出、見、五、四、日、皆、生、欲、乃、死、今、成、及、先、從、明、治、相、亦、令、利、入、後、德、食、小、正

Table 3. The 132 general characters occurring in all 25 Pre-Qin literatures, sorted by the frequency in decsending order

Also top-100 frequent characters	Characters of lower frequences
之、不、也、而、以、其、子、曰、人、者、有、為、則、於、公、君、無、大、故、王、天、所、夫、可、是、國、下、矣、民、與、乎、上、事、行、三、知、一、能、侯、言、必、如、得、若、謂、道、非、將、用、然、在、至、士、命、自、中、師、見、五、四、日、皆、生、死、成、及、先、從、明、治、相、利、後、食、小	年、未、百、立、聞、地、義、善、心、時、受、文、服、長、重、來、惡、取、足、敢、眾、雖、又、亡、周、已、親、政、失、名、寡、家、陳、安、萬、舉、過、復、居、易、止、始、進、致、高、和、興、退、觀、姓、加、順、厚、離、畏、深、要

As we can see from Table 2, without distinguishing the traditional and the simplified Chinese characters, many of the listed characters are also used frequently in modern Chinese. "之zhi" is the most frequent character in almost all literatures, with several exceptions in literatures which have distinctive features(also in Table 1).

As an analogy to "的de" in modern Chinese, "之zhi" in the most frequent character in Pre-Qin literatures, indicating the importance of auxiliary character in both Pre-Qin and modern Chinese. Interestingly, we observed two phenomena:

- Among the most frequent 100 characters, 25 characters surprisingly do not occur in all the literatures. A further analysis tells us it is not quite hard to interpret the absence of most of these characters. For example, "于yu" ranks in the 23rd, which doesn't appear in literatures with smaller size such as *Shang Jun Shu*, *The Art of War* and *Wuzi*. However, it is incredible that "此ci" doesn't appear in The Analects. Whether "此ci" really occur in the book or not need further exploration by bibliographers. As a proverb goes, "It's easier to verify the existence of one thing than to verify the non-existence of it", it tells us that in the traditional research work on linguistics, years of accumulation and solid work on card records are required in order to verify the existence or absence of a character. However, our statistics can help us quickly find out this kind of character which is frequently used but not widely used.
- The general characters are themselves of high frequency, but they are not necessarily distributed uniformly. Basically, the most globally frequent characters such as "之zhi" and "不bu" are the most frequent character in each literature. But the less globally frequent characters are not necessarily frequent in each literature. For example, "要yao" ranks the 700th with a total 296 occurrences. It occurs 48 and 38 times in *Lv Shi Chun Qiu* and *Xunzi* respectively, but only once in both *Daoism* and *The Art of War*. This non-uniform distribution reflects the diversity of these literatures in domains, ages, and the writing styles. For instance, the modal particle "兮xi" is the most frequent character in *Chu Ci*, indicating it is a literature of songs with regional characteristics; while the auxiliary character "也ye" occurs the most frequently in *Gongyang Zhuan* and *Guliang Zhuan*, indicating their special writing styles; "子zi" occurs the most frequently in *The Analects*, which can be attributed to the most frequent pattern "子曰ziyue" indicating that the literature was accomplished by Confucius and his students.

As can be seen in Table 3, "之zhi" is not only frequently used, but also widely used. We examined the frequency of each character in Table 3, and found that the corresponding ranks in Table 1 drop very quickly. The last character "要yao" even ranks 700th. We also found that the 132 general characters range over all main categories of character. However, the list of the general character is small, that's because the number of the character tokens in *Daoism*, *The Art of War*, *Wuzi* and *Xiao Jing* are no more than 10,000. Conversely, given limited literatures, we still collect 132 general characters, signifying the importance of these characters.

3.2 The Entropy of Pre-Qin Chinese Character

The entropy of Chinese character has long been the basic issue in Chinese information processing. In information theory, the entropy is a measure of the uncertainty of a random variable. The higher the entropy, the more the uncertainty of the random

variable. In the case of language, let x be a Chinese character, and $P(x)$ be the probability of x, $P(x)$ is simply estimated as the relative frequency of x, i.e. the occurrences of x in the corpus divided by the total number of characters in the corpus. The entropy is defined as follows,

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x) \ . \quad (1)$$

Based on Eq. 1, the entropy on the Pre-Qin Chinese characters in 25 Pre-Qin literatures is 9. 227, compared the results by the entropy on the modern Chinese character as 9.65[10]. We also compute the entropy of the characters in the corpus 199801 (January of 1998 of People's Daily, containing 1,589,735 Chinese character tokens, 4,574 types), which is 9.655, a score which is quite close to 9.65 given in [10]. It shows that the entropy of Pre-Qin Chinese is a little lower than that of Modern Chinese.

4 Statistics of Words in Pre-Qin Literatures

4.1 Distribution of Frequency

In this section, we give the word distribution of 25 literatures, where the auxiliary word "之zhi" is gain the most frequent word. There are 89 words occurring in all literatures (*general word*), and the number is 43 less than that of the general character. Except a two-character word "天下tianxia", all words are single-character words as can be seen in Table 4. We also see that the verbs make up the largest part, followed by nouns and adverbs. We further count the number of words occurring in 20 and 10 literatures respectively, the number is 654 and 28,780, which indicates a weak homogeneity among these literatures.

Table 4. The 89 general words, sorted by the frequency in decsending order

之/u、不/d、也/y、而/c、之/r、其/r、以/p、曰/v、者/r、有/v、則/c、為/v、於/p、人/n、可/v、無/v、矣/y、所/r、君/n、民/n、是/r、國/n、以/c、必/d、乎/y、能/v、得/v、謂/v、知/v、一/m、如/v、三/m、行/v、事/n、大/a、道/n、天下/n、子/n、用/v、皆/d、見/v、至/v、天/n、從/v、言/v、無/d、聞/v、立/v、未/d、在/v、死/v、受/v、治/v、士/n、四/m、五/m、日/n、生/v、心/n、取/v、地/n、成/v、雖/c、及/v、來/v、相/d、命/n、又/d、失/v、時/n、亡/v、食/v、舉/v、行/n、後/d、家/n、進/v、居/v、致/v、服/v、退/v、興/v、過/v、觀/v、加/v、親/v、死/n、和/v、陳/v

4.2 Distribution of Word Length

Here we give the distribution over word length. Conventional wisdom holds that the single-character words predominate in word vocabulary in ancient Chinese especially the Pre-Qin Chinese. However, the conventional practice confuses word token and word type. Therefore, we list the word length distribution in both the number of word tokens and word types in the 25 literatures as well as the 199801 corpus, as show in Table 5.

Table 5. Word length distribution in 25 literatures and the 199801 corpus

Word length	25 literatures		199801 corpus	
	#Word tokens	#Word types	# Word tokens	#Word types
1	15867	1074180	4547	354253
2	24929	104299	35494	497496
3	2879	8305	10687	48672
4	713	1373	5682	20401
5	19	26	729	2463
>5	0	0	583	926
Average	1.74	1.11	2.39	1.73

As listed above, we see that: 1) The number of the multiple-character word tokens surpasses that of the single-character word tokens. However, the number of the multiple-character word types is less than that of the single-character words. It shows that the multiple-character words dominate the vocabulary as early as Pre-Qin period. Removing the numerals, person names, place names, proper nouns and time-related words from the set of multiple-character words, there are still 17,505 multiple-character word types, which account for more than half of the total word types.

2) The average length of word types is close to 2 characters, and single-character word only takes 35.8%. However, the tokens of single-character word take 90.4% of all tokens, and the average length of word tokens is only 1.1 characters. The single-character word tokens obviously surpass the multi-character word tokens.

The average length of word types of 199801 corpus is 2.39 characters, and the average length of word tokens of it is 1.73 characters. Thus, in modern Chinese, the multi-character words dominated in the number of word tokens and types. On the contrary, in pre-Qin Chinese, the multi-character words only dominated in the number of word types.

4.3 Distribution of Words' Part-of-Speech

The POS of Pre-Qin Chinese words is a heated topic in the studies of ancient Chinese. We designed a tagging standard with 27 POS tags (see Table 6, excluding the tag for punctuations). We do not follow the POS principle that one POS has multiple syntactic functions suggested by [11] which is always used in annotating the modern Chinese texts. Instead, we apply the principle suggested by [12]: one POS corresponds to certain syntactic functions, and the context determines the POS of a word. Then it is convenience for us to annotate the POS tag for each word and easy to see the syntactic function of each word in context. We also designed 6 tags for the 6 typical transfer usage of words, cause to do(vs), think to be(vy), do for(vw), noun as adverb(zn), adjective as adverb(za) and verb as adverb(zv). And there is one tag for the special combined word like 诸(zhu, it to) equals two words 之(it) and 于(to).

We can see 2 points in Table 6. Firstly, in word types, nouns, verbs and person names are the top three most frequent words. The number of prepositional words, classifiers and modal words are both over 100, which means the three kinds of words

have been developed in Pre-Qin. Secondly, in word tokens, verbs, nouns and pronouns are the top three most frequent words. Verbs have more tokens than nouns, although verbs have less word types. Person names do not take great part of word tokens, while the pronouns do. Word tokens of classifiers are so limited that this kind of word is not widely used in Pre-Qin literatures. The number of transfer usage of words and combined words are not large.

Table 6. Distribution of words' part-of -speech in 25 literatures

Tag	POS	Word types	Ratio %	Word tokens	Ratio %	Tag	POS	Word types	Ratio %	Word tokens	Ratio %
a	adj	3226	7.6	39293	3.3	**r**	**pron**	**288**	**0.7**	**104505**	**8.7**
c	conj	223	0.5	64887	5.4	s	onoma-topoeia	95	0.2	134	0.0
d	adverb	1403	3.3	86768	7.2	t	time	396	0.9	9655	0.8
f	loca-tive	108	0.3	11080	0.9	u	aux	104	0.2	37120	3.1
i	affix	25	0.1	425	0.0	**v**	**verb**	**7324**	**17.3**	**346036**	**28.7**
j	com-bined	9	0.0	1522	0.1	vs	cause to do	348	0.8	1016	0.1
m	number	379	0.9	23114	1.9	vw	do for	32	0.1	89	0.0
n	**noun**	**17228**	**40.7**	**312898**	**26.0**	vy	think to be	122	0.3	555	0.0
nr	**person**	**7320**	**17.3**	**41860**	**3.5**	x	other	13	0.0	19	0.0
ns	loca-tion	2418	5.7	21145	1.8	y	modal	163	0.4	52214	4.3
nx	other entity	481	1.1	1410	0.1	za	adj as adv	71	0.2	162	0.0
p	prepo-sitional	111	0.3	45158	3.7	zn	noun as adv	215	0.5	1298	0.1
q	clas-sifier	160	0.4	2156	0.2	zv	verb as adv	64	0.2	313	0.0
--	--	--	--	--	--		**SUM**	42326	100	1204832	100

In table 7, we also give the distribution of multi-POS words in 25 literatures. In order to avoid the influence by annotation errors, a limitation was set for the statistics of multi-POS words. It is demanded that each POS tag should have more than 3 tokens for each multi-POS word. The number of multi-POS word types increases while the number of its POS tags decreases. There are 7895 word types and 1075193 word tokens, which take 17.6% and 89.2% of all words respectively.

Table 7. Distribution of multi-POS words in 25 literatures

# Multi-POS words	Word types	POS word types	Word tokens	Example word
11	2	22	4299	然ran
10	2	20	5652	若ruo
9	5	45	29600	上shang
8	17	136	112892	如ru
7	43	301	72514	小xiao
6	72	432	137173	止zhi
5	175	875	163248	故gu
4	301	1204	171679	夜ye
3	634	1902	182650	罪zui
2	1479	2958	195486	雷lei
Sum	2730	7895	1075193	

Table 8 shows the top 10 most frequent words in 25 literatures. The word 然(ran) and 重(chong, zhong) both have 11 POS tags. These words are frequently used in Pre-Qin literatures and are important for language learners.

Table 8. top 10 most frequent words in 25 literatures

Word type	# POS	# Tokens	POS tags（descending by frequencies）
然 ran	11	2797	r、c、v、i、a、n、d、u、x、y、nr
重 zhong	11	1502	a、v、n、d、q、nr、vy、vs、ns、za、m
若 ruo	10	3755	v、c、p、r、d、nr、n、i、y、a
後 hou	10	1897	d、n、f、v、a、t、c、p、vs、zn
上 shang	9	3960	f、n、v、a、zn、d、m、vs、r
是 shi	9	5559	r、v、u、n、a、d、c、i、p
為 wei	9	12920	v、p、n、c、u、a、r、y、d
于 yu	9	6694	p、u、v、i、y、n、c、d、a
厥 que	9	467	r、i、u、c、v、d、n、nr、y
如 ru	8	4057	v、p、c、i、y、r、d、n

5 Conclusions and Future Work

In this paper, we give the statistical distributions of characters and words based on the Pre-Qin corpus which include 25 classical literatures. And 7 conclusions are made to describe the properties of Pre-Qin Chinese. 1) The literatures differ a lot from each other. The characters and words appear in all the 25 literatures are only 132 and 89

respectively. 2) The most frequent character and word in the corpus is the same one, "之zhi", but it is not the same in every literature. 3) The entropy of characters in the corpus is 9.227, which is a little lower than modern Chinese. 4) The multi-character words take advantage in word types, while the single-character words take advantage in word tokens. 5) The most frequent word types are nouns, verbs and person names, while the most frequent word tokens are verbs, nouns and pronouns. 6) The modal words and classifiers which are the exclusive words in Chinese have been developed in Pre-Qin times. 7) The multi-functional words occur frequently in the corpus, which takes 89.2% in word tokens.

In the future, we have to continue the work of checking errors of the corpus, and release it to public as soon as possible. Second, we will enlarge the corpus by adding other literatures of Pre-Qin. Third, we will try to annotate the sense of each word in the corpus, which would be very helpful to study the development of Chinese in history.

Acknowledgments. We thank anonymous reviewers for their constructive suggestions. This work is the staged achievement of the projects supported by National Social Foundation of China (10&ZD117, 10CYY021) and the research base of Philosophy and Social Sciences for Universities in Jiangsu (2010JDXM023) and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Chen, X.H.: Information Processing of Pre-Qin Chinese. In: The 27th Anniversary of Chinese Information Processing Society of China, Beijing (2008)
2. Shi, M., Chen, X.H., Li, B.: CRF Based Research on a Unified Ap-proach to Word Segmentation and POS Tagging for Pre-Qin Chinese. Journal of Chinese Information Processing 2(24), 39–45 (2010)
3. Zhang, S.D.: Vocabulary Study of Lv Shi Chun Qiu. Shandong Education Press, Jinan (1989)
4. Chen, K.J.: Dictionary of Chunqiu Zuozhuan. Zhongzhou Ancient Books Publishing House, Henan (2004)
5. Che, S.Y.: Vocabulary Study of Hanfeizi. Bashu Publishing House, Chengdu (2008)
6. Ye, Z.B.: Vocabulary Study of Archaic Chinese. The Central Literature Publishing House, Beijing (2007)
7. Academia Sinica Tagged Corpus of Old Chinese, http://old_chinese.ling.sinica.edu.tw/#
8. Pan, Y.Z.: The Formation and Development of Chinese Basic Vocabulary. Journal of Zhongshan University 1, 98–121 (1959)
9. Zhou, J.: Distinction between Basic Vocabulary and General Vocabulary. Journal of Nankai University 3 (1987)
10. Feng, Z.W.: The Entropy of Chinese Characters. Revolution of Chinese Characters, 12–17 (1984)
11. Zhu, D.X.: Lecture Notes on Grammar. The Commercial Press, Beijing (1983)
12. Li, J.X.: The New Chinese Grammar. The Commercial Press, Beijing (1924)

Automatic Acquisition of Chinese Words' Property of Times

Liu Liu[1], Bin Li[1,2], Lijun Bu[1], Tian-tian Zhang[1], and Xiaohe Chen[1]

[1] Research Center of Language and Informatics,
Nanjing Normal University, Nanjing, China, 210097
{liuliu1989,libin.njnu,bljpaulauster}@gmail.com,
799548719@qq.com, chenxiaohe5209@126.com
[2] State Key Laboratory for Novel Software Technology,
Nanjing University Nanjing, Nanjing, China, 210046

Abstract. Words' property of times is an important type of additional meaning which represents the spirit of times. People get the information of times from words by their own experience, but automatic recognition by computers is still difficult. This paper proposes a method of automatic recognition of the property of times based on large-scale corpus, which uses the TF-IDF and TF-IWF values to quantify Chinese words' property of times. Experiments on People's Daily of 54 years show that words' TF-IDF values aided with TF-IWF value outperform words' frequency. Naïve Bayes classifier is also used in for automatic acquisition of words' property of times, and it achieves satisfactory results.

Keywords: Property of Times, Frequency, TF-IDF, TF-IWF, Naive Bayes.

1 Introduction

Property of times, as an important part of additional meanings of words, shows the society's changing and developing of particular times. For example, words like "红卫兵 Red Guard", "帝国主义 imperialism", "解放 liberation" and "反动派 reactionaries" remind us of "the old days" from year 1949 to 1977, while words like "极客 geek", "博客 blog", "蓝牙 Bluetooth" and "笔记本 laptop" remind us of "nowadays". So called "the old days" and "nowadays" are words' properties of times. Property of times is epitomes of the society, closely related to people's daily life. For human, it is very easy and natural to recognize a word's property of times. However, computers lack the knowledge hierarchy and life experience and this causes a great difficulty for computers to get the relations between words and the property of times. The understanding of additional meanings of words is a very hard task in natural language understanding. Property of times contributes an important part of additional meanings of words. To provide a solution that can make the effective automatic acquisition of words' additional meanings will be helpful to the semantic research in nature language processing. And the automatic judgment of a text's times will be possible.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 154–165, 2013.

We use People's Daily (year 1946-1999, 506,131,918 words) to build a large-scale corpus, research in words' frequency, TF-IDF values and TF-IWF values to find the most appropriate method in weighting words' property of times.

2 Related Works

Words' property of times has been researched separately in two main fields. Traditional linguists think "language societies usually embody the language factor that can be perceived directly with timer shaft. For example, people will feel that the old fashioned language factors are outmoded, the latest language factors are fashionable." [1]. Words' property of times can be seen as a kind of color from outside the words. The color of times "is the special atmosphere and flavor of particular times that words reflect, from which we can see the social's changings." The words which have the colors of times "must reflect the important social historical substances." Words colors of times will change as the usages change, and we can learn the variation tendency of times through the changing of these colors of times. The colors of times as a part of the meanings of words are depending on three aspects: the particular times, the high frequency of words and the widely use of words [2]. The color of times has properties of high frequency, timeliness, selectivity, systematicness and assimilation of spoken and written language [3]. Words' property of times should not only based on their rational meanings but also be closely connected to their actual usages. Words' property of times can reflect to property of times themselves [4].

Computational linguists investigate this problem with the third generation corpus, "dynamic and circulated corpus". The new times that based on this kind of corpus which is called the cyber times has a goal to build a "self-actor" which is capable in "the ability of learning, feed backing and controlling of language". This "self-actor" will change as the social changes [5]. The dynamic and circulated corpus is a dynamic large-scale corpus of real language material. This sort of corpus can be used to solve many problems that traditional way cannot deal with. National Language Resources Monitoring and Research Center (NLRMRC) used the major newspapers in China to build a dynamic and circulated corpus, and they started a research of Chinese catchwords based on it [6]. LIVAC Synchronous Corpus is a typical dynamic and circulated corpus which consists of newspapers of Hong Kong, Taiwan, Beijing, Shanghai, Macao and Singapore. It can be used to find new properties of Chinese words [7]. Management System of Broadcast Media Language Corpus is also an open dynamic and circulated corpus. They processed on the resource of 15871 broadcast and TV programs from 2008 to 2010. The corpus can be used for searches on general and special purposes [8]. Google made public an Ngram Viewer database based on Google Book. This database provides the distribution situation of words frequency of several languages including Chinese [9].

We can see from above that traditional linguists had done lots of meaningful researches in words' property of times, but they mostly focused on the property, no more further research on how to use it. The idea of dynamic and circulated corpus can help to solve this problem, but we can only find dynamic and circulated corpus used for simple searches and catchwords findings. No research has focused on the acquisition of words' property of times and that's what we do in this research.

3 The Corpus

Our corpus in the research is People's Daily Corpus based on People's Daily Newspaper resource of 54 years (from year 1946 to 1999, 506,131,918 words). We use ICTCLAS [10] to do the automatic word segmentation and the part-of-speech tagging. Then we count the occurrence frequency of each word like table 1. We then divide the 54 years into 6 times[1], and filter the words with a stop wordlist[2]. This table contains all the words which would have the property of times of each 6 times (see table 2).

Table 1. Word's frequency

Word	Part of Speech	Frequency
欢送 huansong send-off	v	0.002%
恢复 huifu resume	v	0.0193%
回国 huiguo home-come	v	0.0056%
会员 huiyuan member	n	0.0033%

Table 2. Word's frequency of 6 times

Word	Part of Speech	Times	Frequency
生产 shengchan produce	vn	1960s	0.2005%
国家 guojia country	n	1980s	0.2138%
群众 qunzhong the masses	n	1970s	0.3177%
和平 heping harmony	n	1950s	0.1433%

4 Quantification of Words' Property of Times Based on Frequency, TF-IDF and TF-IWF

To acquire words' property of times, we must acquire the words that have the property of times at first. Then we can quantify the information of times that the words contain, the quantified values can be seen as several extents on a temporal dimension, and the extents that have the largest values should be the words' property of times.

4.1 Quantification Based on Highest Frequency

We select 1000 words with highest frequency from table "word frequency of 6 times" and form 6 tables of each times that contain the these words.

[1] 1946-1949, 1950-1956, 1957-1965, 1966-1976, 1977-1988, 1989-1999. We'll call these times 1940s, 1950s, 1960s, 1970s, 1980s, 1990s, for convenient.

[2] The stop words form contains common empty words, numbers, punctuations and error information. Obviously we think these words don't have the property of times, so we should delete them to prevent errors and inconvenience that would be caused by these words in further researches. To make sure that our values in the latter experiment are accurate, we just delete these words but preserve their frequency.

Table 3. Words with highest frequency of the 1950s

Word	Part of Speech	Frequency
生产 shengchan produce	vn	0.2015%
苏联 sulian Soviet Union	ns	0.1791%
计划 jihua plan	n	0.1466%
和平 heping harmony	n	0.1433%

Since words' property of times has the feature of high frequency, we suppose that these words with highest frequency should most likely to have the property of times. We experiment on 100 words with highest frequency to find out whether these words have the property of times or not. The result is not so good (see table 12).

Table 4. Top 100 words with highest frequency of each times

Times	Number of Words	Percentage
1940s	4	4%
1950s	2	2%
1960s	1	1%
1970s	8	8%
1980s	6	6%
1990s	11	11%

We can see from table 4 that the number of words acquired in the highest frequency words is far from what we need to acquire the property of times. Many words that have the property of times may not have high frequency, and some common words that have high frequency do not have the property of times. So the acquisition of the words that have the property of times cannot easily base on words frequency.

4.2 Quantification Based on TF-IDF

TF-IDF (Term frequency & Inverse document frequency) is based on TF (term frequency). TF cannot solve the problem that some high frequency words' abilities of distinguishing the texts are weak, and some low frequency words' abilities of distinguishing the texts are strong. IDF (inverse document frequency) was suggested to solve this problem [11]. The TF-IDF shows that the higher a word's appearance frequency in a text is, the stronger the word's ability of distinguishing the text is(TF) and the more texts a word appears in, the weaker the word's ability of distinguishing

the text (IDF) is. The classical function to calculate the TF-IDF value is shown below:

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_i} \quad . \tag{1}$$

In this function, w_{ij} means the TF-IDF value of a word Ti in text D_j, tf_{ij} means the appearance frequency of the word T_i in text D_j, n_i means the the number of texts that word T_i appears in and N means all the texts that we have.

We see the People's Daily Corpus of 6 times as 6 texts and build 6 tables of each word's TF-IDF value of each times.

Table 5. Word's frequency and TF-IDF value of 1990s

Word	Part of Speech	Frequency	TF-IDF
乡镇企业 xiangzhenqiye township enterprises	n	0.0141%	6.75E-05
邓小平理论 dengxiaopinglilun Deng Xiaoping Theory	n	0.0069%	5.38E-05
经贸 jingmao trade	j	0.0138%	4.15E-05
调控 tiaokong regulation and control	vn	0.0077%	3.7E-05
旅游 lvyou journey	vn	0.0120%	3.61E-05

We can see from the table that words like "乡镇企业 xiangzhenqiye township enterprises/n", "邓小平理论 dengxiaopinglilun Deng Xiaoping Theory/n", "经贸 jingmao trade/j", "调控 tiaokong regulation and control/vn" have high TF-IDF values although their frequency are not high.

We experiment on 100 words of each times with highest TF-IDF values to find out whether these words have the property of times or not. The result turns out much better than the result of the one with highest frequency (see table 13).

Table 6. Words with the property of times (Frequency and TF-IDF) of each times

Times	Frequency	TF-IDF
1940s	4%	7%
1950s	2%	3%
1960s	1%	7%
1970s	8%	15%
1980s	6%	12%
1990s	11%	26%

The TF-IDF way finds 38% more words than the highest frequency way. We can say the TF-IDF Value is much more appropriate for the quantification of words' property of times.

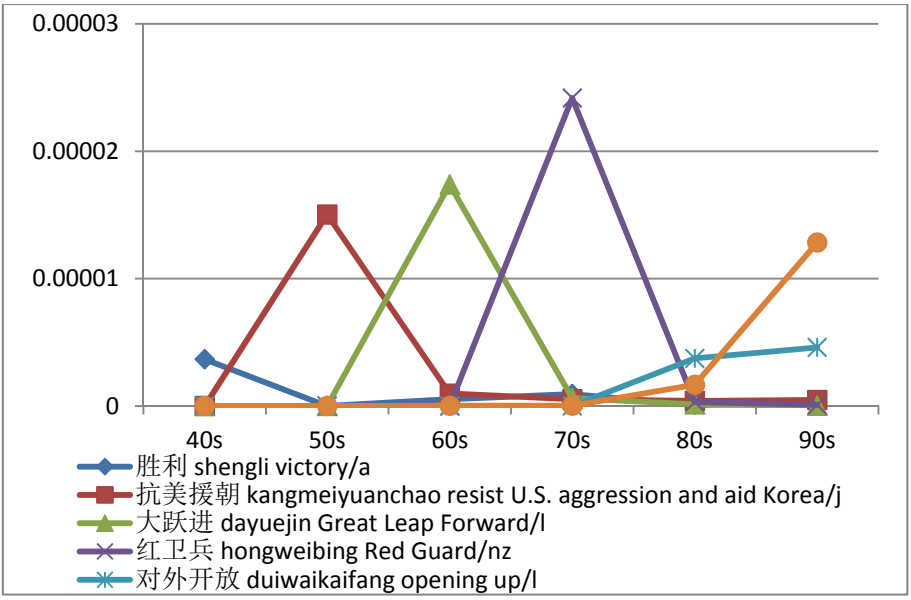


Fig. 1. This figure shows the quantification of words' property of times based on TF-IDF. We can find directly that the property of times of the word "对外开放opening up/l" is 1990s and "红卫兵 Red Guard/nz" is 1970s.

4.3 Quantification Based on TF-IWF

Using TF-IDF value, we can easily acquire the words that are more likely to appear at a particular time. That makes the TF-IDF way our main method to acquire the words which have the property of times and to acquire words' property of times.

But to compare the two ways above, we can find that words like "国民党 Kuomintang", "土地 tudi land", "帝国主义 diguozhuyi imperialism", "修正主义 xiuzhengzhuyi revisionism", "发展fazhan develop", "改革 gaige reform", "市场 shichang market" appear in the acquisition result of highest frequency way but not in the TF-IDF way. This is because our People's Daily Corpus is large that some words may appear in each times. This leads to a word's TF-IDF value may be 0 according to formula (1).

Quantification based on function 1 causes we lose words that appear in each times but still have the property of times. To solve this problem we adopt a way call TF-IWF that is based on TF-IDF as a supplementary way, a way using the logarithm of the reciprocal of a word's frequency to replace the IDF(IWF), and to use the square of IWF [12].

$$w_{ij} = tf_{ij} \times \left[\log\left(\frac{\sum_{i=1}^{M} nt_i}{nt_i}\right)\right]^2 . \tag{2}$$

In this function, M means the number of words. We build 6 tables of each word's TF-IWF value of each times based on formula (2).

Table 7. Word's frequency and TF-IWF value of 1970s

Word	Part of Speech	Frequency	TF-IWF
革命 gemin revolution	vn	0.3362%	0.03093
无产阶级 wuchanjieji proletariat	n	0.2934%	0.03182
斗争 douzheng fight	vn	0.2519%	0.02351
毛泽东思想 maozedongsixiang Maoism	n	0.1152%	0.01565
修正主义 xiuzhengzhuyi revisionism	n	0.1104%	0.01512

Table 8. Words with the property of times (Frequency, TF-IWF and TF-IDF) of each times

Times	Frequency	TF-IWF	TF-IDF
1940s	4%	7%	7%
1950s	2%	2%	3%
1960s	1%	1%	7%
1970s	8%	10%	15%
1980s	6%	6%	12%
1990s	11%	13%	26%

Table 9. The rank of word's frequency and TF-IWF value of 1990s

Word	Frequency	TF-IWF
经济 jingji economy/n	1	1
企业 qiye company/n	4	2
发展 fazhan develop/vn	5	3
建设 jianshe construct/vn	10	9
市场 shichang market/n	13	11
发展 fazhan grow/v	16	13
发展 fazhan grow/v	16	13
技术 jishu technology/n	21	25
国际 guoji International/n	24	17
改革 gaige reform/vn	56	35
生产 shengchan produce/vn	58	*[3]
科技 keji science and technology/n	74	37
管理 guanli administration/vn	*	31
香港 xianggang HongKong/ns	*	58

We experiment on 100 words of each times with highest TF-IWF values to find out whether these words have the property of times. The result shows that words that have the property of times among these 100 words are quite similar to the words acquired in the highest frequency way. The TF-IWF way finds 7% more words than the highest

[3] "*" means the word is not found in the top 100 words (Frequency and TF-IWF value).

frequency way (see table 14). But the TF-IWF way is far from good as the TF-IDF way which still finds 31% more.

To find out why the TF-IWF way can acquire more words that have the property of times than the frequency way, we have another experiment. We tag the rank of numbers on each word acquired in each way. The rank number is the rank of the word among the 100 words with highest frequency or highest TF-IWF values. We can find that words acquired in the TF-IWF way have the smaller rank number than the frequency ones in average.

Among the 30 words that both appear in the highest frequency way and TF-IWF way, 27 of them have the smaller rank number in the 100 TF-IWF value words, which take 90%. This means in a larger scale (1000 words for example), we will acquire much more words that have the property of times.

To sum above, we thought the TF-IWF value is much more appropriate to be the supplementary way to quantify the property of times.

5 Automatic Acquisition of Words' Property of Times

With the quantification results based on TF-IDF value and TF-IWF value, we can try to automatically acquire words that have the property of times.

5.1 Naive Bayes

We use Naive Bayes as the classifier to automatically decide whether a word has the property of times or not.

Naive Bayes is a classical classifier which uses the joint probability between the property items and classes to decide the class of a given text.

$$\mathbf{classify}(f_1, \ldots, f_n) = \underset{cp(C=c)\prod_{i=1}^{n} p(F_i=f_i|C=c)}{\mathbf{argmax}} . \quad (3)$$

5.2 Classification with Naive Bayes

We see each word with their TF-IDF value and TF-IWF value as a text to be classified. We prepared 1000 words (900 as the training set and 100 as the test set), and classified with the Naive Bayes, the result is shown below.

Table 10. The closed test results of classification with Naive Bayes

Class	Precision	Recall	F-Measure
1940s	0.893	1	0.943
1950s	1	1	1
1960s	0.923	0.8	0.857
1970s	0.962	0.595	0.735
1980s	0.5	0.345	0.408
1990s	0.879	0.974	0.924
Weighted Avg.	0.841	0.853	0.839

Table 11. The open test results of classification with Naive Bayes

Class	Precision	Recall	F-Measure
1940s	1	1	1
1950s	1	1	1
1960s	1	1	1
1970s	0.667	0.4	0.5
1980s	0	0	0
1990s	0.756	0.969	0.849
Weighted Avg.	0.651	0.76	0.694

The classified result turns out good enough so the automatic acquisition of words' property of times with Naive Bayes is efficient.

6 Conclusion and Further Works

This paper offers an efficient corpus-based investigation of Chinese words by focusing on the quantification and automatic acquisition of words' property of times in the People's Daily Corpus. Through the experiments on the words of each times from the corpus, we discovered that TF-IDF aided with TF-IWF value performs well to represent the property of times. The Naive Bayes classifier further enhances the results of automatic acquisition of time property. This method should also be useful on acquisition of other words' properties as well.

This is still a starting work that we do on discovering words' property of times. The corpus we use is only consisted of news. And the corpus' scale is not balanced. For example, the corpus scale of 1940s is only 10% of the 1990s. But despite the disadvantages we still offer a valuable investigation of Chinese words' property of times. We intend to improve our work in further researches with the application of a larger and more general corpus, such as Google books Ngram Viewer database. Then, we want to find and apply a more efficient classifier to help us make the acquisition work more efficient and accurate. At last, we will use words' properties of times in deep natural language understanding and generation applications.

Acknowledgments. We are grateful for the comments of the anonymous reviewers. This work was supported in part by National Social Science Fund of China under contract 10CYY021, State Key Lab. for Novel Software Technology under contract KFKT2011B03, China PostDoc Fund under contract 2012M510178, Jiangsu PostDoc Fund under contract 1101065C, Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions and Jiangsu Research and Innovation Program for Postgraduates of Ordinary Colleges and Universities under contract CXLX12_0357.

References

1. Jakobson, R.: Time Factor in Language. Collected Works of Roman Jakobson. The Commercial Press, Beijing (2012)
2. Yang, Z.N.: First exploration of words'property of times. Transactions of Shandong University (edition of philosophy and social science) 3, 102–106 (1988)

3. Shen, M.Y.: Discuss on principal properties of words' colors of times. Transactions of Inner Mongolia Nationality Normal University 3, 24–29 (1991)
4. Wang, J.H.: Words' colors of times and the usages of words. Theory and Modernization, 372–377 (2001)
5. Zhang, P.: On Cybernetics and Dynamic Updating of Language Knowledge. Applied Linguistics 4, 76–82 (2001)
6. National Language Resources Monitoring and Research Center. Broadcast Media Language Branch, http://ling.cuc.edu.cn/RawPub/Default.aspx
7. Research Centre on Linguistics and Language Information Sciences. Hong Kong Institute of Education: LIVAC Synchronous Corpus, http://www.livac.org
8. National Language Resources Monitoring and Research Center, http://cnlr.blcu.edu.cn/news_show.aspx?nid=286
9. Google. Google books.Ngram Viewer, http://books.google.com/ngrams/datasets
10. ICTCLAS, http://www.ictclas.org
11. Jones, S., Karen: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), 11–21 (1972)
12. Basili, R., Moschitti, A., Pazienza, M.: A text classifier based on linguistic processing. In: Proceedings of IJCAI 1999. Machine Learning for Information Filtering (1999)

Appendix

Table 12. Top 100 most frequent words with the property of times

1940s	1980s
苏联 sulian Soviet Union/n	经济 jingji economy/n
介石 jieshi Kai-shek/nr	发展 fazhan develop/vn
国民党 guomingdang Kuomintang/n	企业 qiye company/n
土地 tudi land/n	技术 jishu technology/n
1950s	发展 fazhan grow/v
朝鲜 chaoxian Korea/ns	改革 gaige reform/vn
合作社 hezuoshe artel/n	1990s
1960s	经济 jingji economy/n
帝国主义 diguozhuyi imperialism/n	企业 qiye company/n
1970s	发展 fazhan develop/vn
革命 gemin revolution/vn	建设 jianshe construct/vn
无产阶级 wuchanjieji proletariat/n	发展 fazhan grow/v
斗争 douzheng fight/vn	市场 shichang market/n
毛泽东思想 maozedongsixiang Maoism/n	技术 jishu technology/n
修正主义 xiuzhengzhuyi revisionism/n	国际 guoji International/n
阶级斗争 jiejidouzheng class conflict/l	改革 gaige reform/vn
文化大革命 wenhuadagemin great proletarian cultural revolution/nz	生产 shengchan produce/vn
贫下中农 pinxiazhongnong poor and lower-middle peasants/j	科技 keji science and technology/n

Table 13. Top 100 highest TF-IDF value words with the property of times in 6 times

1940s	乡镇企业 xiangzhenqiye township enterprises/n
苏联 sulian Soviet Union/n	专业户 zhuanyehu family that produces a special product/n
苏维埃 suweiai soviet/n	四化建设 sihuajianshe building of the four modernizations/l
介石 jieshi Kai-shek/n	四化建设 sihuajianshe building of the four modernizations/j
波茨坦 bocitan Potsdam/n	四化 sihua Four Modernizations/j
前线 qianxian battlefront/n	承包责任制 chengbaozerenzhi contract and responsibility system/n
胜利 shengli victory/a	耀邦 yaobang Yaobang/nr
伤俘 shangfu injured prisoners/n	联产承包 lianchanchengbao production-related contracting/l
1950s	经济特区 jingjitequ special economic zone/l
抗美援朝 kangmeiyuanchao resist U.S. aggression and aid Korea/j	第三产业 disanchanye tertiary-industry/l
农业社 nongyeshe agricultural society/n	外向型 waixiangxing export oriented/b
朝鲜战争 chaoxianzhanzheng Korean war/n	1990s
1960s	乡镇企业 xiangzhenqiye township enterprises/n
人民公社 renmingongshe people's commune/l	邓小平理论 dengxiaopinglilun Deng Xiaoping Theory/n
核武器 hewuqi nuclear weapon/n	经贸 jingmao trade/j
大跃进 dayuejin Great Leap Forward/l	调控 tiaokong regulation and control/vn
大跃进 dayuejin Great Leap Forward/nz	旅游 lvyou journey/vn
核战争 hezhanzheng nuclear war/n	发展中国家 fazhanzhongguojia developing country/l
鼓足干劲 guzuganjing strain oneself/i	一国两制 yiguoliangzhi one country two systems/j
多快好省 duokuaihaisheng Better and more economical/l	第三产业 disanchanye tertiary-industry/l
1970s	股份制 gufenzhi joint stock system/n
文化大革命 wenhuadagemin great proletarian cultural revolution/nz	镕基 rongji Rongji/n
贫下中农 pinxiazhongnong poor and lower-middle peasants/j	高新技术 gaoxinjishu high and new technology/n
走资派 zouzipai capitalist-roaders/n	媒体 meiti media/n
红卫兵 hongweibing Red Guard/nz	效益 xiaoyi benefit/n
样板戏 yangbanxi model opera/n	音像 yinxiang audio-visual/n
造反派 zaofanpai rebels/n	软件 ruanjian software/n

Table 13. (*Continued*)

上山下乡 shangshanxiaxiang go and work in the countryside or mountain areas/l	叶利钦 yeliqin Yeltsin/nr
霸权主义 baquanzhuyi hegemonism/n	廉政 lianzheng honest or clean politics/n
不结盟 bujiemeng nonalignment/vn	脱贫致富 tuopinzhifu4/l
牛鬼蛇神 niuguisheshen evil people of all kinds/i	电脑 diannao computer/n
夺权 duoquan seize power/v	开发区 kaifaqu development zone/n
忆苦思甜 Yikusitian5/l	集团公司 jituangongsi group company/n
三自一包 Sanziyibao6/j	影视 yingshi film and video/b
支农 zhinong support agriculture/vn	香港特别行政区 xianggantebiexingzhengqu The Hong Kong Special Administrative Region/n
三支两军 sanzhiliangjun three support's and two military's/j	高速公路 gaosugonglu highway/n
1980s	经济效益 jingjixiaoyi economic benefit/n
四人帮 sirenbang gang of four/j	

Table 14. The additional words with the property of times acquired by the TF-IWF way

1940s	1970s	1990s
解放军 jiefangjun People's Liberation Army/n	革委会 geweihui revolution committee/j	管理 guanli administration/vn
内战 neizhan civil war/n	反革命 fangemin counterrevolution/n	香港 xianggan Hong Kong/ns
陕北 shanbei north of Shanxi Province/ns		

[4] Cast off poverty and become better off.

[5] Call to mind past sufferings and think over the good times.

[6] More plots for private use, more free markets, more enterprises with sole responsibility for their own profit or loss, and fixing output quotas on a household basis.

A Study of English Word Sense Disambiguation Base on WordNet

Deng Pan

Foreign Languages School
Hubei University of Science and Technology,
Hubei, China
Dylandp@163.com

Abstract. WordNet is an important English lexical semantic knowledge base. Focusing on the application of WordNet on corpus-based translation method, this paper presents a hybrid strategy method based on several cyber-translation aids and Word-Net 3.0 to make an investigation of some English words translation versions from the aspect of word class, grammatical feature, understandability and loyalty. Finally, the writer illustrates the authentic usages and application rules of the words and proposes some measures to improve the quality of corpus-based MT in English-Chinese translation. Preliminary experiment indicates that the hybrid strategy improve the quality of the word translation in both general and scientific discourse.

Keywords: WordNet, Statistical Machine Translation, Word sense disambiguation.

1 Introduction

At present the domestic machine translation mainly has two kinds of patterns: one kind is the example-based MT. The basic idea is to reuse examples of already existing translations as the basis for a new translation. The process is broken down into three stages: matching, alignment, and recombination. Another kind is the statistical MT. The essence of the method is first to align phrases, word group, and individual words of the parallel texts, and then to calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in the other language. Neither the example-based nor the statistics-based approaches to MT have turned out to be demonstrably better than the rule-bade approaches, though each has shown some promise in certain cases. As a result of this, a number of hybrid systems quickly emerged.

Along with Google, Baidu, and Netease marching to web-translation market, the online translation market competition is becoming more and more popular, but their translation quality is not satisfactory, the understandability and loyalty of the translation is not good. With the help of WordNet, the translator can internalize the steps of doing professional translation and grasp the skills needed for the future web-translation market.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 166–174, 2013.

Based on the WordNet statistical machine English-Chinese Translation there are two major steps. The first step is the analysis possible senses of the vocabulary, because an English word may have multiple meanings. WordNet concept dictionary will classify different word class and senses for each word, it is means do the classification the meaning of the term of concept granularity; the second step is to model the translation process in term of statistical probabilities and use their examples if we take the input sentence, the amongst the possible translation. The essential feature is the availability of a suitable large bilingual corpus of reliable translation.

This paper firstly describes the domestic and foreign research on applications of WordNet in machine translation; secondly make an experience of two English translations form the level of syntactical and word sense disambiguation; then discussion of the case study is made; finally make the expectation of English word sense disambiguation base on WordNet.

2 Related Researches and Basic Concepts

With the development of language, the senses of vocabulary are becoming more and more colorful and the same time the collocation of the words are increasing, ever since. The importance of lexical sense is been remarked by Wang Yan. AS he mentioned, "Sense is not only the embodiment of thinking but the core of language communication".(Wang Yan,2001) But for the definition of meaning, there is not a comprehensive, widely accepted one. By far the most widely accepted Semantic classification is made by the British linguist Geoffrey Leech. He divided the broadest sense of "meaning" into seven different types: conceptual meaning, connotative meaning, social meaning and emotive meaning, reflective meaning, and thematic meaning. (G.Leech,1974) To learn a language, one only understand words' lexical meaning, but know litter about their associative meanings, who cannot be said to truly master the language, not to mention to grasp the correct use of the language.

2.1 The Introduction of WordNet

WordNet is a semantic dictionary that can convey the conceptual relations. It organizes messages of word through semantic meaning. By using synonymy set to convey concept, semantic relationship has been revealed through conceptions. It organizes English vocabulary into synonymy set and each one of them is marked with a lexical concept. Meanwhile, it tries to establish different standards between concepts and convey different semantic relationships. For example, noun phrases form their concept trees according to hierarchies. For an instance: oak@--->tree@--->plant@---> organism. The lower hierarchy concept reserves the overall property of the higher hierarchy concept. And thus the abstract concept has turning into concept tree and one can carry out connectional reasoning and calculation.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings (Grishman R,1986). However, there are some important distinctions. First, WordNet interlinks not just word forms strings of letters, but

specific senses of words(Graeme，2000).As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

2.2 The Introduction of MT

The term machine translation (MT) refers to computerized systems responsible for the production of translation with or without human assistance. A distinction is common only made between human-aided MT and machine-aided human translation. The latter comprises computer-based translation tools which support translators by providing access to on-line dictionaries, remote terminology databank, transmission and reception of text, stores of previously translated text, and integrated resources, commonly referred to as translation workstations or translator workbenches. The term computer-aided translation (CAT) is sometimes used to cover all these computer-based translation systems.

2.3 The Introduction of Statistical Machine Translation

During the last two decades, statistical machine translation has successfully outweighed other machine translation approaches in many evaluations, which makes it one of the hottest issues in the machine translation domain. Through a statistical analysis of the bilingual corpus, it is able to learn the translation knowledge automatically.

In its pure form, the statistics-based approach to MT makes use of no traditional linguistic data. The essence of the method is first to align phrases, word group, and individual words of the parallel texts, and then to calculate the probabilities that any one word in a sentence of one language corresponds to word or words in the translated sentence with which it is aligned in the other language. This approach is to model the translation process in terms of statistical possibilities: to use their example, if we take the input sentence (1), then amongst the possible translations are sentence (2a) and (2b).

Sentence 1 National Development and Reform Commission (NDRC), the country's top price regulator, has ordered a crackdown on the manipulation of food prices, after several industry associations and firms announced plans to raise prices.

Sentence 2a 国家发展和改革委员会(NDRC)，该国的最高价格调节，操纵粮食价格已下令下调，后几个行业协会和企业宣布提价。(Guó jiā fā zhān hé gǎi gē wěi yuān huì, gāi guó de zuì gāo jià gé tiáo jié, cāo zòng liáng shí jià gé yǐxià lìng xià tiáo, hòu ji gè háng yè xié huì hé qé yè xuān bù tí jià.)(Google's translation)

Sentence 2b 国家发展和改革委员会（发改委），国家最高价格调节，下令扫荡操纵粮食价格，经过几个行业协会和公司宣布计划提高价格。(Guó jiā fā zhān hé gǎi gē wěi yuān huì, (fā gài wěi), Guó jiā zuì gāo jià gé tiáo jié, xià lìng sǎo dàng cāo zòng liáng shi jià gé, jīng guò jǐ gè háng yè xié huì hé gōng sī xuān bù jì huà tí gāo jià gé.) (Baidu's translation)

What should emerge is that the probability that (2a) is a good translation is very high, while the probability for (2b) is low. So for every sentence pair S and T there is a probability P (T/S), i.e. the probability that T is the target sentence, given that S is the source. The translation procedure is a question of finding the best value for P (T/S).

With the development of Google's search engine, statistical machine translation(SMT) has gained increasingly attention because of its excellent constructing efficiency and translation quality. This paper explores the application of SMT in web-based translation teaching and proposes that SMT can not only help learners in information retrieval such as words and proper nouns but also provide facilities in reference translation and translating techniques. Therefore it will expectedly play an increasingly important role in the professional—oriented translation education.

3 Case Study on "Crack"

3.1 Analysis of "Crack's" Senses and Word Classes

We do the word classification based on WordNet 3.0（http://wordnet.princeton.edu/）. The semantics distribution of "crack" in WordNet 3.0:

Noun

Sense 1*….(3)S: (n) crack, cleft, crevice, fissure, scissure (a long narrow opening)

Sense 2…. .(2)S: (n) gap, crack (a narrow opening) e.g. "he opened the window a crack"

……...

Sense 8…..S: (n) crack, crack cocaine, tornado

Sense 9…..S: (n) crack, fling, go, pass, whirl, offer (a usually brief attempt) e.g. "he took a crack at it";

Sense 10…..S: (n) fracture, crack, cracking (the act of cracking something).

Verb

Sense 1*….. (6)S: (v) crack, check, break (become fractured; break or crack on the surface only)

e.g. "The glass cracked when it was heated"

Sense 2….. (4)S: (v) crack (make a very sharp explosive sound) e.g. "His gun cracked"

Sense 3….. (2)S: (v) snap, crack (make a sharp sound)

……..

Sense 12…..S: (v) crack (reduce (petroleum) e.g. to a simpler compound by cracking)

Sense13…..S: (v) crack (break into simpler molecules by means of heat) e.g. "The petroleum cracked"

Adjective

Sense1….(2)S:(adj) ace, A-one, crack, first-rate, super, tiptop, topnotch, e.g. "a crack shot".

Table 1. The word class of "crack"

Word Class	Frequency	Percentage
Noun	10	41.7%
Verb	13	51.2%
Adjective	1	4.1%

From table 1, we know that crack has 23 senses. When crack is used as a verb, it has the most sense number 54.2%. The noun of "crack" ranks the second of the list, with the number of 41.7%. The adjective of "crack" has only one sense. So when are sure when "crack" is used as an adjective in specific context, we can identify its meaning "ace, A one, crack, first-rate, super, tiptop, topnotch, top-notch, tops (of the highest quality).

Table 2. The key senses distribution of "crack" (Noun)

Senses	Frequency	Percentage
Sense 1	3	42.8%
Sense 2	2	28.6%
Sense 4	1	14.3%
Sense 5	1	14.3%

In table 2 the statistics shows: "crack" has 4 main senses when it is used as a noun, among which the first sense "crack, cleft, crevice, fissure, scissure (a long narrow opening)". Crack has 7 key senses, also on the top the sense list is sense1: crack, check, break (become fractured; break or crack on the surface only) "The glass cracked when it was heated."

3.2 Translation of the Contextual Meaning of "Crack"

The word class of "crack" based on WordNet shows that when crack used as a noun in medical context, sense8 seems to be the only sense of "crack". Therefore, we take a paragraph of machine translation on medical context as an example.

(1) Crack baby is a term for a child born to a mother who used crack cocaine during her pregnancy.

1a:对一个在她的怀孕期间使用纯可卡因的母亲生的小孩。(Duì yī ge zài tā de huán yùn qī jiān shǐ yòng chún ké ká yīn de mǔ qīn shēng de xiǎo hái) (Google's translation)

1b:高明的婴孩是孩子的一个期限对在她怀孕期间用可卡因的母亲。(Gāo mīng de yīng hái shì hái zi de yì gè qī xiàn dùi zài tā huán yùn qī jiān yòng ké ká yīn de mǔ qīn) (Yahoo's translation)

1c：裂纹的婴儿是一个长期的一个孩子出世的母亲谁使用可卡因在怀孕期间。(Liè wēn de yīng er shì yí ge cháng qī de yī gè hái zǐ chū shì de mǔ qīn shúi shǐ yòng ké ká yīn zài huán yùn qī jiān)（Baidu's translation）

The first version has no correspondence translations with the first “crack”. The second “crack” is translated to “纯的”(chún de). The second version translates the first “crack” into “高明的”(Gāo mīng de). The error of the third version is the same as version one. Based on the analysis of WordNet, let’s analysis the machine translation result on the point of word class. There are two “crack” in the original version. The first step we should identify the literary style of the context. It’s a medical context. The first “crack” is a noun which means a purified and potent form of cocaine that is smoked rather than snorted; highly addictive.

From the WordNet we know that when “crack” used as a noun in medical context, it is always the sense8. For the second “crack”, we know that it is an adjective. From the semantics distribution of “crack”, we know crack has only one sense when it is used as adjective. It means ace, A-one, crack, first-rate, super, tiptop, topnotch, top-notch, tops (of the highest quality).Therefore ,in this medical context, it should be translated “高纯度的”(gāo chún dù de). So online translation version “高明的婴孩”(Gāo mīng de yīng hái) seems to be ridicule and meaningless.

3.3 Translation of the Collocative Meaning of “Crack”

From the WordNet 3.0, we can find out most of the phrase related to “crack”. The frequency about the crack’s collocation in the WordNet can be very useful for machine translation.

Verb

- [37]S:(v) crack up, crack, crock up, break up, collapse (suffer a nervous breakdown)
- [32] S: (v) crack up (rhapsodize about)
- [29] S: (v) break up, crack up (laugh unrestrainedly)

In machine translation, computer sometimes can identify the word classification, but it always fail to translate the “crack” phrase. Even it is the same collocation of “crack”, different online website offers different versions.

For example: crack up a boat.

2a: 吹捧一只小船。(Chuī pěng yī zhiīxiǎo chuān)(Jin Qiao’s translation)

2b: 打沉了船。(Dǎa chén le chuān)(Google’s translation)

2c：使小船发笑。(Shǐ xiǎo chuān fā xiào)（Yahoo’s translation）

Comparing the three translation versions and the semantics distribution of “crack up”. We can easily discovered that three online website respectively applies one of the semantic distribution without any repetition. However, the version of Jin Qiao website and that of yahoo may make people confused. Only human being can carry out the action of “发笑”(fā xiào/ laughing) and “吹捧”(chuī pěng/flattering).Only the version of Google's can make sense. It will be better to translated it into “击沉了船(Jī chén le chuān). So we can use WordNet to help us to choose the appropriate collocatve meaning of crack in specific context.

3.4 Translation of Idiom of “Crack”

Statistical Machine translation of “crack” always choose the most frequency sense of it in given word class .When a noun of “crack” is inputted to the computer, it is always appears “裂缝”(lèi fèng/ crack)in the result. Comparing to the WordNet3.0, “裂缝” in English “a long narrow opening” is on the top of the least. But the translation of “crack” in the version doesn’t make any sense. In this case, we should bear in mind the general translation principle.

For an instant the translation an idiom ---“Are you on crack?” in the authentic conversation.

The original context as follow:

L: 哟，我的天啊，我车后面的一个轮胎爆了！Michael, 一个轮胎爆了还能开一段路，对不对？(Yōu, Wǒ de tiā na, wǒ chē hòu miàn de yì gè lún tāi bào le! Michael, yīgè lún tāi bào le hái néng kāi yí duàn lù, dùi bú dùi.)

M: Li Hua!? Are you on crack? You can't drive with a flat tire; you'll ruin your car, and probably get in a wreck!

L: 哎哟，我也不想弄坏我的车，也不想出车祸。不过，你刚才问我 Are you on crack？这是什么意思啊? crack 这不是一种毒品吗？on crack不就是吸毒吗？(Ai you, wǒ yě bù xiǎng nòng huài wǒ de chē, yě bù xiǎng chū chē huò. Bù guò, nǐ gāng caí wèn wǒ “Are you on crack?” zhè shìshén me yìsi a? Crack zhè bú shì yī zhǒng dú pǐng ma? “On crack” bú jiù shìi xī dú ma?)

M: Yes, "crack" ---- as in "crack cocaine". "To be on crack" means to be high on cocaine. And when I say, "Are you on crack?" I really mean ----"Are you crazy"?

L: 等等，这得弄清楚。你的意思是: to be on crack可以指吸毒，但是你刚才对我说的不是这个意思。你是说：我要在一个轮胎破了的情况下继续开车，简直是疯啦！哎，你可得解释清楚了，我从来不吸毒的！(Děng deng, zhè denòng qīng chu. Nǐ de yì si shì to be on crack kě yǐ zhǐ xī dú, dàn shì Nǐ gāng cái duì Wǒ shuō de bú shì zhè ge yì si. Nǐ shì shuō: wǒ yào zài yīgè lún tāi pò le de qǐng kuàng xià jì xù kāi chē, jiǎn zhí shì fēng le! Ai, nǐ kě de jiě shì qīng chǔ le, wǒ cóng lái bù xīdú de.)

There’re several versions form SMT. Let‘s make a comparison of the quality of each version form pragmatic view.

Looking back upon “crack(n)” in the WordNet3.0. The highest frequency sense of “crack (n)” is sense1 “a narrow opening”. Machine translation chooses sense1 of “crack(n)”based on computer grammar produce. But actually, in the context, it is impossible to choose sense1.

In the context, we should choose sense 8 of “crack (n)” S8:(n) Crack, crack cocaine, tornado (a purified and potent form of cocaine that is smoked rather than snorted; highly addictive). Based on WordNet3.0, the version should be “你在吸毒吗?”(Nǐ zài xī dú ma / Are you taking drug?) This version can not be accepted by Li Hua, at the same time, it’s not the meaning of “crack” that Michael wants to express in the sentence. In other words, the translation seriously violates functional equivalence theory and leads to misunderstanding between them.

4 Discussion and Conclusion

Both English and Chinese have a rich vocabulary. Every word has a different meaning and language is changing all the time. New words have been created day by day. The change and complexity of words' senses adds difficulties for machine translation. Although natural language processing systems have been successful so far in application such as spelling, style, or grammar checking . However, computers are not able to identify the word class and the style of context. In addition, shortage of linguistics knowledge and social background, the version of machine translation is far from people's expectation. People need other tool to help them to choose the right version of machine translation and the appropriate meaning of given word in the context. this paper presents a hybrid strategy method based on several cyber-translation aids and Word-Net 3.0 to make an investigation of some English words translation versions from the aspect of word class, grammatical feature, understandability and loyalty. Finally, the writer illustrates the authentic usages and application rules of the words and propose some measures to improve the quality of corpus-based MT in English-Chinese translation. Preliminary experiment indicates that the hybrid strategy improve the quality of the word translation in both general and scientific discourse.

WordNet is a large lexical database of English; a semantic machine dictionary composed synonymous sets as well as semantic database interlinked by complicated semantics. With the aid of WordNet, we can not only clarify the word class, key senses, and using frequency of crack, but also build verb semantic knowledge description frame and give comprehensive and multi-level description of verb semantic knowledge. (Wang Zheng, 1999)

With the release of Google's search engine, statistical machine translation(SMT) has gained increasingly attention because of its excellent constructing efficiency and translation quality. This paper explores the application of SMT in web-based translation teaching and proposes that SMT can not only help learners in information retrieval such as words and proper nouns but also provide facilities in reference translation and translating techniques. Therefore, it will expectedly play an increasingly important role in the professional-oriented translation education.

Acknowledgments. This work is supported by Hubei Province Humanistic Social Science Youth Project: Research on the Local University Students' Developmental Approach of Lexical Competence" (Grant No. 2012Q813).

References

[1] Leech, G.: Semantics: The Study of Meaning, 2nd edn. Penguin, Harmondsworth (1981)
[2] Zhiwei, F.: The Current Situation and Problems in Machine Translation, pp. 1–10. Science Press, Beijing (2003)
[3] Graeme, K.: A Introduction to corpus Linguistics. Foreign Language Teaching and Research Press, Beijing (2000)

[4] Grishman, R.: Computational Linguistics: an Introduction. Cambridge University Press (1986)
[5] Liu, Q., Li, S.: Word Similarity Computing Based on Hownet. Zhong Wen Xin Xi Xue Bao (5), 25–29 (2008)
[6] Wang, Y.: Semantic Theory and Language Teaching. Shanghai Foreign Language Press, Shanghai (2001)
[7] Wang, Z., Sun, D.: The Application of Statistical Machine Translation in Web-based Translation Teaching Shanghai. Journal Translation (1), 73–77 (1999)
[8] WordNet, http://www.cogsci.princeton.edy/~wn

The Unified Platform for Language Monitoring Based on the Temporal-Spatial Model of Vocabulary Movement

Wei He, Jinling Zhang, Yu Zou, Yonglin Teng, and Min Hou

Broadcasting Language Research Center, Communication University of China,
100024 Beijing, China
zhangjinling20062006@126.com,
{hewei,zouiy,tengyonglin,houmin}@cuc.edu.cn

Abstract. A unified platform to extract all kinds of vocabulary should be required for vocabulary monitoring. We argue just like other physical movements all words are doing the temporal-spatial movement. The classification of vocabulary is actually based on the types of movement. To model the temporal-spatial movement will lead to the extraction of all kinds of words. Therefore this paper proposes the temporal-spatial model of vocabulary movement with three quantities that is the state function, the state change function and the speed function. Furthermore the two new metrics (the normalized usage and the usage ratio) are proposed to calculate the model quantities. The model has been applied to extract the catchwords, new words, terminologies, commonly used words, and emergency terms. The results show that the temporal-spatial model has good robustness, which can eliminate the influence of the different corpus scale and give a package solution for vocabulary monitoring.

Keywords: language monitoring, the temporal-spatial model of vocabulary movement, the normalized usage.

1 Introduction

Language monitoring means to collect language data, quantify language units and discover linguistic phenomena. The object of Language Monitoring is to reflect the situation of language usage, explore language resources, protect language ecology, and create a harmonious language living. As a world's large-scale language monitor, China's National Language Monitoring and Research Center focus on Chinese and annually issues the Green Paper of Chinese Language Situation Report. In U.S. Global Language Monitor Center is mainly monitoring English and published data in the Global Language Monitor Online [1].

Since vocabulary is the most sensitive and active language element, vocabulary monitoring is definitely an important part of the language monitoring. Language monitoring should cover various kinds of vocabulary appearing in language, such as new words, buzzwords, terminology, emergency terms, and commonly used words. How to extract the specific type of vocabulary attracts more and more attention. For example,

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 175–185, 2013.

many studies have proposed various automatic identification methods for new words [2] [3], while others studied the definition and extraction method [4] [5] of buzzwords, also the terminology extraction work in various fields [6] [7]. However, divide and rule is not appropriate for language monitoring, which focus on all kinds of vocabulary not only one kind. Separately for each type of vocabulary to establish a monitoring system is a huge workload and with too complex structure to monitoring a new type quickly. Therefore, language monitoring is necessary to build a unified platform of vocabulary monitoring, which can give a package solution.

Assume that a vocabulary system can be composed of all words and placed on a timeline. So we will see each word is moving and changing along the timeline, all individual movement together push the evolution of the entire vocabulary system. This is called the temporal-spatial movement of vocabulary.

Although the nature of the temporal-spatial vocabulary movement is difficult to reveal, the representation of the temporal-spatial movement can be observed, that is, each word in the vocabulary system changes in the distribution of space. We often classify those words with the similar temporal-spatial movement characteristics as a type of vocabulary and assign some certain name. For example, the words with the largest spatial distribution during a period of time can be called the common words; new words are those occurred at some point later and began to occupy space in the vocabulary system; buzzwords means a type of vocabulary whose spatial distribution increased much more than before; the terms refer to a class of words limited to one or more of the subspace at a moment, Similar cases are numerous. In this way, all words belong to a vocabulary system, moving in a unified spatial-temporal system. When observed along the time axis you can see the vocabulary movement and change. If perpendicular to the time axis you can see the cross-section of the vocabulary system, which is the spatial distribution of all words at the moment.

Language monitoring should monitor the vocabulary system, which means monitoring all kinds of vocabulary according to the type of temporal-spatial movement.

The following sections in this article will first introduce the temporal-spatial model of vocabulary movement, define the parameters of the model, and then perform the vocabulary monitoring experiments by using the temporal-spatial model to extract several types of vocabulary from two year newspaper data. Finally the conclusions are given at last.

2 The Temporal-Spatial Model of Vocabulary Movement

Since the vocabulary movement is similar to the physical movement, there are certainly some physical quantities to describe the movement, such as the state, displacement and velocity. If the physical quantities are defined, the vocabulary movement will be modeled as the temporal-spatial model of vocabulary movement. Furthermore we give the calculation of the parameters to make the model computable.

2.1 Definition of the Temporal-Spatial Model

If the overall space of the vocabulary system is supposed as 1, the spatial distribution of any word is no more than 1. At the moment T, the spatial distribution of the word

W in the vocabulary system is defined as Φ_T, which is the state function of W at T. From T to $T+t$, the state change function of the word W is defined as $\Delta\Phi_t$, and then the average speed is defined as $V_t = \Delta\Phi_t / \Delta t$. Therefore the property of the temporal-spatial model could be represented by the expression $W_T(\Phi_T, \Delta\Phi_t, V_t)$, which contains the state function Φ_T, the state change function $\Delta\Phi_t$, the speed function V_t. Here the state function Φ_T could be any number less than 1 but never be negative, that is $1 \geq \Phi_T \geq 0$. If the vocabulary system is divided into a number of subspaces indexed with $(1, \cdots, i, \cdots, N)$, Φ_T could also be divided into $(\Phi_{1,T}, \cdots, \Phi_{i,T}, \cdots, \Phi_{N,T})$.

The temporal-spatial model of vocabulary movement can make a very good explanation of various vocabulary types. The words with high speed are relatively active, rapid and sensitive, such as buzzwords, new words, and emergency terms. The words with low speed are relatively stable, slow and insensitive, such as high-frequency words, common words, and rarely used words. Terminology generally refers to the class of words whose distribution of the state functions is very uneven among the various sub-spaces. How to calculate the physical quantities of the vocabulary temporal-spatial model? The key lies in the calculation of the state function. To calculate the model quantities we introduce three metrics: the normalized usage, the usage ratio and the average speed.

2.2 Calculation of the Model Quantities

The distribution space is measured by the usage metric which is the product of the word frequency and the document frequency as $f \bullet df$. However the sum of the usage of all the words in the vocabulary system is not equal to 1. Regarding the assumption that the overall space is 1, the usage expression should be normalized. Therefore the normalized usage [8] is proposed as follows:

$$U = \frac{\mathrm{C_w} \times df_{\mathrm{w}}}{\sum_{w \in V} (\mathrm{C_w} \times df_{\mathrm{w}})} \tag{1}$$

In formula 1, U is the normalized usage, $\mathrm{C_w}$ is the frequency of word W, df_{w} is the document frequency of W, the denominator is the normalization entry and V means all the words. df_{w} is Calculated as follows:

$$df_{\mathrm{w}} = d_w / D \tag{2}$$

d_w is the number of texts containing the word W and D is the total number of texts.

The normalized usage is used to measure the distribution space. For all the words in the vocabulary system the sum of the normalized usage is just 1 in line with the assumption. We argue that the normalized usage can convert the absolute value to relative value, which means the spatial distribution of the same word can be compared between two different vocabulary systems.

By introducing the normalized usage U, the state function Φ_T is defined as $\Phi_T = U_T$ .It should be noted that the real value of the state function never be available in practical applications. Whatever the corpus size is, the corpus is just the sampling of all language facts. Therefore the normalized usage obtained through the corpus statistics can only be the approximate value to the state function. When the larger the corpus, the approximate calculation will be more close to the real value.

Since the state function $\Phi_T = U_T$, the state change function $\Delta\Phi_t$ consequently can be defined as $\Delta\Phi_t = \Delta U_t$. In order to avoid the high-frequency interference we relativise the calculation of ΔU_t as shown in formula 3.

$$\Delta U_t = \frac{U_T - U_{T-t}}{U_T + U_{T-t}} \tag{3}$$

We call ΔU_t the usage ratio. Formula 3 expresses the meaning that the change rate of the normalized usage should be compared with the normalized usage itself. In other words, the absolute magnitude of the change is converted to the relative magnitude of the change. The usage ratio can reduce the effect of the high-frequency interference. As we know the normalized usage of high-frequency words is much greater than those middle or low frequency words. If measured by the absolute value, even a minor fluctuation of high-frequency words will be far more than any change of low-frequency words. Therefore the fluctuation of high-frequency words will mask that of low-frequency words. This phenomenon is named the high-frequency interference. Unfortunately the fluctuation of high-frequency words is usually caused by the corpus noisy, i.e. the change of the corpus size or content. Considering highlighting the vocabulary movement we should try to diminish the effect of the high-frequency interference. With the introduction of the usage ratio, for high-frequency words the relative change in amplitude is very small despite of the high absolute value. Conversely for low-frequency words, although the change in amplitude is small in absolute terms, but the relative changes in amplitude is large. The usage ratio can effectively get rid of the high-frequency interference and highlight the changes of the middle or low frequency words which more likely to be caused by the temporal-spatial vocabulary movement.

If the usage ratio ΔU_t is available, the average speed V could be calculated by formula 4 as follows:

$$V = \Delta U_t / t \tag{4}$$

After introduction of the three metrics the normalized usage U, the usage ratio ΔU_t and the average speed V the three quantities of the temporal-spatial model can be represented respectively.

The state function: $\Phi_T = U_T \qquad 1 \geq U_T \geq 0$

The state change function: $\Delta\Phi_t = \Delta U_t \quad 1 \geq \Delta U_t \geq -1$

The speed function: $V_t = V \qquad 1 \geq V_t \geq -1$

3 The Experiments of Vocabulary Monitoring

In order to test the validity of the temporal-spatial model of vocabulary movement this section will apply the model to extract the commonly used words, buzzwords, new words, terminology, and emergency terms separately. It must be pointed out that the calculation of the quantities for all character strings is required just only once and various type of vocabulary could be extracted by the different thresholds value.

3.1 The Experiments Corpus

Two years data from the Chinese newspaper Beijing Youth Daily have been collected as the experiment corpus. The 2010 year data are used as the background set and the 2011 year data are used as the test set. The background set contains 47384 texts and about 34.9 million words. The test set includes 50604 texts and about 38.5 million words. The newspaper contains 7 news plates: daily news, city news, national news, international news, entertainment & culture news, financial news and sports news. The news plates are classified as four areas, namely, current affairs (daily news, national news, and international news), culture and sports (entertainment & culture news, sports news), finance (financial news), and society (city news). The overview of the corpus is shown in Table 1.

Table 1. Overview of the experiments corpus

		current affairs	culture and sports	society	finance	total
The background set (year 2010)	texts	17157	12477	11882	5868	47384
	Characters (million)	13.9	8.5	7.8	4.7	34.9
The test set (year 2011)	texts	17670	12937	13050	6947	50604
	Characters (million)	15.4	9.5	8.4	5.2	38.5

The unbalanced nature of the corpus distribution can be observed from Table 1. The background set is more than the test set about 3.6 million characters in which 1.5 million characters from the current affairs field, 1 million characters from the culture and sports field, 0.6 million characters from the society field and 0.5 million characters from the finance field. This means compared with 2010 the data size and composition of 2011 have some changes. In addition, there is a wide difference among the various fields. The data of the current affairs field is almost three times more than that of the finance field. For such an uneven corpus the traditional methods are hardly to

be applied. The experiments will show our method based on the temporal-spatial model are very robust on the uneven corpus and can successfully extract various types of vocabulary.

3.2 The Process of Experiments

The entire experimental process is divided into four steps, which are in sequence word segmentation, word string statistics, calculation of model parameters and vocabulary extraction. The details of each step are given as follows:

1) Word segmentation. Since the corpus texts are in Chinese automatic segmentation is required. The corpus data are divided into the four fields and processed with the fine-grained segmentation. The segmentation principle is to split the combination expression possibly, e.g. the name are divided into the family name and given name. Proper nouns such as the place names and organization names are also split.

2) Word string statistics. All possible word strings are counted. Within the window of 7 characters length, each possible combination of the segmentation units is considered to be a string of words. For each word string the number of times, texts and days of its appearance in both the test set and the background set are counted.

3) Calculation of model parameters. According to the formulas mentioned above, the normalized usage should be figure out firstly for every word string in each field of each set. Then the usage ratio could be acquired by comparing the normalized usage of the test set with that of the background set. Finally the usage ratio would be divided by the number of days of the word in the test set, which will make the average speed. Thus the three metrics of each word string's movement model would be obtained.

4) Vocabulary extraction. According to the temporal-spatial model of vocabulary movement various types of vocabulary can be extracted just by the three metrics. For instance the usage ratio of new words should be 1. Emergency terms should have the high velocity. Buzzwords should have the high normalized usage and usage ratio. Both commonly used words and terminologies are with the low speed and high normalized usage. To demonstrate the power of the model metrics we will give the results of vocabulary filtered with the three metrics and without excluding any rubbish string in the next section.

3.3 The Extraction of Various Types of Vocabulary

The thresholds of the normalized usage, the usage ratio and the average speed would be adjusted to extract commonly used words, terminologies, catchwords, new words and emergency terms according to the characteristics of the temporal-spatial vocabulary movement. These types of vocabulary could be separated by the velocity. Commonly used words and terminologies are relatively stable while catchwords, new words and emergency terms are dynamic. Therefore commonly used words and terminologies will be extracted firstly and then catchwords, new words and emergency terms. It should be noted that the catchwords, new words and emergency terms mentioned in the experiment are restricted in 2011 because the test set is from the year of 2011.

Commonly used words

Because the features of commonly used words movement are high distribution and less change, the word strings with the usage ratio ΔU between -0.03 and 0.03 are selected firstly. And then the intersection of the word strings selected from each field is seemed as the candidates set and calculate the average normalized usage. If the average normalized usage $\overline{U}$ is more than the threshold, the candidate will be chosen as the commonly used word. All commonly used words are ranked in a descending order by the normalized usage. Top 15 commonly used words are shown in table 2 and the first 15 high-frequency words of the total data are also listed in table 2 as a contrast.

Table 2. Top 15 of commonly used words and high-frequency words

Commonly used words				High-frequency words	
Words	The normalized usage	The usage ratio	The average speed	Words	Counts
的(de)	0.188763392	-0.000461	-0.000001	的(de)	862202
在(at)	0.049011436	0.002266	0.000006	在(at)	240178
一(one)	0.031775267	0.005968	0.000016	了(le)	179298
了(le)	0.031492964	0.004956	0.000013	是(is)	178411
是(is)	0.027930102	0.000112	0	一(one)	177430
不(not)	0.01385065	0.001465	0.000004	和(and)	136980
有(have)	0.011836995	-0.001365	-0.000003	不(not)	109997
中(in)	0.010609955	0.00907	0.000024	将(will)	92214
这(the)	0.010425953	0.000231	0	有(have)	92095
个(ge)	0.010306202	0.015458	0.000042	这(the)	81804
人(man)	0.008775165	0.016762	0.000045	个(ge)	80427
就(to)	0.005462352	0.011151	0.00003	北京 (Beijing)	79036
后(after)	0.005396072	0.005953	0.000016	中(in)	78890
都(all)	0.004681418	-0.007071	-0.000019	人(man)	73373
一个(one)	0.002713157	-0.001686	-0.000005	也(also)	72980

As shown in table 2, the high-frequency word "北京 (Beijing)", which is obviously caused by the local reports of the newspaper, is excluded from top 15 commonly used words. This phenomenon explains that extraction of commonly used words by the normalized usage and the usage ratio can reduce the interference of the corpus content.

Terminologies

The data from the finance area are selected to extract terminologies. Similar to commonly used words terminologies are also stable and with high distribution, but just in one field rather than all fields. The extraction process is to first select word strings with the usage ratio ΔU between -0.3 and 0.3 from the finance area, and then get rid of those appear in other areas, the remaining candidates are list according to the normalized usage from highest to lowest. The top 10 of the terms in the finance field are shown in table 3.

Table 3. Top 10 terms in the finance area

Words	Counts	The normalized usage	The usage ratio	The average speed
A股(A-share)	2141	0.000500292	0.233905	0.000724
股价(share price)	1375	0.000209148	0.141222	0.0005
大盘(grail)	1404	0.000192667	-0.106804	-0.000373
回落(recede)	1053	0.000155817	0.152067	0.000526
行情(market)	961	0.000128963	-0.17816	-0.000667
券商(broker)	1119	0.000118097	0.084885	0.000349
证监会(Securities Regulatory Commission)	1099	0.000117804	0.068891	0.0003
收盘(close)	802	0.000116686	0.129771	0.000461
范辉(Fan Hui)	914	0.000105025	-0.167611	-0.000698
新股(new share)	1247	0.000102742	0.049564	0.000244

From table 3 we can find there is a trash string "范辉 (Fan Hui)", which is a reporter's name. The rest candidates are the appropriate terms in the finance area and the accuracy is 90%. Compared with the commonly used words, terminologies occupy smaller space with an order of magnitude lower in the normalized usage and are more dynamic with two orders of magnitude higher in the usage ratio.

Catchwords

Catchwords belong to a class vocabulary with a high distribution space after a huge increase and a rapid movement. According to the characteristics of catchwords movement, the word strings with the usage ratio $\Delta U \geq 0.8$ and the average speed $V \geq 0.005$ would be extracted from all the test data. The candidates are ranked in a descending order by the normalized usage, which means the arbitrary top N catchwords would be available. The top 10 catchwords are shown in table 4.

Table 4. Top 10 catchwords

Words	Counts	The normalized usage	The usage ratio	The average speed
0:00:00	17439	0.000759966188743363	1	0.009523
核电站(nuclear power station)	2200	0.00000222653137545694	0.956648	0.005199
福岛(Fukushima)	1267	0.00000119679434963197	0.99937	0.005583
建党(Party building)	832	0.00000091064383582325	0.949745	0.005721
海啸(tsunamis)	868	0.00000081122708583315	0.865248	0.005181
90周年(90 anniversary)	692	0.00000073147178237739	0.993071	0.00649
叙利亚(Syria)	1285	0.00000071928708317791	0.882505	0.005042
瘦肉(lean)	1063	0.00000066674244206486	0.951931	0.006346
校车(school bus)	1862	0.00000066072256093862	0.919445	0.011638
的黎波里(Tripoli)	911	0.00000052587372523292	0.997516	0.005511

Only the first one is a trash string in top 10 and the rest candidates represent the hot news or words in 2011. The results of catchwords extraction are pretty good whereas there is no any trash post-processing. Compared with the commonly used words and terminologies, the usage ratio and the average speed of catchwords are bigger obviously.

New words

For new words the main feature of movement is the new, which means the word doesn't appear before and the usage ratio ΔU should be 1. In addition new words tend to own a small space at the start and the movement of new words is also rapid. Thus the threshold of the usage ratio is set as $\Delta U = 1$ and the normalized usage $U < 0.00000045$ and the average speed $V \geq 0.02$ respectively. The culture and sports field is chosen to extract new words because this field contains more new words. The word strings satisfying the conditions will be extracted firstly, and then the words appearing in the current affairs, society and finance areas of the background set will be eliminated to tackle the sparse data problem. The remaining candidates are ranked in a descending order by the normalized usage. The top 10 new words are shown in table 5.

Table 5. Top 10 new words

Words	Counts	The normalized usage	The usage ratio	The average speed
微电影(micro film)	155	0.00000038213567623027	1	0.025641
卡马乔的(Camacho's)	108	0.00000031063932390332	1	0.021739
比分推介(score recommend)	114	0.00000030112995684505	1	0.022222
NBA停摆(NBA lockout)	105	0.00000028968349649714	1	0.022222
NBA停(NBA stop)	105	0.00000028968349649714	1	0.022222
胜负推介(Outcome of referral)	113	0.00000028522231195129	1	0.023255
老挝(Laos)	149	0.00000027988063045560	1	0.035714
博阿斯(Boas)	110	0.00000020662328422896	1	0.032258
资方(capital)	100	0.00000018783934929906	1	0.032258
范斌(Fan Bin)	152	0.00000014275790546728	1	0.0625

Strictly speaking, there are five trash strings. Both the string "博阿斯(Boas)" and "范斌(Fan Bin)" are person names which are hardly to be distinguished from new words. The strings "老挝(Laos)"and "资方(capital)"are not new words but could be excluded with more background data available. The last trash string "NBA停(NBA stop)" , a fragment of another new word "NBA停摆(NBA lockout)", could be merged with the true one furthermore. The result shows the method of new words extraction has the latent force to develop if getting more data and filtering trash strings.

Emergency terms

The movement of emergency terms is similar to that of new words. Emergency words should have not appear in the previous period and move rapidly. The tiny difference between emergency terms and new words is that for emergency terms it is impossible to occupy a very small space, because emergencies usually attract quite a few of reports. The extraction of emergency terms is executed in the current affairs field. The thresholds are set as the usage ratio $\Delta U = 1$, the average speed $V \geq 0.02$ and the normalized usage $U > 0.0000001$. All the qualified word strings are selected as the candidates and ranked according to the normalized usage. The top 10 emergency terms are listed in table 6.

Table 6. Top 10 emergency terms

Words	Counts	The normalized usage	The usage ratio	The average speed
★	642	0.00000298834093351837	1	0.022727
康菲(ConocoPhillips)	455	0.0000009596757841164	1	0.022727
过渡委(Transitional Council)	410	0.00000092440160789083	1	0.020833
建党90(Party building 90)	174	0.0000006643909274652	1	0.023809
7 • 23	220	0.00000066402727616391	1	0.029411
建党90周年(Party building 90 anniversary)	168	0.0000006109341861749	1	0.025
美美(Meimei)	262	0.00000044780021241415	1	0.027777
塑化剂(plasticizer)	315	0.00000043529060764961	1	0.038461
英拉(Yingluck)	238	0.00000040678034562812	1	0.023809
郭美美(Guo Meimei)	257	0.00000040187105306064	1	0.03125

There are only three trash strings in table 6. The trash strings “建党90(Party building 90)” and “美美(Meimei)” are the fragments of the true emergency terms “建党90周年 (Party building 90 anniversary)”and “郭美美 (Guo Meimei)” respectively. The other trash string “★” is a sign easily to be removed. Regarding with no filtering trash strings, the results are fairly good.

From table 1 to table 6 we can find the average speed is increasing with an order of magnitude. There is also a sharp demarcation between the stable vocabulary and the dynamic vocabulary by the normalized usage. Therefore we can infer that the vocabulary with the middle or small distribution move more quickly than that with the high distribution. The normalized usage can give the candidates a reasonable ranking rather than the counts, which is important for language monitoring to pay attention to the most significant vocabulary. The last thing should be point out is that the method based on the temporal-spatial model is very robust on the uneven corpus. Although the experiment corpus is a middle size corpus with a data disproportion among the four areas, results show that various type of vocabulary can be extracted successfully even for the low frequency words.

4 Summary

The main purpose of this paper is to propose the temporal-spatial model of vocabulary movement and its application. The model is from our long-standing practice of language monitoring. The vocabulary extraction experiments on a small-scale corpus show that the model can be applied in the integrated extraction of various vocabularies. The method based on the temporal-spatial model is very robust on the uneven corpus, which can also greatly enhance the efficiency of language monitoring. We believe that with the in-depth study of the vocabulary movement, the temporal-spatial model will achieve an ideal application.

References

1. Chinese Language Situation in 2011, 1st edn. Department of Language Information Management in the Ministry of Education. The Commercial Press, Beijing (2011)
2. Zhang, H.J., Shi, S.M., Zhu, Z.Y., Huang, H.Y.: Summarize of Chinese New Words Recognition Technology. Computer Science 3, 6–16 (2010)
3. Wu, Y., Yan, P.J., Di, L.F.: New Words Detection Based on a Background Bigram Model. J. Tsinghua University 9, 1317–1320 (2011)
4. He, W., Hou, M., Wen, C.J.: The Temporal-Spatial Model for Catchwords Monitoring. In: Sun, M.S., Chen, Q.X. (eds.) Frontiers of Content Computing: Research and Application, pp. 276–282. Tsinghua University Press, Beijing (2007)
5. Zhang, P.: Study on the Dynamic Tracking and Auxiliary Finding of Catchwords Based on DCC. In: Sun, M.S., Chen, Q.X. (eds.) Computing Language and Content-Based Text Processing, pp. 20–28. Tsinghua University Press, Beijing (2003)
6. You, H.L., Zhang, W., Liu, T.: An Automatic Terms Identification Method Based on the Weighted Voting. Chinese Information Technology 3, 9–16 (2011)
7. Xun, E.D., Li, S.: The Method of New Terms and Definitions Identification by Using the Definitions Mode of the Terms and Multi-characteristics. Computer Research and Development 1, 62–69 (2009)
8. Hou, M.: Research on Language Monitoring and Vocabulary Quantification. In: Cao, Y.Q., Sun, M.S. (eds.) Recent Progress in Chinese Information Processing, pp. 76–83. Tsinghua University Press, Beijing (2006)

Elementary Discourse Unit in Chinese Discourse Structure Analysis

Yancui Li[1,2], Wenhe Feng[2], and Guodong Zhou[1,*]

[1] Department of Computer Science and Technology, Soochow University, Suzhou 215006
yancuili@gmail.com, gdzhou@suda.edu.cn
[2] Henan Institute of Science and Technology, Xinxiang 453003
wenhefeng@gmail.com

Abstract. Elementary Discourse Unit recognition is the primary task of discourse structure analysis. Different theories have different definition of Elementary Discourse Unit. Combined with Chinese discourse annotation practice, this paper defines clause as Elementary Discourse Unit. There is always punctuation at clause boundary in form, and we adopt different methods to deal with different punctuations. We annotate a certain scale of corpus which is labeled with whether punctuations are the boundary of clauses, and the accuracy of automatic recognizes clause is 90.7%. The experimental result shows that our definition of clause corresponds with Chinese character in theory and has strong distinction ability in calculation.

Keywords: Elementary discourse unit, Clause, Chinese discourse structure, Corpus.

1 Introduction

The natural language units can be divided into words, phrases, sentences and finally discourse from small to large. In discourse structure analysis, discourse refers to the whole language unit which contains a series of consecutive clauses, sentences or discourse constitutions. Discourse units not only have internal coherence, but also are to describe the same problems or situations of a relatively complete language as a whole. Clause, sentence or discourse has certain hierarchy and semantic relationship. We must analyze the hierarchy and semantic relation of the discourse in order to have a general grasp on it. Fig. 1 gives the hierarchical structure of example 1:

Example 1. 1a 浦 东 开 发 开 放 是 一 项 振 兴 上 海 ， 建 设
pǔdōngkāifākāifàngshìyīxiàngzhènxìngshànghǎi jiànshè
现 代 化 经 济 、 贸 易 、 金 融 中 心 的 跨 世 纪 工 程 ， 1b
xiàndàihuà jīngjì màoyì jīnróngzhōngxīndekuàshìjì gōngchéng
因 此 大 量 出 现 的 是 以 前 不 曾 遇 到 过 的 新 情 况 、 新
yīncǐ dàliàngchūxiàndeshì yǐqiánbùcéngyùdàoguòdexīnqíngkuàng xīn
问 题 。 1c 对 此 ， 浦 东 不 是 简 单 的 采 取 “ 干 一 段 时 间 ， 等
wèntí duìcǐ pǔdōngbúshì jiǎndāndecǎiqǔ gànyíduànshíjiān děng

* Corresponding author.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 186–198, 2013.

jī lěi le jīngyàn yǐhòu zàizhìdìng fǎguī tiáolì de zuòfǎ érshì jièjiàn fādá guójiā hé shēnzhèn děng
积累了经验以后再制定法规条例”的做法，1d而是借鉴发达国家和深圳等
tèqū de jīngyàn jiàoxùn pìnqǐng guónèiwài yǒuguān zhuānjiā xuézhě jījí jíshídì
特区的经验教训，1e聘请国内外有关专家学者，1f积极、及时地
zhìdìng hé tuīchū fǎguī xìngwénjiàn shǐ zhèxiē jīngjì huódòng yī
制定和推出法规性文件，1g使这些经济活动一
chūxiànjiù bèinàrù fǎzhì guǐdào
出现就被纳入法制轨道。

1a Pudong development and opening up is a cross-century project of vitalization and building a modern economy, trade and financial center. 1b So there are a large number of new situations and new problems that have not previously been encountered. 1c Pudong is not simply adopting "do a period of time, wait accumulation of experience to develop laws and regulations" approach. 1d But learns lessons from developed countries and the Shenzhen Special Administrative Region (SAR). 1e Employ relevant experts and scholars at home and abroad. Actively and timely formulate and launch the legal document. 1f So that economic activities can be incorporated into the legal orbit when they appeared. (chtb_0001)

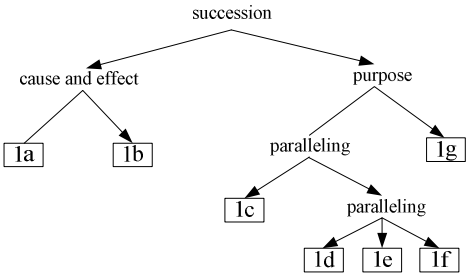


Fig. 1. The discourse structure of example1

Number and letter marks in Fig.1 (such as 1a, 1b, etc.) indicate the Elementary Discourse Units (EDUs). This paper defines EDU as clause. The arrow is pointing to the main clause. The combination of different clauses can be considered as EDUs in a higher level and then the new EDUs can be combined to higher units again. Finally the discourse can be expressed as a structure tree of the combination of EDUs. In discourse structure analysis, the primary task in building up the discourse structure is the identification of EDUs. Each discourse theory has its own specificities in terms of segmentation guidelines and size of units. Hobbs model [1-2] suggested that the discourse structure consist of discourse unit and discourse relation. Discourse unit can be as small as clause and as large as the chapter itself. Givon[3] considered clauses to be the basic unit of the discourse while Sacks et al. [4] thought turns of talk should be the basic unit of the discourse. Polanyi [5] thought “contextually indexed representation of information conveyed by a semiotic gesture, asserting a single state of affairs or partial state of affairs in a discourse world”. Grosz et al. [6] pointed out that the basic unit of the discourse must be acquainted in the context. It is reflected by a certain symbol which can reflect a single state or part of the state of things. Rhetorical Structure Theory of Mann and Thompson [7-8] thought clause should be the basic unit of the discourse, regardless of the clause there is no grammatical marks or lexical marks.

It is very similar with the Hobbs model. The rhetoric structure theory is paying more attention to sentence internal structure compared to the Hobbs model. Discourse unit can be as small as phrase. Regardless of their theoretical stance, all agree that the EDUs are non-overlapping spans of text.

In the field of natural language processing, different discourse corpus has different definition and annotation of EDUs. Carlson et al. [9] combined Grosz and Thompson's theory, considered the vocabulary, syntax, semantics, sentence location and other factors when determining EDUS. Their Rhetoric Structure Theory Discourse Treebank[10] chose the clause as the elementary unit of discourse, used lexical and syntactic clues to help determine boundaries. A few refinements to this basic principle are enumerated below: Clauses that are subjects or objects of a main verb are not treated as EDUs; Clauses that are complements of a main verb are not treated as EDUs; Complements of attribution verbs (speech acts and other cognitive acts) are treated as EDUs; Relative clauses, nominal post modifiers, or clauses that break up other legitimate EDUs, are rated as embedded discourse units; Phrases that begin with a strong discourse marker, such as because, in spite of, as a result of, according to, are treated as EDUs.

For the Penn Discourse Treebank [11], the discourse units are abstract objects such as propositions, facts. Although they are phrased in deferment ways, syntactically these discourse units are generally realized as clauses or built on top of clause, including simple clauses, non-clausal arguments and multiple clauses/sentences (obey minimalism principle)

Natural language processing of Chinese discourse structure analysis are rare. Yue [12-13] defined EDU as sentence, which was separated by period, question mark, exclamation point, semicolons, dashes, colons, ellipsis and paragraph tags. Chen [14] proposed EDUs should be segmented with punctuations (period, commas and semicolons etc.). Xing [15] researched Chinese compound sentences and build Chinese complex sentence corpus guided by his research. But their corpus only contains complex sentences which have connectives. Therefore it's necessary to find an EDU definition method which can embody Chinese character and easy to be calculated. Because there has not been a large-scale Chinese discourse structure corpus, we have not seen result of EDUs recognition.

This paper aims to define the EDUs of Chinese discourse and tries to automatically identify it. The rest of the paper is organized as follows. In Section 2, we present our approach of Chinese EDUs definition. In section 3, we describe our supervised feature learning method and experiment result of recognizing EDUs in CTB6.0 corpus. We conclude in Section 4.

2 Chinese Elementary Discourse Units

The purposes of this paper are to discuss the definition of Chinese EDUs and to annotate Chinese discourse structure corpus. Usually before annotation we must establish a theory of the discourse structure. The theory should reflect both syntax and semantic. On the basis of Chinese complex sentence theory, Chinese sentence group theory, English Rhetorical Structure Theory, PDTB system and other research results, we define Chinese EDUs as clauses based on annotating a certain scale of CTB 6.0

corpus. It will be described in detail below. All examples are from CTB6.0. Parentheses after the examples specified source document number and all examples are segmented manually. EDUs are indicated in the form of number plus a letter (e.g., 1a).

2.1 EDU is Single Sentence

Single sentence can express a relatively complete meaning and have a particular tone. All the words of this sentence only have one center structure, and each structure center can only have one set of composition. Namely clause can only have a subject-predicate structure or a non subject-predicate structure.

Example 2. 2a 外 商 投资企业在改善中国出口商品结构中发挥了显著作用。
wàishāngtóuzīqǐyèzàigǎishànzhōngguóchūkǒushāngpǐnjiégòuzhōngfāhuīlexiǎnzhùzuòyòng

2a Foreign investment enterprises play a significant role in improving China export commodity structure. (chtb_0002)

Example 3. 3a 北海市的崛起，是近年来广西壮族自治区对外开放取得卓著成就的重要标志之一。
běihǎishìdejuéqǐ shìjìnniánláiguǎngxīzhuàngzúzìzhìqūduìwàikāifàngqǔdézhuózhùchéngjiùdezhòngyàobiāozhìzhīyī

3a BeiHai city's rise is the important marks of the Guangxi Zhuang autonomous region opening up in recent years. (chtb_0006)

In Example 2, there is only one punctuation (period, question mark or an exclamation point) in the whole sentence. In this case, we consider the sentence as an EDU. In example 3, although there is punctuation between sentences, it has only one syntactic component, so the sentence is an EDU too.

2.2 EDU is the Clause of Complex Sentence

Complex sentence is constituted by two or more significantly related clauses. Clauses are similar to the structure without complete sentence adjustable grammatical units. Each clause of the complex sentence usually has a pause, formally denoted by a comma, semicolon, or colon.

Example 4. 4a 浦东开发开放是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，4b 因此大量出现的是以前不曾遇到过的新情况、新问题。
pǔdōngkāifākāifàngshìyīxiàngzhènxìngshànghǎi jiànshèxiàndàihuàjīngjì màoyì jīnróngzhōngxīndekuàshìjìgōngchéng yīncǐdàliàngchūxiàndeshìyǐqiánbùcéngyùdàoguòdexīnqíngkuàng xīnwèntí

4a Pudong development and opening up is a cross-century project of vitalization and building a modern economy, trade and financial center. 4b So there are a large number of new situations and new problems that have not previously been encountered. (chtb_0001)

gǔ lǎo de jīng háng dà yùn hé rú jīn bù jǐn zài guàn tōng nán běi
Example 5. 5a 古 老 的 京 杭 大 运 河 如 今 不 仅 在 贯 通 南 北
yùn shū fāng miàn fā huī zhòng yào zuò yòng ér qiě dài dòng qǐ yī tiáo
运 输 方 面 发 挥 重 要 作 用 ， 5b 而 且 带 动 起 一 条
xīn xīn xiàng róng de gōng yè zǒu láng xíng chéng le
欣 欣 向 荣 的 工 业 走 廊 ， 5c 形 成 了 大 运 河 经 济 带 。

5a Now the old Beijing-Hangzhou canal not only plays an important role in the breakthrough transportation between north and south, 5b but also drives up a thriving industrial corridor, 5c forming the grand canal belts. (chtb_0004)

Although example 4 has comma in 4a, it is one EDUs, 4a and 4b is cause-effect relation indicated by “so”. In example 5, 5b and 5c is parallel and conjunction is implicit. 5a and the combination of 5b and 5c have a progressive relationship with connective "not only …..but also".

2.3 Punctuation and EDU

The above methods are the basic principle of judging EDU. As a matter of fact, there are varieties of punctuations in the corpus and the punctuations also have different effect. Inevitable for the EDU boundary, punctuations have great significance for sentence segmentation. The distribution of all different punctuations which could be EDU boundary is shown in Fig. 2:

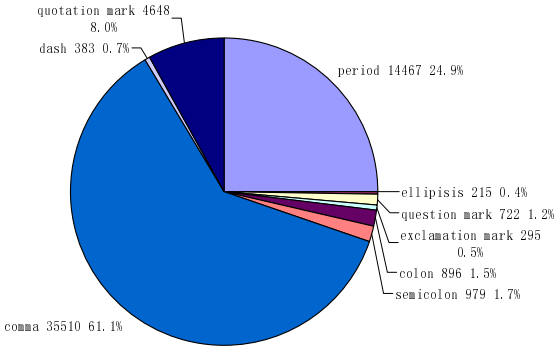


Fig. 2. The punctuations use frequency in CTB6.0

Fig. 2 shows that comma and period appear most frequently (86%), which can be processed according to the above principle. The methods for processing other punctuations are described in the following.

1. Question mark and exclamation mark, the function and treatment is the same as period.

xióng kǒng què shí dài lái lín
Example 6. 6a 雄 孔 雀 时 代 来 临 ？
6a Is the era of male peacocks come? (chtb_1018)

shénmyāoxīndōngxīyěméiyǒuā
Example 7. 7a 什 么 新 东 西也没 有啊!
7a Nothing new things! （chtb_1020）

2. Semicolon indicates a pause between the periods and commas punctuation. Its main usage: complex sentence internal pause between the parallel clauses; the first layer of non-parallel complex sentence; the branch between lists. The proportion of the semicolon is only 1.7% in the corpus, but corpora analysis showed that semicolon separated more than 99% of the sentences. Therefore, this paper treat semicolon as EDU boundary.

nóngmùyè shēngchǎn dàikuǎn bāokuò fúpín dàikuǎn bǐ
Example 8. 8a 农 牧 业 生 产 贷 款 （ 包 括 扶 贫 贷 款 ）比
shàngniánxīnzēngsìdiǎnsānbāyìyuán xiāngzhènqǐyèdàikuǎnzēngfúwéi
上 年 新 增 四 点 三 八 亿 元 ； 8b 乡 镇 企 业 贷 款 增 幅 为
bǎifēnzhīliùshíyīdiǎnbāsān
百 分 之 六 十 一 点 八 三 。

8a Agricultural and animal husbandry production loans (including loans for poverty alleviation) new add 4.38 billion Yuan from a year earlier; 8b Township enterprise loan growth is 61.8 %.(chtb_0007)

dàikuǎn jiāngxiàngnéngyuán jiāotōng diànlì děngjīchǔ
Example 9. 9a 贷 款 将 向 能 源 、 交 通 、 电 力 等 基 础
shèshī chǎnyè qīngxié yóuqí yǐ guówài dàgōngsī zài huá shèlì de
设 施 产 业 倾 斜 ， 9b 尤 其 以 国 外 大 公 司 在 华 设 立 的
dàzhōngxíng qǐyè wéi zhòngdiǎn cǐwài gāojìshù gāokējì gāo
大 中 型 企 业 为 重 点 ； 9c 此 外 ， 高 技 术 、 高 科 技 、 高
chūkǒu gāo lìshuì de qǐyè yě jiāng huòdé zhōngguó yínháng de dàikuǎn
出 口 、 高 利 税 的 企 业 也 将 获 得 中 国 银 行 的 贷 款
zhīchí
支 持 。

9a The loan will be to energy, transport, electricity and other infrastructure industry tilt. 9b Especially focus on large and medium-sized enterprises of foreign big company establishment in China.9c In addition, high technology, high science and technology, high export, high profit tax of the enterprise will also get China bank loan support. (chtb_0007)

In example 8, before and after a semicolon respectively is a sentence. 8a and 8b are clauses. Clauses connected by semicolon have parallel relationship. In example 9, 9a, 9b and 9c are clauses, 9a and 9b have the progressive relationship, 9c and the combination of 9a and 9b are segmented by a semicolon, both before and after semicolon are clauses, indicating parallel relationship.

3. The colon usually prompts a pause after the clues, tips context below and concludes above. Colon in this corpus basically has the following usage: used for discourse organizations, used in headings and subheadings part, used after speaker.

zuòzhě chénbīn
Example 10. 10a 作 者： 陈 彬
chūbǎn shāngxùnwénhuà
10b 出 版： 商 讯 文 化
dìzhǐ táiběishìdàlǐjiē hào
10c 地 址： 台 北 市 大 理 街 132 号

10a Author: ChenBin
10b Published: commercial information culture
10c Address: Number 132, Dali street, Taipei City （chtb_1051）

jiāngzémínzhǔxí zài jiǔjiāng shìchá shí shuō
Example 11. 11a 江 泽 民 主 席 在 九 江 视 察 时 说 ： “ 11b
jiǔjiāng dìchù jīngjiǔ zhōngduàn dìlǐwèizhì hěnhǎo jiǔjiāng
九 江 地 处 京 九 中 段 ， 11c 地 理 位 置 很 好 ， 11d 九 江
qiántúwúliàng
前 途 无 量 。”

11a President Jiang Zemin inspected in Jiujiang said: 11b "Jiujiang is located in the middle of Jingjiu Railway, 11b the geographical position is very good,11c Jiujiang is promising." （chtb_0042）

From examples 10 and 11, it is known that colon has many usages, you can't determine whether it can split sentence, you need to consider the actual situation to distinguish.

4. Quotes are a symbol which marks on quote, focus and special purpose.

qùnián wèishēngbùhéyǒuguānbùménítíchūle yùfángwéizhǔ xuānchuán
Example 12: 12a 去 年 ， 卫 生 部 和 有 关 部 门 提 出 了 “ 预 防 为 主 ， 宣 传
jiàoyù wéizhǔ jīngchángxìng gōngzuò wéizhǔ de àizībìng fángzhì cèlüè bìng dédào
教 育 为 主 ， 经 常 性 工 作 为 主 ” 的 艾 滋 病 防 治 策 略 ， 12b 并 得 到
guówùyuànde
国 务 院 的 确 认 。

12a Last year, the health ministry and the relevant departments put forward the “give priority to prevention, propaganda and education, regular work" HIV/AIDS prevention and control strategy, 12 and get the confirmation of the state council. （chtb_0244）

For the processing of quotes, the principle is that the quotes in a clause internal (as in Example 14) which is not taken into consideration and quotes cited a separate sentence is treated as an independent clause.

5. The dash indicates the change of topic or mood, such as the continuation of voice symbols. It often demarcates a break of thought or some similar interpolation stronger than the interpolation demarcated by parentheses. In CTB6.0 corpus it often indicates subsequent statement is the explanation of some front words. This article thinks dash as EDU boundary except those which often demarcate a break of thought or some similar interpolation stronger than the interpolation demarcated by parentheses

zhègekāifāqūwèiyúzhōngguózhùmíngfēngjǐnglǚyouchéng
Example 13: 13a 这 个 开 发 区 位 于 中 国 著 名 风 景 旅 游 城 ——
hángzhōushìqūnèi shìyījiǔjiǔyīniánguówùyuànpīzhǔnjiànshède
13b 杭 州 市 区 内 ， 13c 是 一 九 九 一 年 国 务 院 批 准 建 设 的
guójiājígāoxīnjìshùchǎnyèkāifāqū
国 家 级 高 新 技 术 产 业 开 发 区 。

13a The development zone is located in the famous Chinese country tourism scenery-13b Hangzhou city center. 13c It is a national high-tech Industrial Development Zone which approved by the state council in 1991. （chtb_0011）

Example 14. 14a 不一会，在一小片开阔地，导游——一位会操5种欧洲语言的女郎——命令车子停下，14b 她让大家趁太阳正露出笑脸向我们致意的时机，赶紧观赏一下圣岳的容颜。

bù yī huì zài yī xiǎo piàn kāi kuò dì dǎo yóu yī wèi huì cāo zhòng ōu zhōu yǔ yán de nǚ láng mìng lìng chē zi tíng xià tā ràng dà jiā chèn tài yáng zhèng lù chū xiào liǎn xiàng wǒ men zhì yì de shí jī gǎn jǐn guān shǎng yí xià shèng yuè de róng yán

14a Soon, in a small piece of open space, the tour guide, a young woman who can speak five European languages, commands the car to stop. 14b She lets everyone hurriedly enjoy the face of the sacred mountain while the sun is revealing the smiling faces to us. （chtb_0207）

6. Ellipsis is a series of marks that usually indicate an intentional omission of a word, sentence or whole section from the original text being quoted. An ellipsis can also be used to indicate an unfinished thought or, at the end of a sentence, a trailing off into silence

Example 15. 15a 这年五月，九七歌仔戏创作研讨会在厦门召开……

zhè nián wǔ yuè jiǔ qī gē zǎi xì chuàng zuò yán tǎo huì zài xià mén zhào kāi

15a In May this year, Jiuqi opera creation seminar was held in Xiamen…（chtb_0836）

Example 16. 16a 处于族群复杂的环境中，邓相扬对族群间的冲突、融合、文化的消失……等情形非常敏感。

chǔ yú zú qún fù zá de huán jìng zhōng dèng xiāng yáng duì zú qún jiān de chōng tū róng hé wén huà de xiāo shī děng qíng xíng fēi cháng mǐn gǎn

16b In an ethnically complex environment, Deng Xiangyang is very sensitive about ethnic conflict, integration, and disappearance of cultures ... etc. situation. （chtb_1011）

For the processing of the ellipsis, if at the end of the sentence (as in example 15) equivalent to the end of sentence label, it is treated as period. If in the middle of sentence (as in example, 16), it is not treated as the boundary of EDUs.

7. Other symbols cannot be used as a clause boundary, so Fig. 2 does not count them. The parentheses are usually used to explain a specific word or phrase. No matter whether parentheses are among sentences, they can't be treated as EDUs boundary (eg Example 16). Pause is Chinese special punctuation indicating the word or phrase pause, it can't be regarded as EDUs boundary no matter where it is in sentence (eg Example 4, Example 16 and Example 17). The title punctuation belongs to the clause internal symbols and do not have to be tackled even if there are other symbols.

yǐ jīngbàomíngcǎifǎngdejìzhěyǒu míng wénzì
Example 17. 17a 已 经 报 名 采 访 的 记 者 有 1 3 2 0 名 （ 文 字 6 0 1
míng shèyǐng míng diàntái diànshìtái míngděng
名 、 摄 影 1 7 1 名 、 电 台 、 电 视 台 4 0 0 名 等 ）。

17a Reporters have already registered to cover are 1320 (text 601, photography 171, radio and television 400 etc.) （chtb_0306）

yī jiǔ jiǔ sì niánbǎn jīngjì báipí shū zhōngguó jīngjì
Example 18. 18a 一 九 九 四 年 版 《 经 济 白 皮 书 ： 中 国 经 济
xíngshì yǔ zhǎnwàng jìnrì yóu zhōngguó fāzhǎn chūbǎnshè chūbǎn
形 势 与 展 望 》， 近 日 由 中 国 发 展 出 版 社 出 版 。

18a The 1994 edition of the《Economic White Paper: China's economic situation and outlook》recently published by the China Development Press. （chtb_0268）

2.4 Special Sentence Pattern Processing

The definition of EDUs above contains the most cases, but there are some special sentences which require special handling. This kind of sentence is a single sentence in whole, but their component is a complex sentence, as in Example 19, 20. According to preliminary statistics, there are total 906 sentences in the top 70 documents of CTB6.0, of which have 114 of such sentences (12.5%).

jù tǒngjì zhèxiē chéngshì qùnián wánchéng
Example 19. 19a 据 统 计 ， 19b 这 些 城 市 去 年 完 成
guónèi shēngchǎnzǒngzhí yībǎijiǔshí duōyì yuán bǐ kāifàng qián de
国 内 生 产 总 值 一 百 九 十 多 亿 元 ， 19c 比 开 放 前 的
yījiǔjiǔyī niánzēngcháng jiǔchéngduō
一 九 九 一 年 增 长 九 成 多 。

19a According to the statistics, 19b these cities had completed more than one hundred and ninety billon Yuan of gross domestic product (GDP) last year. 19c More than ninety percent growth compare with opening before in 1991. （chtb_0003）

tārènwéi zhèwèiměiguó cānyìyuánsuǒwèiqīpiànzhī
Example 20. 20a 他 认 为 ， 20b 这 位 美 国 参 议 员 所 谓 “欺骗” 之
shuōshì háowúgēnjù de bìngduìqífābiǎobùfùzérènde yánlùnbiǎoshì
说 是 毫 无 根 据 的 ， 20c 并 对 其 发 表 不 负 责 任 的 言 论 表 示
yíhàn
遗 憾 。

20a He thinks, 20b the United States senator so-called "cheating" is baseless, 20c and expresses regret for the published irresponsible comments. （chtb_0606）

In example 19, the contents of "statistics" are 19b and 19c. The subject of 19c is the "gross domestic product" in 19b. 21c is the explanation of 21b. In example of 20, 20a is the whole subject and predicate of 20. 20b and 20c are its object. 20b and 20c have a parallel relation. This paper will treat "according to statistics" in the example of 19 as a sentence attachment, in constructing discourse tree it will be attached to the node of 19b and 19c combination result.

Of course, with further research, we may add some special sentences or proper adjustment of these clause judgment rules. In summary, the clauses can be defined as follows in this paper:

Clause is the basic unit of discourse analysis, including traditional single sentence and complex sentence. It contains at least a predicate and expresses a proposition in structure. Clause is not like other clauses structure grammatical components and has any relationship with another. Clause must have punctuation segmentation in form, usually commas, semicolons and period, etc. In the actual corpus, some so-called phrases similar with typical clauses in the structure, function, and form in the specific conditions are treated as clauses.

3 Automatic Identification of the Chinese EDUs

According to the definition of EDUs and punctuations processing mode described in section 2, author manually annotates whether the punctuations (full stop, question mark, exclamation mark, semicolon, colon, comma, ellipsis and dash) can be considered to be the boundary of EDUs in 70 documents (chtb_0001-chtb_0070). 2069 punctuations may be boundary EDUs in total, of which there are 1597 cases are positive (including period, question mark, exclamation mark and semicolon 751, all of them labeled as positive cases) and 472 cases are negative. Since period, question mark, exclamation mark and semicolon must be a clause boundary, this article gives two sets of experiments, the first group uses all punctuations annotation information, and the other uses only the punctuations of the probably clause boundary label information .

We referenced features of recognition whether a comma is sentence boundary in [16], part of the features from Li and Zhou [17], and simple features for this special task. Then we extracted punctuation's the syntax and other feature as follows for experiment:

1 Part of speech tag and string representation before and after punctuation (f1 f2 f3 f4), e.g. f1=NR f2=上海 f3=VV f4=建设

2 The phrase label of the left sibling and the phrase label of their right sibling in the syntactic parser tree, as well as their conjunction (f5 f6 f7), e.g.f5=VP f6=VP f7=VP+VP.

3 The conjunction of the ancestors, the phrase label of the left sibling, and the phrase label of the right sibling (f8).e.g f8=VP+VP+VP

4 Whether there is a subordination conjunction to the left and right of the punctuation. The left search starts from current punctuation and stops at the previous punctuation mark or the beginning of the sentence. The right search starts from current punctuation and stops at the next punctuation mark or the end of the sentence (f9, f10), eg. f9=CS f10=noCS

5 Whether the parent of the punctuation is coordinating IP construction (f11), e.g f11=coordIP

6 The level of the punctuation in place of syntactic tree level (f12). e.g.f12=3

7 The sentence punctuation set (f13), punctuation type (f14).e.g f13=, +, +。 f14=,

8 Whether the length to the left and right segment of the punctuation is small than 5(f15 f16). The absolute left and right segment span of current punctuation length difference is more than 7 (f17). e.g f15<5 f17>7

9 Whether the punctuation parent is NP (f18), whether the punctuation's left and right sibling is NP (f19 f20).e.g f19=NP

10 First word of part of speech and string representation of current punctuation to a previous punctuation or the beginning of the sentence span (f21 f22). f21=NR f22=浦东

The features of the first and second punctuation in example 1 are as follows. The EDUs boundary punctuation instance is indicated by +1, otherwise indicated by -1.

For example: f1=NR f2=上海 f3=VV f4=建设 f5=VP f6=VP f7=VP+VP f8=VP+VP+VP f12=7 f13=, +, +。 f14=, f17>7 f18=yesNP f21=NR f22=浦东 -1

f1=NN f2=工程 f3=AD f4=因此 f5=IP f6=IP f7=IP+IP f8=IP+S+IP f12=1 f13=, +, +。 f14=, f21=VV f22=建设 +1

According to the features shown above, this paper using decision tree, the maximum entropy and Bayesian with Mallet machine learning package separately carried on the experiment to classify the punctuation. The experiment using 10 fold cross validation, the possible EDU boundary punctuations (such as comma, colon etc) recognition accuracy, F-measure score for positive and negative instance as shown in table 1:

Table 1. The results of possible clauses boundary recognition

classifier	Gold-standard Parser			Automatic Parser		
	accuracy	F1(+)	F1(-)	accuracy	F1(+)	F1(-)
MaxEnt	92.4%	93.6%	89.2%	90%	92%	86.8%
C45	**93.2%**	**94%**	**91.8%**	**90.7%**	**92.4%**	**88.2%**
NaiveBayes	91.4%	93.2%	88.8%	86.6%	89.5%	81.7%

Gold-standard parser means parsing result given in the CTB6.0 corpus, automatic means the result produced by Berkeley parser. From table 1 we can see that the best experiment result accuracy is 93.2% for possible EDU boundary punctuations using gold-standard parser, while using automatic parser for possible EDU boundary punctuations the accuracy is 90.7%. Table 1 also gives the F1 measure for positive punctuations and negative punctuations, the results show that the positive F-measure results are better than negative results due to positive train instances are larger than negative and positive instances are relatively easy to recognize.

Table 2. The results of all clauses boundary recognition

classifier	Gold-standard Parser			Automatic Parser		
	accuracy	F1(+)	F1(-)	accuracy	F1(+)	F1(-)
MaxEnt	93.6%	95.9%	87.6%	92%	95%	85.3%
C45	**95%**	**96.7%**	**91.7%**	**93.6%**	**95.8%**	**88.1%**
NaiveBayes	93%	95.5%	87.2%	88.5%	92.3%	80.7%

From table 2 we can see that the best experiment result accuracy is up to 95% for all EDU boundary punctuations using standard parser, while using automatic parser the accuracy is 93.6%. Compare the result of table 1 and table 2 we can find that: First, although there are increase in accuracy and F1(+) for all punctuations, the F1(-) changes little for best classifier C45,while there are decrease using MaxEnt and Naive Byes. The reasons are for all punctuations contain period which is definite positive instance and will enlarge positive training instances and affect the performance of classification. Second, the best classifier for our experiment is C45 for possible punctuations or all punctuations using gold-standard or automatic parser. This shows that EDU boundary punctuations reorganization is relatively simple task which can learn effective rules.

Experiments of this paper show that the result of EDU boundary recognition based on punctuations is feasible and can be calculated.

4 Conclusion

Different theories have different definition of EDU. This article discusses the judgment and recognition of EDUs for Chinese discourse structure analysis from the aspect of Chinese discourse structure analysis. We manually annotate 70 Chinese documents, and on the basis of analysis the corpus defines the Chinese EDU as clause. According to the result of the annotation, this paper concludes the principle for the single sentence, complex sentence and all kinds of punctuations. Finally we give a special kind of sentence processing method. We used many features to train a statistical model using data that we annotate whether punctuation is clause boundary and automatically recognize boundary punctuation. The results show that our definition of clause theory corresponds with Chinese character and has strong distinguished ability in calculation.

The features that this paper used are all plane feature. Zhou etc. [18-19] shows that structured information can improve performance of the related natural language processing tasks. In further study we will use kernel function for our task. In addition, the clause recognition can be converted into a sequence tag problem. The method of sequence tag has obtained good effect in other natural language processing applications [20-21]. So our other job will combine serialized method to improve the performance of clauses recognition. At present the size of corpus is small, we will further increase the corpus size, find problem and improve the principle of judging discourse unit, hoping to find differentiate, consistency and practicability recognition method of EDUs finally.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No.61070123 and No. 90920004, the National 863 Project of China under Grant No. 2012AA011102.

References

1. Hobbs, J.R.: Coherence and co reference. Cognitive Science 1, 67–90 (1979)
2. Hobbs, J.R.: Information, Intention, and Structure in Discourse: A first draft. In: Maratea, Burning Issues in Discourse, NATO Advanced Research Workshop, pp. 41–66 (1993)
3. Givon, T.: Topic continuity in discourse. John Benjamins, Amsterdam (1983)
4. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematic for the organization of turn-taking in conversation. Language 50, 696–735 (1974)
5. Polanyi, M.: Personal knowledge: Towards a post-critical philosophy. Psychology Press (1998)
6. Grosz, B.J., Sidner, C.L.: Attention, intentions, and the structure of discourse. Computational Linguistics 3, 175–204 (1986)
7. Mann, W.C., Thompson, S.A.: Relational propositions in discourse. Discourse Processing 1, 57–90 (1986)
8. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: Toward a functional theory of text organization. Text 3, 243–281 (1988)
9. Carlson, L., Marcu, D., Okurowski, M.: Building a Discourse-tagged Corpus in the Framework of RST. Current Directions in Discourse. Kluwer, New York (2003)
10. RST-DT.: RST Discourse Treebank. Linguistic Data Consortium (2002)
11. PDTB-Group.: The Penn Discourse Treebank 2.0 annotation manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania (2008)
12. Ming, Y.: Annotation and analysis of Chinese financial news commentaries in term of rhetorical structure. Chinese media university PhD dissertation (2006)
13. Yue, M.: Rhetorical Structure Annotation of Chinese News Commentaries. Journal of Chinese Information Processing 4, 19–23 (2008)
14. Chen, L.P.: English and Chinese discourse structure annotation theory and practice, Shanghai International Studies University PhD dissertation (2006)
15. Xing, F.Y.: Research of Chinese complex sentence. Commercial Press, Beijing (2001)
16. Xue, N.W., Yang, Y.Q.: Chinese sentence segmentation as comma classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers. Portland Oregon, pp. 631–635 (2011)
17. Li, J.H., Zhou, G.D.: Unified Semantic Role Labeling for Verbal and Nominal Predicates in Chinese Language. ACM Transaction on Asian Language Information Processing 10(3), Article 13 (2011)
18. Zhou, G.D., Li, J.H., Fan, J.X., Zhu, Q.M.: Tree kernel-based semantic role labeling with enriched parse tree structure. Information Processing and Management 47, 349–362 (2011)
19. Zhou, G.D., Zhu, Q.M.: Kernel-based semantic relation detection and classification via enriched parse tree structure. Journal of Computer Science and Technology 1, 45–56 (2011)
20. Zhou, G.D.: Direct modeling of output context dependence in discriminative Hidden Markov Model. Pattern Recognition Letters 5, 545–553 (2005)
21. Zhou, G.D.: Discriminative hidden Markov modeling with long state dependence using a kNN ensemble. In: COLING 2004 (2004)

A Corpus-Based Study of Epistemic Modality Markers in Chinese Research Articles

Yuyin He[1,2,*] and Han Wang[3]

[1] College of Foreign Languages, Beihang University / 37, Xueyuan Rd, Haidian District, Beijing 100191, China
[2] Broadcast Media Language Branch, National Language Resources Monitoring and Research Center, Communication University of China / 1, Dingfuzhuang Eastern Str., Chaoyang District, Beijing 100024, China
yuyinhe@buaa.edu.cn
[3] China Insurance Regulatory Commission / 15, Finance Avenue, Western City District, Beijing 100140, China
tim10101112@163.com

Abstract. Research on the use of hedging strategies in research articles has received increasing attention. This paper presents a pilot study of a corpus-based study of epistemic modality markers(EMMs) in Chinese research articles. We first study the use of EMMs in Chinese research articles in linguistics, medicine and aerospace, and then compare their frequency use in research articles of Chinese, English, French and Norwegian, with the data for the other three languages from Vold(2006)'s study. The findings are: 1) disciplines will not affect the frequency use of EMMs in Chinese research articles; 2) culture affects the frequency use of EMMs significantly and Chinese research articles are more heavily hedged than the western ones in the sense of statistics. Chinese research articles are characterized by low uncertainty avoidance.

Keywords: epistemic modality markers(EMMs), hedging strategies, uncertainty avoidance, Chinese research articles, comparative studies.

1 Introduction

Epistemic modality markers(EMMs) function as an important and frequently used type of hedges. Zadeh [1], the founder of fuzzy sets, introduced the expression of "hedges" to describe linguistic fuzziness. Lakoff [2] is the first linguist to probe into the semantics of hedges and introduce the term "fuzzy logic". Hyland [3] put forward that academic research writing is characterized by hedging, "one part of epistemic modality", which "indicates an unwillingness to make an explicit and complete commitment to the truth of the propositions". Hyland [3] presented his definition of hedges as "the means by which writers can present a proposition as an opinion rather than a fact: item are only hedges in their epistemic sense, and only then when they mark uncertainty". The use of hedges in scientific discourse reflects Halliday's [4] interpersonal macro function of

* Corresponding author.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 199–208, 2013.

language, for hedging can be considered as a rhetorical device used to convince and influence the reader. Hedging is an argumentative strategy vital in research articles.

Quite a considerable amount of research on hedges and their communicative functions in English academic discourse has been carried out during the last few decades [3], [5-7], but very little research has been done on hedging in Chinese academic discourse, and even less has been focused on the comparison between Chinese and Western languages from the view of the use of hedging strategies in academic texts.

This article aims to find out whether disciplinary affiliation has any influences on the use of hedges in Chinese research articles and to what extent cultural factors can affect the use of hedging strategies in research articles. Section 2 introduces EMMs and semantic explanations and recognitions of EMMs in Chinese research articles. Section 3 describes research design. Section 4 presents the analysis results, and Section 5 concludes the paper with the research findings, limitations, and areas for future research.

2 Epistemic Modality Markers (EMMs)

EMMs expressing uncertainty act as a dominant and basic type of hedges. Vold [7] defined them as "linguistic expressions that qualify the truth value of a propositional content (for example perhaps, probably)". To illustrate her definition, Vold [7] also expounded the truth value of the proposition "smoking causes lung cancer" as follows:

a) It is possible that smoking causes lung cancer.
b) Smoking probably causes lung cancer.
c) We know that smoking causes lung cancer.

She concluded that the proposition was marked as a possibility in a), a probability in b), and a certainty in c).

Our study will employ Vold [7]'s definition of EMMs. In Chinese research articles, they are also used to express the truth of a propositional content, more exactly, the sense of uncertainty. Take the following sentence as an example:

(1) 但海马和小脑内的Cav-1 的表达与我们的实验结果不同，我们认为可能与动物的性别差异有关(dàn háimǎ hé xiáonǎo nèi de Cav-1 de biǎodá yǔ wǒmen de shíyàn jiéguǒ bùtóng, wǒmen rènwéi kěnéng yǔ dòngwù de xìngbié chāyì yǒuguān)(However, the expressions in Cav-1 in sea horses and cerebella are different from those in our experimental results, thus we think this may be associated with sex differences of animals)。

The EMM 可能(kě'néng) (may or could) expresses a kind of possibility of the proposition "但海马和小脑内的Cav-1 的表达与我们的实验结果不同，[这]与动物的性别差异有关(dàn háimǎ hé xiáonǎo nèi de Cav-1 de biǎodá yǔ wǒmen de shíyàn jiéguǒ bùtóng, [zhè] yǔ dòngwù de xìngbié chāyì yǒu guān)(However, the expressions in Cav-1 in sea horses and cerebella are different from those in our experimental results, [which] is associated with sex differences of animals)".

To identify EMMs in academic discourse, our study employs Vold [7]'s following two criteria:

"i) The marker had to explicitly qualify the truth value of a certain propositional content.

ii) The marker also had to be a lexical or a grammatical unit."

In accordance with these two criteria, the lexical verb suggest in English or 表明 (biǎomíng, show/indicate) in Chinese, which in many other contexts would be classified as an EMM, is not classified as so in English example sentence (2) and Chinese example sentence (3), because there is no proposition content to be modified:

(2) … and flow losses at sudden area changes (Floss) are included in the gas-dynamic model via standard engineering correlations as suggested by Groth et al.

(3) 通过一系列的实验表明,我们证实了构建的复制子载体具有自主复制的能力(tōnggguò yíxìliè de shíyàn biǎomíng, wǒmen zhèngshí le gòujiàn de fùzhì zǐ zàitǐ jùyǒu zìzhǔ fùzhì de nénglì)(A series of experiments show that we have verified the constructed replicator carriers are capable of automous replication)……

When using the criteria to identify EMMs, we should note that many polysemous words such as may and 比较(bǐjiào, compare/comparison/basically) do not always function as EMMs. Take the word 比较(bǐjiào, compare/comparison/basically) as an example, it could be used as a verb expressing "to compare"(as in 比较这两组数据 bǐjiào zhè liǎng zǔ shùjù compare these two sets of data), or as a noun "comparison"(as in 通过比较我们发现 tōnggguò bǐjiào wǒmen fāxiàn we find, through comparison, that), or as an epistemic marker, which is the function this paper is interested in (as in 这比较能反映我国腹泻儿童中 HuCV 感染的状况 zhè bǐjiào néng fǎnyìng wǒguó fùxiè értóng zhōng HuCV gánrǎn de zhuàngkuàng this basically reflects the situation of HuCV infection in children with diarrhea in our country). In order to make the analysis and comparison precise, it is necessary to classify the occurrences according to their meaning in a particular context. All the occurrences should be divided into two groups: "those occurrences that were classified as EMMs expressing uncertainty, and other meanings" [7].

In order to classify the occurrences as systematically as possible, we can use a substitution test [7], which means trying to replace the polysemous marker with an intrinsically epistemic marker in cases of doubt. If such a substitution is possible with no great change of meaning, the word can be classified as an EMM; otherwise, it should be classified into "other meanings".

Another useful test is trying to add a clause like but I'm not sure or 但我并不确定 (dàn wǒ bìng bú quèdìng, but I am not sure) after the marker/ proposition and try to formulate a less hedged version of the proposition [7]. When added an uncertain expression, the proposition still sounds natural, and thus the occurrence should be classified as epistemically modal. If this is not possible, the occurrence should be considered non-epistemic.

We can also use syntactic criteria and semantic criteria to classify the occurrence of Chinese polysemous markers. For example, 基本(jīběn, essential/ basically) as an adjective meaning "essential" is not related to epistemically modal senses, while 基本(jīběn, essential/ basically) as an adverb meaning "basically" often functions as an EMM.

3 Research Design

3.1 Research Questions

This paper aims to investigate the following two questions:

1) Is there any influence of the discipline on the use of EMMs in Chinese research articles?

2) Do different cultures affect the use of EMMs in research articles?

3.2 Research Data

A Chinese corpus is compiled of 60 Chinese research articles in the disciplines of linguistics, medicine and aerospace, with 20 articles for each discipline. All of these articles are written by native speakers and are taken from well-recognized and high-quality refereed journals. For this study, the articles come from the sources listed in Table 1.

Table 1. Sources used for this study

Discipline	Sources	No. of articles
Linguistics	现代外语 xiàndài wàiyǔ Modern Foreign Languages	10
	语言教学与研究 yǔyán jiàoxué yǔ yánjiū Language Teaching and Linguistic Studies	10
Medicine	病毒学报 bìngdúxué bào Chinese Journal of Virology	10
	生理学报 shēnglǐxué bào Acta Physiologica Sinica	10
Aerospace	航空动力学报 hángkōng dònglì xuébào Journal of Aerospace Power	10
	航空学报 hángkōngxué bào Acta Aeronautica et Astronautica Sinica	10

The research articles are preprocessed by excluding abstracts, notes, references, quotations, linguistic examples, tables and figures, with merely the body remained. The segmentation and POS tagging tool, which is developed by Broadcast Media Language Branch of National Language Resources Monitoring and Research Center in Communication University of China, is used to process the corpus material so that we can obtain the words for each subcorpus. The POS-tagged files are imported into Antconc, a freeware corpus analysis toolkit, and we find that the linguistics subcorpus has 78,108 words, the medicine subcorpus 43,924 words, and the aerospace subcorpus 43,505 words. Since the number of words is very unevenly distributed over the disciplines, attention should be paid to the relative frequency rather than to the number of occurrences.

3.3 Research Method

To select the markers in Chinese academic discourses, we study the markers' frequencies in an exploratory corpus, which consists of 1/5 articles of the whole Chinese corpus. The exploratory corpus reflects the whole corpus' distribution over different journals and disciplines in its composition to ensure that it represents the whole corpus. In the study of the exploratory corpus, all epistemic markers are written down and studied in research articles for their frequencies and concordances through Antconc. The most frequent EMMs are selected for a quantitative analysis of the whole corpus. These markers are listed in Table 2.

Table 2. The most frequent EMMs in the exploratory corpus

EMMs	Occurrences
较 jiào relatively/quite	96
可能 kě'néng possible/probable/likely/perhaps/may/could	34
比较 bǐjiào basically	25
相对 xiāngduì relatively	19
表明 biǎomíng indicate	16
认为 rènwéi consider	16
基本 jīběn basically	13
一般 yìbān commonly	12
通常 tōngcháng generally	10
显示 xiǎnshì suggest	9
应该 yīnggāi must	6

4 Results and Discussions

4.1 No Statistically Significant Differences of EMMs Cross-Disciplinarily

Table 3 illustrates the results for the Chinese linguistics, medicine and aerospace articles. As mentioned above, many of the markers are polysemous, and all the occurrences therefore have to be checked whether they are the use of EMMs according to their meanings. Both the column of epistemic occurrences and the column of all occurrences are divided into two parts, with one indicating the relative frequency per thousand words(f/1000) and the other giving the exact number of all occurrences(No.). By comparing the two columns, we can find out the extent to which a marker can be seen mainly as an EMM.

Table 3. EMMs in the Chinese research articles

	disciplines	EMM Occurrences		All Occurrences	
		f/1000	No.	f/1000	No.
较 jiào reltively/quite	Linguistics	1.04	81	1.08	84
	Medicine	1.96	86	2.57	113
	Aerospace	2.69	117	2.90	126
可能 kě'néng possible/probable/likely/perhaps/may/could	Linguistics	1.27	99	1.87	146
	Medicine	1.55	68	1.59	70
	Aerospace	0.74	32	0.85	37
比较 bǐjiào basically	Linguistics	0.45	35	0.77	60
	Medicine	0.14	6	1.73	76
	Aerospace	0.64	28	1.65	72
认为 rènwéi consider	Linguistics	0.42	33	1.70	133
	Medicine	0.27	12	0.50	22
	Aerospace	0.09	4	0.18	8
一般 yìbān commonly	Linguistics	0.37	29	0.65	51
	Medicine	0.18	8	0.25	11
	Aerospace	0.28	12	0.41	18
基本 jīběn basically	Linguistics	0.09	7	0.42	33
	Medicine	0.46	20	0.48	21
	Aerospace	0.25	11	0.64	28
应该 yīnggāi must	Linguistics	0.35	27	0.63	49
	Medicine	0.07	3	0.07	3
	Aerospace	0.21	9	0.39	17
通常 tōngcháng generally	Linguistics	0.19	15	0.23	18
	Medicine	0.09	4	0.11	5
	Aerospace	0.18	8	0.23	10
相对 xiāngduì relatively	Linguistics	0.42	33	0.63	49
	Medicine	0.25	11	0.41	18
	Aerospace	0.14	6	0.46	20
表明 biǎomíng indicate	Linguistics	0.26	20	0.81	63
	Medicine	0.50	22	1.89	83
	Aerospace	0.21	9	0.46	20
显示 xiǎnshì suggest	Linguistics	0.17	13	0.78	61
	Medicine	0.18	8	0.25	11
	Aerospace	0.07	3	0.09	4

From Table 3, it is obvious that the markers 较 (jiào, relatively/quite) and 可能 (kě'néng, possible/probable/likely/perhaps/may/could) are the most frequently used EMMs in Chinese research articles in all of the three disciplines, with especially 较 (jiào, relatively/quite) in Chinese aerospace articles, whose frequencies (3.57) account for almost half of those of all the 11 markers. Comparing the frequencies of the

EMMs in these three different disciplines, we can see some interesting differences. For instance, although比较 (bǐjiào, basically) is the third most frequently used markers in linguistics and aeronautics discourses, it is the second least frequently used in medical articles; on the other hand, 基本 (jīběn, basically) is frequently used in medical articles, but the linguistic articles seem to avoid using it. It suggests that discourses of linguistics and aerospace use EMMs in a similar way, while the medical discourse uses them differently.

From Table 3, we can see that in the linguistic discourses, the epistemic modality markers studied constitute approximately 5.02 per thousand words. The corresponding number is 5.65 per thousand words for the medical research articles, and 5.49 per thousand words for aeronautic articles. Although more hedges are used in medical articles than the other two disciplines, the differences are very slight. In order to discover whether the disciplinary affiliation affects the frequencies of EMMs used in Chinese research articles, we employ Z-Test for ratio tests on all data sets to determine whether there are statistically significant differences of the usage frequencies of EMMs among three different disciplines.

Now that $n*p \geq 5$ and $n*(1 - p) \geq 5$(Note: $p = x/n$, and x can be taken as the occurrences of EMMs and n the size of subcorpus in this study.), the three data sets are valid for Two Independent Population Ratio Test among one another. The test results display that Z value for Ratio Test between linguistics and medicine is -1.457, Z value for Ratio Test between linguistics and aerospace -1.105, and Z value for Ratio Test between medicine and aerospace -0.302. At 0.05 significance level, the critical value is 1.645. Apparently, the absolute values of each Z values for Ratio Test among three different disciplines do not go beyond the critical value. The data sets have provided sufficient data to conclude that the usage frequencies of EMMs among three different disciplines are not of statistically significant difference. Different disciplines in Chinese research articles will not bring about frequency differences in using EMMs.

4.2 Chinese Research Articles Much More Heavily Hedged in Comparison with those in Other Languages

Table 4 shows language-specific differences of EMMs in research articles, regardless of disciplines. In order to achieve homogeneity within research data, only the data in Chinese research articles in linguistics and medicine are chosen to use in the table, for the data quoted from Vold [7] of EMMs in French, Norwegian and English research articles are in linguistics and medicine.

Table 4. Language-specific differences of EMMs

Language	f/1000	no. of occurrences	no. of words
Chinese	5.2	640	122,032
French	2.1	268	129,907
Norwegian	3.3	447	133,813
English	3.5	1630	468,909

Vold [7] employed two-tailed Mann-Whitney on language pairs to illustrate language-specific differences among French, Norwegian and English in using EMMs. She found that both the English authors and the Norwegian authors differed significantly from the French authors, with p = 0.001 and p = 0.008 respectively, whereas the English and the Norwegian authors did not differ significantly from each other, with p = 0.465.

Since $n*p \geq 5$ and $n*(1 - p) \geq 5$, Two Independent Population Ratio Test appears to be a suitable test for the four data sets in the above table. The test results indicate that Z value for Ratio Test between French and Norwegian and that between French and English is -6.307 and -8.896 respectively, while Z value for Ratio Test between Norwegian and English is 0.747. Apparently Z values for Ratio Test in our study have produced the same statistic conclusions as two-tailed Mann-Whitney test in Vold [7]'s study.

According to the above table regarding language-specific differences of EMMs, the Chinese texts have the highest frequency of EMMs, with a relative frequency of 5.2. Then follow the English and Norwegian texts, with almost the same relative frequency of 3.5 and 3.3. The French research articles use least epistemic modality, with a relative frequency of 2.1.

We compute Z values for Ratio Test to check whether the differences between Chinese and French Norwegian, and English are statistically significant. The results show that the Chinese authors differ significantly, at the level of 0.05 in a statistical sense, from the French authors, the Norwegian authors and the English authors in using EMMs in writing research articles, with Z values are 13.318, 7.396 and 8.896 respectively.

Now that EMMs are one of the most dominant and frequent type of hedges, it is justified that Chinese research articles are more heavily hedged than western ones.

4.3 Functions of Hedging Strategies and Chinese Traditional Academic Culture

From above analyses, it appears that the most important factor that affects the frequencies of EMMs is language or culture.

As for the reasons why the use of EMMs is language specific, the functions hedging strategies serve should be explored. Two major types of hedges will be discussed here: real hedges and strategic hedges. The former refer to hedges "used to convey real uncertainty" and "serve to give an accurate picture of the level of certainty" [7]. The real hedges are used not to be polite or modest but to be precise. Nevertheless, strategic hedges are hedging associated with tentativeness, cautiousness, politeness and a humble attitude. This kind of hedges does not necessarily express uncertainty; but rather they are part of the conventions for academic writing. These hedges have a variety of functions. They can be used in a context to express possible opinions or interpretations, and thus the author anticipates potential criticism [7]. They may also be used as a politeness strategy in order to cautiously criticize fellow researchers [7]. The hedging strategies also fill other functions. For example, they can be used to tune down

statements and claims in order for the author to be less vulnerable to criticism. And authors sometimes use hedges to encourage the readers to give feedbacks to their discourse.

Vold [7] claims that the use of hedging strategies may be associated with different academic cultures. The Chinese academic culture is greatly influenced by the Chinese tradition, which tends to create a harmonious atmosphere and avoid direct conflicts with others. Thus when Chinese researchers put up a new theory, they tend to mitigate its truth; when they comment on others' theories, they try to do so in a roundabout way to avoid conflicts. Hedging strategies appear to be an important way to achieve their intention.

To check the traditional academic culture, we have studied 梦溪笔谈(mèng xī bǐ tán, Notes at Mengxi Garden) by Shen Kuo, a famous ancient scientist in Song Dynasty. In his book Shen Kuo used many EMMs in his articles of various disciplines. Some cases are listed in the following two examples, in which 疑(yí, may) and 恐(kǒng, perhaps) are EMMs [8]:

* 于定国饮酒数石不乱，疑无此理。（斛斗）[yúdìngguó yǐnjiǔ shù dàn bú luàn，yí wú cǐ lǐ。(hú dòu)][Yu Dingguo won't be drunk after drinking several dans of spirits, which may sound ridiculous. (Dendrobium Measures) (Notes: dan is a unit of measure, which equals l20 jin.)]

* 恐四代之法，容有不相袭者。（不袭古法）[kǒng sìdài zhī fǎ，róng yǒu bù xiāngxí zhě。(bù xí gǔfǎ)][Perhaps some laws of the four dynasties of Yu, Xia, Shang and Zhou have not followed ancient laws. (Not Following Ancient Laws)]

Scholars' different cultural backgrounds are reflected in their linguistic choices in writing research articles. Western scientists, French scientists in particular, tend to be more critical and more authoritarian [9]. Salager-Mayer et al [9] considered France "of all the European countries, the one that has most obstinately clung to its traditions (cultural legacy) by attempting, inter alia, not to succumb to" "'wreck of subjectivity', and to fiercely resist the Anglo-Saxon cultural penetration". French-speaking scientists use fewer EMMs when expressing themselves or offering facts.

As a matter of fact, Chinese research articles are characterized by low uncertainty avoidance. Chinese scholars, deeply influenced by the thought-patterns of Chinese traditional academic culture, tend to be humble in putting forwards their ideas with uncertainty expressions, trying not to imposing their ideas upon readers and colleagues. EMMs assist Chinese authors to show politeness in the construction within academic discourse community. But scholars of other nationalities with other cultural backgrounds might find this style somewhat inappropriate for scientific discourse.

5 Conclusion

In this paper, we examine the use of a selection of EMMs indicating uncertainty in research articles of medicine, linguistics and aerospace in Chinese and also compare the results with the data Vold [7] collected in her research. The results show that no statistically significant differences in the frequency of the selected markers can be detected among disciplines in Chinese research articles. At the same time, Chinese

researchers, deeply influenced by traditional Chinese academic culture, use considerably more hedging strategies than their western colleagues do.

It should be noted that this investigation has some shortcomings. The disciplines involved are linguistics, medicine and aerospace. But there might be bigger differences among some other disciplines, so other disciplines and sub-disciplines should be compared to answer the question of disciplinary variation fully. In addition, the size of research article corpus is a bit small. Thus in the follow-up study, we shall examine EMM s in a large-scale comprehensive corpus of Chinese research articles to get a better understanding of low uncertainty avoidance in Chinese academic language.

Acknowledgements. The research was supported by the Project Funds of Humanities and Social Sciences from Ministry of Education of China under grant 11YJA740030 and Fundamental Research Funds for the Central Universities from Beihang University, China under grant YWF-12-JRJC-028.

References

1. Zadeh, L.A.: A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges. Journal of Cybernetics 2(3), 4–34 (1972)
2. Lakoff, G.: Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. Journal of Philosophical Logic 2(4), 458–508 (1973)
3. Hyland, K.: Hedging in Scientific Research Articles. John Benjamins, Amsterdam, 3, 5 (1998)
4. Halliday, M.A.K.: An Introduction to Functional Grammar, 2nd edn. Edward Arnold, London (1994)
5. Salager-Meyer, F.: Hedges and Textual Communicative Function in Medical English Written Discourse. English for Specific Purposes 13, 149–170 (1994)
6. Lewin, B.A.: Hedging: an Exploratory Study of Authors' and Readers' Identification of 'Toning Down' in Scientific Texts. Journal of English for Academic Purposes 4, 163–178 (2005)
7. Vold, E.T.: Epstemic Modality Markers in Research Articles: a Cross-Linguistic and Cross-Disciplinary Study. International Journal of Applied Linguistics 16, 61–87, 65, 65, 65, 69, 72, 72, 77, 76-78, 81, 81, 82, 82 (2006)
8. University of Science and Technology of China, Notes at Mengxi Garden Commentary and Annotation Group from Hefei Iron and Steel Company. Commentary and Annotation for Notes at Mengxi Garden: Natural Science Section. Anhui Science and Technology Publishing House, Hefei, 128, 230 (1979)
9. Salager-Meyer, F., Angeles M., Ariza A., Zambrano N.: The Scimitar, the Dagger and the Glove: Intercultural Differences in the Rhetoric of Criticism in Spanish, French and English Medical Discourse (1930-1995). English for Specific Purposes 22, 223, 239 (2003)

Rule-Based Computation of Semantic Orientation for Chinese Sentence

Jiang Yang[1] and Min Hou[2]

[1] School of Foreign Studies, Hunan University of Science and Technology, Xiangtan, China
yangjiang@hnust.edu.cn
[2] Broadcast Media Language Branch, Communication University of China, Beijing, China
houmin@cuc.edu.cn

Abstract. Semantic Orientation refers to the positive or negative attitude, standpoint or opinion on a certain person, object or affair. It is of degree diversity and combinability. Rule-based computation of SO for Chinese sentence, after pre-processing the input sentence in lexical and dependency syntax analysis, takes advantage of the syntax analysis results and combines the pre-compiled dictionary resources to apply classification, recognition, combination, computation and disambiguation rules step by step respectively to the tasks of subjectiveness and objectiveness classification, SO discrimination and SO computation. Experiment on this approach achieves an accuracy of 78.25%, proving its effectiveness and validity.

Keywords: semantic orientation, rule, sentence, computation, subjectivity.

1 Introduction

Semantic orientation (SO) computation usually deals with problems of four types: (1) distinguishing a subjective linguistic expression from an objective one, i.e., the classification of subjectiveness and objectiveness; (2) determining the SO of a subjective linguistic expression, i.e., the discrimination of SO; (3) measuring the intensity of a subjective linguistic expression, i.e., the computation of SO; (4) recognizing the holder and topic relating to some SO, i.e., the recognition of associative elements.

The realistic basis of SO computation is linguistic subjectivity. Subjectivity refers to all private states such as standpoint, opinion, attitude, sentiment, etc. expressed in the utterances by the speaker who, by virtue of being the speaker, casts himself in the role of ego and relates everything to his viewpoint. Linguistic subjectivity makes it possible to compute the SOs of different levels of language units.

The feasibility basis of SO computation is what Osgood puts in his *The Measurement of Meaning* the Semantic Differential Theory [1]. Three distinctive findings based on psychological experiments in the theory relate to SO computations: (1) Evaluation is the single most identifiable factor contributing to word meaning; (2) Bipolarity is one of the fundamental characteristics of semantic differential; (3) Congruity shift can depict how words affect each other when they co-occur.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 209–218, 2013.

According to the above mentioned two bases, we propose a rule-based approach for the computation of Chinese sentence SO. The remainder of this paper is organized as follows. Related works are presented in section 2. Section 3 is about SO and its main properties. Section 4 illustrates the method, process and important details of rule-based Chinese sentence SO computation. Section 5 is the experiment results and section 6 is the conclusion.

2 Related Works

Currently, SO computation is generally conducted on three language levels: the word level, the sentence level and the discourse level. The two mainstream approaches applied, especially in real-world applications, are commonly regarded as machine learning based techniques and semantic analysis based techniques. The machine learning approach treats SO analysis as a special kind of classification, the technical key point of which is to adopt suitable algorithm and select effective features. Semantic analysis approach is in nature rule-based, which regards words as the minimal SO carriers and holds that the SOs of bigger language units can be resolved and then calculated from the SOs of words they include by rules. Of the two approaches, most researchers prefer the machine learning approach, mainly because of its relatively higher precision.

However, the rule-based SO computation approach has its own advantages, compared to the machine learning approach. Firstly, it accords with human beings' way of thinking and the modes of how human beings resolve semantics. Secondly, although machine learning classifiers like Support Vector Machine perform very well in the domain they are trained on, their performance drops precipitously when the same classifier is used in a different domain. They are dependent seriously on the training corpora. The rule-based approach, on the other hand, is more domain-general as it simulates a general representation of subjective language. Thirdly, machine learning classifiers are linearly binary, i.e., their outputs are of two classes, representing positive and negative SO. However, SO computation has another task of measuring the intensity of a subjective linguistic expression. In this case, machine learning classifiers cannot do little about it [2-3]. The rule-based approach can combine well the discrimination of SO with the computation of SO. Its outputs are numeric values with positive or negative signs, the values representing the intensity and the sign the classes.

The rule-based approach to SO computation has already made much progress. In early stage, conjunctions are used to infer the SOs of the conjunct words [4]. Later on, typical positive and negative seed words are chosen to analyze the SOs of target words [5]. Two works worth mentioning are [6] and [7], in which lexicon-based and context-sensitive strategies are applied to compute the SOs of English sentences and discourses. For Chinese texts, [8-13] propose various methods to analyze and process the SOs of subjective Chinese texts. But in general there is still much work to do in SO computation for Chinese. This paper focuses on the sentence level and is expecting to enhance previous works.

3 Semantic Orientation and Its Main Properties

Semantic orientation refers to such standpoint, opinion or attitude as approval or opposition, compliment or disparagement, affirmation or contradiction with regard to person, thing or event. From a more generalized perspective, it also covers positive and negative private states such as happiness, anger, sadness and so on. SO is usually classified into three subcategories: positive, negative and neutral, represented by real numbers 1, -1 and 0 respectively.

SO has a property of difference in degree which is a subjective but measurable quantity. Whereas a coarse-granularity description of SO difference in degree can be represented as high, medium and low scales, fine-granularity description may be gradually refined on this basis till at last be replaced by numerical values instead of scales.

SO is combinable, which has two layers of meaning. One is that a complete SO expression is composed of three constituents: the core, the modifier(s) and the conjunction(s). The core constituents are mainly certain words and some phrases combined closely and used steadily. They are the minimal and indecomposable units conveying SO. The modifiers modify core constituents in syntax and make strengthening or weakening effect on SO degree. They are mainly adverbs of degree and negatives. The conjunctions link core constituents or modifiers. The three constituents can make finite combinations to form a certain SO expression and determine the SO value of a complete SO expression. The other is that the SOs of different levels of language units can be computed successively from low level ones to high level ones. The SO value of a relatively high level language unit is the weighted sum of the SO values of all the relatively low level language units it includes. The computation order starts from minimal units (words and some phrases) and successively enlarges to bigger phrases and clauses, ending at the level of sentences. When computing, it always pivots on the core constituents to get individual SO expressions (basic SO units) and then gradually combine these basic SO units according to their semantic relations in a sentence. In this way the SO value of a sentence is obtained.

The SOs of words are polysemous owing to the polysemy in word meaning. A usual case is that a word's SO is uncertain if its polysemous senses belong to different SO subcategories. Under this circumstance it is necessary to "disambiguate" according to the context the word occurs. Therefore, handling the problem of polysemy in SO computation is an inevitable task.

4 Rule-Based Chinese Sentence SO Computation

4.1 Processing Flow

Based on the above understanding, we hold that: (1) SO is measurable; (2) words are the most important minimal units in expressing SOs; (3) SO is combinable and computable; the SO value of a language unit is the weighted sum of the SO values of all constituents it includes; the SO values of different levels of language units can be obtained gradually from basic SO units; (4) context has relatively great effect on the expression of SO.

Therefore, we propose a rule-based approach to Chinese sentence SO computation. The basic idea of this approach is described as follows. The first step is to preprocess the input sentences, including word segmentation, POS tokenization and dependency syntax analysis. The second step is to accomplish the independent tasks of the classification of subjectiveness and objectiveness, the discrimination of SO and the computation of SO according to corresponding classification rules, recognition rules, combination rules and computation rules, using the pre-compiled SO and other related dictionary resources. Figure 1 is the detailed processing flow.

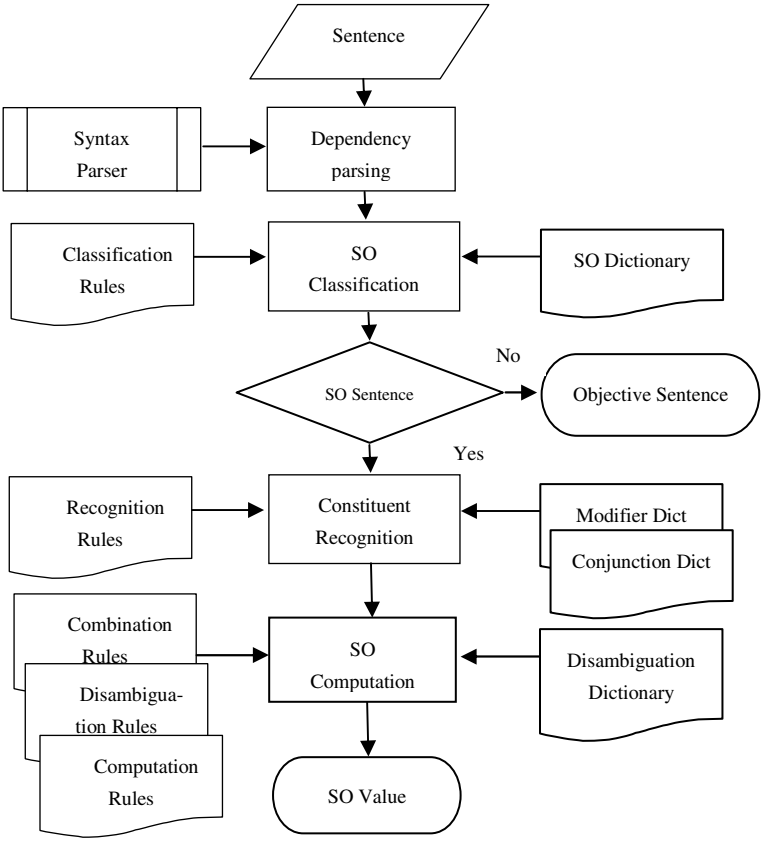


Fig. 1. Processing Flow Chart of Rule-based SO Computation

4.2 The Preparation of Dictionaries

In rule-based approach to SO computation for Chinese sentence, a pre-compiled SO dictionary and several other supporting dictionaries which are called attached dictionaries are required. Our SO dictionary collects words and some phrases combined closely and used steadily with SO. For the need of SO computation, the items of each word or phrase in the dictionary mainly include part of speech and

Table 1. The Information Structure of the SO Dictionary

NO.	Word/ Phrase	POS	Positive SO Value	Negative SO value	Ambiguous
3371	骄傲 (jiāoào,proud)	a.	+0.375	-0.25	Yes

manually annotated SO value. The role of the SO dictionary is used to perform he classification of subjectiveness and objectiveness and provide the core constituents with computable numeric value. Previous verification of the manual annotated results shows that these values are evident in language intuition and with definite difference degree. The information structure of the SO dictionary is as Table 1 shows.

The attached dictionaries include currently the Modifier Dictionary, the Conjunction Dictionary and the Disambiguation Dictionary, among which the Modifier Dictionary is to process modifiers associated with the corresponding core constituents, the Conjunction dictionary is to process conjunctions associated with the corresponding core or modifier constituents and the Disambiguation Dictionary is to disambiguate potential ambiguity in SO. The Modifier and Conjunction Dictionary are compiled and annotated manually while the Disambiguation Dictionary is compiled automatically using reference corpora by extracting collocations and co-occurring words. The information structures of the attached dictionaries are as follows.

Table 2. The Information Structure of the Modifier Dictionary

NO.	Word/ Phrase	POS	Modifying Positive SO Effect	Modify-ing Negative SO Effect
132	非常 (fēicháng,very)	ad.	+80%	+80%

Table 3. The Information Structure of the Conjunction Dictionary

NO.	Word/ Phrase	POS	Pair Word/Phrases	Relation	Conjunction Effect
39	不但 (búdàn,not only)	conj.	而且 (érqiě,but also)；还 (hái,and)	Progres-sive	+25%

Table 4. The Information Structure of the Disambiguation Dictionary

NO.	Word/ Phrase	POS	Positive Collocation/Co-occurring Words	Negative Collocation/Co-occurring Words
11	骄傲 (jiāoào,proud)	a.	成就(chéngjìu,achievement); 自豪(zìháo,proud);值得 (zhídé,worthwhile);令(lìng,let); 深感(shēngǎn,feel deeply);兴 奋(xīngfèn,excited); ……	自满(zìmǎn,pride);防止 (fángzhǐ,prevent);不能 (bùnéng, cannot);惭愧 (cánkuì,ashamed);自大(zìdà, arrogant);不(bù,no); ……

Table 5. An Overview of the Dictionaries

Dict	Item Info				POS Info			
SO	Total	Positive	Negative	Ambi	a.	v.	n.	Other
	19,625	8,899	9,983	743	3,740	4,013	2,742	9,130
		45.34%	50.87%	3.79%	19.06%	20.43%	13.98%	46.53%
Modi	Total	Strengthening	Weakening		ad.	v.	a.	Other
	182	120	62		138	11	7	26
		65.93%	34.07%		75.82%	6.05%	3.85%	14.28%
Conj	Total	Pairs	Non-pairs		conj.	v.	Ad.	other
	160	47	113		92	31	20	17
		29.38%	70.62%		57.5%	19.37%	12.5%	10.63%

4.3 The Compilation and Application of Rules

The input sentence, after pre-processing such as word segmentation, POS tokenization and dependency syntax analysis, is delivered to the SO calculator. The main body of the SO Calculator is a set of rules for SO computation which perform corresponding processing strategies according to the feature of the input sentence. We have compiled so far five types of processing rules, they are classification rules, recognition rules, combination rules, disambiguation rules and computation rules.

The classification rules are applied to classifying SO and non-SO sentences. A lot of formal language features are utilized to accomplish the classification task such as words, POS, punctuations, syntax relations, etc.

The recognition rules are used to recognize different constituents of a SO expression. It can be further divided into three sub-types: the recognition of the SO core constituents, modifier constituents and conjunction constituents. Since most relevant constituents are collected into dictionaries, most of the actions applying these rules are querying the dictionaries and processing the query results.

The combination rules bind together the SO constituents with some syntax relations to be bigger SO units. As is illustrated before, SO is combinable. The combination is always pivoted on the core constituents. Modifiers and conjunctions have some effects on the total unit. Table 6 shows four examples of the combination processes and results.

Table 6. The Combination Processes and Results of Four Examples

NO.	SO Constituents			Result	Realtion
	Conjunction	Modifier	Core		
1		很 (hěn,very)	痛苦 (tòngkǔ,painful)	很+痛苦	adverbial
2	既(jì,not only)		美 (měi,beautiful)	美	parallel
	又(yòu,but also)		科学 (kēxué,scientific)	科学	
3		非常(very)	珍视 (zhēnshì,cherish)	非常+珍视	transition
	尽管 (jǐnguǎn,although)	很(very)	破旧 (pòjiu,worn-out)	很+破旧	
4			致命 (zhìmìng,deadly)	致命+毒药	attributive
			毒药 (dúyào,poison)		

The computation rules are to compute the overall SO of a sentence, taking the combined (bound) units as basic computing units. According to the combination results, for each basic computing unit, it may have two possible computation types. One is the internal computation (Ex.1 and Ex.4 in Table 6) and the other the external computation (Ex.2 and Ex.3 in Table 6). Internal computation means that the computation is conduct on just one original SO expression while the external computation refers to the computation between more than two original SO expressions. Internal computation involves how to process the modifier and its core, especially adverbs of degree and their cores, negators and their cores. For external computation, conjunctions are mainly used to determine the effect of different cores. Considering the page limitation of this paper, details are omitted here. Several examples are presented below.

- 这个方案很无聊(very boring)。 $SO_{很无聊} = -0.25 + (-0.25 \times 70\%) = -0.425$
- 我非常珍视(very cherish)我的中国根。
  $SO_{非常珍视} = 0.625 + (0.625 \times 80\%) = 1.125$
- $SO_{不合格(not\ qualified)} = -0.375$ $SO_{不优秀(not\ excellent)} = -0.0781$
- $SO_{很不优秀(very\ not\ excellent)} = -0.1513$ $SO_{不是很优秀(not\ very\ excellent)} = 0.1875$
  $SO_{不很优秀(no\ very\ excellent)} = -0.125$

Disambiguation rules process SO words like “骄傲(proud)”which may be positive in some contexts and negative in other ones. At present we figure out that it may be possible to use some statistic method to compile a so-called disambiguation dictionary and then make use of it in relevant tasks. Specifically speaking, a word’s SO is explicit in a certain context, or the context determines the SO of a word. If we

simplify the context in which a SO word occurs into its collocations, it may be useful in disambiguation.

5 Experiment Results and Discussions

Experiment data are 400 sentences extracted from an annotated Chinese SO text corpus. Of the 400 sentences, there are 360 SO sentences and 40 non-SO sentences. They are from different genres.

The gold standard, i.e., the SO value of each sentence is given manually by annotators. Precision is adopted as experiment measurement and it is defined as that if the value of the experiment results and the gold standard are within a scale (say, [-0.2, 0.2]) , then the computation result is considered as correct.

Table 7 lists the experiment results.

Table 7. The Experiment Results of Rule-Based SO Computation for Chinese sentence

Correct Sentence No.		Incorrect Sentence No.	
SO	Non-SO	SO	Non-SO
277	36	83	4
313		87	

Therefore the precision of rule-based SO computation for Chinese sentence is:

$$\text{Precision}=\frac{\text{Correct Sentence No.}}{\text{Total Sentence No.}}\times 100\% = 78.25\%$$

In general, the experiment results basically achieve our expectations, showing the effectiveness and validity of rule-based SO computation for Chinese sentence. However, the precision is not satisfying. It is influenced by various factors: (1) the differences between human's language sense and the experiment results; (2) the mistakes and errors in the SO dictionary and its attached dictionaries; (3) the performance of the recognition of SO and non-SO sentences; (4) the problems in rules; (5) the extent of the dependency syntax parsing errors. Take a closer look at the following four incorrect computed sentences:

（1）卷云和卷积云都很高，那里水分少，它们一般不会带来雨雪。
（2）大熊猫小的时候很活泼，喜欢爬上爬下。
（3）一听他要在自己整洁漂亮的小床上睡觉，就哭了起来。
（4）这是一个顶漂亮的城市！

Sentence 1 is discriminated as a SO sentence just because such words as "高(gāo, tall)", "少(shǎo, less)" are in the SO dictionary. Looking back, it is very necessary to put these words into the disambiguation dictionary in case causing more troubles in procedures afterward. Sentence 2 has an experiment result value of +0.8 while the average human annotators give it +0.5, showing that there are differences between

human's language sense and the experiment results. Sentence 3 has an experiment result value of +0.375 which is quite far from the gold standard as the rues cannot process "一(yí, when)……就(jiù, then)……" well. The experiment result of sentence 4 is incorrect just because of the dependency syntax parser makes a mistake on the structure "顶漂亮(dǐng piàoliàng, extremely beautiful)". From the discussion of the incorrect experiment result sentences, we are aware of that for the designed rules it is needed to repeatedly check, verify and polish.

6 Conclusion

Semantic Orientation refers to the positive or negative attitude, standpoint or opinion on a certain person, object or affair. It is of degree diversity and combinability. Rule-based computation of SO for Chinese sentence, after pre-processing the input sentence in lexical and dependency syntax analysis, takes advantage of the syntax analysis results and combines the pre-compiled dictionary resources to apply classification, recognition, combination, computation and disambiguation rules step by step respectively to the tasks of subjectiveness and objectiveness classification, SO discrimination and SO computation. Experiment on this approach achieves an accuracy of 78.25%, proving its effectiveness and validity.

From the perspective of real-world applications, the achieved precision is not satisfying. Therefore, improving the precision of the computation result is the key point of later works, which will focus on deep analysis and conclusion of various language facts, continuously polishing the dictionaries and improving and adding rules. In addition, this work is based on a not-always-true assumption that the SO of a sentence comes from a same holder and targets a same topic. However, this is not always the case. Hence, later works will also have to deal with the problem of solving the association of the SO, its holder and its topic.

Acknowledgements. We thank Harbin Institute of Technology for sharing the LTP package and Mr. Zhendong Dong for sharing the HowNet. This work was supported by grants from the National Social Sciences (No.11CYY032).

References

1. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The Measurement of Meaning. University of Illinois Press, Urbana (1957)
2. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of ACL 2005, Ann Arbor, MI, USA, pp. 115–124 (2005)
3. Wilson, T., Wiebe, J., Hwa, R.: Just how mad are you? Finding strong and weak opinion clauses. In: Proceedings of the 19th National Conference on Artifical Intelligence (AAAI 2004), San Jose, CA, USA, pp. 761–769 (2004)

4. Hatzivassiloglou, V., McKeown, K.R.: Predicting the Semantic Orientation of Adjectives. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, pp. 174–181 (1997)
5. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, USA, pp. 417–424 (2002)
6. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the International Conference on Web Search and Web Data Mining (WSDM 2008), NY, USA, pp. 231–240 (2008)
7. Taboada, M., Brooke, J., et al.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37(2), 267–307 (2011)
8. TSou, B., et al.: Polarity classification of celebrity coverage in the Chinese Press. In: Proceeding of the 2005 International Conference on Intelligence Analysis, pp. 137–142 (2005)
9. Zhu, Y., et al.: Semantic Orientation Computing Based on HowNet. Journal of Chinese Information Processing 20(1), 14–20 (2006)
10. Li, D., et al.: Text Sentiment Classification Based on Phrase Patterns. Computer Science 35(4), 132–134 (2008)
11. Wang, S., et al.: Research on sentence sentiment classification based on Chinese sentiment word table. Computer Engineering and Applications 45(24), 153–155 (2009)
12. Dang, L., et al.: Method of discriminant for Chinese sentence sentiment orientation. Application Research of Computers 27(4), 1370–1372 (2010)
13. Zhao, Y., et al.: Syntactic Path Based Appraisal Expression Recognition. Journal of Softwares 21(8), 1834–1848 (2010)

Studies on Automatic Recognition of Contemporary Chinese Common Preposition Usage

Kunli Zhang, Hongying Zan, Yingjie Han, and Tengfei Zhang

College of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China
{ieklzhang,iehyzan,ieyjhan}@zzu.edu.cn, ztf986@163.com

Abstract. Automatic recognition of prepositional usage is of great significance in parsing and syntax analysis. Many researches have been focused on preposition usage. In this paper, we introduce the triune knowledge base (usage dictionary, usage rule and usage corpus) of Contemporary Chinese preposition that we have finished .On this basis, we firstly adopt rule-based method to automatically annotate the prepositions in the corpus of People's Daily, in which the precision rate achieves 68.68%. Then to the prepositions whose precision rate is less than 80%, we use statistics-based method to annotate them with different models, features and context windows. The best precision rate achieves 90.86%. Experiments show that the statistics-based method can efficiently meet the need of the automatic recognition of prepositions' usage.

Keywords: Usage Automatic Recognition of Preposition, Usage Rule, Conditional Random Fields, Maximum Entropy.

1 Introduction

The number of the preposition is limited, and preposition is mainly attached to other words to constitute prepositional phrase whose syntactic function is adverbials and attributives. Although a preposition itself can't be used alone, the prepositional phrases play an important role in Chinese grammar structure. High frequency in use, diversiform usage, and the individual character are the prominent features of preposition, so the research work on preposition has important significance, such as the following sentences (1) and (2).

(1) 王洪 把 李斌 推到 讲台 上。
Wanghong ba Libin tuidao jiangtai shang
Wanghong Libin push platform
Wanghong pushed Libin to the platform.

(2) 王洪 被 李斌 推到 讲台 上。
Wanghong bei Libin tuidao jiangtai shang
Wanghong Libin push platform
Wanghong was pushed to the platform by Libin.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 219–229, 2013.

The meanings of sentence (1) and sentence (2) are entirely different because of the different prepositions "被 (bei)" and "把 (ba)". The same preposition may indicate different meanings in different contexts and have different usage, such as the following sentences (3), (4) and (5) including preposition "在 (zai in)".

The preposition "在 (zai in)" refers to time in sentence (3), location in sentence (4) and condition in sentence (5). The preposition "在 (zai in)" has different meanings in the above three sentences. So their usages are different. If the different usage can be automatically recognized in different context, that will play an important role in syntax analysis and semantic analysis.

(3) 我 是 在 到 了 上海 之后 才 听说 这 件 事 的。
Wo shi zai dao le Shanghai zhihou cai tingshuo zhe jian shi de
I arrive Shanghai after hear it
I heard about it after I arrived Shanghai.

(4) 疗养院 坐落 在 半山腰。
liaoyangyuan zuoluo zai banshanyao
nursing home locate in mountainside
The nursing home is located in mountainside.

(5) 在 大家 的 帮助 下，把 落水 的 儿童 救 上 了 岸。
zai dajia de bangzhu xia ba luoshui de ertong jiu shang le an
with many people help drowning child rescue bank
With many people's help, the drowning child was rescued to bank.

Based on the thoughts of building the triune knowledge base of contemporary Chinese functional words [1], in this paper we firstly supplement and modify preposition usage dictionary and usage rules base which have been built by Zan et al. [2-3].Then we manually annotate the preposition usage of texts of five months People's Daily (2000.2-2000.6) which have been segmented and POS tagged and adopt them as automatic recognition corpus of preposition usage. We use rule-based method to automatically recognize the preposition usage and the results show that rule-based method is feasible. In addition, we use Conditional Random Fields (CRF) and Maximum Entropy (ME) to realize automatic recognition of preposition usage based on statistics for common preposition.

2 Related Work

In the field of linguistics, studies on preposition focused on several aspects: the preposition scope and classification [4], some cases of preposition [5], distinguishing prepositions from other parts of speech [4], as well as preposition and prepositional functions [6], etc. These studies are all human-oriented.

Since prepositional phrase recognition plays an important part in syntactic analysis, many scholars have tried to use all sorts of methods for automatic recognition of prepositional phrases. Not only some statistics methods but also rule-based methods

were adopted to recognition of prepositional phrases, and Miao et al.[7] mixed semantic information into SVM model to identify prepositional phrase boundary. But for the prepositional phrase recognition, the different usages of prepositions are ignored.

Yu et al. [1] initially put forward the thoughts of building the triune knowledge base of contemporary Chinese functional words and defined functional words as adverb, preposition, conjunction, auxiliary, modality and position words. Liu [8] constructed the basic framework of functional words dictionary and designed the corresponding description attributes for adverbs, prepositions, conjunctions, auxiliary and modal words. Moreover, they summarized and classified the common functional words. Peng [9] built preliminary contemporary preposition dictionary. Zan et al.[2-3] built contemporary Chinese functional words knowledge base which contains usage dictionary, usage rule base and usage corpus. Liu et al. [10] and Yuan et al. [11] discussed rule-based automatic recognition of functional words' usage preliminarily. Zan et al. [12] studied statistics-based automatic recognition of Chinese adverb "就(jiu only)". These studies mainly aim at constructing knowledge base of functional words or automatic recognition of adverb usage. However, studies on auto recognition of preposition usages are limited.

3 The Triune Knowledge Base of Preposition

3.1 Preposition Usage Dictionary

In reference to [2] and [3], the preposition usage dictionary has been built. In this paper, based on the segmented and POS tagged corpus of People's Daily (2000.2-2000.6) which is offered by computational linguistics institute of Beijing University, entries and usages in the dictionary are adjusted and modified. Compared with the original version, in this dictionary, the number of prepositions and usages changed. At present, this preposition usage dictionary contains 141 prepositions and 331 usages. Table 1 shows the distribution of entries and usages. In Table 1, N_1 is the number of usages, N_2 is the number of entries which are in N_1.

Table 1. Entries and usages distribution of prepositions

N_1	1	2	3	4	5	6	7	9	10	11	12
N_2	66	30	23	7	4	5	1	2	1	1	1

As an example, preposition "在(zai in)" has 5 senses and 9 usages. Table 2 shows its partial description in usage dictionary including ID, sense and usage. The detailed presentation of "usage ID" and the content in angle brackets can refer to Zan et al.[2]

Table 2. ID, sense and usage of preposition "在(zai in)"

Usage ID	Sense	Usage
p_zai4_1a	Express time.<x>	"~…"used before verb, adjective or subject.<b>
p_zai4_1b	Express time.<x>	"~…"used behind verb. The single syllable verbs are limited to "生(sheng born)|死(si die)|定(ding determine)|处(chu stay)|改(gai change)|放(fang place)|排(pai arrange)" and so on, and the bisyllable verbs are limited to "出生(chusheng born)|诞生(dansheng born)|发生(fasheng happen)|出现(chuxian appear)|发现(faxian find)|布置(buzhi arrange)|安排(anpai arrange)|确定(queding determine)|固定(guding set)" etc..<b>
p_zai4_1c	Express time.<x>	Often followed by words "过程(guocheng process)|会议(huiyi conference)|比赛(bisai game)|活动(huodong activity)|斗争(douzheng struggle)|接触(jiechu contact)" which can express duration time.<z>
p_zai4_2a	Express location. Refer to the place where action happens or things exist. <b>	"~…"used before verb, adjective or subject.<b>
p_zai4_2b	Express location. Refer to the birth, occurrence, produce, residence place or the place which the movement arrives. <b>	"~…"used behind verb.<b><z>
p_zai4_3a	Express scope.<b>	"~…"used before verb, adjective or subject.<b> Constituted prepositional phrases such as "~…方面(fangmian aspect)|问题上(wentishang problem)|实践上(shjianshang practice)|生活中(shenghuozhong life)|生活上(shenghuoshang life)|领域(lingyu domain)|工作上(gongzuoshang work)" which express scope. <z>
p_zai4_3b	Express scope.<b>	"~…" used behind verb. Constituted prepositional phrases such as "~…上(shang on)|中(zhong among)|以内(yinei within)|以外(ciwai except for)|以下(yixia under)|之间(zhijian among)|之中(zhizhong among)|之外(zhiwai except for)" which express scope. <x><z>
p_zai4_4	Express condition.<b>	Constituted prepositional phrase such as "~+gerund phrase+下(xia under)"and used before verb or subject.<b>
p_zai4_5	Express subject of behavior. <b>	Often used before person noun or pronoun.<z>

3.2 Preposition Usage Rule Base

On the basis of the dictionary of preposition usage, each preposition usage was investigated in People's Daily corpus, the operable judgment conditions character was extracted with description preposition usage rules with orderly BNF paradigm. In this way, the preposition usages rule base was constructed. In rules, capital letters represent usage features, and lowercase letters represent POS and Chinese characters represent word forms. There are 6 usage features: F, M, L, R, N and E. F stands for head of the sentence contains the target preposition, M stands for before the target preposition but not contiguous, L stands for before the target preposition and contiguous, R stands for behind the target preposition and contiguous, N stands for behind the target preposition but not contiguous, E stands for the end of the sentence contains the target preposition. The detailed presentation of usage rules can refer to Zan et al. [2] and other symbol description can refer to Yuan et al. [11].

According to the usage description shown in Table 2 and the corpus, rules of "在(zai in)" are built as Fig. 1.

```
$在
@<p_zai4_5>→N ^N→看来|来说|而言|说来|来看|来讲
@<p_zai4_3b>→L ^L→控制|限制|保持|维持|稳定|表现|体现
@<p_zai4_3a>→N ^N→方面|问题上|实践上|生活中|生活上|领域|工作上
@<p_zai4_4>→N ^N→(v|<vn>)<下/f>|(条件|前提|情况|情形|形势|背景|原则|努力)下|基础上
@<p_zai4_1c>→N ^N→过程中|活动中|活动上|会议中|会议上|会上|会中|赛中|赛上|斗争中|接触中|实践中
@<p_zai4_1a>→N ^N→(年|月|日|天|号|星期|世纪|期间|初|时|秒|之后|之前|之际|夜晚|同时|t)*v
@<p_zai4_1b>→LN ^L→v ^N→年|月|日|号|天|星期|世纪|期间|初|时|秒|之后|之前|之际|夜晚|t
@<p_zai4_2a>→N ^N→(<ns>|s)*v
@<p_zai4_2b>→LN ^L→v ^N→n|f
@<p_zai4_2a>→N ^N→(<ns>|s)
@<p_zai4_1a>→N ^N→(年|月|日|天|号|星期|世纪|期间|初|时|秒|之后|之前|之际|夜晚|t)
```

Fig. 1. Rules of "在(zai in)"

As is described in table 1, usage "p_zai4_3a" has unambiguous right collocations and using them can distinguish "p_zai4_3a" from other usages. So the rule of "p_zai4_3a" is constructed with enumerating these collocations in feature "N". Usage "p_zai4_3b" also has unambiguous right collocations, but these collocations can also appear in sense 1 and sense 2. Though investigating each sentence which contains "在(zai in)" in corpus, we find that the usage is usually confirmed for "p_zai4_3a" when the word before "在(zai in)" is "控制(kongzhi control)|限制(xianzhi restrict)|保持(baochi keep)|维持(weichi keep)|稳定(wending keep)|表现(biaoxian display)|体现(tixian reflect)". So the rule of "p_zai4_3a" is described as shown in figure 1. While

in “p_zai4_1a” usage right collocation (in clause) can be a verb or not, so “p_zai4_1a” can be described by two rules, which is same to “p_zai4_2a”.

Due to the fact that recognition of preposition usage largely depends on prepositional object, a majority of preposition rules are described with right collocation feature “N”. It is also a typical feature of preposition rules.

Rules with high appearing frequency or rules which have good effect in automatic recognition should have a higher priority in rule ordering. So they are put to a prior position of rules queue to ensure that each preposition is annotated by the best usage rule in automatic recognition based on rules. From rules of “在(zai in)”, we can see that one usage can have one or more rules .At present there are 383 rules in preposition usage rule base corresponding to 331 usages in preposition usage dictionary.

3.3 Preposition Usage Corpus

After we adjusted and modified preposition usage dictionary and usage rule base, we first use the rule-based method to automatically annotate the preposition usage in corpus of People’s Daily (2000.2-2000.6). Then the corpus was artificially proofread with one more people cross to form standard corpus of recognition of usage. The actual usages distributions are shown in Table 3. In Table 3, N_w is the number of entries appearing in corpus, N_f is the number of prepositions appearing in corpus, N_e is the number of prepositions with error segmenting or error POS tagging in corpus, and N is the number of preposition with normal usage annotation.

Table 3. Preposition distributions in corpus

	N_w	N_f	N_e	N
2000.2	96	38,260	228	38,032
2000.3	96	49,059	157	48,902
2000.4	98	47,366	104	47,262
2000.5	94	43,694	91	43,603
2000.6	97	49,116	85	48,931
2000.2-6	105	227,495	665	226,730

4 Automatic Recognition of Preposition Usage

4.1 Rule-Based Automatic Recognition of Preposition Usage

Rule-based automatic recognition system parses the rules of the preposition rule base to determine which usage to be employed. The specific steps are as follows:

1. Initialize annotating corpus and usage rule base. Split corpus into sentences which are read into memory with dynamic array when reading corpus. Usage rules are read into memory with hash table.
2. Read the whole sentence, find all prepositions which need to be annotated and corresponding rules. Then dispose the entire sentence to get corresponding word list and original sentence, as well as the location of prepositions which need annotating in word list and original sentence.
3. Search for the rules of prepositions which need to be annotated, read their usage rules in sequence. Then determine the kind of verifier to analyze and match usage rules and to determine the annotating result. At last, output the entire sentence and read the next sentence when all the prepositions of the current sentence are annotated.
4. Repeat step 2 and 3, until all prepositions in the corpus are annotated.

Automatic annotating system used six types of verifier to satisfy different requirements of feature attributing from finding ranges of dictionary and matching result. Yuan et al. [11] introduced design requirements of all kinds of verifier and realization of automatic annotating system functional words usage based on rules.

4.2 Statistics-Based Automatic Recognition Results of Preposition Usage

Empirical method based on statistics obtains language knowledge from training data automatically or semi-automatically, set up effective statistical language model and optimize according to the actual situation of training data constantly. On the contrary, rationalism method based on rules is difficult to adjust according to actual data. So rule method is less good than empirical methods based on statistics in some ways.

Considering usage of preposition is related to context and context sequence, this paper chooses two statistical models with wide application and good effect in Machine learning. They are Conditional Random Fields (CRF) and Maximum Entropy (ME).

CRF is an undirected graph model which calculates conditional probability of output node under the condition of giving input node. It inspects conditional probability of annotating sequence corresponding to input sequence and its training target is that maximize conditional probability [13]. ME is a statistical model which is widely used in classification problem. Its basic concept is to excavate potential constraint conditions in a known event set, then select a model which must satisfy known constraint conditions and distribute unknown events which may happen as evenly as possible[14].

This paper adopts CRF++ toolkit[1] and MaxEnt (Zhang)[2] as automatic annotating tools.

[1] CRF++: Yet Another Toolkit[CP/OL]. http://www.chasen.org/~taku/software/CRF++

[2] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

5 Experiments and Results Analysis

5.1 Rule-Based Automatic Recognition Results of Preposition Usage

In this paper, the segmented and POS tagged corpus of five months' People's Daily (2000.2-2000.6) is adopted as automatic recognition corpus. First, automatic annotating system is used to annotate all prepositions in experimental corpora to get machine tagging results. If machine annotating results and the data of the standard corpus having been manually proofreaded are the same, machine annotating results would be considered as correct ones. The experiment counts on annotating results of all prepositions in corpus of every month.

We use precision rate (P) as metric of experimental result, as is indicated in formula (1). The experiment results of preposition usage based on rules are shown in Table 4. The distribution of recognition results is also shown in Table 4. In Table 4 N is the number of prepositions with normal usage annotation, N_3 is the correct number, $N_{P>0.8}$ is the number of usages whose precision rate is greater than 0.8, $N_{0.5\leq P\leq 0.8}$ is the number of usages whose precision rate is less than or equal to 0.8, greater than or equal to 0.5, $N_{P<0.5}$ is the number of usages whose precision rate is less than 0.5, and N_{na} is the number of usages which haven't appeared in corpus.

$$precision\ rate\ (P) = \frac{the\ correct\ number\ in\ corpus\ (N_3)}{the\ number\ of\ preposition\ with\ normal\ usage\ labeling\ in\ corpus\ (N)} \quad (1)$$

Table 4. The results of rule-based and distribution of precision rate

	N	N_3	P (%)	$N_{P>0.8}$	$N_{0.5\leq P\leq 0.8}$	$N_{P<0.5}$	N_{na}
2000.2	38,032	25,305	66.53	125	33	48	125
2000.3	48,902	34,653	70.86	129	37	37	128
2000.4	47,262	32,243	68.22	122	34	54	121
2000.5	43,603	29,599	67.88	129	33	55	114
2000.6	48,931	33,919	69.32	127	35	48	114
2000.2-6	226,730	155,719	68.68	134	49	64	84
Ratio of usage (%)	——	——	——	40.48	14.80	19.34	25.38

In terms of results shown in Table 4, the overall precision rate is 68.68% and the number of usages whose precision rate are higher than 80% is 134, accounting for the 247 appearing usages in corpus of 57.25%. These all show that the rule-based method has some effect in automatically recognition of preposition usage. Furthermore, the precision rate of common prepositions (frequency is more than 1000 and the number of usages is more than 2) in corpus is shown in Table 5. In Table 5, N_1 is the number of usages, and N_4 is the number of frequency.

According to the data in Table 5, a total of 21 prepositions meet the requirements, of which there are 7 preposition rate is higher than 80%. And we also found that the number of usage and whether the descriptions of the rules are easy to formalize are the factors that affect the rule-based recognition precision rate. At the same time, the features mutually exclusive and exhaustive which can't be meet lead to a poor precision rate.

Table 5. The rule-based results of high frequency prepositions

Preposition	N_1	N_4	P (%)	Preposition	N_1	N_4	P (%)
按照(anzhao according to)	3	1,719	**92.44**	根据(genju on the basis of)	3	1,989	65.61
自(zi since)	4	1,129	**91.85**	把(ba)	11	8,531	64.75
按(an according to)	3	1,265	**91.54**	对(dui to)	4	23,172	64.06
以(yi according to)	6	10,553	**89.60**	在(zai in)	9	63,595	62.74
比(bi than)	6	23,82	**86.52**	被(bei)	10	6,528	57.64
据(ju according to)	3	49,81	**81.93**	由(you from)	5	6,040	56.36
为了(weile in order to)	3	24,55	**80.86**	给(gei for)	6	3,443	56.08
和(he with)	3	1,537	78.59	从(cong from)	6	10,896	53.74
对于(duiyu toward)	3	1,686	75.80	通过(tongguo via)	3	3,954	48.96
同(tong wtih)	3	2,501	70.73	为{wei4}(wei in order to)	6	12,177	28.27
于(yu at)	10	6,281	70.45				

5.2 Statistics-Based Automatic Recognition Results of Preposition Usages

We choose 14 prepositions whose precision rate is less than 80% in Table 5 to do experiment based on statistics-based method. The experiment corpus is the same as the one in subsection 5.1. We know statistics-based method is concerned with features template, statistical methods and selection of context window, therefore we mainly set different context window size and use different feature template to do experiments. The best results of experiments and the corresponding window size using CRF model and ME model are shown in Table 6. We choose words and their part of speech as features and use unsymmetrical window size in CRF experiment, choose words as features and use symmetrical window size in ME experiment. In "context" column in Table 6 "L" and the number following it stand for the left window size, "R" and the number following it stand for the right window size.We use ten fold cross-validation to obtain average cross-validation results of each preposition.

Table 6. The result of rule-based and the best results of statistics-based

Preposition	CRF		ME	
	P (%)	Context	P (%)	Context
和(he with)	**84.39**	L0R5	81.04	L3R3
对于(duiyu toward)	**69.87**	L0R3	62.03	L7R7
同(tong wtih)	**92.03**	L0R2	92.02	L2R2
于(yu at)	**90.24**	L1R3	87.85	L1R1
根据(genju on the basis of)	**98.39**	L0R2	98.39	L2R2
把(ba)	**87.03**	L2R10	79.57	L10R10
对(dui to)	**87.38**	L0R10	68.09	L9R9
在(zai in)	**92.47**	L1R5	77.01	L10R10
被(bei)	**85.50**	L1R5	67.11	L5R5
由(you from)	**85.91**	L1R5	70.45	L5R5
给(gei for)	**89.06**	L1R5	82.60	L6R6
从(cong from)	**93.09**	L1R3	76.29	L7R7
通过(tongguo via)	**93.77**	L0R4	73.62	L7R7
为{wei4}(wei in order to)	**89.78**	L2R6	70.27	L3R3

From Table 5 and Table 6, we can see that for the fourteen prepositions the Micro average precision rate of rule-based method is 59.46%, of CRF model is 90.26%, of ME model is 75.40%.The precision rate of statistical recognition results are respectively 30.80% , 15.94% higher than the rule's.

6 Conclusions

Based on the triune knowledge base of contemporary Chinese preposition, we first realize rule-based automatic recognition of preposition usage. Experimental results show that to common prepositions the overall precision rate of rule-based method is 68.68%. Furthermore, we use statistical method including CRF model and ME model to annotate common prepositions whose precision rate of rule-based method is less than 80%, the precision rate can be improved by 30.80%.

In the future, we will continue to improve preposition usage dictionary, usage rule base and build a comprehensive accurate knowledge base of contemporary Chinese functional words targeted to natural language processing. Besides, we will try to process weighted rules and adopt combination of rules and statistics method to improve automatic recognition precision rate of preposition's usage according to relative frequency of preposition's usage distribution, or adopt Ensembles of

Classifiers method to improve precision rate. Meanwhile, we will apply results of automatic recognition of preposition's usage to other fields of natural language processing, such as automatic syntactic analysis and machine translation, in the hope of a better application effect.

Acknowledgments. This work was supported by a grant from the Natural Science Foundation of China (No.60970083), Ministry of Education and the Outstanding Young Talents Technology Innovation Foundation of Henan Province (No.104100510026), the Open Projects Program of National Laboratory of Pattern Recognition.

References

1. Yu, S.W., Zhu, X.F., Liu, Y.: Knowledge-base of Generalized Functional Words of Contemporary Chinese. Journal of Chinese Language and Computing. Vol.13 No.1: 89-98(2003)(In Chinese)
2. Zan, H.Y., Zhang K.L., Chai, Y. M., Yu, S. W.: Studies on the Functional Word Knowledge Base of Modern Chinese. Journal of Chinese Information Processing. Vol.21 No.5:107-111(2007)(In Chinese)
3. Zan, H.Y., Zhu, X.F.: NLP oriented studies on Chinese Functional Words and the Construction of Their Generalized Knowledge Base. Contemporary Linguistics, 11(4):124-135(2009)(In Chinese)
4. Zhang, Y.S.: Contemporary Chinese Functional Words. East China Normal University Press, Shanghai(2000)(In Chinese)
5. Chen, C. L.: Prepositions and Prepositional Function. Anhui Education Press, Hefei (2002)
6. Jin, C.J.: Chinese Preposition and Prepositional Phrases. Nankai University Press, Tianjin(1996)(In Chinese)
7. Wen, M.M, Wu, Y.F.: Feature-rich Prepositional Phrase Boundary Identification based on SVM .Journal of Chinese Information Processing2009, Vol.23 No.5:19-24. (2009)(In Chinese)
8. Liu, Y.: The Building of Knowledge Database of Contemporary Chinese Functional Words. Postdoctoral Report. Peking University, Beijing(2004)(In Chinese)
9. Peng, S.: The Building of Knowledge Base of Contemporary Chinese Prepositions and Related Research. Postdoctoral Report. Peking University, Beijing(2006)(In Chinese)
10. Liu, R., Zan, H.Y., Zhang, K.L.: The Automatic Recognition Research on Contemporary Chinese Language[J], Computer Science, Vol.35 No.8A:172-174(2008)(In Chinese)
11. Yuan, Y.C., Zan, H.Y., Zhang, K.L., Zhou, Y.H.: The Automatic Annotation Algorithm Design and System Implementation Rule-base Function Word Usage. In: The 11th Chinese Lexical Semantics Workshop, pp. 163-169. Suzhou (2010) (in Chinese)
12. Zan, H.Y., Zhang, J.H., Zhu, X.F., Yu, S.W.: The Studies on the Usages and Their Automatic Recognition of Chinese Adverb JIU. Recent Advances of Chinese Lexical Semantics, pp.37- 43. (2010)
13. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In: Proceedings of the 18th ICML-01, pp.282-289.(2001)
14. Berger, A.L., Della Pietra, V.J.: Della Pietra S.A. A maximum entropy approach to natural language processing. Computational Linguistics, Vol.22 No.1:39-71.(1996)

Automatic Extraction of Chinese V-N Collocations

Xiaofei Qian

College of Liberal Arts, Shanghai University, Shanghai, China
qierflying@163.com

Abstract. Chinese V-N collocations have two possible structural relations: verb-object relation and attributive-head relation. Both of them are widely used in Chinese language processing tasks, but long distance and low frequency collocations are often difficult to extract. A weighted mutual information (WMI) model and a rule-based method were designed to acquire V-N collocations by taking more syntactic structure features into consideration. The WMI model extracted verb-object collocation within clauses. It reduced the interference of illegal collocates and highlighted the weight of long distance collocates, by giving different weights to collocates in different locations. The rule-based method used part of speech patterns to extract verb-object and attributive-head collocations, and inferred implicit collocations. The experiments show that, the WMI model optimizes evaluation scores of long distance collocations, while the rule-based method is more accurate in extracting and distinguishing the two kinds of collocations, including low frequency collocations.

Keywords: verb noun collocation, automatic extraction, weighted mutual information, rule.

1 Introduction

Collocations are useful resource for language teaching, lexicography, and natural language processing. Extracting collocations automatically from large-scale corpus is the main way to obtain collocation knowledge currently.

Unsupervised statistical method is often used to extract collocations. Church used pointwise mutual information to evaluate the combined capacity of two words [1], and Smadja advanced a dispersion formula to measure the strength of two words by introducing information of location [2]. Similarly, Sun established a quantitative assessment system of collocations, which contained three statistical indexes of intensity, dispersion and peak [3]. Some studies merged unsupervised statistical method and rule method. For example, Lin extracted collocations by correcting dependency parsing errors with statistic data [4], and Bai combined rules and the collocation probability to extract Chinese Verb-Verb collocations [5]. Some other researches attempted to obtain collocations by chunk annotation using supervised statistical method [6].

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 230–241, 2013.

The previous studies have made great progress on algorithm design, but there are still some inadequacies. First, there is restriction on observation window, many of which were restricted to a sliding window, for instance, in the range of [-5, 5]. It may result in failure to extract some long distance collocations which are useful for specific applications. For example, in maximal noun phrase (MNP) recognition task, 15 percent of MNPs are longer than 4 words, and they are more likely to distribute in object slots. Second, insufficient attention was paid to the structural characteristics of language. For example, in the sequence of "v n1 n2 n3", there is little chance for n2 to form verb-object collocation with v in Chinese syntax, but it was often treated as equally as n3. Third, most studies concerned about the syntactically realized collocations, instead of implicit collocations. Moreover, Chinese V-N collocation contains two possible syntactic relation between collocates — verb-object relation and attributive-head relation. Verb-object collocations were focused in most studies, while attributive-head collocations were often ignored.

In order to solve the problems mentioned above, this paper presented two methods to extract the two types of V-N collocations, by taking into account more features of the syntactic structure. One is the weighted mutual information model and the other is the rule-based method based on POS patterns. The results of the study indicated that they improved the extracting precision, especially for low frequency collocations, and the system performance on extracting some long distance collocations.

2 Types of Collocation

2.1 Lexical Collocation and Syntactic Collocation

From the perspective of linguistics, collocations can be categorized into two different types. One is collocations as chunks (lexical collocations), and the other is collocations realized in syntax (syntactic collocations). The two types of collocations are not in the same language level. Lexical collocation is stored in long-term memory, in which collocates are closely tied with each other. It can present in the form of syntactic collocations in a sentence. Examples are as follows:

① 他 喜欢 早晨 读 报纸 。
Ta xihuan zaochen du baozhi .
He likes reading newspapers in the morning.

In this sentence, the binary group (读 du read, 报纸 baozhi newspaper) is a lexical collocation, and “ 读du read ” is firmly combined with “报纸 baozhi newspaper”.

Syntactic collocations are multiword chunks satisfied certain syntactic relations in a sentence. They can be temporary combinations, or realized by lexical collocations. For example:

② 想念你的笑，想念你的外套，想念你白色袜子和你身上的味道。

Xiangnian ni de xiao, xiangnian ni de waitao, xiangnian ni de baise wazi he ni shenshang de weidao.

Miss your smile, miss you coat, miss you white socks and your smell.

In the sentence, the binary groups, e.g. (想念 xiangnian miss, 外套 waitao coat), (想念 xiangnian miss, 袜子 wazi sock), (想念 xiangnian miss, 味道 weidao smell) are all syntactic collocations. And the nouns here, including "外套waitao coat ", "袜子 wazi sock" and "味道 weidao smell", cannot satisfy the typical semantic configuration requirements of the verb "想念 xiangnian miss". Thus, they are temporary combinations.

Lexical collocations are the result of chunked syntactic collocations repeatedly occurring in sentences. Loose syntactic collocations and lexical collocations are at both endpoints of a continuum. And the extraction task prefers the end near lexical collocations.

2.2 Explicit Collocation and Implicit Collocation

Collocations in Chinese sentences also can be classified into two types from the perspective of syntactic realization. One is explicit collocation, in which the relation between collocates can be annotated in the syntactic tree. It is widely distributed in texts. The other is implicit collocation, in which the relation between collocates is not realized in surface structure, but hidden in deep structure of the phrase [7-8]. It is low frequency phenomena compared with the explicit collocation, and it mainly refers to the V-N lexical collocation which has no direct syntactic relation between verb and its nominal argument in a sentence. For example:

③ 他吃了一个苹果。

Ta chi le yi ge pingguo.

He ate an apple.

④ 他吃的那个苹果很甜。

Ta chi de na ge pingguo hen tian.

The apple he ate is very sweet.

The binary group (吃 chi eat, 苹果 pingguo apple) in sentence ③ is realized collocation, while in sentence ④ is potential collocation.

Thus, in a Chinese sentence, there are five situations about the relation between two words, including:

(1) With neither lexical collocation nor syntactic relation.
(2) With no lexical collocation, but syntactic relation.
(3) With lexical collocation, but no (realized) syntactic relation.
(4) With lexical collocation, and implicit syntactic relation.
(5) With both lexical collocation and realized syntactic relation.

All of the situations may be extracted by statistical methods. It is correct in language sense to obtain situation (3), but actually you will find it a lucky hit while looking into corpus. Relatively speaking, the implicit collocation relation in case (4) is a low frequency phenomenon in Chinese language. It is difficult for statistical methods to identify it. However, an explicit collocation and an implicit collocation may appear one after another in texts. So extracting implicit collocations also helps extracting low-frequency explicit collocations. Therefore, it is valuable to study how to extract low-frequency implicit collocations, as well as how to reduce the probability of obtaining illegal combinations.

3 Syntactic Relations and POS Tagging of V-N Collocations

V-N collocation contains two possible relations, which are verb-object relation and attribute-head relation. Verb-object relation describes the constraints of verb and the head of the nominal object in context, for instance, in the phrase "处理chuli handle内部 neibu internal矛盾 maodun contradiction", the binary combination (处理chuli handle, 矛盾 maodun contradiction) forms a verb-object collocation. Attribute-head relation depicts the constraints of attributive verb and its noun head in context, for instance, (工作 gongzuo working, 人员 renyuan personnel) forms an attribute-head relation collocation.

The two types of collocations not only compete with illegal combinations, but often interfere with each other. To distinguish the internal relations of V-N collocations precisely, a word segmentation and POS tagging job is necessary. Chinese lexical analyzer has been able to identify most of the verbs in the position of attributives and heads. In the POS tag set of Peking University, the verb and the noun in verb-object relation are tagged in accordance with their original parts of speeches, "v" and "n", while the verb and noun in attributive-head relation are tagged as "vn n", which means, the verb has been recognized as an attributive [9-10]. However, POS tagging task is influenced by various factors, which may lead to the failure in distinguishing syntactic relations between two words.

(1) Syllables. According to the POS tag set of Peking University, only two-syllable verbs in attributive positions are tagged as nominal verbs "vn", and single-syllable verb and multi-syllable verb in the same position are still given the tag "v", such as "醉/v 罗汉/n zui luohan drunken arhat", "现代化/v手段/n xiandaihua shouduan modernized means".

(2) Syntactic position. The tag set of Peking University illustrates that, whether a verb should be labeled as "v" or "vn" is determined in the range of the smallest structure the verb belongs to. It can make the same verb in similar V-N relations have different tags, for example, "这些/r 介绍/vn 机构/n zhexie jieshao jigou these agencies" and "职业/n 介绍/v 机构/n zhiye jieshao jigou employment agency". Although the tagging rule is reasonable for tagging task, we still need to solve the problem, and make the identification of the attributive-head collocations more accurate.

(3) Tagging errors. It is important despite of some problems for a tagging system to decide whether a V-N combination is a verb-object collocation or an attributive-head collocation. For example, in the phrase "这些 介绍 机构zhexie jieshao jigou these agenciyes", "介绍 jieshao introducing " may be given a wrong tag "v" instead of the correct tag "vn" in a different context. An important reason is that the second factor has a negative impact on statistical models.

In addition, deep syntactic relations are hard to be recognized by tagging system. Many phrases, such as "老师/n 讲授/v 的/u 知识/n laoshi jiangshou de zhishi the knowledge that teacher taught", "知识/n 的/u 传授/vn zhishi de chuan shou teaching of knowledge", "学习/v 的/u 对象/n xuexi de duixiang learning objects", are structures with implicit verb-object relations or attributive relations, which cannot be expressed in realized syntactic level. If the implicit relations can be identified, we will acquire more knowledge from limited corpus.

4 Extraction of V-N Collocations

A statistical model based on mutual information is designed to extract VO collocations. Mutual information reflects the correlation between the two events set, and can be used to measure the tightness of inter-linkages of two collocates. If the probabilities of two words w_1 and w_2, are $P(w_1)$ and $P(w_2)$, their mutual information can figure out according to the following formula:

$$I(w_1;w_2) = \log \frac{P(w_1,w_2)}{P(w_1)P(w_2)} . \quad (1)$$

Since lexical collocations are repeatedly presented syntactic pattern, theoretically, mutual information should be calculated on the basis that w_1 and w_2 have certain syntactic relation. However, it is not easy to acquire Chinese tree corpus of high quality, so collocation data are usually extracted from segmented or tagged texts, and the mutual information is calculated approximately. Therefore, a weighted mutual information model is advanced to get collocations and its mutual information data more accurately.

Furthermore, statistical methods based on co-occurrence are hard to resolve ambiguity of V-N relations, and extra linguistic knowledge is needed to solve the problem. In this paper, some rules are presented to acquire collocations too, and the extraction process is divided into two stages. Verb-object collocations are extracted in first stage, while attributive-head collocations are acquired based on the output of first stage.

4.1 Language Resources Preparation

A grammatical information dictionary of verbs is prepared for both stages. It is expected to help the extraction models to have better performances on accuracy and reducing the mutual inference of the two relations. There are two sets of useful information in the dictionary.

(1) Valence of verb. Verbs in the dictionary are classified into three categories: monovalent verbs, bivalent verbs and trivalent verbs. This kind of information is useful to decide whether a verb can have an object and form a verb-object collocation. For example, a monovalent verb in V-N combination is more likely to be attributive instead of object.

(2) Subcategorization of verb. Modal verbs, such as "会 hui might", "敢 gan dare", and “能 neng can”, as well as verbal object verbs, such as “打算 dasuan intend”, “希望 xiwang hope”, “觉得 jue de / feel ”, cannot be followed by nominal object, and cannot function as attributives. Thus, they cannot form verb-noun collocations. So the information is useful to distinguish whether a verb can form a V-N collocation.

4.2 Extraction of Verb-Object Collocation Based on Weighted Mutual Information Model

In order to obtain more long distance verb-object collocation and reduce the opportunity of extracting the wrong instance, we designed a new model based on weighted mutual information to replace the conventional MI model with a fixed window in which the frequencies of different collocates are equal.

Suppose the clause S is divided into different blocks by verbs and prepositions, for each verb in S, the last noun of each block after the verb is called first class candidate collocate, and the other nouns in the block are called second class candidate collocates. Collocations within the clauses are obtained in four steps.

Step 1: Use the grammatical information dictionary of verbs to mark the validity of each verb to form a collocation in sentence. Bivalent verbs and trivalent verbs which can have objectives will be identified as valid verbs, except the verb adjacent to the structural auxiliary word "的 de of" on the right. For each valid verb, step2 to step4 will be performed.

Step 2: Identify the first class candidate collocates, and initialize the frequency of the nearest collocate of the verb to 1. As a starting point, the frequencies of the first class candidate collocates on the right will be multiplied by the weighting factor f1 one by one. When there is structural auxiliary word "的 de of", the weight of the first class candidate collocate will be reset to 1.

Step 3: Take each first class candidate collocate as a starting point, and scan the sentence from right to left to get the second class collocates, with their frequencies multiplied by the weighting factor f2 one by one.

Step 4: Calculate the WMI value for each collocation, and sort the collocations.

Table 1 shows the weighted data of the part of speech sequence "v1 n n v2 n n", the two target verbs v1 and v2 are given two weights: weight (v1) and weight (v2).

Table 1. Examples about weighting POS sequence

wordID	POS	weight(v_1)	weight(v_2)
1	v_1		
2	n	$1.0*f_2$	
3	n	1.0	
4	v_2		
5	n	$1.0*f_1*f_2$	$1.0*f_2$
6	n	$1.0*f_1$	1.0

The weighted mutual information model considers the left recursion characteristics of Chinese noun phrase. It can give higher WMI value to long distance collocations, especially to the V-N collocations in the structure "verb + noun phrase including 的 (de of)". Adjusting the weighting factors f1 and f2 can control the weights of first class collocates on the right end of clause, as well as the decay rate of the homogeneous composition from right to the left.

4.3 Extraction of Verb-Object Collocations Based on Rules

The method based on mutual information is suitable to high frequency collocation extraction, and the rule method can extract low frequency collocations, as well as distinguish verb-object relation and attributive-head relation more accurately. In order to improve the precision, the rule method mainly extracts the collocations located at the end of clauses with three steps.

Step 1: Use the grammatical information dictionary of Verb to mark the validity of each verb to form a collocation in sentence. Bivalent verbs and trivalent verbs which can have objectives will be identified as valid verbs, except the verb adjacent to the structural auxiliary word "的 de of" on the right. For each valid verb, step2 to step3 is performed.

Step 2: Recognize nouns in sequence "n + conjunctive | modal particle | punctuation" as head noun N.

Step 3: Identify verb-object collocations based on five rules. These rules are mainly based on part of speech sequences, including:

Rule 1: For sequence "v + 了 (le) | 着 (zhe) | 过 (guo) + x + N", if there is no interrupting constituent, such as "的 de of", verbs and prepositions in x, identify "v N" as verb-object collocation. For sequence "v + 了 (le) | 着 (zhe) | 过 (guo) + x + 的(de of) + x1 + N", if there is no interrupting constituent, such as "的 de of", verbs and prepositions in x and x1, identify "v N" as verb-object collocation.

Rule 2: For sequence "v v1 N", if "v N" or "v1 N" is verb-object collocation, the frequency of collocation plus 1.

Rule 3: For sequence "[vi x]+ 的(de of) N", if "vi N" is verb-object collocation, the frequent of collocation plus 1.

Rule 4: For sequence “v x 的(de of) N”, if there is no interrupting constituent in x, identify “v N” as verb-object collocation.

Rule 5: For sequence “v x N”, if there is no interrupting constituent in x, identify “v N” as verb-object collocation.

Among the rules, the symbol "|" means “or”, the brackets "[]" means “optional”, the symbol “+” or space means “adjacent to”, the symbol “+” means the item appears once or many times previously. These symbols are of the same meaning below.

4.4 Extract Attribute-Head Collocations Based on Rules

Some rules of high reliability are made to acquire attributive-head collocations on the basis of verb-object collocation extraction. The process of program is the same with rule-base extraction of verb-object collocation. There are five main rules, including two identification rules (Rule 1-2) and three inference rules (Rule 3-5).

Rule 1: For sequence “v 的(de of) N”, identify “v N” as attributive-head collocation.

Rule 2: For sequence “的(de of) v N”, identify “v N” as attributive-head collocation.

Rule 3: For sequence “v v1 N”, if “v N” is verb-object collocation, identify “v N” as attributive-head collocation.

Rule 4: For sequence “v x [的(de of)] v1 N”, if “v N” is verb-object collocation, identify “v1 N” as attributive-head collocation.

Rule 5: For sequence “v x v1 x n1 的(de of) N”, if “v N” is verb-object collocation, identify “v1 n1” as attributive-head collocation.

In the rules mentioned above, verb-object collocations are restricted to the ones whose WMI value≧3 and frequency≧5, and the ones extracted by rules. The entire rule-based extraction tags the linkage of two words in sentence, and makes sure that they are not acquired repeatedly by differently rules.

5 Experiments and Analyses

The experiment uses the Beijing Youth Daily (2004) as test corpus to obtain collocations. It calls the ICTCLAS2009 interface developed by the Institute of Computing Technology Chinese Academy of Sciences (CAS) for segmentation and part of speech tagging, using Peking University tag set. The tagged corpus is put into use without a manual proofreading.

5.1 Experiments and Analyses of Verb-Object Collocation Extraction

Since the corpus scale is large, we choose the collocations with the frequency≧10 to do evaluation. In the weighted mutual information model, set f1=f2=0.8, and select five random samples, with each sample containing 100 combinations. Each combination is evaluated manually whether it is a collocation. The experimental results are as follows:

Table 2. Results of verb-object extraction based on WMI model

Sample	1	2	3	4	5	ave
Precision (%)	62.00	67.00	66.00	65.00	71.00	66.20

The data analysis shows that, the weighted strategy produce beneficial changes to the MI value, and particularly it makes tight collocations of long-distance obtain a higher evaluation value. Table 3 reports the evaluation values of some long-distance collocations acquired by WMI model and classic MI model (the observation window is [0, 5]).

Table 3. Changes of MI value in WMI model

ID	V	N	Pinyin	Translation	WMI	MI
1	提高	认识	tigao renshi	raise awareness	4.970712	3.487450
2	位于	处	weiyu chu	locate at	6.128926	3.766732
3	平	纪录	ping jilu	equal the record	7.111156	5.149114
4	运用	方法	yunyong fangfa	use … method	7.149187	5.919477
5	抽查	产品	choucha chanping	make a random inspection of products	7.169726	5.791281
6	审议	方案	shen yi fang an	review… plan	7.484358	5.910167
7	经历	考验	jingli kaoyan	experience test of	7.557571	6.277409
8	听取	报告	tingqu baogao	listen to a report	9.476893	7.657483
9	奠定	地位	dianding diwei	lay … position	10.18848	8.365789
10	开创	先河	kaichuang xianhe	initiate the beginning of	13.98664	12.63321

As the V-N collocation contains two different relations, verb-object relation and attribute-head relation, the ability of different verbs and nouns to form collocations of the two relations is various. We evaluate the collocations containing the noun “能力 nengli ability” which has good binding ability with verbs. When frequency≧5, the results of extraction experiment are as follows:

Table 4. Extraction of collocations containing “能力 nengli ability” based on WMI model

RealRel	ExtrVO (type)	Total(type)	Percent (%)
VO	75	153	49.02
DZ	50	153	32.68
ALL	125	153	81.70

In table 4, “RealRel” means the correct relation of the extracted collocations, and “ExtrVO” means the number of collocations extracted as verb-object relation. The precision of verb-object collocation extraction is 49.02%. 32.68% of the extracted

combinations are attributive-head collocations, and 18.30% are illegal combinations. It is clear that mutual information is not sensitive to distinguish the syntactic relation of V-N collocation.

In order to distinguish the two relations more accurately, we use rule-based method to acquire verb-object collocations. Compared with statistic-based approach, the rule based method has lower recall rate, but it is more effective in distinguishing the two relations. Five random samples are selected from extraction results through a random sampling procedure, with each sample containing 100 combinations, and each combination is evaluated manually whether it is a collocation. The experimental results are as follows:

Table 5. Evaluation of verb-object collocations extracted by rules

Sample	1	2	3	4	5	ave
Precision(%)	82.00	89.00	81.00	87.00	88.00	85.4

Most of the extraction errors are illegal collocations, and few of them are attributive-head collocations. The most common error type is "Vc + N" where "Vc" stands for a complement verb. Most of "Vc + N" sequences are part of "V Vc N" where "Vc" mainly makes up of directional verbs.

The rule based method is also effective in extracting low frequency verb-object collocations. Select five random samples when the collocation frequency $f \in [1, 4]$, with each sample containing 100 combinations. The experimental results are as follows:

Table 6. Evaluation of low frequency verb-object collocations extracted by rules

Sample	1	2	3	4	5	ave
Accuracy (%)	83.00	83.00	82.00	78.00	80.00	81.20

By checking the frequency data supplied by the weighted mutual information model, we find that a large number of low frequency collocations extracted by rule based method and WMI model overlap with each other, and many collocations captured once by rules also have very low frequency in corpus. In this situation, the MI value is hard to measure the intensity of collocation. So the rule-based method and the WMI model are complementary to each other in those cases.

5.2 Experiment and Analysis of Attributive-Head Collocation Extraction

With the support of POS tagging, the rule based method obtains a better result by careful rule designation and rule-based reasoning. A selection of five random samples with each sample containing 100 combinations leads to the following experiment results:

Table 7. Evaluation of attributive-head collocations extracted by rules

Sample	1	2	3	4	5	ave
Accuracy (%)	89.00	86.00	91.00	85.00	85.00	87.20

More than half of the errors are verb-object collocations, and some errors are incurred by segmentation and part of speech tagging. Besides, a superficial description of sub-categorization of verbs affects the accuracy rate too. Some verbs, such as "告诉 gaosu tell", which cannot be a attributive, are extracted. These errors can be avoided by elaborating the sub-categorization of verb in more details in dictionary.

The rule-based method has a good performance on low frequency collocations as well. Five random samples are selected when the collocation frequency $f \in [1, 4]$, with each sample containing 100 combinations. The precision is close to the evaluation of overall sample:

Table 8. Evaluation of low frequency attributive-head collocation extracted by rules

Sample	1	2	3	4	5	ave
Precision (%)	87.00	85.00	88.00	86.00	85.00	86.40

6 Conclusions

The goal of collocation extraction task is to extract lexical collocations, including explicit or implicit in sentence. V-N collocation has two different syntactic relations between collocates, which are verb-object relation and attributive-head relation. It is necessary to extract collocations of both the types. The classical statistical model usually sets a fixed observation window, and assigns the same frequency to all collocates. In this paper, we give more considerations to structural factors, and advance a weighted mutual information model and a rule-based method to obtain verb-object collocations and attributive-head collocations. The former is a flexible way to obtain long distance verb-object collocations, and it can optimize the MI evaluation. The latter is more accurate in extracting and distinguishing verb-object collocations and attributive-head collocations, and it is especially useful to acquire the low frequency collocations, as well to extract targeted implicit collocations. Still, the current study merits further research in the following two aspects: to explore the combination rules and subcategorization rules of higher validity and to improve the recall rate of rules.

Acknowledgments. This work was supported by Science Foundation for the Young College Teachers of Shanghai Municipal Education Commission (No. shu11053) and Innovation Project of Shanghai University (No. 2010CXYB03).

References

1. Church, K.W., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1), 22–29 (1990)
2. Smadja, F.: Retrieving Collocations from Text: Xtract. Computational Linguistics 19(1), 143–177 (1993)
3. Sun, M.S.: Quantitative Analysis of the Chinese Collocations. Studies of Chinese Language 256(1), 29–38 (1997)
4. Lin, D.K.: Extracting Collocations from Text Corpora. In: Proceedings of 1st Workshop on Computational Terminology, pp. 57–63. MIT Press, Montreal (1998)
5. Bai, M.Q., Zheng, J.H.: Study on Ways of Verb-Verb Collocation. Computer Engineering and Applications 40(27), 70–72 (2004)
6. Wang, X.: A Study on the Automatic Acquisition of Verb-object Collocations in Chinese. Applied Linguistics (1), 137–143 (2005)
7. Zhu, D.X.: "de" Phrase and Judgment Sentence ("的"字结构和判断句). Studies of Chinese Language (1), 23–27 (1978)
8. Zhu, D.X.: "de" Phrase and Judgment Sentence ("的"字结构和判断句). Studies of Chinese Language (2), 104–109 (1978)
9. Yu, S.W., Duan, H.M., Zhu, X.F., Sun, B.: The Basic Processing of Contemporary Chinese Corpus at Peking University Specification. Journal of Chinese Information Processing 16(5), 49–64 (2002)
10. Yu, S.W., Duan, H.M., Zhu, X.F., Sun, B.: The Basic Processing of Contemporary Chinese Corpus at Peking University Specification. Journal of Chinese Information Processing (continued) 16(6), 58–64 (2002)

Identification on Semantic Orientation of Adjectives in Nominal Compound Phrases AN_1N_2

Zan He[1,2] and Caijun Li[1,2]

[1] College of Chinese language and literature, Wuhan University, China
[2] Center for Study of Language and Information, Wuhan University, China
Whu-hezan@foxmail.com, Licaijun160@gmail.com

Abstract. Nominal compound phrase AN_1N_2 is composed of an adjective and two nouns. According to the difference of semantic orientation of adjectives, AN_1N_2 structure can be classified into four semantic hierarchical models: A- (N_1) -N_2, (A-N_1) -N_2, A- (N_1-N_2), A-N_1/N_2. Based on the distinctive features of the four types, operable recognition models can be established which could make it possible for computers intelligently identifying the semantic orientation of the AN_1N_2 structure.

Keywords: Nominal compound phrase, AN_1N_2, Adjective, Semantic orientation, Semantic hierarchical model.

1 Introduction

There is complex semantic information under the nominal compound phrases, and the semantic explanation on which is a challenge issue in the field of Natural Language Processing. On the computer semantic comprehension, it's complicated about the semantic orientation between words. For example:

1) xīn zhígōng sùshè (new quarters for staff /quarter for new staff)
2) xīn kuǎnshì fúzhuāng (new styles of clothes)

In phrase 1), the semantic of the adjective *xīn (new)* could be oriented to the noun zhígōng *(staff)*, or be oriented to the noun *sùshè (quarter)*. Ambiguity exists. But in phrase 2), the semantic of the adjective *xīn (new)* is definitely oriented to the noun *kuǎnshì (style)*. It demonstrates the great importance of accurately distinguishing the semantic relations between words of the nominal compound phrase to the computer semantic comprehension.

Nominal compound phrase is a kind of noun phrase that generally composed of numbers of nouns and verbs or adjectives. Chinese language scholars have done some research on semantic analysis of NVN [1], NNN [2], N_1AN_2 [3], N_1N_2V [4], and other forms of nominal compound phrases. Unfortunately, research on AN_1N_2 is still insufficient. In this article, with the analysis on the semantic orientation of adjectives, AN_1N_2 structure is classified into four semantic hierarchical models: A- (N_1) -N_2, (A-N_1) -N_2, A- (N_1-N_2), A- N_1/N_2. Based on the distinctive features of the four types, operable recognition models can be established which could make it possible for computers intelligently identifying the semantic orientation of the AN_1N_2 structure.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 242–250, 2013.

2 Principles on Corpus Selection

The corpora in this article are all selected from Corpus Query System of the State Language Commission. Nominal compound phrase AN_1N_2 generally meets the following conditions:

1) There are no structural words such as *de (of)*, *hé (and)*, *yǔ (and)*, *huò (or)* and so on between A and N_1 or N_1 and N_2.

2) N_1, N_2 are pure nouns.

3) AN_1N_2 structure is independent and complete. It is not part of the other structure.

For example, in the sentence *jiākuài quán xìtǒng chǎnpǐn jiégòu de tiáozhěng (speed up the adjustment of the system-wide product architecture)*, though the phrase *quán xìtǒng chǎnpǐn (system-wide product)* is in line with the AN_1N_2 structure, but it is actually part of the noun phrase *quán xìtǒng chǎnpǐn jiégòu (the system-wide product architecture)*. It is incomplete semantically. So the phrase *quán xìtǒng chǎnpǐn (system-wide product)* is an unqualified corpus.

4) Corpus containing monosyllabic nouns is generally not selected, because it is difficult to make a judgment about whether the monosyllabic noun is a word or morpheme, such as *gāo tǒng xuē (high boots)*, *dú tái xì (independent drama)*.

3 Classification on Semantic Orientation of Adjectives in AN_1N_2

3.1 Methods to Judge the Semantic Orientation of Adjectives

In the syntactic structure, the semantic relationship with a certain direction and a certain target between the syntactic elements is called semantic orientation [5]. The semantic orientation of adjectives in AN_1N_2 is such semantic relationship between A and N_1 or A and N_2. Zhou pointed out that the semantic compatible principle is the first principle to determine the semantic orientation. Based on this principle, we use the following two methods to determine the semantic orientation of adjectives.

1) Combination test. If the adjective can be combined with N_1 or N_2, the two words are in the semantic relationship. For example, in the phrase *yánzhòng pífū shīzhěn (severe dermatological eczema)*, *yánzhòng (severe)* and *shīzhěn (eczema)* can be combined into *yánzhòng shīzhěn (severe eczema)*. But *yánzhòng (severe)* cannot be combined with *pífū (dermatological)*. It refers that the semantic orientation of the adjective is oriented to N_2 instead of N_1.

2) *De (of)* test. If *de (of)* can be added between A and N_1, it refers that A is in closer semantic relationship with N_2. If *de (of)* can be added between N_1 and N_2, it refers that A is in closer semantic relationship with N_1. For example, in the phrase *zhùmíng móshù biǎoyǎnzhě (famous magic performer)*, *de (of)* can only be placed between *zhùmíng (famous)* and *móshù (magic)*, instead of being placed between *móshù (magic)* and *biǎoyǎnzhě (performer)*. It refers that *zhùmíng (famous)* is in closer semantic relationship with *biǎoyǎnzhě (performer)*.

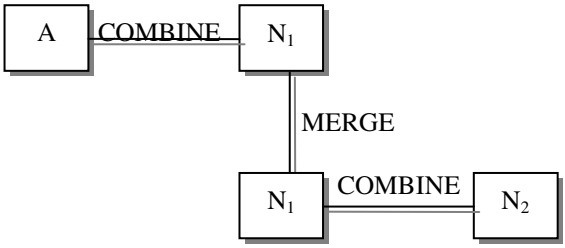
3.2 Classification on Semantic Orientation of Adjectives

The meanings of nouns come from the objects, and the meanings of adjectives come from the characteristic of objects or the state of affairs. Adjectives appeal to nouns. Thus, the adjective in AN_1N_2 is oriented to one noun at least. Adjective has three potential orientations: adjacent orientation, non-adjacent orientation, double orientation.

3.2.1 Adjacent Orientation

Adjacent orientation means that the adjective is oriented to N_1. In the structure AN_1N_2, A-N_1 can be combined, but A-N_2 cannot be combined. According to the combination of N_1-N_2 collocation, adjacent orientation is divided into two types.

1) N_1 and N_2 can be combined. By merging the same word N_1, A-N_1 collocation and N_1-N_2 collocation are combined into A- (N_1) -N_2. It's named for snap-fit series semantic model.

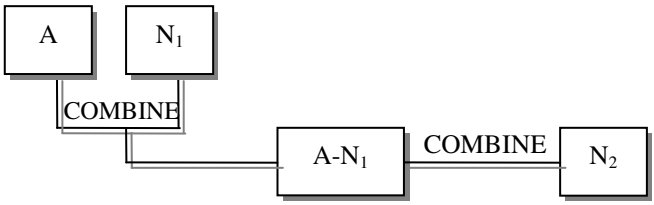


For example, in the phrase *gāosù jiānjījī fēixíngyuán (high-speed fighter pilot)*, *gāosù (high-speed)* and *jiānjījī (fighter)* is combined into the phrase *gāosù jiānjījī (high-speed fighter)* first. *Jiānjījī (fighter)* and *fēixíngyuán (pilot)* is combined into the phrase *jiānjījī fēixíngyuán (fighter pilot)* at the same time. And then, by merging the identical word *jiānjījī(fighter)*, these two phrases are combined into *gāosù jiānjījī fēixíngyuán (high-speed fighter pilot)*. In other words, the phrase *gāosù jiānjījī fēixíngyuán (high-speed fighter pilot)* is short for the phrase *gāosù jiānjījī de jiānjījī fēixíngyuán (fighter pilot of high-speed fighter)*.

Similar examples are as follows:

guóchǎn shāngpǐn zhìliang (the quality of domestic goods)
dà qǐyè fùzé'rén (person-in-charge of the large enterprises)
zhěnggè zhànzhēng mùdi (the purpose of the entire war)
xiǎo hútong shēnchù (deep in the small alley)
yí'nán jíbìng huànzhě (patients with incurable diseases)
zuìgāo wěiyuánhuì wěiyuán (members of the supreme council)

2) N_1 and N_2 cannot be combined. The adjective and N_1 are combined into A-N_1 first. And then A-N_1 semantically associated with N_2 as a whole. The hierarchical model is (A-N_1) -N_2. It's named for progressive semantic model.



For example, in the phrase *dà miàn'é rénmínbì (large-denomination RMB)*, *dà (large)* and *miàn'é (denomination)* are firstly combined into the phrase *dà miàn'é (large-denomination)*. And then *dà miàn'é (large-denomination)* semantically associated with *rénmínbì (RMB)*. If testing by adding *de (of)*, *de (of)* can only be placed between A and N_2, instead of being placed between A and N_1.

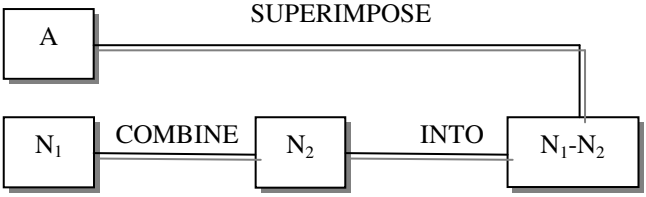
Similar examples are as follows:

héping lìliang liánméng (alliance of peaceful force)
gāo gōnglǜ fādòngjī (high-power engine)
dī nénghào chǎnpǐn (low power loss products)
gāo chǎnliàng zuòwu (high-yield crops)
dà píngyuán shàngkōng (over the great plains)
duō ménlèi chǎnpǐn (multi-category products)

In $(A\text{-}N_1)\text{-}N_2$ progressive semantic hierarchical model, the $A\text{-}N_2$ cannot be combined theoretically. However, examples exist in the actual corpus that $A\text{-}N_2$ can be combined. For example, in the phrase *xīn kuǎnshì fúzhuāng (new styles of clothes)*, *xīn (new)* and *fúzhuāng (clothes)* can be combined into the phrase *xīn fúzhuāng (new clothes)"*. With analysis, we found that *xīn (new)* has two meanings at least. One meaning is about extent, and the other is about time. *Xīn (new)* in the phrase *xīn kuǎnshì fúzhuāng (new styles of clothes)* only has the meaning about time. That's why *xīn (new)* can only be oriented to *kuǎnshì (styles)* instead of *fúzhuāng (clothes)*. Actually $A\text{-}N_2$ still cannot be combined.

3.2.2 Non-adjacent Orientation

Non-adjacent orientation means that the adjective is oriented to N_2. $A\text{-}N_2$ can be combined, but $A\text{-}N_1$ cannot be combined. If testing by adding *de (of)*, *de (of)* can only be placed between A and N_1, instead of being placed between A and N_2. In this structure, $N_1\text{-}N_2$ is bound to combine. Firstly, N_1 and N_2 are combined into $N_1\text{-}N_2$, and then the semantic of the adjective is superimposed on $N_1\text{-}N_2$. The hierarchical model is $A\text{-}(N_1\text{-}N_2)$. It's named for stacking semantic model.



For example in the phrase *zhòngduō tǐcāo míngjiàng (a number of gymnastics champion)*, *tǐcāo (gymnastics)* and *míngjiàng (champion)* is combined into *tǐcāo míngjiàng (gymnastics champions)* first. And then the semantic of *zhòngduō (a number of)* is superimposed on *tǐcāo míngjiàng (gymnastics champion)*.

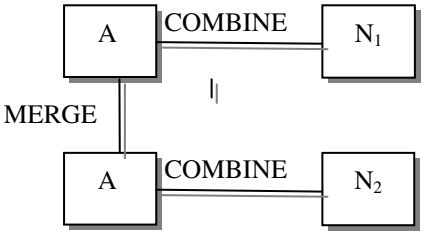
Similar examples are as follows:

guóchǎn pēnqìshì jiānjījī (domestic jet fighter)
zhōngděng zhuānyè xuéxiào (specialized secondary schools)
yánzhòng pífū shīzhěn (severe dermatological eczema)
cháng yùndòng yīkù (long exercise clothes)
dīwā yánjiǎn dìqū (low-lying saline areas)
zhòngdà guójì wèntí (major international issues)

In A- (N_1-N_2) stacking semantic model, A-N_1 cannot be combined theoretically. Similarly, examples also exist in the actual corpus that A-N_1 can be combined. For example in the phrase *zhùmíng móshù biǎoyǎnzhě (famous magic performer)*, *zhùmíng (famous)* and *móshù (magic)* can be combined into *zhùmíng móshù (famous magic)*. However, testing by adding *de (of)*, we found that *de (of)* can only be placed between *zhùmíng (famous)* and *móshù (magic)*, instead of being placed between *móshù (magic)* and *biǎoyǎnzhě (performer)*. This demonstrates that *zhùmíng (famous)* and *biǎoyǎnzhě (performer)* are in closer relationship semantically. Actually A-N_1 still cannot be combined.

3.2.3 Double Orientation

Double orientation means that the adjective is oriented to N_1 and N_2 at the same time. N_1 and N_2 are generally in parallel relationship. By merging the same adjective, A-N_1 collocation and A-N_2 collocation are combined into A-N_1/N_2. It's named for snap-fit parallel semantic model.



For example, in the phrase *dàliàng rénlì wùlì (a lot of human resources and material resources)*, *dàliàng (a lot of)* and *rénlì (human resources)* is combined into the phrase *dàliàng rénlì (a lot of human resources)*. At the same time, *dàliàng (a lot of)* and *wùlì (material resources)* is combined into the phrase *dàliàng wùlì (a lot of material resources)*. And then, by merging the same adjective *dàliàng (a lot of)*, these two phrases are combined into *dàliàng rénlì wùlì (a lot of human resources and material resources)*. Generally, *hé (and)* or *yǔ (and)* can be placed between N_1 and N_2 in such phrases. Similar examples are as follows:

bùshǎo gōngchǎng qǐyè (many factories and enterprises)
jùtǐ fāngfǎ bùzhòu (the specific method and procedure)
zhǔyào zhèngcè cuòshī (major policy and measure)

zhěnggè guójiā mínzú (the whole country and nation)
guǎngdà dúzhě guānzhòng (the majority of readers and viewers)
xǔduō gōngrén nóngmín (many workers and peasants)

3.2.4 Ambiguity

Ambiguity means that there are two mutually exclusive situations on the orientation of the adjective. When the adjective is oriented to N_1, the phrase belongs to A- (N_1)-N_2 snap-fit series semantic model. When the adjective is oriented to N_2, the phrase belongs to A- (N_1-N_2) stacking semantic model. Both of the two situations are fit for the principle of semantic compatible, so the exact meaning should rely on the context.

For example, in the phrase *xīn zhígōng sùshè*, when *xīn (new)* is oriented to *zhígōng (staff)*, the phrase *xīn zhígōng sùshè* is short for *xīn zhígōng de zhígōng sùshè (quarter for new staff)*. But when *xīn (new)* is oriented to *sùshè (quarter)*, the semantic of *xīn (new)* is superimposed on *zhígōng sùshè (staff quarter)*, and the whole phrase means *new quarters for staff*.

Similar examples are as follows:
dàxíng qìchē tíngchēchǎng (parking lot of large car /large parking lot)
xīn xuéshēng shítáng (cafeteria for new students /new cafeteria for students)
yīliú chéngshì biāozhǔn (standards of class city /class standards of city)
héfǎ shāngpǐn jiāoyì (transaction of legitimate goods /legal commodity trading)

4 Semantic Hierarchical Models of AN_1N_2 and Its Identification

4.1 Internal Semantic Combinations of AN_1N_2

By analyzing the orientation of the adjective in AN_1N_2, AN_1N_2 structure can be classified into four semantic hierarchical models: A- (N_1)-N_2 snap-fit series semantic model, (A-N_1)-N_2 progressive semantic model, A- (N_1-N_2) stacking semantic model and A-N_1/N_2 snap-fit parallel semantic model. The semantic combinations of A-N_1, A-N_2 and N_1-N_2 are as follows (+ means that can be combined; - means that cannot be combined):

	A- (N_1)-N_2	(A-N_1) -N_2	A- (N_1-N_2)	A-N_1/N_2	Ambiguity
A-N_1	+	+	—	+	+/—
A-N_2	—	—	+	+	—/+
N_1-N_2	+	—	+	+	+

4.2 Procedure on Semantic Hierarchical Model Identification of AN_1N_2

The semantic hierarchical model of AN_1N_2 can be identified by confirming the semantic combinations of A-N_1, A-N_2 and N_1-N_2. The specific steps are as follows (the flow chart is in the Appendix):

1) To determine whether N_1-N_2 can be combined. If not, the nominal compound phrase is (A-N_1)-N_2 progressive semantic model. If yes, to enter the next step;

2) To determine whether N_1 and N_2 are in parallel relationship. If yes, the nominal compound phrase is A-N_1/N_2 snap-fit parallel semantic model. If not, to enter the next step;

3) To determine whether A-N_2 can be combined. If not, the nominal compound phrase is A- (N_1)-N_2 snap-fit Series semantic model. If yes, to enter the next step;

4) To determine whether A-N_1 can be combined. If not, the nominal compound phrase is A- (N_1-N_2) stacking semantic model. If yes, enter the next step;

5) To determine whether "de (Of)" can be added between N_1 and N_2. If not, the nominal compound phrase is A- (N_1-N_2) stacking semantic model. If yes, the nominal compound phrase is ambiguous structure.

4.3 Verification and Analysis

To verify the accuracy of the classification of semantic hierarchical models and the reliability of the identification methods, we selected corpora in the State Language Commission of Contemporary Chinese Corpus to analyze. Corpus theme is "newspaper". Time ranges from 1947 to 1992. The search keywords string is "a n n". 2051 pieces of corpora were hit. Removing repeated corpora, 1688 pieces were left. With the corpus selection principles in 2.2, 688 pieces of corpora were eligible after filtering.

Judged with the procedure in 4.2, the quantity and proportion of the four semantic hierarchical models in the corpora are in the following table:

	A- (N_1)-N_2	(A-N_1)-N_2	A- (N_1-N_2)	A-N_1/N_2	Ambiguity
Quantity	86	92	383	25	102
Proportion	12.5%	13.37%	55.67%	3.63%	14.83%

The results showed that the proportion of A- (N_1-N_2) stacking semantic model is much higher than the other models. A- (N_1)-N_2 Snap-Fit Series semantic model, (A-N_1)-N_2 progressive semantic model and ambiguity are of similar proportion. A-N_1/N_2 Snap-Fit parallel semantic model is in the lowest proportion. This shows that A- (N_1-N_2) stacking semantic level model is the main semantic hierarchical model of AN_1N_2.

5 Further Study

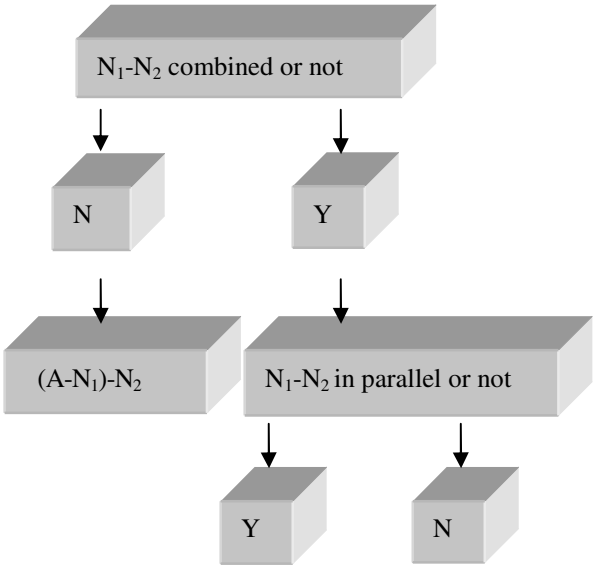
In establishing semantic hierarchical models of AN_1N_2, the identification on semantic orientations of adjectives is simplified into the determination on semantic combinations of A-N_1, A-N_2 and N_1-N_2, which is more operational and objective. Therefore the remaining problem to be solved is how to determine the validity of A-N combination and N-N combination. This needs to do a more detailed classification and description of the properties of adjectives and nouns. Some scholars have already conducted some research on this issue. Zhao Chunli has made an classification of adjectives and nouns, and built four types of semantic combination model of A-N: adjectives of human - nouns of human, adjectives of affair - nouns of affair, adjectives of objects - nouns of objects, evaluation adjectives - nouns of human/ affair / objects / logic [6]. All combinations of the four models are effective semantic combinations. Zhu Yan has established a noun grid system for semantic analysis that contains 22 categories of nouns grid [7]. According to the noun grid system by Zhu Yan, Zhou Ri'an has described 18 common semantic relations of the N-N combination [8]. We will continue to do further study in the future.

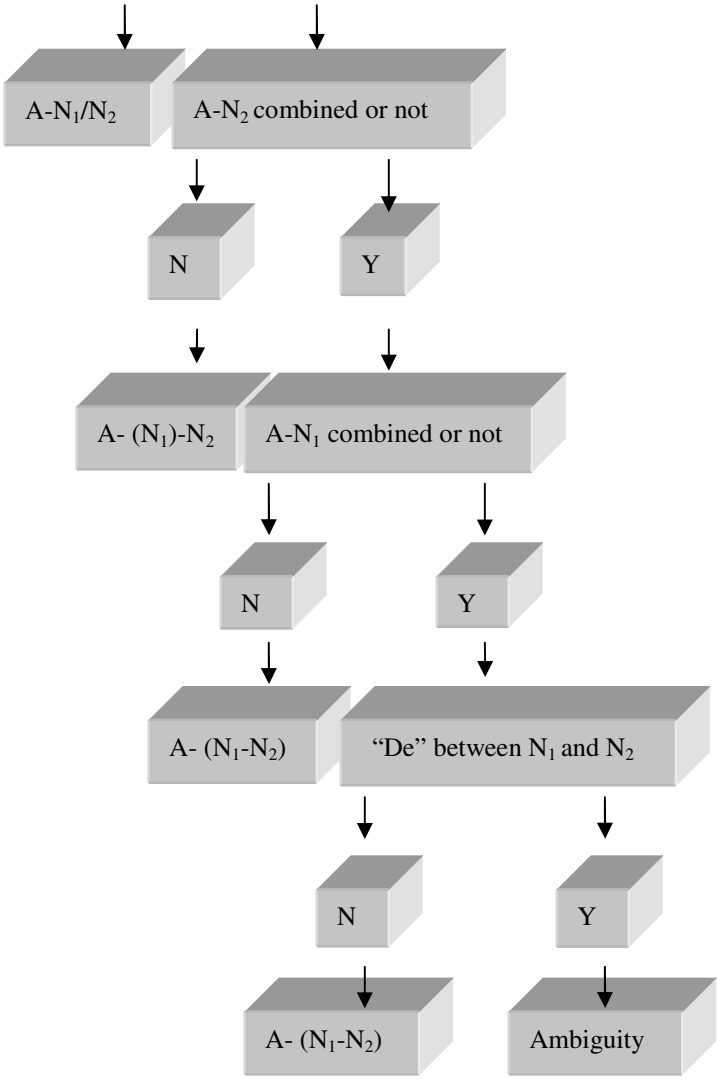
Acknowledgements. This work has been supported by the Major Projects of Chinese National Social Science Foundation (11&ZD189) and the National Natural Science Foundation of China (61173095). All errors are ours.

References

1. Xing, F.Y.: NVN Structure and Its Omitted Form NV | VN. Language Study 2, 1–12 (1994) (in Chinese)
2. Yao, W.: A Syntactic Ambiguity study on "noun+nonu+noun" Structure of Modern Chinese for Chinese Information Processing, pp. 2–8. Central China Normal University (2007) (in Chinese)
3. Feng, W.H., Ji, D.H.: Judgment and Analysis on Semantic Orientation of Adjectives in Compound Noun Phrase N_1AN_2. In: 11th Chinese Lexical Semantic Workshop, pp. 250–256 (2010) (in Chinese)
4. Xiao, S.B., Zhao, H.G.: Noun Phrase of "N_1+N_2+V" Structure in Search Engine Query Logs. Guangxi Normal University (Natural Science) 3, 116–122 (2011) (in Chinese)
5. Zhou, G.G.: On the Principle and Method of Semantic Orientation Analysis. Language Science 4, 41 (2006) (in Chinese)
6. Zhao, C.L., Shi, D.X.: A Semantic Construction Model between Adjectives and Nouns in Chinese. Journal of Chinese Information Processing 9, 13–16 (2009) (in Chinese)
7. Zhu, Y.: Study on Semantic Word Formation of Chinese Compound Words, p. 125. East China Normal University (2003) (in Chinese)
8. Zhou, R.A.: A Syntactic and Semantic Study on Noun-Noun Constructions, pp. 126–130. Jinan University (2007) (in Chinese)

Appendix: Flow Chart of Identification Steps of the Semantic Hierarchical Model of AN_1N_2




A-N1/N2
A-N2 combined or not
N
Y
A- (N1)-N2
A-N1 combined or not
N
Y
A- (N1-N2)
"De" between N1 and N2
N
Y
A- (N1-N2)
Ambiguity

Measuring the Semantic Relevance between Term and Short Text: Using the Concepts of Shortest Path Length and Relatively Important Community

Hua Yang[1,2,*], Donghong Ji[2,3], Mingyao Zhang[1,4], Bo Chen[3], and Hongmiao Wu[4]

[1] College of Chinese Language and Literature, Wuhan University, Wuhan, 430072, China
yanghuastory@263.net, myzhang@whu.edu.cn
[2] School of Mathematics and Computer Science, Guizhou Normal University, Guiyang, 550001, China
[3] School of Computer, Wuhan University, Wuhan, 430072, China
donghongji_2000@yahoo.com, cb9928@gmail.com
[4] School of Foreign Languages and Literature,Wuhan University, Wuhan, 430072, China
hongmiao23@163.com

1 Introduction

Performance of Information Retrieval (IR) can be improved by query expansion (QE) [1-2]. Current research of Chinese QE focuses mainly on expanding single term, which assumes that the user query is a short and complete term. However, user query may be a complex query, i.e., a query expressed in natural language, such as "列举全球变暖的危害 (lie ju quan qiu bian nuan de wei hai, List the damages resulting from global warming)". Little QE work focus on complex query and the prevalent approaches go like this: segment the original Chinese query Q into the vector Qa composed of multiple terms and expand each term in Qa respectively. These approaches ignore some valuable information, such as term combination and term concurrence in the complex query. Take the query "列举全球变暖的危害" for example, this query can be segmented as {列举(lie ju, list), 全球(quan qiu, global), 变暖(bian nuan, warming), 危害(wei hai, damage)}. Intuitively , "全球变暖(quan qiu bian nuan, global warming)" expressed the users' intention better than "全球(quan qiu, global)" and "危害(wei hai, damage)"; moreover, the combined term "全球变暖(global warming)" has turned into a semantic unit with more special connotation than single term "全球(quan qiu, global)" or "变暖(bian nuan, warming)".

In this paper, we build key term concurrence network (KTCN) based on large-scale corpus to expand complex query. Key step of our method is to acquire terms highly related to complex query. In this process, our method considers the information of term combination and concurrence of terms. The task aims at measuring the relevance between a given term and a given short text. For the sake of clarity, section 2 describes the process of generating KTCN, describes the method of acquiring sub-network (community) of KTCN related to a complex query, and measures the

* Corresponding author.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 251–258, 2013.

relevance between a term and the complex query. Section 3 takes the QE task as testing method to test our method of measuring the relevance between a short text (i.e. the complex query expressed in natural language and a term (i.e., each expanded term), and analyzes the experimental results in the end.

```
Input: large-scale document set corpus
Output : Key term network KTCN
 1) KTCN=empty graph ∅ ;
 2) for each document D in corpus
     acquire D's key terms list LD
     for each paragraph P in D
         for every pair of terms (K1, K2) that K1, K2∈ LD
            if K1 and K2 co-occurs in P
                if edge (K1, K2) does not exist in KTCN
                    create a new edge (K1, K2), and
  weight this edge1;
                else
                   increase weight of edge (K1, K2) by 1;
 3) MaxWeight = maximum of all edge weight in KTCN;
 4) for each edge e in KTCN
         e's new weight=MaxWeight-e's original weight
  +1;  //Note 1
```

Fig. 1. The process of constructing KTCN

2 Query Expansion Based on the Concepts of Shortest Path Length and Relatively Important Community

Our method involves two conceptions in graph and complex network theory. The first conception is the Weighted Shortest Path Length (WSPL), which is defined as the length of a shortest path in a weighted network; another conception, community, is the concept in complex network theory which will be explained soon. The main idea is as follows: 1) build KTCN; 2) acquire the terms with lowest WSPLs to each term in the complex query; 3) acquire the community of the terms which is relatively important to the query; 4) employ the clustering coefficient as parameter to measure the relevance between expanded term and the complex query; 5) acquire the expanded terms and weight them.

2.1 Network Construction

The construction process of KTCN is shown in Fig.1, L, the list of key terms of the corpus is also acquired. After the operation of Note 1 in Fig.1, the smaller the value of an edge weight, the more times the two terms corresponding to the two nodes of the edge in KTCN co-occurred in the Corpus.

2.2 Discussion of How to Measure Relevance between Terms in KTCN

Some research has been carried out about how to measure relevance between two terms. The methods include mainly 3 categories, for more detailed information, see reference [3]. However, method based on real corpus is beyond references we collected so far. Although not clearly proposed as an issue in graph theory, the issue of measuring relevance between two nodes in a network deserves more research. Specifically in KTCN, the issue of measuring relevance between two terms is corresponded to the issue of measuring intimacy between a pair of nodes in KTCN. Furthermore, we do not aim at measuring relevance between two terms, but measuring relevance between a term and a short text, and this problem is transformed as measuring relevance between a term and a term set, e.g., Qa. Shortest path length may be a good parameter to measure the relevance between terms: this theory assumes that the smaller the length of the shortest path between two nodes, the more semantically related between the two involved terms. However, this idea has some limitation because of the fact of data sparseness of corpus, which is a common phenomenon in corpus. In KTCN, data sparseness is exhibited as scale-free distribution of edge weight. High co-occurrence frequency may result from the fact that too many documents are related to some specific topic in the corpus instead of real high semantic relevance between terms. As a result, while computing WSLPs, some lower WSPL may result from the very low weight of some edges. The concept of community in complex network theory can help overcome the data sparseness problem to some extent. Communities are subgraphs of a graph; nodes in each community have higher probabilities to be connected than those between communities. In references community is a concept with no absolute definition, but it is commonly approved that community has noticeable property of small world. A node's coefficient is defined as the probability of its neighbor to be linked. A network's clustering coefficient is defined as the average of the network's all nodes' coefficients. A network's average path length l is defined as the average of short path length (note that this definition is not based on non-weighted network, compared with WSPL) between all pairs of nodes. In this paper, clustering coefficient is the only parameter used to express the small-world property of a community for the sake of application of community.

We use the concept of community to acquire terms closely related to Qa and use the terms for QE. After acquiring Qae, whose elements are terms which have lowest WSPLs to the source term (each term in Qa), we use the concept of community to improve the performance of QE. The idea goes like this: obtain sub-network composed of node in Qae, i.e., the community; and, in this community, we use each node's coefficient as the basis to weight the term corresponding to the nodes. This idea has some evidence in linguistics. Reference [6] discusses the method of computing lexical cohesion in single text based on thesaurus. Inspired by the idea and concepts in [6], e.g., lexical chain, we speculate that terms in multi-documents highly-related to a specific topic in a large corpus tends to cluster together. This is a cross-document lexical clustering effect which corresponds to the concept of community. Namely, corresponding terms may constitute a community with high clustering coefficient in KTCN. Most references about community detection focus on how to partition a network into multiple communities, but the relatively important community (RIC) is also an important concept which means the community in which the nodes are most related to a given specific node. In this paper, RIC means a community in which nodes are important to a small group of nodes, i.e., Qa, instead of single nodes.

```
Input: Qae, which is set of expanded term;
output: the vector of relevance between expanded terms
 and original query, with the elements form as <word,
 clustering coefficient>, the elements are sorted by
 clustering coefficient non-descending;
1) Graph Gsub=empty graph ∅;
2) for all the terms in Qae, add them to Gsub as
 Gsub's nodes;
3) for each pair of nodes (N1, N2) in Gsub
      if (N1, N2) ∈ the edge set of KTCN
          create a new edge (N1,N2) in Gsub;
   // Note: Gsub is the final community acquired.
4) initiate a set Qae=empty;
5) for each node Nsub in Gsub
      calculate the clustering coefficient CNsub;
      new an element Eqaes=<term corresponding to
 Nsub,CNsub>
      add Eqaes to Qaes;
6) sort the elements by CNsub in Qase non-descending;
7) eliminate the elements whose clustering coefficient
 dose not exist in the range of p%×|Qase| in Qase.
 //Note: p is assigned 30 in our experiment
```

Fig. 2. Strategy for detecting RIC to a complex query and ranking expanded terms

2.3 Acquiring Expansion Candidate Corpus of Original Query

For a complex query Q expressed in natural language, we dissolve it as follows: for each word W in L, if W is included in word string L, add W to Qa. So, Qa is the initial query analyzing result. For example, query question Q= "列举全球气候变暖的危害" can be dissolved into Qa={变暖，列举，气候，气候变暖(qi hou bian nuan, climate warming)，全球，全球气候(quan qiu qi hou, global climate)，全球气候变暖(quan qiu qi hou bian nuan, global climate warming)，危害}. We do not use the method of Chinese word segmentation which depends on large-scale word list because it ignores some valuable information such as word combination. For example, "全球气候" and "全球气候变暖"express more specific connotation than single terms "全球" and "变暖".

2.4 Strategies for Detecting Relatively Important Community and Ranking Expanded Terms

After acquiring Qa, we apply Dijkstra algorithm to extract K (=200) words which have the lowest WSPLs to each Tqa (term in Qa) and add these terms into Qae which

```
Input: Qaes (vector of terms in which terms are ordered
 by clustering coefficient non-descending)
Output: Qaesw (vector of expanded terms and their
 weights, with the form of <term, the term's weight>)
for each element Eqaes in Qaes
{
  Lqaes=length of Qaes;
  for(i=1;i<=L;i++)
  {
      Term=qaes [i];
      TermWeight=(L-i+1)/L;
     qaesw[i].Term=Term;
      qaesw[i].Score= TermWeight;
  }
}
```

Fig. 3. Strategy for measuring relevance between document and query

is the set of expanded terms. Note that Tqa∈Qae with the WSPL 0 to Tqa itself. Fig.2 describes the procedure in which the community is established by the nodes that are corresponding to elements in Qae.

2.5 Retrieval Process and Strategy for Weighting the Expanded Terms

This paper is not intended for query expansion but for the method of measuring the relevance between a term and a short text. Since the goal of QE is to test the validity of our measuring method which is based on complex network. So we do not use the common method relevant in most Chinese retrieval system: create inverted indexing for unigram (single Chinese character) or bigram. Instead, key terms of the documents are indexed. The standard of indexing is simple: if a term T is a key term of a document D, create an indexing item <T, D> in inverted indexing file FInvIdx for this "Term-Document" pair. The strategy for measuring the relevance between an expanded term and a Qa is shown in Fig. 3. The retrieval process, with the scoring strategy for measuring relevance between original query and a document is shown in Fig. 4.

3 Experimental Results

IR4QA (Information Retrieval for Question Answering) task in NTCIR-7 [7-9]is used to test the performance of our methods. KTCN constructed on Chinese corpus for IR4QA have node number of 713218 and edge number of 19042384, which establish a large-scale network. Average Precision (AP), Q-measure (Q)[10] and nDCG[11] are used to test the performance of system for IR4QA task.

```
Input: Qaesw; inverted index file FInvIdx•the factor
 beta used to ensure the priority of terms in original
 query. //after large amount of experiments, the
 experiment results reach the best when beta=11.
Output: retrieval result vector Vdoc, with the elements
 form as < DocID, relevance between document and q
 DocScore>
1) vector of documents Vdoc=empty;
2) for each element Eqaesw=<Term,TermWeight> in Qaesw
       find indexing item <Eqaesw.Term, DocID> in FIn-
 vIdx
       if (found)
          find the element Evdoc with the document
 ID=DocID in Vdoc;
     else
        add element<DocID, 0> to Vdoc;
       if(Term∈ QA)
          Vdoc [DocID].Score= Vdoc [DocID].DocScore
 +beta *Eqaesw.TermScore;
     else
       Vdoc [DocID] .Score = Vdoc [DocID]. DocScore +
Eqaesw.TermScore;
3) sort the elements in Vdoc non-descending ;
4) for each element EVdoc in Vdoc
       EVdoc.DocScore= EVdoc.DocScore / EVdoc
 [1].DocScore;  //normalization
```

Fig. 4. Retrieval Process and Strategy for scoring the relevance between the Documents and the original user query

Table 1 lists the experimental results and related comparison. Conclusions can be drawn as follows (only AP is concerned here): Compared to A, C's AP is increased by 470.6%; which shows that our strategy of weighting the expanded terms is effective; Compared to B, C's AP increased by 150.7%, which shows that the expanded words are of high quality. Case C, the best score we obtained, is close to the average score of systems submitted to NTCIR-7 IR4QA but still have a long distance from F (the best performance of CS-CS in NTCIR-7 IR4QA).

However, because our result is not submitted to NTCIR-7 for pooling process, our system should perform better than case C shown in Table 1. Unfortunately, we do not know how much we can be better actually; see reference [9] for the detailed reasons for this. Furthermore, after observing the result of each topic with respect to each QA, it is found that initial analysis of query is poor for those topics that we get low AP. For example, for the topic "谁是本拉登？(shui shi ben la deng, Who is Bin Laden?)", the vector of initial analysis is empty, so AP score for this topic is 0; for the topic "列举中俄之间发生的事情(lie ju zhong e zhi jian fa sheng de shi qing, List the events happened between China and Russia)", Qa={列举（list），俄之间(e zhi jian, a meaningless string composed of Chinese Characters here)，发生(fa sheng, happen),

发生的事情(fa sheng de shi qing, events happened), 事情(shi qing, event), 之间(zhi jian, between)}, elements in Qa do not include valuable term expressing intention of original QA, and AP of this topic is 0.004. The above cases demonstrate that the bottleneck of our system is the initial analysis for original query. This fact brings a bad message and a good message: the bad one is that our IR performance depends heavily on topic analysis; the good one is: the performance for detecting expanded terms and the strategy for weighting the expanded terms perform well, which, conforms to the aim of this paper.

Table 1. Experimental result and comparison with related cases

Case	AP	Q	nDCG	Case Notation
A	0.0830	0.09636	0.1973	adding expanded terms to Qa, but not weight the terms.
B	0.1889	0.1968	0.3168	adding no expanded terms
C	0.4736	0.4819	0.7654	the best results we have acquired so far.
D	0.6337	0.6490	0.8270	the best performance of CS-CS run of NTCIR-7 IR4QA
E	0.4121	0.4156	0.60128	average performance submitted by all the CS-CS groups in NTCIR-7 IR4QA

Acknowledgement. This paper is supported by Natural Science Foundation Project (61070243, 61133012, 61070082, 61173062, 61202193), Major Project of Invitation for Bid of National Social Science Foundation (11&ZD189), Guizhou High-level Talent Research Project (TZJF-2010-048), Guizhou Normal University PhD Start-up Research Project (11904-05032110011), and Governor Special Fund Grant of Guizhou Province for Prominent Science and Technology Talents (identification serial number "黔省专合字(2012)155号").

References

1. Van, R.C.: A new theoretical framework for information retrieval. In: Proceedings of 1986 ACM Conference on Research and Development in Information Retrieval, pp. 194–200. ACM, New York (1986)
2. Baeza-yates, R., Ribeiro-neto, B.: Modern information retrieval. Addison-Wesley Harlow, England (1999)
3. Mohammad, S., Hirst, G.: Distributional measures as proxies for semantic relatedness (2005), http://www.cs.toronto.edu/compling/Publications
4. Watts, D.J., Strogatz, S.H.: Collective dynamics of'small-world' networks. Nature (London) 393(6684), 440–442 (1998)
5. Sole, R.V., Murtra, B.C., Valverde, S., et al.: Language Networks: their structure, function and evolution. Trends in Cognitive Sciences (2006)
6. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics 17(1), 21–48 (1991)
7. Mitamura, T., Nyberg, E., Shima, H., et al.: Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access. In: Proceedings of the Seventh NTCIR Workshop Meeting, Tokyo, Japan (2008)

8. Tetsuya Sakai, E.: Overview of the NTCIR-7 ACLIA IR4QA Subtask (2008)
9. Sakai, T., Kando, N., Lin, C., et al.: Overview of the NTCIR-7 ACLIA IR4QA Task. In: Proceedings of the Seventh NTCIR Workshop Meeting, Tokyo, Japan (2008)
10. Sakai, T.: Evaluating information retrieval metrics based on bootstrap hypothesis tests. IPSJ Digital Courier 3(0), 625–642 (2007)
11. Kek, J.K.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 20(4), 422–446 (2002)

Rapid Increase of the Weighted Shortest Path Length in Key Term Concurrence Network and Its Origin

Lan Yin[2,3], Hua Yang[1,2,*], Donghong Ji[2,3], Mingyao Zhang[1,4], and Hongmiao Wu[4]

[1] College of Chinese Language and Literature, Wuhan University, Wuhan, 430072, China
yanghuastory@263.net, myzhang@whu.edu.cn
[2] School of Mathematics and Computer Science, Guizhou Normal University, Guiyang, 550001, China
yindew@gmail.com
[3] Computer School, Wuhan University, Wuhan, 430072, China
donghongji_2000@yahoo.com
[4] School of Foreign Languages and Literature, Wuhan University, Wuhan, 430072, China
hongmiao23@163.com

Abstract. In previous work, we constructed a Key Term Concurrence Network (KTCN) based on large-scale corpus with an attempt to apply weighted shortest path length to measure semantic relevance between terms. The parameter was tentatively used for query expansion in Information Retrieval task directed to complex user query expressed in natural language. The data obtained from the experiment demonstrated improved performance in the task. However, we also found that as more new expanded terms are appended to the vector of original query, the performance decreases drastically after reaching a peak. This paper respectively explains the causes of this phenomenon from two perspectives: the property of complex network property and corpus linguistics. Based on this conclusion, future work is directed towards how to improve our work.

Keywords: semantic relevance, complex network, weighted shortest path length, scale-free distribution.

1 Introduction

Complex network theories have been initially applied in Natural language Processing (NLP) tasks. Erkan used centrality of sentence feature vector to measure the salience of the sentence in a text and then used it in text summarization [1]; Antiqueira et al. represent texts by complex networks, and used network properties positively related to text quality to assess text quality automatically [2]. Complex network techniques also used for evaluation of summary quality [3]. And other typical work using complex network to accomplish NLP tasks mainly include: automatic summarization [4], information retrieval and related tasks[5-7], analysis of text sentiment [8]. As for Chinese language, Reference [9] clusters words by using word concurrence network, uses the word clusters for text segmentation, and analyzes the topics of Chinese text.

* Corresponding author.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 259–267, 2013.

In our previous work, we constructed a key term concurrence network (KTCN) based on large-scale corpus. Weighted shortest path length (WSPL) obtained from KTCN was employed to expand complex query and achieved good performance. However, when the number of expanded terms was greater than 25, the performance of our information retrieval system declined rapidly. For 642 nodes corresponding to terms in the original query (source node), we obtained 200 nodes which had the shortest WSPLs to each source node, and observed the WSPLs to each source node. For each source node, we found that the WSPLs increased rapidly from the minimum WSPL to the maximum WSPL. This implies that, the limitation of using WSPL for measuring semantic relevance is exhibited noticeably with the increase of WSPLs. This paper aims at analyzing the origin from the perspective of network properties and corpus linguistics respectively, and points out the limitations of using WSPL for measuring the semantic relevance between terms, as well as the direction to improve our previous work.

This paper is organized as follows: Section 2 describes the method of generating KTCN and discusses the storage of large-scale term network and application of algorithms based on KTCN. Section 3 presents the phenomenon of rapid increase of WSPL in KTCN. Section 4 explains the origin of the rapid increase from the perspectives of complex network properties and corpus linguistic respectively. Conclusions are drawn in section 5 and the direction for improving our previous work is implied in section 6.

2 Construction of KTCN and Calculation of WSPLs

2.1 Construction of KTCN

In most references, the term complex networks are constructed with single word as node, concurrence in sentences or syntax relations as edge without weight [10-13]. In previous work, we constructed key term co-occurrence network (KTCN) as a network in which edges are assigned weights: nodes are key terms extracted from corpus obtained with the algorithm mentioned in reference [14]; edges represent the co-occurrence relation in paragraphs of documents; and each edge is weighted the times that the two involved terms satisfy our definition of edge. The key term of a document is not limited to word from strictly linguistic perspective and may include combination of words. Take “武汉大学(wu han da xue, Wuhan University)” as an example, if it is meant to describe the entity of “Wuhan University”, it will not be dissolved into “武汉(wu han, Wuhan)” and “大学(da xue, University)”. Algorithm in Ref [14] is employed to obtain the key terms of the documents in the corpus. L, the list of all key terms of all documents, is also obtained after key terms are obtained for each document. The fundamental goal for constructing KTCN, is to reflect semantic relevance between Chinese terms instead of syntactic relation or other relations. For example, relevance between “政府(zheng fu, government)”and “ling dao, 领导(lead)”, relevance between “冠军(guan jun, Champaign)” and “金牌(jin pai, gold medal)” can be intuitively sensed from common sense but can not be described precisely. Furthermore, the relevance between “三峡(san xia, Three Gorges)”与 “移民(yi min, migration)” comes from realistic events.

Fig. 1 shows the construction process of KTCN. The operation of note 1 is to reverse the edge weight. After the operation, the lower the edge weight, the more times the two involved terms co-occur in the corpus, the higher probability that the two involved terms are semantically related. Large-scale Chinese corpus including 545162 documents used for IR4QA[16] in NTCIR-7[15] is used to construct KTCN.

```
Input: large-scale document set corpus
Output: Key term network KTCN
1) KTCN=empty graph ∅;
2) for each document D in corpus
      acquire D's key terms list LD
      for each paragraph P in D
          for every pair of terms (K1, K2) that
K1∈LD and K2∈LD
              if K1 and K2 co-occur in P
                 if edge (K1, K2) does not exist in
KTCN
                     create a new edge (K1, K2), and
weight this edge1;
                 else
                      increase weight of edge (K1, K2)
by 1;
3) MaxWeight = maximum of all edge weight in KTCN;
4) for each edge e in KTCN
  e's new weight= MaxWeight - e's original weight
  +1;  //Note 1
```

Fig. 1. The construction of KTCN

2.2 Discussion: Storage of Large Scale Network and Computation of Weighted Shortest Path Length

Triplet (N, E, W) is used to describe KTCN's scale where N, E, W are defined as KTCN's node number, edge number and sum of all edge's weights respectively. For KTCN obtained in our experiment, the scale is (713218, 19042384, 71188915), with the maximum edge weight MaxWeight=10337. Compared with other term networks generated in references, KTCN has very large scale. Furthermore, KTCN is so large and sparse that some traditional algorithms with polynomial complexity are difficult to be actually implemented. So, storage structure and actual applicability of graph algorithm deserve further discussion. As far as construction of specific data structure and implementation of specific algorithm are concerned, the following two points deserve attention.

(1) Storage structure: for large-scale and sparse unidirectional network, the useful techniques include: compressing storage of symmetrical matrix; compressing storage of sparse matrix; Hash techniques (for locating element with specified row number and column number in the matrix with low complexity). All of the above techniques do not lose network information. If KTCN's scale is too large to be stored with the above-mentioned techniques, edges with weight higher than specified threshold can be eliminated.

(2) Calculation of WSPL: the classic algorithms for computing WSPL include Floyd algorithm and Dijkstra algorithm, with the complexity of $O(n^3)$. These two algorithms have similar complexity. However, as far as method of obtaining a few nodes with lowest WSPLs to a given node, Floyd algorithm is difficult to implement while Dijkstra algorithm is still applicable [17] on large networks.

3 Phenomenon of Rapid Increase of the Shortest Path Length

Take "奥运会(ao yun hui, the Olympic Games)"as the source node, 200 terms with the lowest WSPLs to the source node, and the corresponding WSPLs are obtained by Dijkstra algorithm, as shown in Fig. 2. Phenomenon of Rapid Increase of the Shortest Path Length.

奥运会(ao yun hui, Olympic games), 0; 奥运(ao yun, Olympic), 6899; 金牌(jin pai, golden medal), 7945; 比赛(bisai, contest), 8024; 选手(xuan shou, Contestant) , 8025; 北京(bei jing, Beijing), 8260; 悉尼(xi ni, Sydney), 8668; 参加(can jia, take part in), 8732; 冠军(guan jun, championship), 8990; 运动员(yun dong yuan, Athlete), 8990; 世界(shi jie, world), 9151; 悉尼奥运会(xi ni ao yun hui, Sydney Olympic games), 9155; 项目(xiang mu, project), 9182; 申办(shen ban, Bid for), 9258; 国际奥委会(guo ji ao wei hui, IOC), 9259; 获得(huo de, achieve), 9303; 体育(ti yu, physical training), 9313; 成绩(cheng ji, score), 9386; 运动 (yun dong, sports), 9487; …… 主办(zhu ban, sponsor), 10056; 举行(ju xing, hold), 10058; 委员(wei yuan, committee), 10061; 印尼(yin ni, Indonesia), 10065; 主席(zhu xi, Chairman), 10081; 参赛(can sai, attend competition), 10081; 我们 (wo men, we), 10087; 队员(dui yuan, team member), 10090; 代表团(dai biao tuan, delegation), 10090; 萨马兰奇(sa ma lan qi, Samaranch), 10090; 米自由泳(mi zi you yong, here is a Chinses Character string composed of "meter" and "free-style",which is not a term with sensible meaning), 10262; 女子举重(nv zi ju zhong, weight lifting for female), 10264; 北京奥运会(bei jing ao yun hui, Beijing Olympic Games), 10264 罗马尼亚(luo ma ni ya, Romania), 10265; 射箭(she jian, shoot-ing), 10265; 国际奥委会委员(guo ji ao wei hui wei yuan, IOC Member), 10266; 昨天(yesterday) , 10267; 协会(association) , 10267; 波兰(Poland) , 10267; 世界锦标赛(world championship), 10267; 雅典奥运会(Athens Olympic Games) , 10267; 国家体育总局(General Administration of Sport of China) , 10268; 北京申奥(Beijing Olympic Bid) , 10268; 历史(history) , 10269; 排球(volleyball) , 10269; 打破(break/ set a new record), 10269; 大阪(Osaka) 10269

Fig. 2. Example: 200 terms with lowest WSPLs to the source node "奥运会(the Olympic Games)" and the corresponding WSPLs

Fig.2 is an example that "奥运会(ao yun hui, the Olympic Games)"as a source node, with WSPLs to the source node in KTCN calculated by Dijkstra algorithm.

As mentioned above, the maximum edge weight in KTCN is 10337. From the above example, with the increase of the WSPLs, the values of WSPLs approximate to 10337 rapidly. Obviously, in the above example, terms following "大阪(da fan, Osaka)" possess higher WSPL values than "大阪", which has WSPL of 10269 to the

source node. The trend of the curve is shown in Fig. 3. If the curve is influenced by the source node, for length of path from "Olympic game" to itself is defined as 0. The curve is shown again in Fig. 4, where the source node is eliminated. the phenomenon of rapid increase of WSPLs still exists.

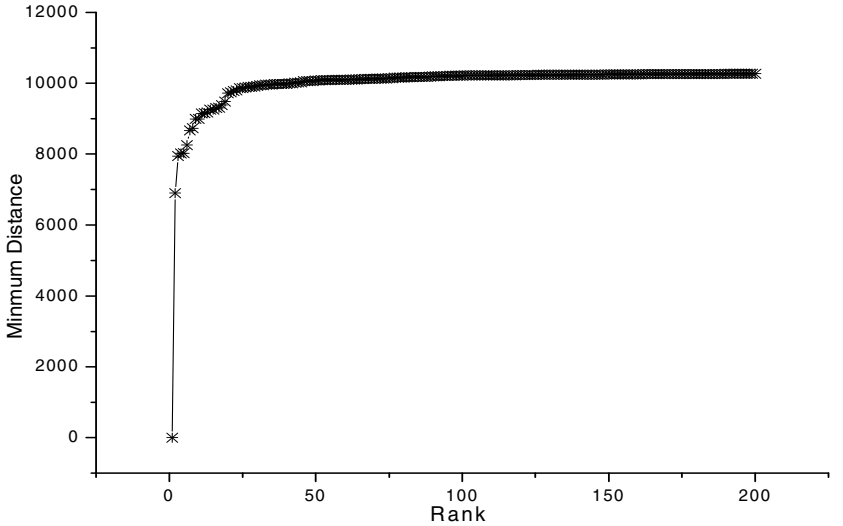


Fig. 3. WSPL distribution of 200 terms with the lowest WSPLs to the source node "奥运会(ao yun hui, the Olympic Games)"

Therefore, Question is raised about our previous work: if the above regularity is universal, what is the origin? Another question is concerning whether the application of WSPL as measurement of semantic relevance and using it in query expansion are of high limitation? What method can be taken to overcome the defect?

4 Origin of FISPL

In the original KTCN (the KTCN without the operation reversing edge weight), which will be called KTCN_O, we investigated the node num distributions in the range of [6, 29], [30, 246], [247, 573], [574, 820] respectively, as shown in Fig.5, Fig.6, Fig.7, Fig.8. For the remainder data, the edge in the area of [821,10337] contains 1063 types of edge weight value, and 642 of them are weighted as 1, all the other weight values are below 10, the lowest one is 1, the highest is 8, the average is 1.76199, the standard deviation is 1.19981, the median is 1. It is obvious that edge weights of KTCN obey scale-free distribution described in [18]. The concept of scale-free network refers to the distribution of node degree in network in most references. Investigation of distribution of edge weight is beyond references we have collected. Therefore, during the process of using Dijistra algorithm to obtain the node with the lowest WSPL, the second lowest WSPL, etc., if a WSPL passes an edge of very low weight, the other

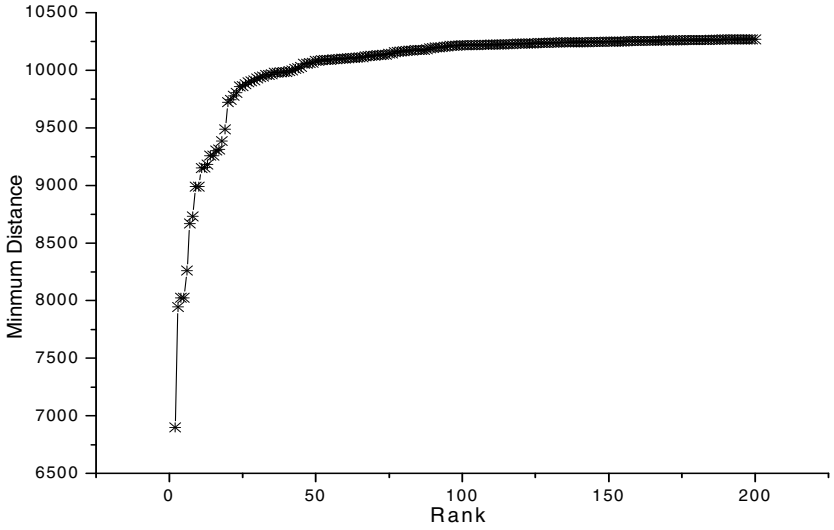


Fig. 4. WSPL distribution of 199 terms with the lowest WSPLs to the source node "奥运会(ao yun hui, the Olympic Games)"

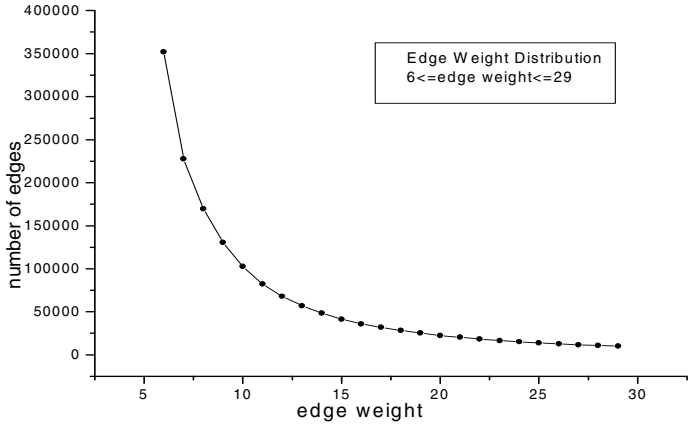


Fig. 5. Distribution of edge weights in the range of [6, 29]

WSPLs will pass this edge with high probability. So, from the perspective of network property, it can be inferred that FISPL originated from the scale-free distribution of edge weight of KTCN-O.

As mentioned above, FISPL is the origin of the scale-free distribution of edge weight of KTCN-O, but what is the origin of the scale-free distribution? In fact, this origin is well-known data sparseness problem in corpus linguistics: high edge weight in KTCN-O may result from high co-occurrence frequency in corpus instead of real high semantic relevance between terms. FISPL is similar to Zipf law which describes the distribution of term frequency while FISPL describes the distribution of co-occurrence frequency of key terms in documents.

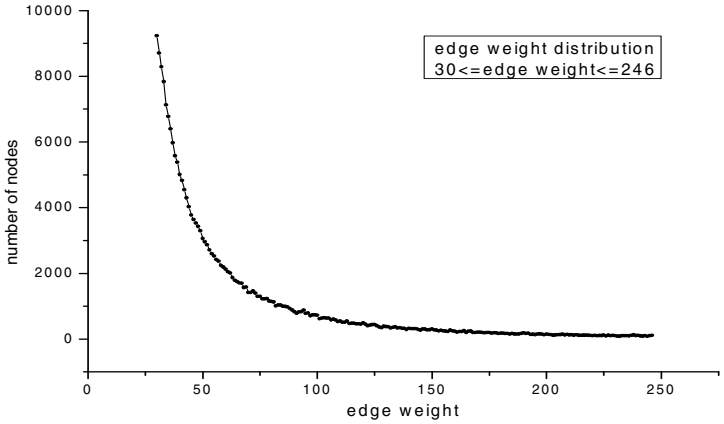


Fig. 6. Distribution of edge weights in the range of [30, 246]

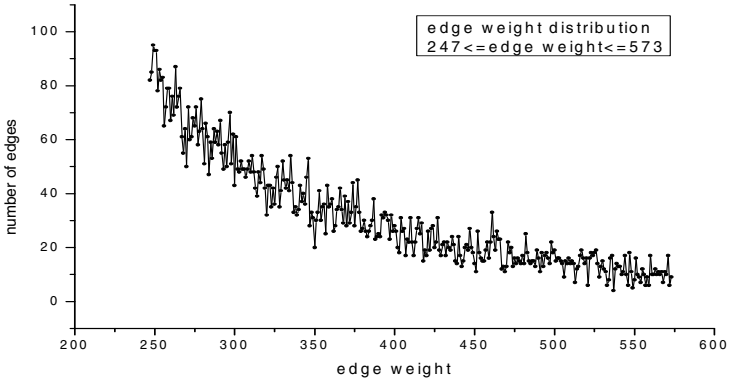


Fig. 7. Distribution of edge weights in the range of [247, 573]

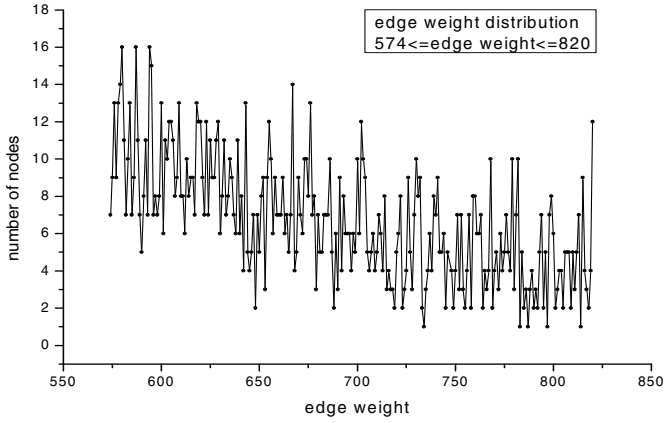


Fig. 8. Distribution of edge weights in the range of [574, 820]

5 Conclusion

It is universally acknowledged that the sparse data problem of corpus is a commonplace practice in NLP tasks, and it is difficult to avoid this problem in computation method. That's why many researchers in NLP field prefer method based on rules or manual resources to method based on corpus. Moreover, the data sparse problem can not be resolved by simply increasing the scale of corpus. Therefore, it is of limitation of our previous work which takes WSPL to measure semantic relevance between terms: negative influence of data sparse cannot be avoided while computing the WSPLs, and then the obtained result of semantic relevance is partial.

6 Further Work

Though the method we employed to measure the semantic relevance between terms using WSPL is of limitation, its advantages are still noticeable: 1) the rich information of global text is considered; 2) the process of obtaining the new node with Dijkstra algorithm considers nodes and corresponding WSPLs already obtained, which, conforms to the psychological process of human association; 3) the semantic relevance can be measured even if the involved nodes are not linked directly in KTCN.

In order to overcome the disadvantages caused by the data sparseness to some extent while maintaining the advantage of our previous idea, there are at least two directions: 1) Semantic relevance between terms is a conception of high subjectivity, the intuition of different people with different background may be different, and the measuring method should conform to context in which the terms are used. 2) More network properties, such as node clustering coefficient, may be a better parameter for measuring relevance because it is less influenced by data sparseness.

Acknowledgement. This paper is supported by Natural Science Foundation Project (61070243, 61133012, 61070082, 61173062, 61202193), Major Project of Invitation for Bid of National Social Science Foundation (11&ZD189), Guizhou High-level Talent Research Project (TZJF-2010-048), Guizhou Normal University PhD Start-up Research Project (11904-05032110011), and Governor Special Fund Grant of Guizhou Province for Prominent Science and Technology Talents (identification serial number "黔省专合字(2012)155号").

References

1. Erkan, G., Radev, D.R.: LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 22, 457–479 (2004)
2. Antiqueira, L., Nunes, M.G., Oliveira, J.O., et al.: Strong correlations between text quality and complex networks features. Physica A: Statistical Mechanics and its Applications 373, 811–820 (2007)
3. Pardo, T.A., Antiqueira, L., Nunes, M.G., et al.: Using complex networks for language processing: The case of summary evaluation. In: Proceedings of the International Conference on Communications, Circuits and Systems (ICCCAS 2006) Special Session on Complex Networks, pp. 2678–2682 (2006)

4. Mihalcea, R.: Language independent extractive summarization, pp. 49–52. Association for Computational Linguistics, Morristown (2005)
5. Page, L., Brin, S., Motwani, R., et al.: The pagerank citation ranking: Bringing order to the web. Technical Report. Stanford InfoLab (1998)
6. Kurland, O., Lee, L.: PageRank without hyperlinks: structural re-ranking using links induced by language models. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 306–313 (2005)
7. Otterbacher, J., Erkan, G., Radev, D.R.: Using random walks for question-focused sentence retrieval, pp. 915–922. Association for Computational Linguistics, Morristown (2005)
8. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, pp. 271–278 (2004)
9. Shi, J., Hu, M., Dai, G.Z.: Topic Analysis of Chinese Text Based on Small World Model. Journal of Chinese Information Processing 21(003), 69–75 (2007)
10. Dorogovtsev, S.N., Mendes, J.F.: Language as an Evolving Word Web. Proceedings: Biological Sciences 268(1485), 2603–2606 (2001)
11. Ferrer, I., Cancho, R., Sole, R.V.: The small world of human language. Proceedings of the Royal Society B: Biological Sciences 268(1482), 2261–2265 (2001)
12. Heyer, G., Quasthoff, U., Wittig, T.: Text Mining: Wissensrohstoff Text Konzepte, Algorithmen, Ergebnisse. W3L-Verl. (2006)
13. Ferrer i Cancho, R., Solé, R.V., Köhler, R.: Patterns in syntactic dependency networks. Physical Review E Phys. Rev. E 69, 051915 (2004)
14. Yang, L.P., Ji, D.H., Li, T.: Chinese information retrieval based on terms and ontology (2004)
15. Kando, N.: Overview of the Seventh NTCIR Workshop. In: Proceedings of NTCIR-7 Workshop Meeting, Tokyo, Japan (2008)
16. Sakai, T., Kando, N., Lin, C., et al.: Overview of the NTCIR-7 ACLIA IR4QA Task. In: Proceedings of the Seventh NTCIR Workshop Meeting, Tokyo, Japan (2008)
17. Yang, H.: The application of complex network in natural language processing. Wuhan University, Wuhan (2009)
18. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509 (1999)

VO Verbal Compounds and the Realization of Their Objects

Huibin Zhuang[1], Zhenqian Liu[2,*], and Yuan Zhang[3]

[1] Henan University, Kaifeng, China
[2] Shandong University, Jinan, China
School of Foreign Languages and Literature
[3] Shandong Normal University, Jinan, China
{huibinzhuang,zhenqianliu,cassie0848}@yahoo.com.cn

Abstract. The objects of VO verbal compounds are realized in a special way. Owing to the lack of Case, they cannot appear in situ, but have to combine with a preposition (so as to get an oblique Case) or appear in the attributives of Os. VO verbal compounds should be distinguished from VO-verbs, which can assign Cases directly to their objects. Besides, affected by prosody, some adverbials of VO-phrases will leave out the prepositions, appearing like objects, and forming another kind of pseudo-VO verbal compounds.

Keywords: VO verbal compound, object realization, Case-assignment, prosody.

1 VO Verbal Compound Constraint on Object

In the current study, VO verbal compounds refer to the complex verbal constructions that consist of verbs and object-like NPs, with both of them contributing to the meanings of the complex verbal constructions. VO verbal compounds are commonly treated as idioms. According to Huang [1] and Her [2], they are interesting in several aspects, of which two are most prominent:

First, a VO verbal compound seems to consist of a V(erb) and an O(bject). It is interesting to note that this O is not a real object, but part of the verb, as shown in (1):

(1) a. 泼 冷 水
po leng shui
pour cold water
'throw cold water on'

b. 吃 醋
chi cu
eat vinegar
'become jealous'

c. 拍 马 屁
pai ma pi
pat horse rear
'flatter'

d. 炒 鱿鱼
chao youyu
stir-fry sleeve-fish
'fire (sb)'

e. 打 主意
da zhuyi
hit idea
'try to get (sth/sb)'

f. 占 便宜
zhan pianyi
possess cheap
'take advantage of'

* All correspondence please address to: Zhenqian Liu.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 268–279, 2013.

g. 露 马 脚
lou ma jiao
show horse foot
'show the cloven hoof '

h. 开 玩笑
kai wanxiao
open joke
'make a jest'

i. 做 工作
zuo gongzuo
make job
'persuade'

Second, an object semantically required by VO complex verb (if there is) cannot be placed after the VO complex verb. It either appears as the possessor of O, thus becoming a "possessive object" [3], or turns up in a PP (or *BA*-construction) before the verb[1], as shown in (2) and (3) respectively:

(2) a. 泼 张三 的 冷 水
po Zhangsan de leng shui
pour Zhangsan *de* cold water
Lit: 'pour Zhangsan's cold water'
'pour cold water on Zhangsan'

b. 吃 李四 的 醋
chi Lisi de cu
eat Lisi *de* vinegar
Lit: 'eat Lisi's vinegar'
'be jealous of Lisi'

c. 拍 王五 的 马 屁
pai Wangwu de ma pi
pat Wangwu *de* horse rear
Lit: 'pat Wangwu's horse-rear'
'lick Wangwu's boots'

d. 炒 赵六 的 鱿鱼
chao Zhaoliu de youyu
stir-fry Zhaoliu *de* sleeve-fish
Lit: 'stir-fry Zhaoliu's sleeve-fish'
'fire Zhaoliu'

(3) a. 向 张三 泼 冷 水
xiang Zhangsan po leng shui
to Zhangsan pour cold water
'pour cold water on Zhangsan'

b. 为 李四 吃 醋
wei Lisi chi cu
for Lisi eat vinegar
Lit: 'eat vinegar for Lisi'
'be jealous of Lisi'

c. 给[2] 王五 拍 马 屁
gei Wangwu pai ma pi
give Wangwu pat horse rear
Lit: 'pat the horse-rear for Wangwu'
'lick Wangwu's boots'

d. 把 赵六 炒 鱿鱼
ba Zhaoliu chao youyu
BA Zhaoliu fry sleeve-fish
Lit: 'fry Zhaoliu as a sleeve-fish'
'fire Zhaoliu'

This characteristic of VO verbal compounds is referred to as VO verbal compound Constraint on Object (hereafter referred to as VOCO). Early in the 1960s-1980s, VO verbal compounds have attracted much attention from scholars [3-6]. Later, many accounts from different perspectives have been put forward to explain this phenomenon, such as Generative Grammar [7-8], Prosodic Grammar [9-11], Cognitive Grammar [12], Construction Grammar [13], etc. Scholars have realized that this phenomenon is a syntax-semantics mismatch, but none of them could reveal the real reason. Until now, no satisfactory explanation has been proposed.

2 A Syntactic Explanation on VOCO

In fact, the VOCO can be explained by means of standard θ-theory and Case theory, to be more specific, θ-Criteria and Visibility Condition. According to Chomsky, the θ-Criterion

[1] Note that not all the objects of VO verbal compounds can be realized in both of the two forms.

[2] In Chinese, *gei* 'give' can be treated as a preposition.

states that "Each argument bears one and only one θ-role, and each θ-role is assigned to one and only one argument" [14]. Visibility Condition is defined in Chomsky as a condition which states that an element must be Case-marked in order to be visible for θ-marking (which in turn is required by the θ-criterion) [15]. In this part, two issues will be discussed: how the O in VO verbal compounds satisfies the θ-Criterion and the demand for Case, and how the objects of VO verbal compounds are realized.

2.1 Os in VO Verbal Compounds: Quasi-arguments

It is obvious that the Os in VO verbal compounds are important in conveying their idiomatic meanings, which cannot be derived (compositionally) from the combination of the literal meanings of their individual constituents. Now let us consider the examples below (Note that = indicates that the literal reading is available, # the idiomatic reading is available):

(4) a.张三 喜欢 吃 醋(=, #) b.李四 擅长 拍 马 屁 (=, #)
Zhangsan xihuan chi cu. Lisi shanchang pai ma pi.
Zhangsan like eat vinegar Lisi be good at pat horse rear
'Zhangsan got jealous easily.' 'Lisi is good at flattering others.'

Obviously, the idiomatic meanings 'be jealous of' and 'flatter' are not from the meanings of the individual words in (4a) and (4b). In these cases, obviously, *cu* 'vinegar' and *ma pi* 'horse-rear' do not refer to entities in the real world, i.e., "sour liquid made from malt, wine, cider, etc. by fermentation and used for flavoring food and for pickling" and "the back part of a horse." The fact that they are not referential expressions in the idiomatic reading implies that they are not arguments. However, the presence of *cu* 'vinegar' and *ma pi* 'horse-rear' is necessary for the meanings of the VO verbal compounds to be conveyed. Otherwise, they will become unacceptable or lose the idiomatic meanings, as shown in (5) and (6). In (5), *cu* 'vinegar' and *ma pi* 'horse-rear' are absent, and in (6) they are replaced by other expressions.

(5) a. 张三 喜欢 吃 (=) b.?李四 擅长 拍 (=) [3]
Zhangsan xihuan chi. *Lisi shanchang pai.*
Zhangsan like eat Lisi be good at pat
'Zhangsan likes eating.' 'Lisi is good at patting'

(6) a. 张三 喜欢 吃 酱(=) b.?李四 擅长 拍 牛 屁(=)[4]
Zhangsan xihuan chi jiang *Lisi shanchang pai niu pi*
Zhangsan like eat sauce Lisi good at pat niu rear
'Zhangsan likes sauce.' 'Lisi is good at patting on the rears of oxen.'

[3] Someone may argue that (5b) and (6b) might have the idiomatic readings in certain contexts. But here these sentences appear in isolation, it is hard to say that they have idiomatic readings.

[4] It seems that this reading does not make sense, even though it is literal. Imagine a group of children are competing who is better in patting the rear of an ox, and the result is that Lisi is the best.

Although *cu* 'vinegar' and *ma pi* 'horse-rear' in idiomatic reading are non-referential expressions, they function as special arguments of the verbs *chi* 'eat' and *pai* 'pat' in their idiomatic uses. In other words, the verbs *chi* 'eat' and *pai* 'pat' assign special θ-roles to *cu* 'vinegar' and *ma pi* 'horse-rear' respectively.

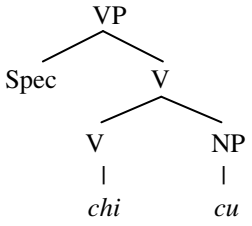
In the literature, Her treats VO verbal compounds as compounds and argue that they are lexical categories [2], as shown in (7). (8) shows the structure of the literal reading:

(7) *Zhangsan xihuan* [V *chi cu*]. (#)
(8) *Zhangsan xihuan* [VP *chi cu*]. (=)

This certainly helps to distinguish the status of VO verbal compounds. However, under this hypothesis, neither the ambiguity, nor the presence of (fake) attributives in (2) can be accounted for. Therefore, Her suggests that the Os in VO verbal compounds are referential, although in a metaphorical way. According to Ouhalla [16], this type of expression is called quasi-arguments. Quasi-arguments which are responsible for the idiomatic meanings are assigned special θ-roles.

Assuming this to be the case, *cu* 'vinegar' and *ma pi* 'horse-rear' must get Cases in order to become visible before they get their θ-roles. Therefore, it is reasonable to argue that the Os in VO verbal compounds occupy a place properly governed by the Vs, as shown below:

(9)



2.2 Objects of VO Verbal Compounds and Their Realization

From the above discussion, it is not difficult to tell that a VO verbal compound should be treated as a whole in which the O is not a real object but part of this verb; and that if this complex verb requires an object, the object cannot appear after the VO verbal compound but before the verb as the possessor of O or in a PP (or *BA*-construction). A natural question arises then as to why this object cannot be realized in situ. This question is not difficult to answer if θ-Criteria and Visibility Condition are taken into consideration. The VO verbal compound, as a verb, should assign a θ-role to its object; however, it cannot assign a Case to its object, because the only Case-marked position, as shown in (9), is occupied by the O of the VO verbal compound. Being not able to get a Case in situ, the object, in order to satisfy the demand for Case, has no choice but to move to other positions to get a Case.[5]

[5] Actually, "possessive object" can be found not only in Chinese, but also in other languages. For example, in English, we can find "pull my leg," "watch your head," "take advantage of me," etc.

3 Pseudo-VO Verbal Compounds

The above explanation seems to be very nice. However, problems appear again if more examples are taken into consideration. For instance, (10-13) show that the objects of the VO verbal constructions, in contrary to our expectations, are realized in situ:

(10) 着 手 此 事
zhuo shou ci shi
put hand this matter
'put one's hand to the plough'

(11) 得 罪 王五
de zui Wangwu
obtain blame Wangwu
'offend Wangwu'

(12) 关 心 李四[6]
guan xin Lisi
close heart Lisi
'care about Lisi'

(13) 负 责 此 事[7]
fu ze ci shi
bear duty this matter
'be responsible for this matter'

What is the possible reason? For several decades, scholars have carried out many studies, trying to figure out the reason(s) from different perspectives, such as history, grammar, prosody, lexicalization, semantics, valency, rhetoric, pragmatics, etc. Among them, the most successful one until today should be the V^0-movement account by Feng [11].

3.1 Feng's Account

Having adopted the Larson's VP-shell hypothesis, Feng [11] endeavors to explain the phenomenon that VO constructions take objects, illustrated by (14-15).

(14) 收 徒 山神庙
shou tu shanshenmiao
accept prentice temple of the mountain god
'take an apprentice in the temple of the mountain god '

(15) 讲 学 中南海
jiang xue Zhongnanhai
present lecture Zhongnanhai
'present a lecture/lectures at Zhongnanhai'

Feng assumes that prepositions in Chinese, such as *zai* 'at', *cong* 'from', and *wei* 'for', are light verbs, i.e., V (P), and that the verbal VOs below can be raised to be incorporated into them. Thus, the form of a VO construction taking an object is derived, as shown by (16).

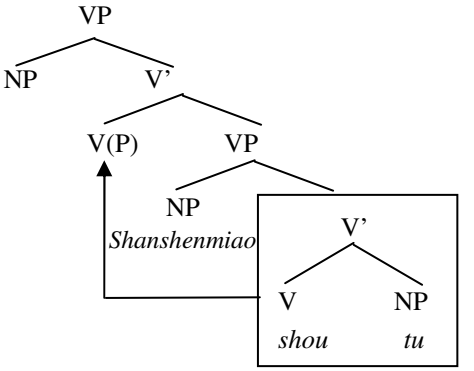
Through the interaction between prosody and grammar, Feng even distinguishes words and phrases, pointing out the possible movements and the impossible ones.

[6] For detailed discussion of the historical development of 关心*guan xin* 'care', see Cui [17].

[7] Someone may argue that 负责*fuze* 'be responsible for' can appear as a VO verbal compound, such as负此事的责 *fu ci shi de ze* 'be responsible for this matter'. This shows exactly that there is an overlap between VO verbal compounds and VO verbs. The same is true of nouns and verbs in Chinese.

For example, (17) is illegitimate, because only those two-syllable VO constructions can precede their objects, while those of more than two syllables cannot. This explains why (17) is not acceptable.

(16)



(from Feng [11])

(17) * 收 徒弟 山神庙
shou *tudi* *shanshenmiao*
accept prentice temple of the mountain god
'take an apprentice in the temple of the mountain god'

However, in this way, Feng makes the matter more complex:

First, the V0-movement proposed by Feng certainly can explain the phenomena illustrated by (10-15), but at the same time, it opens up opportunities for the objects of VO verbal compounds to be realized after VOs, predicting all the following examples are correct. (Note that all the readings given in (18) are intended readings.)

(18) a.*告 状 他/张三
gao *zhuang* *ta/Zhangsan*
file lawsuit him/Zhangsan
'sue him/Zhangsan'

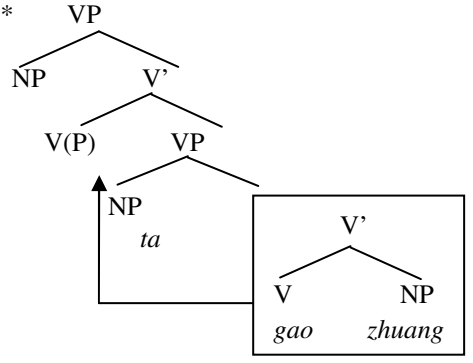
b.*将 军 他/张三
jiang *jun* *ta/Zhangsan*
command army him/Zhangsan
'outfox him/Zhangsan'

c. *扫 兴 他/张三
sao *xing* *ta/Zhangsan*
sweep excitement him/Zhangsan
'cast a chill over him/Zhangsan'

d. *吃 醋 他/张三
chi *cu* *ta/Zhangsan*
eat vinegar him/Zhangsan
'become jealous of him/Zhangsan'

Take *gao zhuang* 'sue' as an example. Its derivation can be shown as (19).

(19) *



The derivation, obviously, follows Feng strictly. However, the result is unacceptable. Why is it possible for *guan xin* 'care', *fu ze* 'be responsible', *de zui* 'offend', *qu xiao* 'laugh at', etc. to be raised as a whole, but not possible for *gao zhuang* 'sue', *chi cu* 'be jealous', *jiang jun* 'outfox', etc.? Someone may argue that words such as *guan xin* 'care' are more lexicalized than words such as *gao zhuang* 'sue'. However, *shou tu* 'take an apprentice' in (14) and *jiang xue* 'present a lecture' in (15) are obviously not lexicalized, but they can be raised to be incorporated into the light verb, according to Feng. Therefore, the V^0-movement explanation does not hold water.

Second, even if the light verb is assumed to be in existence and VO verbal constructions can be raised as a whole, problems still exist. Take *taoyan* 'dislike/loathsome' as an example. *Taoyan* can be found in the historical documents from the Ming Dynasty (In documents of the Yuan Dynasty, *tao ge yan jian* 'ask for a snub' can be found.). For example:

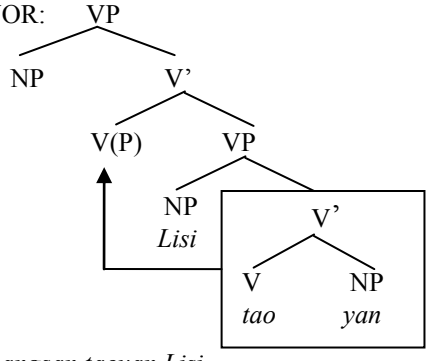
(20) 取 憎 讨 厌， 龌龊 不 洁
qu zeng tao yan, wochuo bu jie,
ask.for hate ask for disgust dirty not clean
'(A person who looks) unpleasant, repulsive and dirty (is a vulgarian).'

The fact that *tao yan* in (20) appears in juxtaposition with *qu zeng* whose structure is VO indicates that the structure of *tao yan* is VO. If, according to Feng's account, *taoyan* 'dislike' in Modern Chinese is assumed to be derived through V^0-movement, its D-structure should be like (21):

(21) 张三 从 李四 那里 淘 （了 许多） 厌
Zhangsan cong Lisi nali tao le xuduo) *yan*.
Zhangsan from Lisi there ask for ASP many disgust
Intended reading: 'Zhangsan was disliked (a lot) by Lisi.'

If the VO construction in (21) rises, (22) should be legitimate.

(22) a. DS: *Zhangsan cong Lisi nali tao yan*.
b. VOR:



c. SS: *Zhangsan taoyan Lisi*
d. PF: *Zhangsan taoyan Lisi*

The derivation itself seems to be correct. The meaning of the (22c & d), however, is totally different from (21) and (22a). This result, obviously, is not predicted by Feng.

Actually, previous studies (e.g. Xing [18]) also show that the phenomena like (14) and (15) can be found in historical documents, as shown in (23) (taken from *Shiji* (or *Records of the [Grand] Scribe*)) and (24) (taken from the Song Dynasty):

(23) 以 古 法 议， 决 疑 大 狱
yi *gu* *fa* *yi*, *jue* *yi* *da* *yu*
according to old law discuss judge doubt big lawsuit
'(He) made judgement on big lawsuits in accordance with doctrine.'

(24) 借 路 平 栾 州 归
jie *lu* *ping* *luan* *zhou* *gui*…
borrow road Ping Luan city return
'(The army) can return by way of the cities Pingzhou and Luanzhou.'

It is necessary to note that two thousand years ago when *Shiji* was compiled by Sima Qian, PPs could not appear before verbs as is the case nowadays. This seems to indicate that the prepositions which are treated as light verbs by Feng [11] today did not develop at that time. Without the light verb V(P), how can a verbal VO be raised? Where can its landing site be?

3.2 Classification of Pseudo-VO Verbal Compounds

Since VO constructions in (10-13) and (14-15) cannot be treated as VO verbal compounds, I name them pseudo-VO verbal compounds. For the sake of discussion, they can be further divided into two sorts: VO-verbs, as in (10-13), and VO-phrases, as in (14-15).

VO-Verbs.

Although VO-verbs' internal structures are VOs, they are lexical verbs. Their VO-like structures are not derived syntactically, but are directly determined in lexicon. VO-verbs are syntactically different from the VO verbal compounds in several aspects.

First, the V in VO verbal compounds can take an aspectual marker, such as *-le, -zhe, -guo*. This, however, is not possible for those in (10-13), as shown below:

(25) a. 吃 着/了/过 （李小姐） 的 豆腐
chi *zhe/le/guo* *Lixiaojie* *de* *doufu.*
eat ASP Miss Li *de* tofu
'flirting/flirted with (Miss Li)'
b. *关 着/了/过 心 （李小姐）
guan *zhe/le/guo* *xin* *Lixiaojie*
concern ASP heart Miss Li
Intended reading: 'care about (Miss Li)'

Second, the O in VO verbal compounds can be modified by certain categories, but it is not possible for those verbs in (10-13), as shown in (26).

(26) a. 吃 了 （李小姐） 不少 的 豆腐
chi le Lixiaojie bushao de doufu.
eat ASP Miss Li much *de* tofu
'flirted with Miss Li a lot.'

b. *关 不少 （的） 心 （李小姐）
guan bushao de xin Lixiaojie
concern much *de* heart Miss Li
Intended reading: 'care about Lisi very much'

Third, the objects of verbs in (10-13) cannot be realized as the attributive of O, as shown below.

(27) *着 此 事 的 手
zhuo ci shi de shou
put this matter *de* hand
Intended reading: same as (10).

(28) *得 王五 的 罪
de Wangwu de zui
obtain Wangwu *de* blame
Intended reading: same as (11).

(29) *关 李四 的 心
guan Lisi de xin
concern Lisi *de* heart
Intended reading: same as (12).

(30) *负 此 事 的 责
fu ci shi de ze
bear this matter *de* duty
Intended reading: same as (13).

Last but not least, if those verbs in (10-13) take aspectual markers, they will appear after O. For example, if *guan xin* 'care about' takes an aspectual marker, it appears after *xin*, that is, *guanxin zhe/le/guo*.

These four points obviously indicate that although VO-verbs look like VO verbal compounds, they are actually perfect, inseparable, and true verbs. They assign Cases as well as θ-roles to their objects directly, and their objects are realized in situ. It is undoubtedly inappropriate to mix them with VO verbal compounds.

VO-Phrases

The reason why VO-phrases are named "phrases" is that their syntactic behaviors are identical with common VO phrases in most cases, although their adverbials of place appear as NPs instead of PPs. In fact, this kind of examples can be found in abundance in historical documents, as shown in (23-24). More examples are given below:

(31) 八月 蝴蝶 黄
bayue hudie huang,
August butterfly yellow
双 飞 西 园 草
shuang fei xi yuan cao.
pair fly west garden grass
'The paired butterflies are already yellow with August
Over the grass in the West garden.' (By Li Bai)

(32) 孤 舟 蓑 笠 翁
gu zhou suo li weng,
lone boat reed cloak reed hat old man

独 钓 寒 江 雪
du diao han jiang xue.
alone fish cold river snow
'In a lone boat, rain cloak and a hat of reeds,
An old man's fishing the cold river snow.' (By Liu Zongyuan)

In the literature, many scholars believe that the NPs following VO-phrases come from PPs, through a process referred to as *yu*-ellipsis (e.g., Ling [19]). Why is it possible for *yu* to be omitted? The answer to this question is related to the nature of *yu*. Let us first consider its historical origin.

When it comes to the origin of the preposition *yu*, three hypotheses exist in the literature: (1) from the verb *yu* [20-23]; (2) from a harmonic which existed in Old Chinese [24-25]; (3) from the Case-auxiliary of proto-Chinese [26]. The first hypothesis enjoys a wide acceptance, but many scholars cast doubts on its reliability constantly (See, for instance, Xuan [27]). As to the second and the third hypotheses, it is hard to judge which one is more proper until more data from Old Chinese are taken into consideration, as shown by the following examples:

(35) 王 学(敦) 众 伐 于 髦 方。
wang xiao zhong fa yu mao fang
King train subject attack YU Mao tribe
'The king trains his subjects to attack Mao.' (oracle-bone inscription)

(36) 帝 弗 缶 于 王。
di fu fou yu wang
God not bless YU king
'God will not bless the king.' (oracle-bone inscription)

(37) 室 于 怒 市 于 色 者
shi yu nu shi yu se zhe
room YU anger street YU color ZHE
'When he gets angry at home, he takes it out on others outside in the street.'
(*Zuo Zhuan*)

The position where *yu* appears is not stable. For example, in (35) and (36), *yu* precedes the accusative-complement; in (37), however, it appears after the adverbial of place. Obviously, only the second hypothesis can explain this naturally.

Suppose that the analysis above is on the right lines, the ellipsis of *yu* can be explained naturally, as shown below:

(38) a. 王其 田 莒，无 灾？
wangqi tian Ju, wu zai
king hunt Ju no disaster
'The king will hunt at Ju. No disaster?' (oracle-bone inscription)

b. 王其 田 莒，不 小 雨？
wangqi tian Ju bu xiao yu
king hunt Ju not small rain
'The king will hunt at Ju. No light rain?' (oracle-bone inscription)

In Middle Chinese and Modern Chinese *yu*-ellipsis can also be found, but it is mainly caused by the prosodic constraint. For example, in (31-32), the ellipsis of *yu* is due to the requirement of poetic rhythm. In (14-15), the secret lying behind this phenomenon is that their rhythm is consistent with the Chinese natural metric feet. According to Feng, "the natural way to read a five-syllable cluster is [σσ # σσσ], but not [σσσ # σσ]…. In Chinese, rhythm like [σ # σσ # σσ] and [σσ # σσ # σ] does not exist" [9]. That is to say, the natural metric feet in Chinese is [σσ # σσσ], which exactly explains the *yu*-ellipsis in (14-15).

If the above analysis is correct, then the O in a VO-phrase must be its object. The NP following it is not its object but an adverbial. What causes the adverbial NP to be positioned immediately after a VO is that *yu*, which links them, is omitted. It is safe to assume that the NP following a VO is not an object, but a fake object.

4 Conclusion

The objects of VO verbal compounds are realized in special ways. Owing to the lack of Case, they cannot appear in situ, but have to be combined with a prepositions (so as to get oblique Cases) or become the possessors of Os. VO verbal compounds should be distinguished from two kinds of pseudo-VO verbal compounds. One is VO-verb, which can assign Case directly to its object. The other is VO-phrases, which is derived via preposition-omission under the effect of prosody.

Although VO verbal compounds are syntactically special, they can be explained by the current syntactic theories. In fact, the framework of Generative Grammar has already provided adequate explanation for them. The main reason why the previous studies are faced with a lot of problems is that they blurred the boundary between true VO verbal compounds and pseudo-VO verbal compounds, with the aim to give a united account, which, of course, is far beyond our ability.

Acknowledgments. The study is jointly supported by the National Social Science Foundation of China (12BYY049), Humanities and Social Sciences by the Ministry of Education (12YJC740144) and Social Science Foundation of Shandong Province (11CWZJ59).

References

1. Huang, C.R.: A Unification-Based LFG Analysis of Lexical Discontinuity. Linguistics 28, 263–307 (1990)
2. Her, O.-S.: Grammatical Representation of Idiom Chunks. In: International Association of Chinese Linguistics 8th Annual Conference, Melbourne, Australia (1999)
3. Chao, Y.R.: A Grammar of Spoken Chinese. University of California Press, Berkeley (1968)
4. Li, C.N., Thompson, S.A.: Mandarin Chinese: A Functional Reference Grammar. University of California Press, Berkeley (1981)
5. Zhu, D.: Lectures on Grammar. The Commercial Press, Beijing (1982) (in Chinese)

6. Huang, G.: On de: its syntax and semantic function. Studies in Language and Linguistics (1), 101–129 (1982) (in Chinese)
7. Huang, C.T.J.: Wo Pao De Kuai and Chinese phrase structure. Language 64, 274–311 (1988)
8. Tang, S.W.: Word Order in Natural Languages and the Theory of Phrase Structure. Contemporary Linguistics 2, 138–154 (2000) (in Chinese)
9. Feng, S.: The Prosodic Syntax of Chinese. Shanghai Education Press, Shanghai (2000) (in Chinese)
10. Feng, S.: Prosodically Determined Distinctions between Word and Phrase in Chinese. Studies of The Chinese Language (1), 27–37 (2001) (in Chinese)
11. Feng, S.: On the Interface between Prosodic Morphology and Prosodic Syntax. Studies of the Chinese Language (6), 515–524 (2002) (in Chinese)
12. Cai, S.: Semantic Properties and Syntactic Constructing Process of the Specific Dative Construction "V+X+de+O". Chinese Teaching in the World (3), 363–372 (2010) (in Chinese)
13. Li, G.: On the Possessive Object Construction "VN de O". Chinese Language Learning (3), 63–69 (2009) (in Chinese)
14. Chomsky, N.: Lectures on Government and Binding. Foris, Dordrecht (1981)
15. Chomsky, N.: Knowledge of language: its nature, origin and use. Praeger, New York (1986)
16. Ouhalla, J.: Introducing Transformational Grammar: From Principles and Parameters to Minimalism. Edward Arnold, London (1999)
17. Cui, X.: On *kaixin* (开心) and *guanxin* (关心). Studies of The Chinese Language (5), 410–418 (2009) (in Chinese)
18. Xing, G.: On a kind of sentence pattern that is likely to become popular. Language Planning (4), 21–23 (1997) (in Chinese)
19. Ling, D.: An Analysis on Verb-object Type+ Object. Chinese Language Learning (5), 9–13 (1999)
20. Guo, X.: On the Origin and Development of the Preposition *yu*. Studies of The Chinese Language (2), 131–138 (1997) (in Chinese)
21. Mei, T.L.: The Source of the Preposition *yu* in Oracle Bone inscription and in Sino-Tibetan. Studies of The Chinese Language (4), 323–332 (2004) (in Chinese)
22. Luo, G.: On the Usage of Verb "yu". Research in Ancient Chinese Language (6), 73–75 (2007) (in Chinese)
23. Zhang, Y.: On the Origin of "yu". Chinese Linguistics (4), 16–22 (2009) (in Chinese)
24. Zhao, Z.: On the Prepositions in Ancient Chinese "yu" (于), "yu" (於) and "hu" (乎). Journal of Sun Yatsen University (4), 98–109 (1964) (in Chinese)
25. Zhu, X.: An analysis of "yu" in the Shangshu. Journal of Language and Literature Studies (2), 39–42 (1988) (in Chinese)
26. Shi, B.: On the Origin and Development of Preposition "yu". Studies of The Chinese Language (4), 343–347 (2003) (in Chinese)
27. Xuan, J.: Concerning "yu" Used as a Verb in Oracle Bone Inscriptions. Research in Ancient Chinese Language (1), 33–38 (2009) (in Chinese)

Specification for Segmentation and Named Entity Annotation of Chinese Classics in the Ming and Qing Dynasties

Dan Xiong[1], Qin Lu[1], Fengju Lo[2], Dingxu Shi[3], Tin-shing Chiu[1], and Wanyin Li[1]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{csdxiong,csluqin,cstschiu}@comp.polyu.edu.hk,
csclaireli@gmail.com
[2] Department of Chinese Linguistics & Literature, Yuan Ze University, Taiwan
gefjulo@saturn.yzu.edu.tw
[3] Department of Chinese & Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong
ctdshi@polyu.edu.hk

Abstract. The quality of text segmentation and annotation plays a significant role in Natural Language Processing especially in downstream applications. This paper presents the specification for word segmentation and named entity annotation targeted for novels in the Ming and Qing dynasties. The purpose of this work is to build the foundational work for computer-aided lexical semantic analysis of classical Chinese literature, especially the transition of Chinese literature from its traditional forms such as traditional verses and vernacular styles to modern Chinese. To assist in literature study, an elaborate named entity annotation scheme is specially developed for classical Chinese. Computer-aided segmentation and named entity annotation are conducted on some famous Ming and Qing Chinese classics. The specification for the segmentation and annotation is produced based on the studies of the morphology and semantics differences as well as similarities between classical Chinese and modern Chinese with reference to the existing standards for modern Chinese processing widely used in Mainland China and Taiwan.

Keywords: segmentation and PoS principles, named entities, novels in the Ming and Qing dynasties, computer-aided annotation, semantic analysis.

1 Introduction

With the rapid development of information technology and the digitization of documentation, different kinds of annotated text corpora are established to assist in natural language processing applications. The project "Building a Diachronic Language Knowledge-Base" aims to build a comprehensive knowledge base of Chinese language in different eras ranging from traditional verse (韻文) and vernacular literature (語體文) to modern Chinese. This involves the integration of texts of different styles in different historical periods with annotated information to assist in the understanding

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 280–293, 2013.

and computer processing of lexical meaning, semantic marking and classification of semantic concepts, as well as the description of grammatical knowledge. In this way, the knowledge-base is able to make use of computer technology to help semantic, syntactic and discourse analysis as well as subject extraction and classification.

This paper describes the development of the specification for segmentation and named entity annotation for the novels in the Ming and Qing dynasties. The novels of this period, written in vernacular style, are regarded as important evidence in the transition from classical to modern Chinese. The establishment of the specification for word segmentation and named entity annotation is vital for building a high-quality natural language corpus to shed light on the gradual change of Chinese writing and possibly provide links between ancient Chinese and modern Chinese. The specification is designed based on the analysis of existing segmentation and annotation principles established for modern Chinese as well as the differences and similarities between the Ming and Qing novels and modern Chinese. In accordance with the specification, a corpus of four classical novels in the Ming and Qing dynasties, namely, *A Dream of Red Mansions* (《紅樓夢》), *Romance of the Three Kingdoms* (《三國演義》), *Water Margin* (《水滸傳》), and *The Golden Lotus* (《金瓶梅》), will be annotated through a computer-aided method followed by manual review. The result of the work is expected to offer valuable insights into lexical semantic analysis of classical literature in the Ming and Qing dynasties and possibly provide basis for the studies of the evolution of Chinese from ancient time to modern era.

This paper is organized as follows. Section 2 presents the design concepts of the specification and briefly explains the differences between this work and a previous work done by Academia Sinica. Section 3 and 4 describe the principles of word segmentation and named entity annotation targeted for the Ming and Qing novels, respectively. Section 5 discusses the quality assurance process and the test results of annotation evaluation in different stages. Section 6 gives a conclusion.

2 Design Concepts of the Specification

In the development of this specification, the existing standards for modern Chinese segmentation are used as references, including the specification for corpus processing developed by Peking University in Mainland China [1] and the standard of Chinese segmentation used in Taiwan [2]. But, in view of the features of the Ming and Qing novels, it is necessary to establish a segmentation and annotation specification applicable to this kind of classical literature. Compared to modern Chinese, the most distinctive features of the Ming and Qing novels include frequent use of single-character words, the chapter-by-chapter style, lexical meaning, and the use of named entities, which are the key elements to be considered during the design of the specification.

First, the Ming and Qing novels, appearing more compact, use much more single-character words than multi-character words. According to the statistics on the four classical novels mentioned in Section 1, the number of single-character words is 13 times that of two-character words, 32 times that of three-character words, and 65 times that of other multi-character words. The chapter-by-chapter style is another

feature of the Ming and Qing novels, in which each chapter is headed by a couplet implying the main plot of this chapter and also usually ends with a couplet or poem. The novels contain a large number of poems, couplets, prose poems, lyrics, etc., which will be treated later according to the specification for segmenting traditional verse. It is obvious that many lexical entries used in the Ming and Qing novels are partially or completely different from those in modern Chinese. For instance[1], in the Ming and Qing novels, "一心一計"[2] (yī xīn yī jì, wholeheartedly) is used more frequently, but in modern Chinese the equivalent is "一心一意" (yī xīn yī yì).

In addition to the linguistic features, named entities are another distinguishing feature of the Ming and Qing novels. Taking personal names as an example, in ancient China, an adult male may have a courtesy name (字, zì) and one or more art names (號, hào) besides his given name. After death, an honorable person may be awarded a posthumous name (諡號, shì hào), and an emperor is usually given a temple name (廟號, miào hào). Therefore, a great diversity of personal names can be found in the Ming and Qing novels, which are also used in other ancient Chinese literature. Annotation of such entities in great details enables Chinese literature works in different periods to be linked through extraction and reference of annotation tags. China's system of geographic units and administrative divisions also changes over time, which makes place name, organizational name, and official titles in the Ming and Qing novels different from those in modern Chinese. All of these features should be taken into consideration so that the specification can be applied to this kind of literature.

In general, the specification is designed based on the features of the Ming and Qing novels with reference to the existing word segmentation and annotation principles for modern Chinese and has been established iteratively as the tagging work progresses.

It should be pointed out that there was a related annotation work performed by Academia Sinica on Chinese classics [3]. The whole collection, called Tagged Corpus of Early Mandarin Chinese (hereinafter referred to as the Corpus), has included some classics of the Ming and Qing dynasties. Different from this work, the Corpus which focuses on grammatical tagging from the perspective of word category is fully parsed and annotated to facilitate filtering, searching, analyzing, and statistics of PoS. However, most of the named entities in the Corpus are put into one category and tagged as proper nouns only. In this case, semantic analysis would not be sufficient to link entities of literature in different historical periods which is part of our project objectives. In terms of words, the Corpus considers them more from the lexical side whereas we have more emphasis on semantic integrity. For example, the Corpus treats the segmentation of "桌/上"(zhuō/shàng, on the table) and "心/上" (xīn shàng, in one's heart) in the same way. But in our project, "心上" (xīn shàng, in one's heart) is not segmented because here "上" (shàng) is not a directional indicator.

---

[1] The examples given in this paper are all cited from *A Dream of Red Mansions* and *Romance of the Three Kingdoms*.

[2] In *A Dream of Red Mansions*, "一心一計" (yī xīn yī jì, wholeheartedly) is found in chapters 6, 65, 69, 79, and 101, and "一心一意" (yī xīn yī yì, wholeheartedly) is only found in 98. Their meanings are exactly the same.

3 Principles of Word Segmentation

Based on the above analysis performed on the differences between the language of the Ming and Qing novels and modern Chinese, this work has formulated the basic segmentation rules applicable to the Ming and Qing novels.

First of all, a segmentation unit is defined as the smallest linguistic unit that "has specific semantic or grammatical functions", which is defined in GB13715, i. e., China's national segmentation standard for modern Chinese information processing [4]. The principles of word segmentation for the Ming and Qing novels are established from the perspectives of both semantics and syntax. From the perspective of semantics, the fundamental rule is to segment words into the smallest units without loss of semantic information, distortion, and ambiguity. A set of principles are also established from the perspective of syntax as complementary guidelines.

3.1 Word Segmentation Principles from the Perspective of Semantics

From the perspective of semantics, the basic segmentation principle is to segment character strings into the smallest units while there is no risk of meaning loss, misinterpretation, and ambiguity. For instance, "一/疾/而/終" (yī/jí/ér/zhōng, fell ill and died) is segmented into four units since each individual word has an independent semantic meaning or grammatical function. The following describes the main segmentation principles from the perspective of semantics accompanied by examples.

Handling of Functional Words. In the Ming and Qing novels, the following words are frequently used: "之" (zhī, usually used as a pronoun or a marker of subordination between nouns), "了" (le, a particle usually following a verb to indicate a completed action), "的" (de, a particle with flexible usages, such as attached to a pronoun or noun to indicate possession, to an adjective for description and emphasis, and to a verb for nominalization), "於" (yú, a preposition usually used to indicate time, place, or direction), "眾" (zhòng, used before nouns to indicate plural), "們" (men, used after nouns or pronouns to indicate plural), "只" (zhǐ, an adverb that means only, just, or simply), "被" (bèi, used as a marker of passive voice), "也" (yě, used as an adverb similar to "also", or used in the end of a sentence as a particle implying affirmation), "亦" (yì, an adverb that means "also"), "所" (suǒ, usually used in relative clauses and passives), "而" (ér, usually used as a conjunction to indicate parallel connection, temporal sequence, contrastive or concessive relations, etc.), "得" (dé, usually used as a particle following a verb to indicate the status of being able to), "時" (shí, commonly used as an adverb in the sense of "while" or "at that time"), "者" (zhě, usually used as a marker of nominalization). Since these words are commonly used in combination with different kinds of words to form various structures, they are regarded as independent segmentation units, for example, "用/了" (yòng/le, used) and "投降/者" (tóu xiáng/zhě, surrenders). However, fixed expressions are not segmented, for example, "我們" (wǒ men, we or us) and "作者" (zuò zhě, the author).

Handling of Demonstrative Pronouns. Demonstrative pronouns such as "這" (zhè, this), "那" (nà, that), "此" (cǐ, this), "某" (mǒu, a certain), etc. and words of inclusion or restriction such as "每" (měi, every), "各" (gè, each), "諸" (zhū, all), etc. are treated as segmentation units, for example, "這/石" (zhè shí, this stone). However, the lexical entries containing a pronoun referring to one or more unspecified beings are usually not segmented. This should be justified based on the context. For example, "閑步/至/此" (xián bù/zhì/cǐ, came out for a stroll and stopped here) and "事/已/至此" (shì/yǐ/zhì cǐ, indicating that something cannot be changed) are treated differently. In the former example, "此" (cǐ) refers to the place the character in the novel is arriving at, which is definite; while in the latter example, what "此" (cǐ) refers to is indefinite. So the same lexical entry is treated differently in different contexts.

Handling of Directional Words. The phrases in combination with words of locality are segmented if they indicate location or direction, for example, "門/前" (mén/qián, at the gate). Words of locality include "前" (qián, front), "後" (hòu, back), "左" (zuǒ, left), "右" (yòu, right), "上" (shàng, up), "下" (xià, down), "裏" (lǐ, in), "中" (zhōng, within), "內" (nèi, inside), "外" (wài, outside), "畔" (pàn, side), "旁" (páng, side), "邊" (biān, side or edge), etc. However, if the word groups have new meanings not associated with location or direction or there are no corresponding antonyms, they are not segmented, for example, "心下/乃/想" (xīn xià/nǎi/xiǎng, thinking to herself). The lexical entry "心下" (xīn xià, in one's heart) is not segmented because here "下" (xià) does not function as a directional indicator. In the Ming and Qing novels, "心上" (xīn shàng), "心下" (xīn xià), and "心中" (xīn zhōng) are all used, but they convey the same meaning "in one's heart" and neither "上" (shàng) nor "下" (xià) implies a direction.

Handling of Negation. The phrases in combination with negatives such as "不" (bù), "沒" (méi), "非" (fēi), "無" (wú), "勿" (wù), etc. are segmented, for example, "不/可" (bù/kě, should not). The context becomes a major factor for judging whether a phrase of this kind should be segmented. There are also some special cases:

- If a word group has new meaning or does not convey the negative meaning, it is not segmented. For example, "不但" (bù dàn, not only) is not segmented because it is not a negative.
- A fixed expression is not segmented. For example, although "不消" (bù xiāo, need not) is a negative, it is not segmented because it is used as a fixed expression.

Handling of Repetition. Duplicated words are not segmented, for example, "隱隱" (yǐn yǐn, half hidden). However, those inserted by a word such as "一" (yī, once) and "了" (le, a particle usually following a verb to indicate a completed action) are segmented, for example, "享/一/享" (xiǎng/yī/xiǎng, enjoy).

Handling of Words with Similar or Opposite Meanings. The phrases composed of two words with similar or opposite meanings are not segmented because they are usually used as fixed expressions, for example, "悲歡" (bēi huān, the joys and sorrows) and "抄錄" (chāo lù, copy). To help future study, a list of such lexical entries is appended to the specification document.

Handling of Suffixes. In Chinese, suffixes do not have independent semantic or grammatical functions, so they are not segmented from the words to which they are attached in this work. For example, "花兒" (huā ér, flower) is not segmented because it has the same meaning as "花" (huā, flower). The most commonly used suffixes in the Ming and Qing novels include "兒" (ér, mainly attached to nouns and verbs, used in dialects or informal situations), "子" (zǐ, mainly attached to nouns, sometimes also attached to verbs and adjectives for nominalization), "著" (zhe, mainly attached to verbs to indicate the unchanging state of an action), "些" (xiē, usually attached to verbs, adjectives, or pronouns to indicate indefinite amount or degree), and "然" (rán, mainly attached to adjectives and adverbs).

Handling of Idioms. Idioms, set phrases and the word groups with particular meanings different from the literal combinations of individual words are not segmented, for example "連二連三" (lián èr lián sān, in turn). It worth noting that there are some phrases used in Ming Qing novels, which may not be used in modern Chinese any longer, for example, "四下裏" (sì xià lǐ, everywhere).

3.2 Word Segmentation Principles from the Perspective of Syntax

Some rules are also formulated from the perspective of syntax as complementary guidelines. The following describes the main principles with examples.

Handling of Verb-Object Structure. The verb-object word groups are segmented if both the verb and the object have independent semantic functions, for example, "理/朝廷" (lǐ/cháo tíng, regulate the government). However, there are many special cases:

- Fixed expressions and the verb-object word groups with a particular meaning that cannot be inferred from the meaning of each separate word are not segmented. For example, "嚼舌根" (jiáo shé gēn, describe sb. as foul-mouthed) is not segmented because it has a new meaning not associated with "嚼" (jiáo, chew) and "舌根" (shé gēn, tongue).
- The same lexical entry that has more than one grammatical function is treated differently in different contexts. For example, when "回書" (huí shū) functions as a predicate which means "to reply to one's letter", it is segmented into two units; however, when it is used as a noun with the meaning of "a letter in reply", it should not be segmented.

- The same lexical entry that has the same grammatical function may have different semantic meanings in different contexts. For example, "下馬" (xià mǎ) functions as a predicate but carries more than one meaning, including "to dismount from a horse" and "to assume a post", the two most common meanings in the Ming and Qing novels. When it means "to dismount from a horse", it is segmented. When it means "to assume a post", it is not segmented because it has a particular meaning that is different from the literal sense of the individual words, that is, not associated with "下" (xià, down) and "馬" (mǎ, horse).

Handling of Adjective-Noun Structure. The adjective-noun word groups are segmented if both the adjective and the noun have independent semantic functions, for example, "奇/物" (qí/wù, something special).

Handling of Subject-Predicate and Predicate-Complement Structure. The word groups of this kind are usually segmented because the elements of these structures normally have independent semantic functions, for example, "吃/盡" (chī/jìn, finish off).

Handling of Combinations of Number, Quantifier, and Noun. The word groups of number, quantifier, and noun are segmented because the elements have independent semantic functions. These structures include "marker of ordinal numerals + number + quantifier", "number + quantifier + noun", "number + noun", etc. Here are two examples: "第/二/日" (dì/èr/rì, the next day) and "幾百/株/杏花" (jǐ bǎi/zhū/xìng huā, hundreds of apricot trees).

Handling of Adverb-Verb Structure. The adverb-verb word groups used as fixed expressions are not segmented, for example, "嚎哭" (háo kū, wail).

Handling of Combinations of Directional Verb and Directional Complement. In the Ming and Qing novels, many main verbs consist of a directional verb and a directional complement. Usually used as fixed expressions, they are not further segmented, for example, "上來" (shàng lái, come up) and "下去" (xià qù, go down). Directional verbs include "上" (shàng, up), "下" (xià, down), "過" (guò, cross over), "回" (huí, back), "進" (jìn, in), "出" (chū, out), "起" (qǐ, get up), "歸" (guī, go back), "到" (dào, get to), "走" (zǒu, walk), etc. Directional complements include "來" (lái, come), "去" (qù, go), "入" (rù, enter), etc.

4 Segmentation and Annotation of Named Entities

In this work, named entities most commonly used in the Ming and Qing novels are classified into six categories: personal name (人名), term of address (人物稱謂),

name of official position (官職) and title of nobility or honour (爵位、封號), place name (地名), building name (建築名), and organizational name (組織名).

In the Ming and Qing novels, there are many compound named entities. For instance, various terms of address are generated by different combinations of any form of a person's name with a title indicating the person's rank or position. Any personal name, term of address, or title may also be used as a part of a place name or a building name. That is why a variety of compound named entities are formed. This study seeks to establish a set of annotation rules for different kinds of named entities to ensure that they can be segmented and tagged in a consistent and flexible way. The establishment of these rules is based on researches on Chinese literature, analysis and statistics in combination with experience in actual annotation. In general, the approach of Peking University for named entity tagging [1] is followed: square brackets ([]) are used to enclose compound named entities and labels are marked by the slash sign (/). The units in the square brackets are segmented and tagged according to the unified specification stated in this paper.

To facilitate future research and application in literature study, this work also distinguishes real persons and places from fictional ones. Even though persons and places in literature works are usually fictional, there are many references to real historical figures and places in Chinese classics due to the rich history of China. For example, in the novel *Romance of the Three Kingdoms*, there are many references to real persons and places which can be found in the official records. To make a distinction and enable easier identification for the study of Chinese classics, personal names, terms of address, place names, and building names are further classified into four categories as described below:

- Real entities: labeled with "#", for example, "蘇軾/nr3#" (Sū Shì/nr3#, one of the major poets of the Song dynasty). In this work, the entities recorded in *The Twenty-Four Histories* (《二十四史》) are considered as real ones.
- Mythical entities: labeled with "*", for example, "灌愁海/ns3*" (Guàn Chóu Hǎi/ns3*, the Sea of Brimming Grief).
- Fictional entities cited from other literature works: labeled with "&", for example, "紅娘/nr6&" (Hóng Niáng/nr6&, a maid in *Romance of the West Chamber* 《西廂記》) which is mentioned in a dialogue of *A Dream of Red Mansions*.
- Fictional entities in the novel being processed: this is default with no special symbol required.

Once the named entities are properly categorized and tagged, not only the correlation between the entities within the novels but also the connection between the entities in the novels and those in other literature works can be established.

4.1 Personal Names

Various forms of names in the Ming and Qing novels are categorized into six types: surname, given name, surname + given name, courtesy name, surname + courtesy

name, and alternative name. This section describes the principles for tagging personal names in details.

Surname. The most common surname contains only one character, but there are also multi-character surnames in China. It is unnecessary to segment a surname no matter how many characters it contains. Any surname referring to a specific person is tagged with "/nr1", for example, "薛/nr1 林/nr1 二/人" (Xuē/nr1 and Lín/nr1, referring to Xue Baocai and Lin Daiyu, two main characters in *A Dream of Red Mansions*).

Given Name. It is unnecessary to segment a given name. Any given name referring to a specific person is tagged with "/nr2", for example, "黛玉/nr2" (Dàiyù/nr2).

Surname + Given Name. A complete personal name is composed of a surname name plus a given name. Any "surname + given name" referring to a specific person is tagged with "/nr3". There are different cases:

- If the surname contains only one character, it is the default case and there is no need to separate it from the given name because the computer will regard the first character as the surname in this case, for example, "林黛玉/nr3" (Lín Dàiyù/nr3).
- If the combination of "surname + given name" contains a multiple-character surname, the surname is separated from the given name with "//", for example, "司馬//相如/nr3#" (Sīmǎ//Xiàngrú/nr3#, an official of the Western Han Dynasty well-known for his prose poems).
- Sometimes, when mentioning a woman, people may add her husband's surname before her own surname. If a compound name contains more than one surname, the surnames are separated with "//" and they are also separated from the given name with "//". In this way, different kinds of combinations can be treated consistently and flexibly, such as "two-character surname//given name", "two-character surname//one-character surname//given name", and "one-character surname//two-character surname//given name". Thus, both surnames and given names can be recognized by the computer easily.

Courtesy Name (字, Zì). In traditional Chinese culture, an adult male usually selects or acquires from other people a courtesy name as a symbol of adulthood and respect. As another form of a given name, it commonly consists of one or two characters. It is unnecessary to further segment a courtesy name and the whole courtesy name is tagged with "/nr4", for example, "孔明/nr4#" (Kǒngmíng/nr4#, the courtesy name of Zhuge Liang, a famous strategist during the Three Kingdoms period of Chinese history).

Surname + Courtesy Name. Any "surname + courtesy name" referring to a specific person is tagged with "/nr5" with other rules similar to that of "surname + given

name", for example, "諸葛//孔明/nr5#" (Zhūgě//Kǒngmíng/nr5#, the combination of surname and courtesy name of Zhuge Liang).

Alternative Name. Besides given name and courtesy name, people may have some other alternative names, including milk name, nickname, pen name, art name (號, hào, an alternative courtesy name most commonly three or four characters in length), posthumous name (諡號, shì hào, a honorary name selected after a person's death), temple name of an emperor (廟號, miào hào), etc. All these alternative names are put into one category. Any alternative name referring to a specific person is tagged with "/nr6", for example, "顰兒/nr6" (Pín ér/nr6, a nickname of Lin Daiyu). There are some special cases:

- Since most names of foreign nations and races in the novels are translated or transliterated from foreign languages, they are treated as alternative names if the surname and the given name cannot be identified, for example, "金環三結/nr6" (Jīn Huán Sān Jié/nr6, one of the chiefs of the tribesmen).
- A person's name may be changed for different reasons. For example, in ancient China, the names of the emperors, elders, and people of higher rank are regarded as taboos, so a person's name may have to be changed. Whatever the reason is, the new name is regarded as an alternative name.
- If a compound alternative name contains a surname with only one character, it is unnecessary to separate it from the following alternative name. If a compound alternative name contains one surname with two or more characters, or it contains more than one surname, the principle is the same as that for "surname + given name".

4.2 Terms of Address

A variety of address terms can be found in the Ming and Qing novels. If they are not tagged in the corpus, word sense ambiguity may be caused. For example, 公 (Gōng) has many meanings when used as a common adjective, such as public, fair, etc. In the Ming and Qing novels, it is also used as a term of address in respectful term or as the title duke. If the named entities are not properly tagged, this kind of information cannot be identified and extracted efficiently. Terms of address in the Ming and Qing novels fall into two categories. In the first category, it is used alone without being combined with personal names whereas in the second category, it is used in combination with other names given in Section 4.1.

Terms of Address Used Alone. A term of address may indicate a person's gender, marital status, kinship, social class, occupation, religious belief, etc. Any address term of this kind referring to a specific person is tagged with "/na2", for example, "老爺/na2 說/了" (lǎo yé/na2 shuō/le, the master says). Terms of address are segmented according to the word segmentation principles stated earlier. Square brackets are used to enclose those consisting of more than one segmentation unit, for example, "[二/小姐]/na2" ([èr/xiǎo jiě]/na2, the Second Young Lady). Only the term of address

referring to a particular person that can be identified from the context is tagged. For example, the following terms of address are not tagged: "一/個/小/丫頭/扶/了" (yī/gè/xiǎo/yā tóu/fú/le, leaning on a young maid's arm), "老爺/們" (lǎo yé/men, the masters).

Terms of Address Combined with Any Form of Name or Title. A term of address may be combined with a surname, given name, title, etc. to form a compound one. Any address term of this kind referring to a specific person is tagged with "/na1". The whole compound address term is enclosed in square brackets, in which the units are segmented and tagged according to the principles stated in this paper, for example, "[政/nr2 老爺]/na1" ([Zhèng/nr2 lǎo yé]/na1, Lord Zheng). This principle ensures that all kinds of complicated compound address terms are treated in a consistent way. If a compound address term contains more than one surname, the surnames are separated with "//" and they are also separated from the term of address with "/", for example, "[張//王/nr1 氏]/na1" ([Zhāng//Wáng/nr1 shì]/na1, in which Wáng is the surname of a woman, Zhāng is her husband's surname, and here shì is used to address a married woman).

4.3 Names of Official Position and Titles of Nobility or Honour

A name of official position is tagged with "/nu1". There are also many compound ones used together with an organization or a place. In this case, they are also treated in the same way as other compound named entities, for example, "[太醫院/nt 正堂]/nu1" ([Tài Yī Yuàn/nt zhèng táng]/nu1, the director of the Academy of Imperial Physicians) where "/nt" refers to an organization.

The system of nobility and honorific titles is an important feature of Chinese culture in the imperial age. Of course titles of nobility vary from dynasty to dynasty, but the most common five ones are used almost throughout China's whole imperial history: 公 (Gōng , duke), 侯 (Hóu, marquis), 伯 (Bó, earl), 子 (Zǐ, viscount), and 男 (Nán, baron). A title of nobility or honour may be granted to members of the imperial house and their blood relatives and in-laws. It may also be bestowed on persons of high merits or heroes who have made great contributions. In addition, there is also a well-developed system of titles for female members of the aristocracy. All of these noble and honorific titles are put into one category and tagged with "/nu2", for example, "[保齡/侯]/nu2", ([Bǎolíng/Hóu]/nu2, Marquis of Baoling). Sometimes, the name of the land granted to a noble person is attached to the title of nobility or honour. In this case, it is also treated in the same way as other compound named entities, for example, "[烏程/ns2# 侯]/nu2" ([Wūchéng/ns2# Hóu]/nu2, Marquis of Wucheng).

4.4 Place Names

Places are generally classified into four categories:

- **Country:** A country name is tagged with "/ns1", for example, "暹羅國/ns1#" (Xiān Luó Guó/ns1#, Siam, the former name of Thailand).

- **Prefecture, city, county, township, village, street, road, etc.:** This kind of place name is tagged with "/ns2", for example, "金陵/ns2#" (Jīnlíng/ns2#, the present-day Nánjīng).
- **Mountain, grassland, river, lake, sea, island, etc.:** This kind of place name is tagged with "/ns3", for example,"灌愁海/ns3*" (Guàn Chóu Hǎi/ns3*, the Sea of Brimming Grief).
- **Other places:** Other place names are tagged with "/ns4", for example, "虎牢關/ns4#" (Hǔláo Guān/ns4#, Hulao Pass, a mountain pass which is the site of many historical battles).

4.5 Building Names

Buildings in the Ming and Qing novels are broadly divided into three categories:

- **Palace, mansion, official residence, private garden, pavilion, etc.:** This kind of building name is tagged with "/nv1", for example, "大觀園/nv1" (Dà Guān Yuán/nv1, the Grand View Garden).
- **Temple, pawnshop, restaurant, teahouse, and some other public spaces:** This kind of building name is tagged with "/nv2", for example, "葫蘆廟/nv2" (Hú Lu Miào/nv2, the Gourd Temple).
- **Other buildings:** Other building names are tagged with "/nv3", for example, "翠煙橋/nv3" (Cuì Yān Qiáo/nv3, the Green Mist Bridge).

Sometimes, a building name is combined with a surname, alternative name, title etc. In this case, it is treated in the same way as other compound named entities: the whole compound building name is enclosed in square brackets, in which the units are segmented and tagged accordingly. The following gives two typical examples: "[榮/nu2 府]/nv1" ([Róng/nu2 fǔ]/nv1, the Rong Mansion, which is the residence of the Duke of Rongguo and his descendants), "[[忠靖/侯]/nu2 史/nr1 府]/nv1" ([[Zhōngjìng/Hóu]/nu2 Shǐ/nr1 fǔ]/nv1, the residence of Marquis of Zhongjing whose surname is Shǐ).

4.6 Organizational Names

The name of a particular organization is tagged with "/nt", for example, "太醫院/nt" (Tài Yī Yuàn/nt, Academy of Imperial Physicians). The compound organizational name containing more than one segmentation unit is treated in the same way as other compound named entities.

5 Quality Assurance and Annotation Evaluation

To ensure the quality of the annotated corpus, the project takes the approach of computer-aided annotation rather than using purely computerized or purely manual approach. The process of segmentation and tagging mainly consists of three stages:

system training and automatic processing, manual annotation and review, and post-processing.

In the first stage, literature dictionaries of some classics and a small amount of named entities which are manually identified are integrated into the existing segmentor/tagger [5]. In this way, the segmentor/tagger, which was originally developed to process modern Chinese, is trained to process the Ming and Qing novels. Then the software is used to segment and tag the text automatically. In the second stage, named entities are tagged and the whole text is manually reviewed by one annotator at least twice in strict accordance with the specification stated in this paper. In the third stage, a software tool is used to check inconsistency. Even though this process works in stages, iterations may be required from time to time as new cases in manual review and consistency check may make it necessary to slightly revise the specification and update the dictionaries.

To evaluate the quality of the output produced by the segmentor/tagger, ten chapters, 67,181 characters in length, were randomly selected from *A Dream of Red Mansions* for evaluation. The output compared to the result achieved after the 3rd stage shows that the automatically processed text can achieve 90.14% in precision and 94.48% in recall without human intervention. To evaluate the quality of the result in the 2nd stage, ten paragraphs were randomly selected from *A Dream of Red Mansions*, and the differences generated during the cross-check were discussed with the annotator to set an agreed golden answer. The results show that the precision is 99.44%. About 0.31% of all the errors are caused by different interpretations of different people on some lexical entries or on the specification. Once the specification is refined, these errors can be further reduced. The rest of errors, about 0.25% of the total, are human errors which are difficult to be eliminated completely. According to the specification, some lexical entries in the dictionaries integrated into the system should be further segmented manually. During manual review, it is possible to ignore a few entries. If necessary, the dictionaries will be updated accordingly. After correction of the errors, the specification document and the dictionaries will also be updated to improve the performance of automatic segmentation and annotation. The target is to ensure that the final corpus can reach a precision of 99.5%. In view of the quality assurance measures taken throughout the whole process, this goal is achievable, which is proven by the previous test data.

6 Conclusion

This study presents the specification for segmentation and annotation of Chinese novels in the Ming and Qing dynasties from the perspectives of semantics and syntax based on the standards for modern Chinese segmentation widely accepted in Mainland China [1] and in Taiwan [2]. This specification is designed in consideration of the unique characteristics of the novels in that period compared to modern Chinese, including frequent use of single-character words, the chapter-by-chapter style, the combinations of compound named entities, idiomatic expressions, etc. In terms of segmentation, the fundamental principle is to segment text strings into the smallest

linguistic units with independent semantic or grammatical functions while there is no risk of meaning loss, distortion, misinterpretation, and ambiguity. Different types of named entities, especially the compound ones flexibly grouped by various elements, are segmented and tagged in a consistent and flexible way. In the end, this paper also shows the quality control process, the measures taken in practical work, and a preliminary evaluation of the annotation quality in different stages of this work. The annotated corpus built in accordance with the specification can be used in different fields of study such as linguistics, literature, history, teaching of Chinese, and even information technology. The results of this study are expected to lay a foundation for computer-aided semantic analysis and named entity tagging of Chinese literature works in the Ming and Qing dynasties and more ancient times.

Acknowledgments. This work is partially supported by the Chiang Ching-kuo Foundation for International Scholarly Exchange under the project "Building a Diachronic Language Knowledge-base" (RG013-D-09).

References

1. Yu, S.W., Duan, H.M., Zhu, X.F., Swen, B., Chang, B.B.: Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation. Journal of Chinese Language and Computing 13(2), 121–158 (2003) (in Chinese)
2. Segmentation Principle for Chinese Language Processing (CNS14366). National Bureau of Standard, Taiwan (1999) (in Chinese)
3. Wei, P.C., Thompson, P.M., Liu, C.H., Huang, C.R., Sun, C.F.: Historical Corpora for Synchronic and Diachronic Linguistics Studies. International Journal of Computational Linguistics & Chinese Language Processing 2(1), 131–145 (1997) (in Chinese)
4. Liu, Y., Tan, Q., Shen, X.K.: Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology (GB13715). Qinghua University Press, Beijing (1994) (in Chinese)
5. Lu, Q., Chan, S.T., Xu, R.F., Chiu, T.S., Li, B.L., Yu, S.W.: A Unicode based Adaptive Segmentor. Journal of Chinese Language and Computing 14(3), 221–234 (2004)

A Survey on the Adjective in Learner's Dictionary

Aili Zhou and Shiyong Kang

The School of Chinese Language and Literature, Ludong University, 264025 Yantai, China
{lovely-zhou,kangsy64}@163.com

Abstract. The column of adjective synonyms in the learner's dictionary, not only helps learners to master a variety of synonymous expressions, but also helps to expand the vocabulary of the learners. It is worth noticing that if a synonym is listed in the microscopic structure, the macro-structure is supposed to list the word, which is called the closed-loop principle. That will make learners track the synonyms in process of studying. This survey extracts some adjective words and their synonyms to build corpus, using the method of quantitative and qualitative analysis, making a research on the words of reception, part of speech tagging and the closed-loop principle based on the Dictionary of Chinese Usage 8000 Words, in an attempt to provide reference with the revision and compilation of learner's dictionary.

Keywords: learner's dictionary, adjective, synonym, part of speech, closed-loop principle.

1 Introduction

Teaching Chinese as a Foreign Language (TCFL) in China has been two thousand years of history. Nevertheless, the learner's dictionary for foreigners started late, which began about the Ming and Qing Dynasties. At that time, a lot of foreign missionaries came to China to teach the doctrine. In order to communicate with others, they had to learn Chinese and began to compile a series of dictionaries. Until the early 20th century, the main form of learner's dictionary was still Chinese-English dictionary. After the founding of new China, teaching Chinese as a foreign language and the lexicography of learner's dictionary for foreigners have ushered in a new situation. A series of diverse types of learner's dictionaries sprang up, promoting the development of lexicography career.

With the improvement of Chinese comprehensive national power and international status, more and more foreigners put energy into learning Chinese. So the TCFL sector also presents an excellent situation. To those foreigners who have no obvious language advantage, owning a high-quality learner's dictionary becomes their urgent needs. After observing the various versions of learner's dictionaries which emerged in recent years, we can easily find the original and innovative one is rare.

In the long-term practice of codification, we have gradually worked out a series of principles and lexicography methods[1]. Are the theories of lexicography really effectively applied to the codification of the learner's dictionary practice? We intend to

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 294–301, 2013.

examine a learner's dictionary, extracting a part of adjective, in order to make a survey on the received words, part of speech tagging and the closed-loop principle respectively. And then we analyze the results, in order to provide some references for the revision and codification of the learner's dictionary.

2 The Construction of Corpus

This survey extracts a part of the synonyms of the adjective (A—H) from the "the Dictionary of Chinese Usage 8000 Words" (hereinafter referred to as "the 8000 Words Dictionary") to establish the closed domain of synonyms group.

There are some considerations that we select "the 8000 Words Dictionary" to make this survey:

1. The dictionary is compiled and published by the HSK Center of Beijing Language and Culture University, which is based on the vocabulary outline of the Chinese Proficiency Test (HSK) – the level of Chinese vocabulary and character grade outline, receiving a total of A, B, C D four commonly used words of 8822[2] .It is characterized with a high scientific and authoritative feature.

2. The dictionary is a comprehensive reference book which reflects the vocabulary and grammar points that the Chinese learners should master. The compilation principles are scientific, normative and practical. One of the most important features of the dictionary is to label the commonly used level of the words. That is to say, the words in this dictionary are more abundant than other existing learner's dictionaries of teaching Chinese as a foreign language. So, when we study the words of reception, part of speech and the closed-loop principle, the systematic character will be superior.

Some adjectives in the dictionary have multiple synonyms to correspond, so we list each group of synonyms separately. For example, "chengken (sincere)-chengzhi (sincere), zhencheng (sincere), chengshi (honest), kenqie (sincere)", we list them as four groups of synonyms, "chengken (sincere)-zhencheng (sincere), chengken (sincere)-chengzhi (sincere), chengken (sincere)-chengshi (honest), chengken (sincere)-kenqie (sincere)". In this investigation, we call the "chengken (sincere)" as the preceding item (detonated as "a"), while its synonyms "zhencheng (sincere), chengzhi (sincere), chengshi (honest), kenqie (sincere)" are called the latter item (detonated as "b"). According to the statistics, 630 groups of synonyms are our investigation subject. There are many adjectives in the formation of synonyms group is one-to-multiple relationship, a total of 611 valid adjectives enter our survey.

In this thesis, the focus of investigation is the closed-loop principle of synonyms. But in the course of the investigation, we find there are a few words as the latter items of synonyms don't appear in the form of the head word in the 8000 Words Dictionary, even also don't appear in the "the Modern Chinese Dictionary" (the 5th edition) (hereinafter referred to as "the Modern Chinese"). So we make a survey on the words reception on the basis of the 630 synonym groups. We also find that the POS of preceding and latter item of many synonyms are not consistent. Therefore, we also compared the part of speech of 611 valid words between the two dictionaries.

3 Investigation and Analysis

3.1 The Investigation and Analysis of Received Words

The words that have the same or similar meaning are called synonym. The synonym can make the expression more precise and rigorous. It makes the stylistic feature more distinctive. It also can make the sentences more vivid and varied. The last but not the least, the synonyms used in conjunction can strengthen the language potential, and enrich the semantics [3]. It can be seen that they play an important role in communication and expression, which is a big challenge for Chinese learners. Moreover, the meaning of synonym is similar, but that doesn't mean the same usage characteristics. There are some minor differences among them, such as rational meaning, additional color, which is even more difficult for Chinese learners during they learn synonyms. Therefore, during the compilation, the learner's dictionary must consider the basic and comprehensive characteristics in receiving words.

On reception of words, "the 8000 Words Dictionary" takes the Chinese proficiency test vocabulary guideline as the basis, receiving a total of 8822 words which are classified as A, B, C and D four commonly used words, and covering the basic needs of daily communication. As the authority of the inward language dictionary, the Modern Chinese settles a wide range of words, receiving a large number of words, so the reception of words between these two dictionaries is not too large. By analyzing 611 adjectives, only seven words appear in the 8000 Words Dictionary as the latter item of synonyms, but they don't find in the 8000 Words Dictionary as the head word. Also, the corresponding entries are not found in the Modern Chinese. They are: jibo (barren), beipo (force), tingzhi (straight), xianghe (consistency), butuo (inappropriate), yuanyou (original), fanchou (bothersome). The remaining words are found at least more than one dictionary.

We have to reflect on this phenomenon deeply. The words which have not been included may have two reasons: different types of dictionaries and different source of words. Therefore, in the revision of the two dictionaries, it is recommended to use the appropriate diverse types of corpus, and the method of word frequency statistics, in order to control the words of reception more scientifically and rationally.

3.2 Investigation and Analysis on the Part of Speech Tagging

"The Modern Chinese" has made it clear on tagging part of speech in its preface. It points out that, "the word entries that can become words are tagged the parts of speech, while those cannot become words, such as morphemes or characters, are not tagged the parts of speech."[4] This principle about tagging part of speech is also positive for externally oriented learner's dictionary.

Synonym group means the preceding and the latter item must be words. But, during making an analysis in our survey, we find some morphemes or non-morpheme words are taken as headwords, labeling the parts of speech, and even list the corresponding synonyms. It may cause some adverse effects in teaching Chinese as a foreign language. For example, "hun(faint)" labeled as an adjective, Chinese learners

may create such a sentence: the lights are hun (faint). Virtually, it increases the burden of the distinction between morphemes and words in teaching. In this survey, there are 13 morphemes are annotated part of speech or listed synonyms, they are: e (evil), chi (late), han (cold), liang (good), mo (ink), xuan (black), chi (red), lie (inferior), xu (gently), huan (slowly), hun (faint), qi (extraordinary) and kuo (wide). It is worthy noticing that, as a learner's dictionary, the commercial press learner's dictionary of contemporary Chinese is excellent in the aspect of tagging part of speech. It distinguishes each prefix as a word or morpheme. The word will tag the parts of speech, while the morpheme is not tagged.

Theoretically, synonyms are word-word semantic field. Some words in certain context may form the phrase or idiom to correspond, but we must be prudent to list the phrase or idiom in the synonyms column. In this survey, there are 4 phrases or idioms which are annotated the parts of speech or listed as synonyms. They are: wusuo weiju (fearless), bushao (many), budeliao (extremely) and wuke naihe (helpless). The words "bushao (many)", "budeliao (extremely)" are annotated as adjectives in the 8000 Words Dictionary, while in the Modern Chinese dictionary they are labeled as the phrases.

In order to investigate the tagging of part of speech in the 8000 Words Dictionary, we make a comparison with the Modern Chinese. A total of 66 words are not consistent in the part of speech. The specific results can be seen in the Table 1 below. (Due to the paragraph limitations, we only list a part of data).

Table 1. The comparison between the 8000 Words Dictionary and the Modern Chinese Dictionary

headword	part of speech	synonym	part of speech
chengxin (deliberately)	adv.	guyi (intentionally)	adv.
dai (foolish)	adj.	fadai (trance)	v.
dai (foolish)	adj.	faleng (daze)	v.
feikuai (quick)	adj.	feisu (fast)	adv.
feikuai (quick)	adj.	huosu (fast)	adv.
gebie (individual)	adj.	fenbie (respectively)	adv.
gongtong (joint)	adj.	yiqi (together)	adv.
gongtong (joint)	adj.	yitong (together)	adv.
guodu (deuced)	adj.	guoyu (excessively)	adv.
guyi (intentionally)	adv.	youxin (deliberately)	adv.
guyi (intentionally)	adv.	chengxin (deliberately)	adv.
guyi (intentionally)	adv.	youyi (purposely)	adv.
haoduo (many)	adj.	haoxie (many)	num.
haoxie (many)	num.	haoduo (many)	num. / pro.
heshi (proper)	adj.	shihe (fit)	v.
huangmang (hurried)	adj.	jimang (hastily)	adv.
huangmiu (ridiculous)	adj.	cuowu (mistake)	n.

In the dictionary, the editors state that the set of synonym section is to make the dictionary user distinguish the subtle differences among synonyms and master the usage of these synonyms. Although the editors strive to do best in the reception of

synonyms as far as possible, and make the part of speech consistent between the headword and synonym. There will inevitably be few defects. Based on the corpus statistics, in terms of POS tagging, a total of 46 groups of synonyms are not consistent, which accounts for 7.3% of all the synonyms groups. In the 611 effective adjectives, inconsistent POS tagging between "the Modern Chinese" and "the 8000 Words Dictionary" are 66, accounting for 10.80% of the total.

In the case of POS tagging inconsistence, the inconsistencies of adjectives and verbs, adjectives and adverbs are the majority. Former is 17, accounting for 25.76% of the total. The latter is 24, accounting for 36.36% of the total. In addition, such as "xuduo (many/much)", "haoxie (many/much)", "haoduo (many/much)" in "the Modern Chinese" are tagged as numeral or pronoun, while "the 8000 Words Dictionary" marked them as adjective.

3.3 Investigation and Analysis of the Closed-Loop Principle

"The macro-structure is a close system, which is the meaning of structure."[5] Every word in the whole dictionary should be seen in this structure. The microscopic structure is the specific information that is arranged systematically. Mr. Huang points out that the specific information includes the spelling or wording, phonetic transcription, part of speech, etymology, interpretation, the word cases, the special justice, encyclopedic information, phrases, idioms, synonyms, antonyms and so on. Of course, that is not to say every dictionary must contain all of the information listed in the above. As a matter of fact, the information will depend on the nature and size of the dictionary. However, we must bear in mind that the microstructure has the characteristic of stability, that means once the editor decides to provide certain information, and then he should keep consistent with all of entries. If some words are provided with certain micro-information, while some words are not, that may disrupt the stability and balance of the structure. That is to say, the information of the macro-structure and microstructure should be a mutual mapping relationship.

The closed-loop principle refers to the words used in the interpretation of the entry must be reflected in the macro-structure of the dictionary to build a closed network of the interpretation in the dictionary, making the users query cyclically. This principle applies not only to the interpretation and examples of the microscopic structure, but also the other microscopic information, such as synonyms, antonyms. In other words, the synonyms of the microscopic structure should appear in the macro-structure to reflect the mapping and closed-loop relationship of macro-structure and microstructure in the dictionary. Based on the above considerations, we make a survey on the closed loop principle of adjective synonyms group.

In 630 groups of synonyms, the latter item as the microscopic information appear in the macro-structure, namely as the lexical entries are 182, accounting for 28.89% of the total. This is the perfect closed-loop nature of our survey, such as, the synonyms "canbao (brutal)" is "canku (cruel)", while in the entry word "canku (cruel)", its synonyms appear the word "canbao (brutal)". There are some similar examples, beiai (sad)-beishang (sad), chengshi (honest)-chengken (sincere), gudan (alone)-gudu (alone) and so on.

There are 224 synonyms groups which preceding items appear latter items, while the latter items are not received in the dictionary. This is the typical case of non-closed-loop in the dictionary, accounting for as high as 35.56%. The remaining 224 groups of synonyms are not one-to-one correspondent. 186 synonyms groups are the case: the preceding item (a) lists b, while the synonym column of b has not listed a, accounting for 83.04%. The remaining 38 synonyms groups are: the synonym column of a lists b, while b has no the synonym column. The specific survey results can been seen in table 2. (Due to the paragraph limitations, we only list a part of data)

It is noteworthy that, "buxing (unfortunately)-xingyun (lucky)", "buxing (unfortunately)-zouyun (lucky)" are treated as synonyms, this should be proofread mistakes, we have excluded them in the survey.

Table 2. B has no synonym's column to correspond to a in the 8000 Words Dictionary

a	b	a	b
anjing (quiet)	ningjing (quiet)	gan (sweet)	tian (sweet)
anwen (stable)	pingwen (stable)	ganjing (totally)	guang (solely)
anggui (expensive)	gui (expensive)	ganzao (dry)	gan (dry)
ci (female)	mu (female)	guding (fixed)	wending (stable)
danchun (simple)	chun (pure)	guguai (strange)	qiguai (strange)
elie (evil)	dilie (inferior)	hanhu (vague)	yinyue (ambiguous)
ewai (extra)	lingwai (another)	hao (great)	miao (indigenous)
fanmang (busy)	manglu (busy)	hefa (legal)	fading (statutory)
fanmen (boredom)	fannao (worry)	heshi (suitable)	shidang (proper)
fangbian (convenient)	shiyi (proper)	hen (ruthless)	du (poisonous)
fennu (anger)	qifen (anger)	hongda (grand)	juda (huge)
fenhen (hate)	tonghen (angry)	hun (faint)	an (dark)

As can be seen by the above data, the closed-loop principle of the synonym group in the 8000 Words Dictionary is not satisfactory. In order to master a variety of ways of the synonymous expression, the learners will find the synonyms by the dictionary. But when they go in for some subtle differences between synonyms, they may not find the correspondent entry word, or even if they refer to the appropriate interpretation of information, but the synonyms are not listed in the column of that adjective. All of these conditions will have certain adverse effects for the learners.

From the point of view of lexicography, the mapping relationship between macro-structure and the micro-structure is an important manifestation of system and closed-loop in a dictionary. From the above data, it does not arouse the enough attention of the lexicographers.

4 Conclusion

It can be seen from the above survey that the set of synonym section provides the learners with synonymous expressions to expand their vocabulary. But it also reflects the current situation of learner's dictionary compilation that the macro structure and micro-structure mapping is not so perfect. Many synonyms, which are listed in the microscopic structure, are not been reflected in the macro-structure.

This is due to the relevant characteristics of the learner's dictionary itself. The lexicography of learner's dictionary aims to allow learners to master a language. While the users have different characteristics, so the dictionary is therefore divided into primary, intermediate and advanced level. A learner's dictionary can not cover all Chinese vocabulary. Therefore, on the closed-loop principle, the correspondence may happen between macro-structure and micro-structure. Many domestic scholars have begun to try to limit the interpretation of the words and attempt to identify the basic word to solve the above problems.

In fact, as early as in 1920s, the research on basic vocabulary has been carried out in many foreign countries. The Longman dictionary of contemporary English is an outstanding representative of contemporary dictionary in the field of definition with limited words. Since its publication in 1978, the dictionary has revised several times, but it has been adhering to the principles that the explanation must be interpreted by 2000 commonly used words. The basic vocabulary which is determined by the method of survey analysis is not only able to explain the meaning of words concisely, but also make the learners grasp the meaning and usage of the new words more easily. More importantly, to a large extent, using the basic vocabulary can maintain a systematic, closed-loop nature of the dictionary.

There are some achievements on basic vocabulary in domestic, such as "the Basic Words of the Modern Chinese Interpretation". But the definitional word extraction in Chinese dictionary starts late, a lot of theoretical research has not been used in the practice of dictionary compilation[6]. From the above passage, we know that the learners can use the collection of basic words to extend their vocabulary and to learn more common words to communicate. Therefore, it will be profound to identify a set of scientific, systematic basic-words in lexicography. The learners can use them to understand the interpretation of lexical entries and synonyms, antonyms, and adopt the method of associative memory to remember new words. Also, they can take advantage of the pooling of such words to extend their vocabulary, making it easier to refer to the dictionary. Of course, this is a very difficult and voluminous job, and we need to push on more deep studies and surveys.

Part of speech is one of the most important grammatical information in the dictionary, which directly affects the use of the word. But from the survey results, we can see, as a far-reaching learner's dictionary, "the 8000 Words Dictionary" needs to do a better modification work on the part of speech processing, in order to fit the true grammatical features of words.

This thesis only extracts the synonyms of parts of the adjective in the dictionary to investigate. The purpose of the investigation is not to accuse the flaws of the dictionary, on the contrary, to make the careers of teaching Chinese as the second language develop more vigorously. However, due to our shallow knowledge, the describing of the survey are the dominant, the analysis is also not so sufficient, and the theoretical level is limited. We are welcome the community of experts and scholars to criticize and correct.

References

1. Zhang, Y.H., Yong, H.M.: Contemporary Lexicography. Commercial Press, Beijing (2007)
2. Liu, L.L.: The Dictionary of Chinese Usage of 8000 Words (HSK Chinese Proficiency Test Vocabulary Guideline). Beijing Language and Culture University Press, Beijing (2000)
3. Huang, B.R., Liao, X.D.: Modern Chinese (the fourth addition edition). Higher Education Press, Beijing (2007)
4. Institute of Chinese Academy of Social Sciences, Room of Dictionary Editing. The Modern Chinese Dictionary (the fifth edition). Commercial Press, Beijing (2005)
5. Huang, J.H.: On Dictionaries. Shanghai Lexicographical Publishing House, Shanghai (2001)
6. An, H.L.: The Primitive Word Study of Modern Chinese Interpretation. China Social Sciences Press, Beijing (2005)

Chinese Idiom Knowledge Base for Chinese Information Processing

Lei Wang[1,2], Shiwen Yu[1], Xuefeng Zhu[1], and Yun Li[3]

[1] Key Laboratory of Computational Linguistics of Ministry of Education
[2] Department of English at Peking University Beijing 100871
[3] Institute of Linguistics, Chinese Academy of Social Sciences Beijing 100732
{wangleics,yusw}@pku.edu.cn, liyun@cass.org.cn

Abstract. In the vocabulary of Chinese language, idioms are a distinctive part for its fixed constitution grammatically and metaphorical meaning semantically. This paper introduces the entries and fields selected for the Chinese Idiom Knowledge Base (CIKB) built by the Institute of Computational Linguistics at Peking University (ICL/PKU) as well as its construction methods, research and applications so far. As a new member of the Comprehensive Language Knowledge Base (CLKB) constructed by ICL, our idiom knowledge base represents the new development of CLKB and will play a major role in NLP tasks such as machine translation, computer-aided translation, cross-language search, linguistic research, teaching Chinese as a foreign language and even as an electronic tool for preserving this non-material Chinese cultural and historical heritage.

Keywords: idiom knowledge base, Chinese information processing, construction and applications.

1 Introduction

Idioms are concise in structure, vivid in expression and profound in meaning. For idioms usually crystallize a nation's history, society and culture, they serve as the most outstanding part of a national language. A Chinese idiom refers to the expression of fixed compositionality formed in the long period of time of usage in the language of Han nation in China. Its construction is more complex than a word but it usually serves the function of a word syntactically. Chinese idioms are generally consisted of four characters, which enable them to be in perfect agreement with the Chinese way of expression. Meanwhile, Chinese idioms are often closely related with the fables, stories or historical accounts from which they were originally born and usually have a history of thousands of years and have always been regarded as national treasure. Moreover, characteristic of Chinese language, Chinese idioms have a quite large number and high frequency of usage.

Since fixed expressions represented by idioms play an important role in Chinese, research on idiomatic expressions, proverbs and set phrases will become valuable to the development of language teaching and culture studies[1], lexicography[2], Natural Language Processing (NLP)[3-5], and so on. A large-scale, high quality, lexically accurate language knowledge base of idiomatic expressions is especially in bad need.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 302–310, 2013.

Professor Shiwen Yu remarks, "With a frequent appearance, understanding idioms, including correct translation, is an indispensible part of text understanding. Though large in number, they have a limit; though difficult to understand, they can be found in most lexicons. If a knowledge base for idioms can be built, understanding idioms will not be that difficult." It is to realize the importance of idioms that he proposes an idiom knowledge base be constructed. In 2004 his idea was practiced in a grant from the 973 National Basic Research Program of China (No. 2004CB318102).

2 A Brief History of CIKB and Its Content

2.1 The History of CIKB

For the necessity of a language knowledge base in an NLP system, its scale and quality determine the success of the system. For the importance of idioms in Chinese language and culture, an idiom bank with about 6,790 entries were included in the most influential Chinese language knowledge base – the Grammatical Knowledge Base of Contemporary Chinese (GKB)[6] completed by the Institute of Computational Linguistics at Peking University (ICL/PKU), which has been working on language resources for over 20 years and building many knowledge bases on Chinese language. Later these knowledge bases are integrated into the Comprehensive Language Knowledge Base(CLKB)[7]. Based on it, the Chinese Idiom Knowledge Base (CIKB) had been constructed from the year 2004 to 2009 and collects more than 36, 000 idioms with more semantic and pragmatic properties added. The idioms are classified into three grades in terms of appearance in texts and complexity of annotation. The most commonly used 3,458 idioms serve as the Core idioms according to the statistics obtained from the corpus of People's Daily, a newspaper that has the largest circulation in China. Another 11,705 idioms are selected into a category named as Basic idioms (fully annotated in almost every field) and the total 38,117 forms the Complete knowledge base. Its hierarchical structure can be seen in Figure 1.

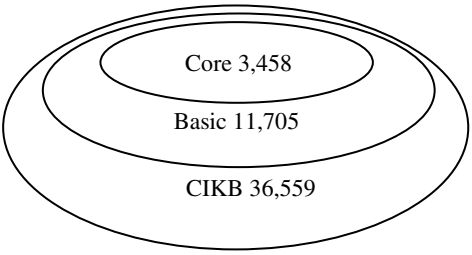


Fig. 1. The hierarchical structure of CIKB

2.2 Entries and Fields of CIKB

Zhou[8] believes that an important criterion to justify an idiom should be its "classicness", and proposes "Idioms should be originated from authoritative works such as

The Thirteen Scriptures, officially or privately written chronicles, masterpieces from Confucian works and collections; therefore they possess 'classicness'. Other idiomatic expressions, such as proverbs, *xiehouyu*[1], set phrases, do not possess 'classicness' since they are not authorized and mostly created colloquially and casually from unofficial expressions."

In spite of the fact that ancient classics serve as the sources of idioms, quite a number of idioms originate from contemporary popular works. From the historical perspective of language development, although most contemporary popular Chinese literary works are written in spoken language, this is mainly due to the historical transformation from ancient Chinese to modern Chinese in the 1920s, but not because their authors intended to abandon ancient Chinese and adopt modern spoken Chinese language. Therefore we believe that the idioms formed in those works also possess "classicness".

As we can see from the field "Era" in Table 1, Han Dynasty is the time when most Chinese idioms came into being. This also proves that notions such as "Han Language" and "Han Nation" are from that period. From the perspective of idiom

Table 1. Top 10 classic literary works that idioms appeared

ID	Name	No. of idioms	Era	Length(chars)	Proportion
1	Historical Records	1,567	Western Han(93 B.C.)	533,505	0.28%
2	History of the Han Dynasty	1,181	Eastern Han(83 A.D.)	742,298	0.15%
3	History of the Late Han Dynasty	1,067	Northern and Southern Dynasty(488 A. D.)	894,020	0.12%
4	Chronicle of Zuo	941	Late Spring-autumn Period(770 B.C.-476B.C.)	196,845	0.48%
5	History of the Jin Dynasty	872	Sui-Tang Dynasty(579-648A.D.)	1,158,126	0.07%
6	Chuang-tzu	847	Western Jin(265-316A.D.)	80,400	1.05%
7	The Analects	833	Warring States Period(475B.C. - 221B.C.)	21,683	3.84%
8	Mencius	698	Warring States Period(290 B.C.)	34,685	2.01%
9	The Book of Songs	634	Western Zhou(1046-771B.C.)	41,500	1.53%
10	The Book of Rites	625	Western Han(202B.C.-9A.D.)	99,010	0.63%

[1] A two-part allegorical saying, of which the first part, always stated, is descriptive, while the second part, sometimes unstated, carries the message.

generation, The Analects ranks top in terms of "classicness" in Chinese literary history, followed by Mencius, The Book of Songs and Chuang-tzu, which makes it obvious that Confucian tradition has great influence on Chinese culture.

150,850 characters are used in all the idioms in CIKB, with a total number of 5,103 different characters. We list the 10 most commonly-used characters and least-used characters in Table 2:

Table 2. 10 most commonly-used characters and 10 least-used characters of idioms

Most commonly-used	No. of times	Frequency	Least-used	No. of times	Frequency
不	3132	2.0762	傯	1	0.0007
之	1685	1.1170	驺	1	0.0007
一	1595	1.0573	诹	1	0.0007
无	1339	0.8876	菹	1	0.0007
人	1203	0.7975	诅	1	0.0007
心	1147	0.7604	攥	1	0.0007
天	887	0.5880	晬	1	0.0007
风	853	0.5655	撙	1	0.0007
如	708	0.4693	昨	1	0.0007
大	685	0.4541	薤	1	0.0007

In CIKB, 34,709 idioms with four characters account for 95% of the total, the largest proportion of all. This fact is in accordance with the linguists' notion as Lv and Zhu[9] believe that most Chinese idioms have four characters and are composed in a pattern of antithesis, such as远走高飞(yuan zou gao fei, fly high and go far) and摩肩接踵(mo jian jie zhong, shoulder to shoulder and closely upon heels). As for the selection of entries in CIKB, we find that there are idioms with two characters in some dictionaries such as 陵迟(ling chi, weak) and 横失(heng shi, to criticize boldly). In fact it is quite disputable whether these expressions meet the criteria of idioms. However, what they have in common is that they are rarely used in modern Chinese; therefore we do not include them in CIKB. It is rather difficult to make definite decision about three-character expressions because some dictionaries include them while others do not. Therefore there do not exist definite and widely-accepted standards[8]. Based on the principle of "classicness", we abandon expressions such as卷铺盖(juan pu gai, pack up and quit) and 耍嘴皮(shua zui pi, pay lip service) but keep those that have definite sources in our CIKB such as 借东风(jie dong feng, borrowing east wind) while considering their practical use (mostly frequency in corpora). Other than the four-character idioms, the distribution of idioms is seen in Figure 2:

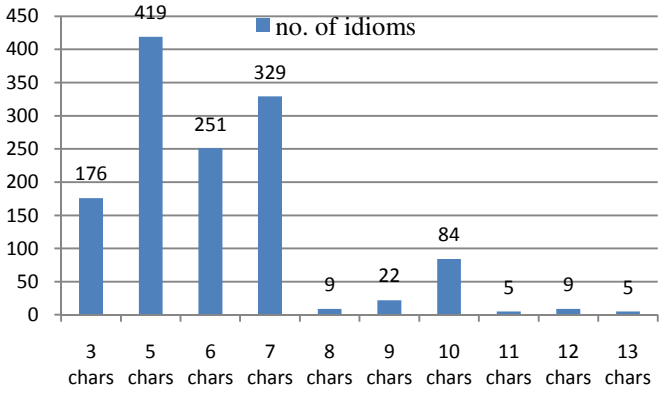


Fig. 2. No. of idioms other than four-character idioms

Basically the properties of each entry in CIKB can be classified into four categories: lexical, semantic, syntactic and pragmatic, each of which also includes several fields in its container -- the SQL database. Table 3 shows the details about the fields.

Table 3. Property categories of CIKB

Categories	Properties
Lexical	idiom, Pinyin[2], full Pinyin[3], bianti, explanation, origin
Semantic	synonym, antonym, literal translation, free translation, English equivalent
Syntactic	compositionality, syntactic function
Pragmatic	frequency, emotion, grade

There are three fields of translation as we can see in Table 3. In spite of the fact that a literal translation of an idiom will not reflect its metaphorical meaning generally, it will still be of value to those who expect to get familiar with the constituent characters and may want to connect its literal meaning with its metaphorical meaning, especially for those learners of Chinese as a foreign language. "Bianti" refers to a

[2] Pinyin (拼音, literally "phonetics", or more literally, "spelling sound" or "spelled sound"), or more formally Hanyu Pinyin (汉语拼音, Chinese Pinyin), is currently the most commonly used Romanization system for standard Mandarin. The system is now used in mainland China, Hong Kong, Macau, parts of Taiwan, Malaysia and Singapore to teach Mandarin Chinese and internationally to teach Mandarin as a second language. It is also often used to spell Chinese names in foreign publications and can be used to enter Chinese characters on computers and cell phones.

[3] full Pinyin, a form of Pinyin that replaces the tone marks with numbers 1 to 5 to indicate the five tones of Chinese characters for the convenience of computer processing.

variant form of the idiom that was caused by random misuse, literary malapropism, etc. For instance, 山盟海誓(shan meng hai shi) can also be written as 海誓山盟. The field "Frequency" is an integer that shows the frequency of the idiom in 55-year (1947-2002) People's Daily corpus. In the field of "Emotion", we define the emotion types as "appreciative (A)", "derogatory (D)" and "neutral (N)".

The syntactic category aims at NLP tasks like automatic identification or machine translation. Compared with English idioms, the identification of Chinese idioms is not so difficult for its fossilized structure, i.e. continuity in a text. To build a lexicon like CIKB will complete the task successfully. As for machine translation, however, it is completely another story because the compositional complexity of Chinese idioms enables them to function as different syntactic constituents with variable part-of-speech (POS). We classify them into nine categories according to its compositional relations of the morphemes and into seven categories according to its syntactic functions that they may serve in a sentence, as is shown in Table 4.

Table 4. Compositionality and syntactic functions of idioms

ID	Compositionality	Tag	ID	Syntactic function	Tag
1	modifier-head construction	pz	1	as a noun	IN
2	subject-predicate phrase	zw	2	as a verb	IV
3	coordination	bl	3	as an adjective	IA
4	predicate-object phrase	db	4	as a complement	IC
5	predicate-complement	dbu	5	as an adverbial	ID
6	predicate-object-complement	dbb	6	as a classifier	IB
7	serial verb	ld	7	as a modifier	IM
8	pivotal verb	jy			
9	repetition	fz			

3 The Construction Methods of CIKB

The main content of CIKB is composed mainly by human labor. In the process of construction, we refer to many authoritative idiom dictionaries, from which we select, sort and edit to make sure that the idioms we collect are commonly-used, complete, both representative and accurate. Since CIKB aims to natural language processing, in order to save human labor and acquire information more accurate, we also apply some NLP techniques and adopt a method of combining human labor and automatic processing.

Take the "English Equivalent" field as an example. At present there are three fields that provide translation information of the idioms in CIKB, i.e. "Literal Translation", "Free Translation" and "English Equivalent". The 11,705 idioms in the Basic base all have the "Literal Translation" and "Free Translation" fields while the "English Equivalent" is mostly empty. This is mainly due to the fact that to fill up this field will be both time-consuming and costly. In order to complete the missing information in this field automatically, we use an available electronic dictionary of English idioms to find the equivalents for the Chinese idioms, combined with human work of correcting and editing.

Our electronic dictionary has 2,717 entries, all of which have two fields: English idiom and Chinese translation. In the dictionary basically two translation methods are adopted: One is free translation, i.e. trying to translate the English idiom into a Chinese equivalent idiom. For instance, the English idiom "You cannot make a crab walk straight." , which is translated into the Chinese idiom "江山易改，本性难移。(jiang shan yi gai, bing xing nan yi. You can change a dynasty but not a personality.)". The other is literal translation. For instance, the English idiom "You cannot have two forenoons in the same day." does not have a Chinese equivalent; therefore it is translated directly into "一日之中不可能有两个上午。(yi ri zhi zhong bu ke neng you liang ge shang wu)"

Our goal is to automatically find the candidates for a Chinese idiom as the English equivalent idiom by the information from the electronic dictionary and recommend to the human editor responsible for the field to select the best one for the entry in CIKB. For the information of the electronic dictionary is quite limited, we have to rely on the Chinese translation information to compute its similarity to the relevant information of Chinese idioms, i.e. the idiom and its Chinese explanation. We adopt three algorithms: the number of same characters, edit distance and Longest Common String(LCS)[10].

For each entry in the Basic base, we match and compute its edit distance x with each entry of the electronic dictionary. Given the lengths of Chinese entry *IdiomC* (or Chinese explanation *IdiomCE*)and the translation of English idiom *IdiomET* as m and n, then the normalized formula to compute the score of the similarity between the two entries is:

$$Match(IdiomC/CE, IdiomET) = 1 - \frac{x}{max(m,n)} \tag{1}$$

When the two entries are the same, x=0 and the score is 1; when the two entries are totally different, the edit distance is $max\ (m,n)$ and the score is 0.

Given the number of the same characters is y, the length of LCS is z, the lengths of Chinese entry *IdiomC*(or Chinese explanation *IdiomCE*) and the translation of English idiom *IdiomET* as m and n, then the normalized formula to compute the score of the similarity between the two entries is:

$$Match(Idiom\ C/CE, IdiomET) = \frac{y(or\ z)}{min(m,n)} \tag{2}$$

Suppose *Match(IdiomC, IdiomE)* is the final comprehensive score of the three methods, the formula for computing it is:

$$Match(IdiomC, IdiomE) = \frac{w_1\delta\left(Match\left(\frac{IdiomC}{CE}, IdiomET\right)\right)+w_2\delta(Match(IdiomC/CE, IdiomET))+w_3\delta(Match(IdiomC/CE, IdiomET))}{w_1+w_2+w_3} \tag{3}$$

Where w_i (i=1, 2, 3) is the weights of the three methods and $\boldsymbol{\delta(x)} = \frac{1}{1+e^{-5(x-0.5)}}$ is the smoothing function. After computation, the match results of Chinese idioms and English idioms are seen in Table 5:

Table 5. Match results of Chinese idioms and English idioms

ID	Chn. idioms	Eng. idioms	Translation of Eng. idioms	score
1	山雨欲来	Coming events cast their shadows before them	山雨欲来风满楼	0.9241418
704	放任自流	Let things take their course	听其自然	0.9241418
344	沉默寡言	Better say nothing than nothing to the purpose	与其说话不中肯，不如一言不发好	0.7544851
65	公正无私	Business is business	公事公办	0.7544851
83	自暴自弃	All one's geese are swans/As the tree falls, so shall it lie/…	自吹自擂/自作自受/…	0.7544851/0.7544851/…
725	分文不取	Do not all you can, spend not all you have; believe not all you hear; and tell not all you know.	不要为所能为，不要花尽所有，不要全信所闻，不要言尽所知	0.7544851

We can see from the results in Table 4 that good matches are made when the score is greater than 0.9; while complexity is generated when the score is 0.7544851. For instance, although Chinese idioms No. 344 is different from No. 65, its explanation is quite close to the translation of English idiom No. 65 in form; therefore a good matching result is obtained. No. 83 matches more than an entry that is close to it in form but not related to it at all semantically. For idioms like No. 725 have many characters after translation, which increases the possibility of matching greatly, idioms as such match many Chinese idioms. In practice, we delete these entries from the match results.

4 Research and Applications Based on CIKB

As for research and applications, Li[11] completes the paper *Investigations on the Frequency and Distribution Consistence of Idioms in People's Daily Corpus*. In it the frequency, direct current frequency and alternating current frequency of idioms are thoroughly compared and analyzed in 55-year (1947-2002) People's Daily corpus and a diachronic curve is depicted in authentic text. Bai[10] completes her thesis for bachelor degree *Strategies and Implementation of the Auto-mapping between Chinese Idiom and Allusion Knowledge Base* based on CIKB and Allusion Knowledge Base constructed by Taiwan Yuanzhi University. Through this work we hope to establish a good platform for Chinese Information Processing and Chinese teaching, and thereby the content of CIKB is further enriched. Wang[12] selects 1,000 most commonly-used idioms based on the "Frequency" field of idioms and edit the book *Learning 1,000 Practical Chinese Idioms* for Teaching Chinese as a Foreign Language(TCFL). The author expects the book will enable Chinese learners to improve their level by using more idioms in their speaking and writing and become even more interested in Chinese long history and rich culture. Moreover, CLKB, in which CIKB serves as an important part, won the second-place National Award for Science and Technology Progress (Certificate No. 2011-J-220-2-02) in 2011.

5 Future Work

At present we are making further plans to conduct metaphor research of idioms and classification of idioms targeted to event depiction (i.e. the context of an idiom in use) based on a perfected CIKB. In this project we hope to add more linguistic features of idioms so as to better understand the influence of morphemes on expressions with fixed compositionality such as Chinese idioms.

Acknowledgements. Our work is supported by National Natural Science Foundation(Grant No. 61170163), Open Project foundation of National Key Laboratory of Pattern Recognition(No. 201001116), 863 Project(No. 2012AA011101) and Chiang Ching-kuo Foundation for International Scholarly Exchange(2009). The construction of CIKB is benefited from the work on fundamental resources of language data at ICL/PKU for many years. Special thanks should go to Ms Huiming Duan, Professor Fengxin Wang, Dr Zhimin Wang and many other contributors.

References

1. Shi, S.: A Study on Chinese Idioms. Sichuan People's Press, Sichuan (1979)
2. Lo, W.H.: Best Chinese Idioms, vol. 3. Hai Feng Publishing Co., Hong Kong (1997)
3. Cook, P., Fazly, A., Stevenson, S.: Pulling Their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Prague, pp. 41–48 (2007)
4. Fellbaum, C.: Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies (Research in Corpus and Discourse). Continuum International Publishing Group Ltd., London (2007)
5. Lin, D.K.: Automatic Identification of Noncompositional Phrases. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Maryland, pp. 317–324 (1999)
6. Yu, S.W., Zhu, X.F., Wang, H.: A Complete Specification of Grammatical Knowledge Base of Contemporary Chinese, 2nd edn. Tsinghua University Press, Beijing (1998)
7. Yu, S.W., Sui, Z.F., Zhu, X.F.: Comprehensive Language Knowledge Base and Its Prospect. Journal of Chinese Information Processing 25(6), 12–20 (2011)
8. Zhou, J.: On the Classicness of Idioms. Journal of Nankai University 1997(2), 29–35 (1997)
9. Lv, S.X., Zhu, D.X.: Speech on Grammar and Rhetoric. China Youth Press, Beijing (1979)
10. Bai, Y.: Strategies and Implementation of the Auto-mapping between Chinese Idiom and Allusion Knowledge Base. Thesis for Bachelor's Degree, pp. 5–20. Peking University Library (2011)
11. Li, Y.: Investigations on the Frequency and Distribution Consistency of Idioms in People's Daily Corpus. In: Proceedings of The 7th Chinese Language Semantics Workshop, pp. 264–270. National Chiao Tung University of Taiwan (2005)
12. Wang, L.: Learning 1,000 Practical Chinese Idioms. Peking University Press, Beijing (2011)

Developing Mongolian Phrase Information Resources

Dabhurbayar and Bayarmend

School of Mongolian Studies, Inner Mongolia University. 010021,
Hohhot, P.R. China
dabhvrbayar@163.com, mli@imu.edu.cn

Abstract. Developing modern Mongolian phrase information resources is an important research topic in Mongolian language information processing. The authors discussed the related issues to the phrase definition and classifications, formally analyzed modern Mongolian sentences in light of the theory of phrase structure grammar and developed three kinds of linguistic resources for the application and research on Mongolian language and scripts: modern Mongolian syntactic Treebank, phrase structure information database and phrase structure rule bank. The process of developing these resources and their structures are covered in detail in this paper. The authors concluded the current research and mentioned the follow-up research tasks in the end.

Keywords: Mongolian phrases, formal analyses, syntactic Treebank, phrase structure information database, phrase structure rule bank.

1 Introduction

Developing modern Mongolian phrase information resources is an important research topic in Mongolian language information processing. This research covers building a phrase-based modern Mongolian syntactic Treebank, making a regulation of tagging phrase information in the Treebank, developing a database of modern Mongolian phrase structure information and a Mongolian phrase structure rule bank in which the phrase structure types and their features are formally described. Mongolian syntactic Treebank and a regulation of tagging phrase information in the Treebank become the basic linguistic resources for the research and application of Mongolian language and scripts in the information age. These resources certainly promote the process of regulating and informationizing Mongolian language and scripts. The phrase structure information database and the phrase structure rule bank can directly support a series of applications in developing and implementing machine translation, intelligent input system, automatic proofreading in Mongolian language. It also enhances the development of sentence processing in Mongolian language information processing.

2 The Related Issues of Mongolian Phrase

There are three kinds of viewpoints about the definition of phrase in linguistics. (1) Phrase is a lexical-syntactic unit. It is composed of two or more content words in a

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 311–321, 2013.

sentence, but does not express a complete thought [1]. There are certain grammatical relations between the constituents of a phrase. In this kind of definition [2], there are two aspects: one is that phrase is similar with word in expressing certain concept and being an element in sentence-making. For example, *CAGAN T0L0GAI*[1] (alphabet), *GAJAR-VN JIRVG* (map); the other is that phrase, not as a lexical unit, is different from word in being a syntactic unit which consists of certain grammatical and semantic relations. For example, *JAHIDAL BICIHU* (write letters), *GILBAGAN GEREL SACVRAGSAN* (a flash of light shined); (2) Phrase is a kind of grammatical structure. At least, it includes one word, but does not include subject-predicate structure. In certain grammatical frame, single word can be a phrase which is the smallest example of a phrase. For example, single noun word can be taken as the smallest example of a noun phrase. (3) Phrase is a language unit consisting of two or more words which are combined in certain way [3]. Phrase is a number of words which can be a collocation grammatically and semantically. It has not any intonation like a sentence and is bigger than a word and cannot be a sentence [4].

There is not a strict standard for the number of words in a phrase in Natural language processing. Many books and papers argue that a single word can be a phrase [5]. According to phrase structure grammar, words constitute phrases and phrases constitute sentences. However, it is not hard to find that there are a lot of sentences consisting of only two words in fact. It is well-known that *S→NP VP* is a basic formula in phrase structure grammar. So, a single word can be a phrase in the phrase structure grammar. Actually, it is the result of the classification according to the function that a single word can be a phrase. For word and phrase, as a grammatical unit, the meaning of function is more important than the meaning of entity. The difference in form between word and phrase can be ignored in some times [6].

There are different viewpoints about the POS features of the words constituting a phrase[2] in many books and papers on Mongolian grammar. A few of them suggest that phrase should include the combination of content word and empty word [2], [7-8] while Most of them argue that it is only the combination of two or more content words [9-11]. In fact, a study of Mongolian language information processing should cover all kinds of combinations of words, for example, the combination of content words, the combination between content word and empty word, and the combination of empty words and so on. At present, we named the combination of words as *HELHICE* in Mongolian linguistics and Mongolian language information processing[3].

In our research, we define *HELHICE* as the combination of two or more words [7]. We argue that it can be a single word in certain formal grammar. We define *H0LB0G_A UGE* as the combination of two or more content words. According to such definitions, *H0LB0G_A UGE* is the part of *HELHICE*. In other words, it is a kind of *HELHICE*.

---

[1] It is the way of Romanization for the Mongolian language in our corpus.

[2] Generally, phrase is named as *H0LB0G_A UGE* in traditional Mongolian linguistics.

[3] We also use the same English word of "phrase" translate it into English. In this paper, phrase means *HELHICE* the combination of words.

There are two kinds of classifications for Mongolian phrases. One is by the structure of phrases; the other is by the function of phrases. The first one is the inward classification focusing on the structural types of phrases; the second one is the outward classification focusing on the functional types of phrases, which means that phrases are classified according to the ability of being the syntactic constituents in a bigger language unit. The former study of phrases in Mongolian linguistics seems to pay more attention to the structural classifications, for example, phrases are classified by structure as the form of coordination and the form of dominance; then, classify the form of dominance into the adverbial-predicate structure, the attribute-head structure and so on. In brief, the different classifications have the different target. Because of the different target, there appear the different standards of classifications. As the result, there will be the different classifications [4]. The target of classifying Mongolian phrases for language information processing is to provide the necessary knowledge of phrases for the practical software system of Mongolian language information processing.

We do some research on the classifications of the Mongolian phrases in order to extract and use the detailed information of the phrases as far as possible. On one hand, we classified the phrases on the same level according to the following features: (1) the POS feature of their head word, (2) the internal structural relations between the different components in the phrases, (3) the word number in the phrases; on the other hand, taking the noun phrases and the verb phrases for examples, we classified them on the different levels according to the different features [12-14]. In reference to the related research and the analyses of examples, we classified the Mongolian phrases into 12 classes according to the POS features of their head word: noun phrase(NP), verb phrase(VP), adjective phrase(AP), pronoun phrase(RP), numeral phrase(MP), measure word phrase(QP), spatial word phrase(OP), time word phrase(TP), adverb phrase(DP), postposition word phrase(GP), particle word phrase(SP), modal word phrase(XP). The structural relations in phrases are classified into 8 classes: subordinate(s), adverbial-predicate (b), object-predicate (t), subject-predicate (u), coordinate (h), attribute-head (d), parity (j), summarization (x). We also classified the phrases into 10 classes according to their word number. We developed a set of corresponding tags for these classifications. For example, the tag of “NP3d” means that it is a noun phrase which consists of 3 words and has the internal structural relations of attribute-head. In the Mongolian grammar, phrases are classified into free phrase and fixed phrase. Dr. Chengelt did a series of detailed research on the Mongolian fixed phrases in the past [15]. Hence, our research focuses on the free phrases in Mongolian language.

3 Formal Analyses of Phrases in Mongolian Sentences

Phrase structure grammar is one of the popular formal grammar theories in language information processing. In light of the theory of phrase structure grammar, phrases are

the immediate constituents of a sentence. We think phrase as a basic unit for the sentence analyses and analyzed a sentence by phrases step by step. It is possible to describe the result of the formal analyses of phrases in a sentence with some formal patterns like trees or brackets. We analyzed the phrases in the real Mongolian sentences using the theory of phrase structure grammar and discussed the process and methods in the manual analyses and the automatic analyses.

All the sentences are chosen from the modern Mongolian corpus. They can be divided into two classes of the context-sensitive sentences[4] and the context-free sentences according to their context information. The former is all the sentences with real context information in a junior textbook of Mongolian language and a novel of the Sun of the Spring Rises from Beijing, the latter is the result of automatically extracting the examples of sentences[5] by the length from the Corpus of Modern Mongolian Language. They are analyzed manually step by step as follows. Take the Mongolian sentence of *HOMOS-UN Ne2*[6] *NEYIGEM-UN Ne2 AHVI Ne2 B0L Sd HOMOS-UN Ne2 UJEL=SANAG_A-YI Yn SIIDBURILE/DEG Ve1 .* (People's social backgrounds decide their thoughts.) as an example:

(1) {HOMOS-UN Ne2 NEYIGEM-UN Ne2 AHVI Ne2 B0L Sd HOMOS-UN Ne2 UJEL=SANAG_A-YI Yn SIIDBURILE/DEG Ve1 .}S
(2) {{HOMOS-UN Ne2 NEYIGEM-UN Ne2 AHVI Ne2 B0L Sd HOMOS-UN Ne2 UJEL=SANAG_A-YI Yn SIIDBURILE/DEG Ve1}VPu .}S
(3) {{{HOMOS-UN Ne2 NEYIGEM-UN Ne2 AHVI Ne2 B0L Sd}NPs {HOMOS-UN Ne2 UJEL=SANAG_A-YI Yn SIIDBURILE/DEG Ve1}VPt}VPu .}S
(4) {{{{HOMOS-UN Ne2 NEYIGEM-UN Ne2 AHVI Ne2}NPd B0L Sd}NPs {HOMOS-UN Ne2 UJEL=SANAG_A-YI Yn SIIDBURILE/DEG Ve1}VPt}VPu .}S
(5) {{{{HOMOS-UN Ne2 {NEYIGEM-UN Ne2 AHVI Ne2}NPd}NPd B0L Sd}NPs {HOMOS-UN Ne2 UJEL=SANAG_A-YI Yn SIIDBURILE/DEG Ve1}VPt}VPu .}S
(6) {{{{HOMOS-UN Ne2 {NEYIGEM-UN Ne2 AHVI Ne2}NPd}NPd B0L Sd}NPs {{HOMOS-UN Ne2 UJEL=SANAG_A-YI Yn}NPd SIIDBURILE/DEG Ve1}VPt}VPu .}S

On one hand, we established the regulation of the formal analyses of the phrases in the sentences; on the other hand, we developed some computer aided software for the tasks of extracting and classifying the sentences, the automatic format-checking of the results of sentence analyses (including the brackets-checking and the spaces-checking). We split one phrase into two sections on one level according to its structure in a sentence. Hence, it is possible to develop some robust parsers in the Mongolian language analyses using the results of this manual analyses and the certain popular algorithm based on the bisections for natural language analyses.

[4] There are 1501 such sentences.
[5] There are the 14534 extracted examples of sentences. But, some of them are no correct sentence in Mongolian language. So, we need to proofread them and correct them one by one.
[6] It is the tag of part of speech for the Mongolian word of "HOMOS-UN (people's)".

4 Mongolian Syntactic Treebank and Some Related Software

The current state of the art in developing the syntactic information resources shows that the current syntactic Treebanks are grouped into two classes. One is the Treebank like U-Penn Treebank based on the analyses of phrases; the other is the Treebank like the PDT Treebank based on the analyses of dependency. Drawing lessons from the similar research and applications, we built the Mongolian syntactic Treebank (MTB) with a computer aided approach. MTB is based on phrase analyses according to the theory of phrase structure grammar. The thoughts of developing MTB are as follows. Firstly, we established the regulation of the formal analyses of the phrases for the manual analyses of the sentences. Secondly, under the restraint of this regulation, we analyzed and tagged all the phrases in 12710 real sentences. Thirdly, we designed and wrote the related programs for the format checking in the sentence analyses.

The MTB Treebank includes two kinds of sentences. One is the context-sensitive; the other is the context-free. The detailed information about the MTB Treebank is showed in the following table.

Table 1. The detailed information about the MTB Treebank

	classes			number	total
the MTB Treebank	The context-sensitive sentences			1501	12710
	The context-free sentences	Total sentences		11209	
		subsets	Sentence of 1 word(S1)	69	
			Sentence of 2 words(S2)	491	
			Sentence of 3 words(S3)	542	
			Sentence of 4 words(S4)	987	
			Sentence of 5 words(S5)	1461	
			Sentence of 6 words(S6)	1619	
			Sentence of 7 words(S7)	1870	
			Sentence of 8 words(S8)	1741	
			Sentence of 9 words(S9)	1463	
			Sentence of 10 words(S10)	966	

For the building and application of the MTB Treebank, we developed the application software in the MTB Treebank, the software for the phrases query and statistics and the software of extracting the information of Mongolian phrases. The software take an important role in the following aspects such as sentence extraction, sentence classification, tagging the number of words in a sentence, format-checking, extracting the structural rules, extracting the information of phrase structure. Take the following sentences from the MTB Treebank for the examples:

(1) {HVRDVLA Ve1 ! W!}S
(2) {{NAMVR-TVR Ne1 VYIDHARLA/BAI Ve2}VPb . W.}S
(3) {{{YAMAR Ra AMITAN Ne1}NPd IRE/BE Ve2}VPu ? W?}S

(4) {{{JARIM Ri NI Sf}NPs {EBEDCIN-IYER Ne2 UHU/DEG Ve2}VPt}VPu . W.}S
(5) {{{{EGEL Ac ARAD-VN Ne2}NPd {DEGJIREL-UN Ne2 TAGARCVG-TV Ne1}NPd}NPd BILHARA/G Ve2}VPt ! W!}S
(6) {{{ENE Rj B0L Sb}RPs {BAYARLA/GVSITAI Ve2 {{UJEGDEL Ne1 MON Sb}NPu GE/BE Vx}VPs}VPb}VPu . W.}S
(7) {{{JARIM Ri NI Sf}RPs {{{ORGEN Ac DALAI Ne1}NPd DEGER_E Oa}OPd {BEY_E-YI Ne1 NOGCIGE/DEG Ve2}VPt}VPb}VPu . W.}S
(8) {{{{HOMOS-UN Ne2 {NEYIGEM-UN Ne2 AHVI Ne2}NPd}NPd B0L Sd}NPs {{HOMOS-UN Ne2 UJEL=SANAG_A-YI Yn}NPd SIIDBURILE/DEG Ve1}VPt}VPu . W.}S
… …

5 A Database of Mongolian Phrase Structure Information

The information of phrase structure is one of the knowledge sources for the practical software system in Mongolian language information processing. It is of great importance for the research on the Mongolian language and the development of the practical software system in Mongolian language information processing to extract efficiently all kinds of information of phrases and store them in the form of easy computation. We developed a database of Mongolian phrase structure information by extracting the information from the MTB Treebank. For this task, we specifically designed and developed the software of extracting the information of Mongolian phrases. The extracted information of the phrases is stored in the form of the database of Microsoft Office Access 2007 and in the form of text separately for the follow-up research and application.

At the present, there are 68773 phrases with their information in this database. There are 13 kinds of information for each phrase like phrase, type, word number, structural relations, the first component, the second component, the POS feature of the first component, the POS feature of the second component, the morphological features of the first component, the morphological features of the second component, the subcategorization of the first component, the subcategorization of the second component, the frequency of the current phrase. In the form of text, each of the 13 information are expressed with the format of double quotation marks like “phrase”, “type”, “number” and so on. In the form of the database of Microsoft Office Access, there formed two dimensional concatenations between each phrase and the related information. The field names for the information of phrases are showed in the following table.

In addition, we classified all the 68773 phrases into 10 classes according to their length (word number of a phrase). Each class of them is stored in a separate database. It is of use to the research on the relations between the length of a sentence and the structural changes of phrases in a sentence.

Table 2. Field names in the database of Mongolian phrase structure information

Field name	meaning	Field name	meaning
ID	Order of a phrase	Pos1	the POS feature of the first component
Phrase	The current phrase	Pos2	the POS feature of the second component
Type	Type of a phrase	Morph1	the morphological features of the first component
Number	Word number in a phrase	Morph2	the morphological features of the second component
Relation	Structural relations of the two components in a phrase	Subcat1	the subcategorization of the first component
Comp1	The first component of a phrase	Subcat2	the subcategorization of the second component
Comp2	The second component of a phrase	Frequency	the frequency of the current phrase

Table 3. All the separate databases in the database of Mongolian phrase structure information

class	meaning	class	meaning
S1	A database including all the information of all the phrases in one-word sentence	S6	A database including all the information of all the phrases in six-word sentence
S2	A database including all the information of all the phrases in two-word sentence	S7	A database including all the information of all the phrases in seven-word sentence
S3	A database including all the information of all the phrases in three-word sentence	S8	A database including all the information of all the phrases in eight-word sentence
S4	A database including all the information of all the phrases in four-word sentence	S9	A database including all the information of all the phrases in nine-word sentence
S5	A database including all the information of all the phrases in five-word sentence	S10	A database including all the information of all the phrases in ten-word sentence

6 Mongolian Phrase Structure Rule Bank

The phrase structure rules in the phrase structure grammar shows the creativity of natural languages at the most. In general, it is necessary for the linguists to discover

and use these phrase structure rules from the phrases and their internal structural relations in a sentence. The phrase structure rules in accordance with the real features of the Mongolian language come from the real corpus of the Mongolian language. They are very useful for the development of the practical software systems in Mongolian language information processing like a parser.

Each sentence in the MTB Treebank is the result of analyzing formally the phrases in the sentence. It can be expressed in many different forms, for example, the form of brackets, the form of tree, and the form of rewriting rules. From what they expressed, they are same. In order to develop a rule bank of the Mongolian phrase structures, we designed the software for extracting the phrase structure rules from the MTB Treebank. The rules are expressed in the form of rewriting rules and stored in the form of text for the follow-up research and application.

The extractions of the rules took two kinds of format. One is with the related sentence while the other is not with the related sentence. The examples of the first format are as follows:

{{{{ILEGUU Ac HVDALDV/GSAN Ve1}VP2b AYIL Ne1}NP3d GAR/BA&V Ve2}VP4u ? W?}S4

S4→VP4u W?

VP4u→NP3d Ve2

NP3d→VP2b Ne1

VP2b→Ac Ve1

Ac→ILEGUU

Ve1→HVDALDV/GSAN

Ne1→AYIL

Ve2→GAR/BA&V

W? →?

The examples of the second format are as follows:

S4→VP4u W?

VP4u→NP3d Ve2

NP3d→VP2b Ne1

VP2b→Ac Ve1

Ac→ILEGUU

Ve1→HVDALDV/GSAN

Ne1→AYIL

Ve2→GAR/BA&V

W? →?

According to the relations between the rules and the vocabulary, the phrase structure rules can be classified into two classes in computational linguistics [16]. One is named as the structural rules; the other is named as the lexical rules. The examples of the structural rules are as follows:

⑴S5→VP5u W.
⑵VP5u→NP3s VP2b
⑶NP3s→NP2s Sh
⑷NP2s→Yn Sf
⑸VP2b→Ne2 Ve2

The examples of the lexical rules are as follows:

⑹Yn→DAGV=ANIR
⑺Sf→NI
⑻Sh→CV
⑼Ne2→GVTVRAL-TAI
⑽Ve2→S0N0SDA/N_A
⑾W.→.

We classified and stored the structural rules and the lexical rules according to the length of sentences. The result of statistics for two kinds of rules is showed in the following table.

Table 4. The statistics of the rules in the Mongolian phrase structure rule bank

Mongolian phrase structure rule bank					
The structural rules			The lexical rules		
name	meaning	quantity	name	meaning	quantity
R12-1	From one-word sentence	21	R12-2	From one-word sentence	60
R22-1	From two-word sentence	226	R22-2	From two-word sentence	719
R32-1	From three-word sentence	657	R32-2	From three-word sentence	1596
R42-1	From four-word sentence	1584	R42-2	From four-word sentence	3346
R52-1	From five-word sentence	2960	R52-2	From five-word sentence	5921
R62-1	From six-word sentence	4537	R62-2	From six-word sentence	8915
R72-1	From seven-word sentence	6950	R72-2	From seven-word sentence	12184
R82-1	From eight-word sentence	8792	R82-2	From eight-word sentence	15225
R92-1	From nine-word sentence	10566	R92-2	From nine-word sentence	17683
R102-1	From ten-word sentence	11815	R102-2	From ten-word sentence	19280

We also calculated the frequency of each rule in the rule bank and tagged it after the related rule. For example:

AP2u→Rb Ac 1
AP2b→Ac Ac 7
AP2b→Ac Sx 2
AP2b→Dq Ac 2
AP2b→Dq Ya 1
AP2b→Dx Ac 5
AP2b→H Ac 2
AP2b→J Ac 1
... ...

Such rule bank with the statistical frequency is very useful to the development of the statistics-based application software in Mongolian language information processing.

7 Conclusion

The modern Mongolian phrase information resources are the basic resources for the research and application of the Mongolian language and scripts. At the present, we developed three kinds of syntactic information resources with formal descriptions, that is, the Mongolian syntactic Treebank, the database of Mongolian phrase structure information and the Mongolian phrase structure rule bank. We are going to fulfill the following tasks in our follow-up research: (1) to expand the current resources; (2) to label the sentence constituents and the semantic roles; (3) to develop efficient and effective parsers for the syntactic-semantic analyses. All the results of this research will certainly provide the necessary knowledge for the many aspects of research and applications of the Mongolian language and scripts, for example, the development of Mongolian language resources, the regulations of Mongolian language and scripts, the teaching in Mongolian language, machine translation, automatic proofreading and so on. This research will certainly promote the research on the Mongolian language and the development of the Mongolian language information processing. It takes important roles in keeping the Mongolian language information processing in our country to stay on top in the similar research on the world.

Acknowledgement. This research is supported by NSSF (07CYY024) and the program of talents in science and technology in Inner Mongolia Autonomous Region (2012).

References

1. Gao, M.K., Shi, A.S.: An Introduction to Linguistics. The Inner Mongolia Education Publishing House, Hohhot (1983)
2. Gonchogsurung: An Introduction to Linguistics. The Inner Mongolia Culture Publishing House, Hohhot (1998)

3. Cai, F.Y., Guo, L.S.: A Dictionary of Language and Scripts. The Beijing Education Publishing House, Beijing (2001)
4. Huang, B.R., Liao, X.D.: Modern Chinese Language. The High Education Press, Beijing (2002)
5. Allen, J.: Natural Language Understanding. The Benjamin/Cummings Publishing Company, Inc., London (1995)
6. Zhan, W.D.: Determining Boundaries and Constructional Relations of Verb Phrases in Contemporary Chinese. Master Degree Thesis in Peking University, Beijing (1996)
7. Dabhurbayar: A Study of the Structural Rules of Mongolian Verb phrases for Language Information Processing. The Inner Mongolia People's Publishing House, Hohhot (2009)
8. Choijongjab: Some Issues of Mongolian Phrases, pp. 1–18. Journal of Inner Mongolia University, Hohhot (1963)
9. The Mongolian Language Institute in Inner Mongolia University: Modern Mongolian Language. The Inner Mongolia People's Publishing House, Hohhot (1964)
10. Poppe, N.: Grammar of Written Mongolian, Otto Harrassowitz, Germany (1954)
11. Chenggeltei: A Grammar of Modern Mongolian Language. The Inner Mongolia People's Publishing House, Hohhot (1999)
12. Dabhurbayar: A Computer-aided Approach for Mongolian Phrase Ambiguity Resolution. In: Xiao, G. (ed.) Recent Advance of Chinese Computing Technology, pp. 101–109. COLIPS Publications, Singapore (2007)
13. Dabhurbayar: Research on the Structural Rules of [X+BVP] Phrases in Mongolian Language. In: Altai Hakpo, vol. 18, The Altaic Society of Korea (2008)
14. Dabhurbayar, Bayarmend: Recognizing Basic Verb Phrases in Mongolian Corpus. In: ALTAI HAKPO, vol. 20, The Altaic Society of Korea (2010)
15. Chengelt: Grammatical Information Dictionary of the fixed phrases in Modern Mongolian Language. The Inner Mongolia Education Publishing House, Hohhot (2005)
16. Feng, Z.W.: On Potential Nature of Ambiguous Construction. Journal of Chinese Information Processing, 14–24 (1995)

Constructing Chinese Sentiment Lexicon Using Bilingual Information

Yan Su and Shoushan Li

Natural Language Processing Lab, Soochow University
1 Shizi Street, Suzhou, China 215006
{yansu.suda,shoushan.li}@gmail.com

Abstract. Currently, sentiment analysis has become a hot research topic in the natural language processing (NLP) field as it is highly valuable for many practical usages and theoretical studies. As a basic task in sentiment analysis, construction of sentiment lexicon aims to classify one word into positive, neutral or negative according to its sentiment orientation. However, when constructing a sentiment lexicon in Chinese, there are two major problems: 1) Chinese words are very ambiguous, which makes it hard to compute the sentiment orientation of a word; 2) Given the related research on sentiment analysis, available resource for constructing Chinese sentiment lexicons remains weak. Note that there are several corpus and lexicons in English sentiment analysis. In this study, we first use machine translation system with bilingual resources, i.e., English and Chinese information, then we get the sentiment orientation of Chinese words by computing the point-wise mutual information (PMI) values with English seed words. Experiment results from three domains demonstrate that the lexicon generated with our approach reaches an excellent precision and could cover domain information effectively.

Keywords: Sentiment Analysis, Sentiment Lexicon, Bilingual, PMI.

1 Introduction

The advancement of Web 2.0 technologies have led to the explosive growth of online opinion data. In order to automatically process these large-scale text information, sentiment analysis has recently received considerable interests in the Natural Language Processing (NLP) community [1-2]. In sentiment analysis, the task of sentiment lexicon construction is considered as a basic task which aims to detect whether a word is positive, negative or neutral, i.e., the semantic orientation of the word. This task is useful in many cases. First, a sentiment lexicon can provide important prior knowledge to improve the classification of higher-level text (document, sentence). For example, most unsupervised document-level sentiment classification methods are based on sentiment words [3]. Second, word-level sentiment analysis also has important significance for the word semantic understanding and disambiguation. As pointed by Wiebe [4], sentiment orientation can be associated with word's definition and contribute to the traditional word sense disambiguation task. Third, sentiment lexicon

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 322–331, 2013.

provides an important foundation for many real-life applications, such as text classification, automatic summarization, and text filtering.

However, up to now, no universally fine sentiment lexicon exists for Chinese sentiment analysis. Basically, it is difficult to build a good sentiment lexicon because many words are ambiguities when expressing the sentiment. That is to say, the polarities of words are sometimes sensitive to the topic domain or the context. Even worse, the same word may indicate different polarities with respect to different topic and context. For example, the Chinese word "圆滑" (yuan hua, slyness) has the meaning of "smooth" or "cunning". In sentence (1) as shown in the following, the word "圆滑" (yuan hua, slyness) is positive while being negative when appearing in sentence (2). In addition, traditional method of building sentiment lexicon is to expand by existing electronic dictionary or word knowledge base. Unfortunately, since the study on Chinese sentiment analysis research starts very late, the available resources on sentiment lexicon are extremely rare. Therefore, designing an efficient algorithm of constructing Chinese lexicon becomes a quite challenging and emergency task.

(1)：该笔记本电脑外形边角处理十分圆滑。(The notebook computer shape corner handling is very smooth.)

(2)：他变得圆滑，只能选择一再躲避现实。(He becomes cunning, and only selecst repeatedly to escape reality.)

In contrast, the study on English sentiment analysis starts much earlier and have got a number of related corpus and resources which could provid many seed words with correct sentimental categories. In this paper, we propose a novel method for computing Chinese word's sentiment orientation which combines bilingual corpus and English seed words to construct a Chinese sentiment lexicon. Our approach adopts machine translation system (Google Translate[1]) to eliminate the barriers between both Chinese and English languages. A source language and the corresponding translation language comment is seemed as an entire document, and then we compute point-wise mutual information (PMI) between each Chinese word and English positive (negative) seed words. Finally, we get a Chinese sentiment lexicon with semantic orientation value weights. Our method not only employs the polarity information in the English seed words, but also combines the bilingual constraint information in different context. Experiments across three domains show that our method could get a Chinese sentiment lexicon with a high precision and also could cover different context information.

The remainder of this paper is organized as follows. Section 2 overviews the related work in constructing sentiment lexicon. Section 3 presents our approach of combining bilingual resource and English seeds words to generate a Chinese sentiment lexicon. Section 4 evaluates the experimental results. Finally, Section 5 draws the conclusion and outlines the future work.

[1] www.google.com

2 Related Work

According to the granularities of the concerned text, the tasks of sentiment analysis can be broadly divided into three main groups: document-level [5-7], sentence-level [8], word-level [9-11]. Among these tasks, the word-level sentiment analysis has been considered as one basic task for sentiment analysis. The main objective of such tasks is to automatically construct a sentiment lexicon. Generally, the main methods for constructing a sentiment lexicon can be categorized into three types: (1) Use existing electronic dictionary or word knowledge data base to generate sentiment lexicon. In English study, the resource of Word Net is popularly used [12-13], while in Chinese study, the resource of HowNet is often used [14]. The main idea of this kind of approach is to find sentiment words in dictionary which have the similar semantics with the unknown word and then infer the sentiment orientation of the unknown words. However, this method cannot cover the context of the words. (2) Apply unsupervised machine learning method: the co-occurrence frequency is often employed in corpus to infer the word's close connection with some kind of polarity categories. Some representative papers include: [15-17]. For these approaches, the initial seed words play a central role in the success of constructing a high-precision sentiment lexicon. (3) Use human-annotated corpus: It inferred the sentiment tendency of one word according the co-occurrence relationships or semantic relations on the basis of annotate sentiment classification corpus. This kind of method needs lager amount manual annotated corpus. Some representative papers include [18-20].

Unlike all above studies, the proposed method in this study do not rely on the Chinese sentiment seed words, but make use of bilingual corpus and the English seed words when build a Chinese sentiment lexicon. This is a first attempt to construct a sentiment lexicon by using bilingual information. Our approach combines the English seed words and bilingual statistics resource to get a Chinese lexicon which could cover better contextual information in a specific domain.

3 Construction Method for Chinese Sentiment Lexicon with Bilingual Resources

3.1 Combine English Seed Words and Bilingual Constrain Information

Traditional methods of constructing a sentiment lexicon includes: one is to expand lexicon by existing electronic dictionary or word knowledge base, this method is obviously dependent on the number of Chinese seed words. The other is based on manual annotated corpus, this need large-scale Chinese corpus. However, Chinese sentiment analysis starts late, the available Chinese corpus and Chinese seed words resources are rather limited. In contrast, English sentiment analysis related resources are rich and easy to get. Therefore, in our approach, we collect bilingual corpus across

three domains and some English seed words to construct Chinese sentiment lexicon. Specifically, with the help of Google Translate, we translate the English comments into the Chinese comments and also translate the Chinese Comments into the English reviews. Therefore, the contents of each comment is consist of the source language text and the corresponding translated text, so Chinese sentiment words and English sentiment words with the same meaning will appear in a single document. Then we calculate the mutual information (PMI) between each unknown Chinese word and the English positive (or negative) seed words. Finally we get a lexicon with confidence weights according to the PMI value of positive (negative) category. The proposed method can take full advantage of the bilingual constraint information. The framework of the entire algorithm is illustrated in Figure 1.

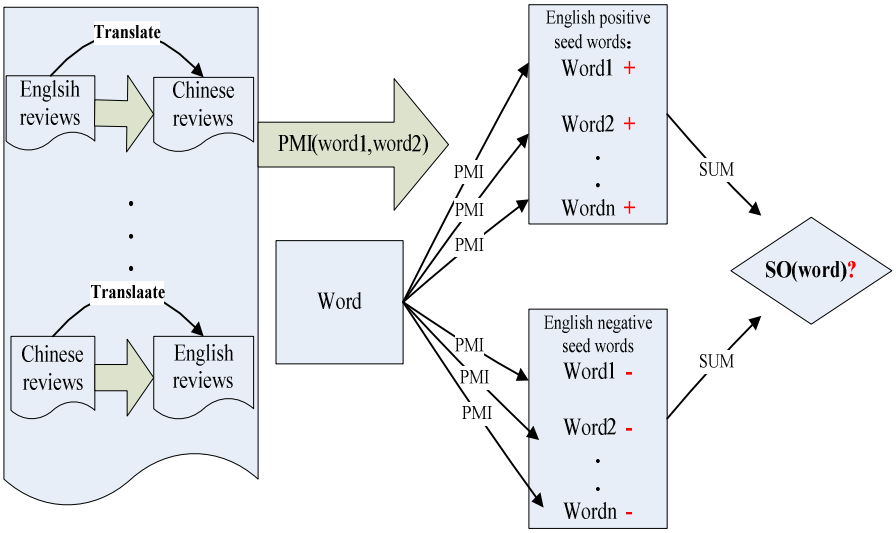


Fig. 1. The framework of Chinese sentiment lexicon construction based on bilingual resources Similarity Computation between Each Two Words

3.2 Similarity Computation between Each Two Words

The PMI-IR algorithm uses mutual information as a measure of the strength of semantic association between two words. In our approach, the point-wise mutual information (PMI) are utilized to compute the similarity between Chinese word and the English positive (or negative) seed words word. The PMI values from both the positive and negative sides can be used to determine the word's sentiment polarity. The computation formula of PMI is shown in Equation (1) [16]:

$$PMI(w_1, w_2) = \log_2 \left(\frac{p(w_1 \,\&\, w_2)}{p(w_1)p(w_2)} \right) \tag{1}$$

Where $p(w_1 \& w_2)$ denotes the co-occurrence probability of the two terms, and $p(w_1)$, $p(w_2)$ denote the occurrence probability of two words respectively. The log of this ratio is the amount of information that we acquire about the presence of one of the words in the same document. Here, co-occurrence means the two words appear in the same document.

Given the English seed words, we first calculate the PMI value of the Chinese word in corpus with all English positive seed words and then calculate the PMI value of the Chinese word with all English negative seed words. Finally, we utilize the discrepancy of two PMI values to determine the polarity of the Chinese word. The specific computation formula is given as follows:

$$SO(w_{Chinese}) = POS_j \times \left[\lambda \times \sum_{i=0}^{N+} PMI(w_{Chinese}, w^i_{English+}) - \sum_{k=0}^{N-} PMI(w_{Chinese}, w^k_{English-}) \right] \quad (2)$$

Where $w^i_{English+}$ is a word in the set of positive seed words in English, $w^k_{English-}$ is a word in the set of negative seed words in English. *N*+ is the size of English positive seed word set, and *N*- is the size of English negative seed word set. The parameter λ is related with the size of two seed word set. In our experiment, we find that the size of positive seed word set is smaller than the size of negative seed word. Thus, the parameter λ is set to be larger than 1, exactly to be 1.2 in this study. In addition, since the sentiment words are more likely to be adjectives and verbs, we set the parameters POS_j of these terms with corresponding priority. In particular, adjectives, verbs, and other word, corresponding priority is set to 3, 2, and 1 respectively. Assuming that the word "good", "excellent", "perfect" as positive seed words, 'bad "," poor "," disappointed "as negative seed words. More specifically, a Chinese word is assigned a numerical rating *SO* by taking the mutual information between the given context and the word see like “excellent” and subtracting the mutual information between the given context and the word set like “poor”. In addition to determining the direction of the word’s semantic orientation (positive or negative, based on the sign of the rating), this numerical rating also indicates the strength of the semantic orientation.

3.3 Our Algorithm

In our approach, we first utilize machine translation system translate Chinese (or English) comments into another language. Consequently, each comment is represented by two languages. Then, we use the English seed words to calculate PMI value between each unknown Chinese word and English positive (negative) seed words according to formula (1). Finally, formula (2) is employed to get a Chinese sentiment lexicon with polarity weight. The detailed algorithm is given in Figure .

Our Algorithm

Input:

English reviews U_{en}, Chinese reviews U_{cn};

English seed words L_{en};

Output:

A Chinese sentiment lexicon with polarity weight L_{cn};

Process:

1. Initialize the bilingual reviews $U = \varnothing$;
2. Translate every comment in U_{en} into Chinese comment, combine the English review and translated review as an entire bilingual comment, add the bilingual comment into U;
3. Translate every comment in U_{cn} into Chinese comment, combine the English review and translated review as an entire bilingual comment, add the bilingual comment into U;
4. In the feature vector set of U to calculate for each Chinese word between positive seed word and negative seed word by mutual information (PMI);
5. According to formula (2) to calculate the polarity of Chinese word, the symbol can represent the polarity of the word, while the absolute value represents the intensity.
6. Sort the lexicon according to the polarity strength of word;

4 Experiment

In this section, we systematically evaluate our approach on the data collection as described in Section 4.

4.1 Experimental Setting

The data collection contains three domains: Electronic, Beauty, and Software. In our approach, we collect both English corpus and Chinese corpus from http://www.amazon.cn/. Table 1 reports the distribution of English documents and Chinese documents in each domain. English seed words come from the lexicon provided by [21]. This lexicon contains 2000 positive and 4000 negative seed words. For the Chinese text, we use the Chinese text segmentation tool ICTCLAS to obtain the word list.

Table 1. The number of English and Chinese documents in three domains

	Electronic	Beauty	Software
English (positive)	2000	1000	1000
English (negative)	2000	1000	1000
Chinese (positive)	2000	1000	1000
Chinese (negative)	1850	1000	700

4.2 Experimental Results

Note that too many words exist in our corpus and manual annotating all the word is too time-consuming. Thus, in the preliminary experiment, we only get part of words with high scores for manual annotation.

In our approach, we adopt precision to evaluate the effect of classification, which is defined as follows:

$$Precision = \frac{number\ of\ correctly\ classified\ words}{total\ number\ of\ all\ labeled\ words} \tag{3}$$

Table 2 shows the precision of the Top 100, 200 words with highest scores and some instances of positive, negative Chinese word. From this table, the precision in top 100 words is very high and only a very small number of words are miss-predicted. In detail, there are 5, 3, and 5 words are not correctly predicted in Electronic, Beauty and Software respectively. As far as the top 200 words are concerned, the precisions of predicting the sentiment category in three domains are 84.5%, 83.5%, 89.0% respectively. The good performance of our approach is mainly due to its both utilizing the polarity of the English seed words and taking full advantage of the bilingual constraints information to sentiment word in the context environment. Therefore, the obtained Chinese sentiment lexicon not only covers the context information of the words in the special environment, but also achieves higher precision.

To check the results, we find that the errors mainly occur in the following two cases: 1) some words appear in the context of a particular sentimental category but contain no sentiment. For example, the Chinese word “造成”(zao cheng, cause) is neutral in the general sense, but it often appear in negative comment sentences. Our method depends on the context information of the words, which makes the word is predicted as a negative word. Another example is the word “加剧”(jia ju, exacerbate) which also often appear in a negative context. 2) Polarity shifting is another popular phenomenon to lead classification error for the polarity detection. For example, “帮助”(bang zhu, help) is positive in general sense, but it also appear in polarity shifting sentence like “There is not too much help for me”. As a result, our approach calculates the negative weight of the word will be much more possible than positive.

Table 2. Classification precision of the Top 100, 200 words with highest score and some positive, negative instances

	Electronic	Beauty	Software
Precision (Top 100)	95.0%	97.0%	95.0%
Precision (Top 200)	84.5%	83.5%	89.0%
Positive instances	柔和、清晰	精致、优雅	简洁、神奇
Negative instances	糟糕、退款	气愤、坏	浪费、沮丧

In addition, the framework we proposed is quite general and applicable for sentiment lexicon construction in any domains. It is capable of incorporating different sources of available information for the automatic construction of a domain- oriented sentiment lexicon. For instance, the Chinese word "圆滑"(yuan hua, slyness) has the same meaning of "smooth" and "cunning". It seems to be positive when is used to descript a computer, but it's considered to negative when decrypting a person. However, in our approach we make full advantage of the bilingual constraint information. "圆滑"(yuan hua, slyness) often appears in the same sentence with its translation "smooth" in electronic domain, so it is more likely to be inferred as a positive word. While the word often appears in a negative context with its translation words "cunning", so it is more likely to be inferred with a negative semantic orientation in this context.

Given an English seed lexicon, a simple method of building a Chinese lexicon might be to translate directly from English lexicon by machine translation systems. However, the direct translation has a lot of ambiguity, and cannot cover a lot of sentiment words in a specific domain. We calculate and compare the coverage ratio of top 100, 200 words with high score in the translated Chinese lexicon across three domains.

Table 3. Coverage ratio of top 100,200 words with high score in the translated Chinese lexicon across three domains

	Electronic	Beauty	Software
Coverage (100)	66.0%	60.0%	65.0%
Coverage (200)	57.0%	51.0%	54.0%

It can be seen from Table 3, the Chinese lexicon obtained by the direct translation has a low coverage of true sentiment words in corpus. Therefore, only use translation cannot get the emotional polarity of the many words in particular domain. Instead, our approach could provide a good supplement to capture many other Chinese sentiment words in a specific domain.

5 Conclusion

In this paper, we employ both the bilingual corpus and English seed words to construct a Chinese sentiment lexicon. In particular, we calculate PMI values between each unknown Chinese word and English positive (negative) seed words. The proposed framework is general and is applicable for lexicon construction in any domain.

It is capable of incorporating different sources of available information for the automatic construction of a context-aware sentiment lexicon. Experiment results from three domains demonstrate that the lexicon generated with our approach reach an excellent precision and could get many sentiment words in a special domain.

For the future work, we will consider the label of annotation corpus and polarity shifting phenomenon to improve the performance of sentiment lexicon construction. Furthermore, we plan to apply our approach to construct sentiment lexicon in other languages.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of EMNLP, pp. 79–86 (2002)
2. Li, S., Huang, C., Zong, C.: Multi-Domain Sentiment Classification with Classifier Combination. Journal of Computer Science and Technology, 25–33 (2011)
3. Kennedy, A., Inkpen, D.: Sentiment Classification of Movie Reviews using Contextual Valence Shifters. Computational Intelligence 22(2), 110–125 (2006)
4. Wiebe, J., Mihalcea, R.: Word Sense and Subjectivity. In: Proceeding of ACL-COLING, pp. 1065–1072 (2006)
5. Hatzivassiloglou, V., McKeown, K.: Predicting the Semantic Orientation of Adjectives. In: Proceedings of ACL, pp. 174–181 (1997)
6. Wiebe, J.: Learning Subjective Adjectives from Corpora. In: Proceedings of AAAI, pp. 735–740 (2000)
7. Cui, H., Mittal, V., Datar, M.: Comparative Experiments on Sentiment Classification for Online Product Reviews. In: Proceedings of AAAI, pp. 1265–1270 (2006)
8. Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In: Proceedings of ACL, pp. 271–278 (2004)
9. Kim, S., Hovy, E.: Determining the Sentiment of Opinions. In: Proceedings of COLING, pp. 1367–1373 (2004)
10. Li, S., Huang, C., Zhou, G., Lee, S.: Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In: Proceedings of ACL, pp. 414–423 (2010)
11. Li, S., Wang, Z., Zhou, G., Lee, S.: Semi-Supervised Learning for Imbalanced Sentiment Classification. In: Proceedings of IJCAI, pp. 1826–1831 (2011)
12. Andrea, E.: Determining the Semantic Orientation of Terms through Gloss Classification. In: Proceedings of CIKM, pp. 617–624 (2005)
13. Hassan, A., Radev, D.: Identifying Text Polarity Using Random Walks. In: Proceedings of ACL, pp. 395–403 (2010)
14. Zhu, Y., Min, J., Zhou, Y., Huang, X., Wu, L.: Semantic Orientation Computing Based on HowNet. Journal of Chinese Information Processing, 14–20 (2006)
15. Hatzivassiloglou, V., Wiebe, J.: Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In: Proceedings of ACL, pp. 299–304 (2000)
16. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL, pp. 417–424 (2002)
17. Popescu, A., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of HLT/EMNLP, pp. 339–346

18. Akkaya, C., Wiebe, J., Mihalcea, R.: Subjectivity Word Sense Disambiguation. In: Proceeding of EMNLP, pp. 190–199 (2009)
19. Li, S., Huang, C.: Word Sentiment Orientation Computing with Feature Selection Methods. In: CLSW (2009)
20. Lu, Y., Castellanos, M., Dayal, U., Zhai, C.: Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach. In: Proceedings of WWW, pp. 347–356 (2011)
21. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In: Proceedings of HLT/EMNLP, pp. 347–354

Constructing Chinese Opinion-Element Collocation Dataset Using Search Engine and Ontology

Tianfang Yao and Mosha Chen

Department of Computer Science and Engineering, Shanghai Jiao Tong University,
800 Dong Chuan Road, Shanghai 200240, China
yao-tf@cs.sjtu.edu.cn, cms914@gmail.com

Abstract. In this paper, we present a novel approach of constructing an opinion-element collocation dataset for Chinese language. The opinion-element collocation is a collocation whose composition words contain opinion/sentiment element. The dataset is useful for opinion mining task in many aspects. A search engine is used as a fundamental tool mainly because it could help us to seek both domain-specific and domain-independent collocation pairs, and at the same time, an ontology is used as a resource because it can offer rich semantic information to help us to classify collocations into domain-specific or domain-independent type. The tool and resource are combined to build a smart system that can automatically crawl data from the Internet and analyze extracted collocations. In order to ensure the quality of extracted collocations, we evaluate it manually. The experimental results on the COAE2008's public corpus have proved the success of this approach on the four domains.

Keywords: Opinion-Element Collocation, Search Engine, Ontology, Opinion Mining.

1 Introduction

Opinion mining is a hot research task in recent years [1-6], which deals with the traditional NLP research area but contains many advanced sentimental analysis technologies. According to Kim and Hovy's definition [1], an opinion contains four elements: claim, holder, topic and sentiment. In a claim, a holder (may be not existing) has sentimental comments on some topic (or more than one topic). Since the final goal of opinion mining is to get the sentimental polarity of a specified topic, and the polarity of a topic is determined by the sentiment that modifies it. Therefore, topic and sentiment play an essential role in opinion mining task and most of the recent work [1], [3], [4] and [7-10] focus on topic and sentiment, including both English and other languages.

A sentence (separated by a period) may contain multiple topics and sentiments, and sometimes a sub-sentence (separated by a comma) may also contain multiple topics and sentiments, depending on the different cases of a sentence. Therefore, it is necessary to point out which sentiment modifies which topic, which is also a key sub-task in opinion mining task. Among all the work involving the analysis of topic polarity,

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 332–343, 2013.

there is a methodology that analyzes topic and sentiment from the perspective of collocation [7] and [11]. A collocation is a pair of two words that co-exist in both verbal and written language, for example, when the word "rich" comes to you, the first thing that comes into your mind is some a Hartawan, like "Bill Gates". An opinion-element collocation is one that contains opinion element in its composition words, and in this research we limit its composition to be a topic word and a sentiment word, like "smart" and "system", they can form an opinion-element collocation. An opinion-element collocation is more suitable for opinion mining task because sometimes it could help determine and validate the correct matching of a given topic and its corresponding sentiment especially when there are multiple topics and sentiments in a sentence.

In this paper, a novel approach is applied to construct such an opinion-element collocation dataset. It is constructed in the form of a pipeline: firstly we use an online search engine to crawl and collect the intended data from the Internet; secondly, after data cleaning and preprocessing, the crawled data is transferred to the next stage to be analyzed by a dependency parsing tool Deparser[1]; finally, HowNet[2], functioning like WordNet[3], is used to mine the semantic information behind collocation pairs as well as classify the pairs into domain-specific and domain-independent ones. It is necessary that the system adopts human involvement and intervention, for example, human judgment is used to check whether the potential collocation extracted by Deparser is correct or not. Section 4 gives the details of the whole system. The system is established to construct an opinion-element collocation resource for more advanced research and is served as a basic component. This paper focuses on how to construct the system and some design details about the system, more application can be found in another research [11]. The experiments are conducted on the COAE2008 (Chinese Opinion Analysis Evaluation 2008)'s public corpus[4] to compare results before and after using the knowledge of opinion-element collocation and the experimental results have proved that the collocation dataset can help a lot in our approach.

The following of this paper is organized as follows: Section 2 talks about the related work. Section 3 introduces the tools and resource we utilize. Section 4 gives the overall architecture of the system and gets into the details of each component. Section 5 presents the experimental results. The conclusion is given in Section 6.

2 Related Work

There are many topics discussing about corpus construction in recent years, there is a trend that researchers begin to facilitate the online resource to conduct their experiments. For example, Wikipedia[5] is a good resource for constructing ontology, and

[1] http://ir.hit.edu.cn/demo/ltp
[2] http://www.keenage.com/
[3] http://wordnet.princeton.edu/
[4] http://ir-china.org.cn/coae2008.html
[5] http://en.wikipedia.org/

Google[6] is famous search engine for large-scale data retrieval. Many researchers have used them to conduct experiments and test their ideas on massive data. For Chinese, there are also many function-like online resources, like Sogou Lab[7] and Baidu[8], of which the former offers many useful corpora extracted from the Internet and the latter is a good search engine especially for Chinese language processing. These two researches [12] and [13] have shown that search engine can be a tool for related collocation researches. Therefore, Internet will play a more and more important role in both research and life.

3 Tools and Resource

In this section, we briefly introduce the tools and resource utilized in the system.

3.1 Search Engine

Here we mainly use search engine to get the search snippet for each search item in search result lists. Figure 1 gives a screenshot of a search result using Baidu. In the figure the content in green box represents snippet and the words in red mean matched keyword. The content in red box represents the URL where the search result comes from, the URL domain will be counted and the frequently existing ones will be kept for future use, i.e. they may be domain-specific sites.

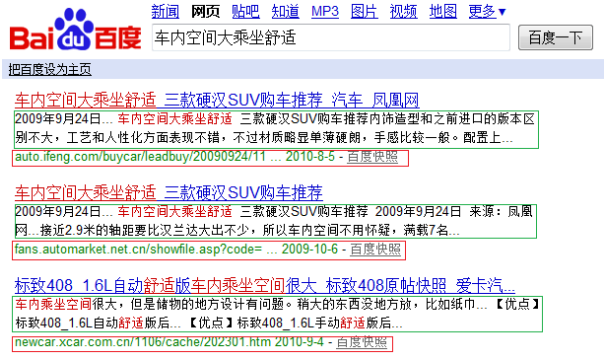


Fig. 1. Search result for "车内(chenei the inner of a car)空间(kongjian space)大(da large)乘坐(chengzuo ride)舒适(shushi comfortable) (the inner space of this car is large and it sits comfortable)"

The terminology "snippet" can be interpreted as the abstract for page content, which tries to summarize and cluster similar and related sentences for meeting the users' purpose. So it is reasonable that a snippet contains useful and meaningful

[6] http://www.google.com/
[7] http://www.sogou.com/labs/
[8] http://www.baidu.com/

information for given key words. In our research, we input a sentence from the testing corpus and want to get similar context, it is effective to use snippets for such a task.

3.2 Deparser

Deparser is a NLP tool that can be used for many Chinese NLP tasks, for example, word segmentation, POS(Part-of-Speech) tagging, dependency parsing and so on. In this section, the dependency parsing is introduced because we depend on this tool to extract potential opinion-element collocations. Its algorithm takes some strategies to extract the most probability modifying-modified dependency relations in a sentence, so it is also an intuitive idea to get the intended opinion-mining collocation from dependency pair plus the POS and dependency type information [8], if such opinion-element collocation does exist in the analyzed sentences and satisfies certain conditions [8]. Detailed algorithm is given in Section 4. Figure 2 gives a simple example of dependency result from which you could get a direct impression of how Deparser works.

Fig. 2. Dependency analysis result for “我(wo I)喜欢(xihuan like)游泳(youyong swimming) (I like swimming)”

In the example, the words “我” and “喜欢” form a SBV (subject-verb) dependency, the words “喜欢” and “游泳” form a VOB (verb-object) dependency. Thus, the words “喜欢” and “游泳” constitute an opinion-element collocation, because the word “喜欢” is a topic element and the word “游泳” is a sentiment element.

3.3 HowNet

HowNet is an online common-sense ontology unveiling inter-concept relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. It offers rich information for the lexicon, including some semantic information, like the different meaning of a word and the occasions when it is used in. Take the word “品尝(pinchang taste)” for example, it has the following property: DEF=attribute|属性(shuxing attribute),taste|味道(weidao taste),&edible|食物(shiwu food). The DEF property gives the concept and attributes of a word (lexicon). In the scenario “taste some foods”, the word “taste” plays as a verb; in another scenario “the taste is not bad”, however, the word “taste” plays as a noun. Another example is that

the word "doctor" and the word "patient" are belong to a more general concept "human being". We will take advantage of the hierarchy information supplied by HowNet to classify opinion-element collocation into domain-specific or domain-independent type.

4 System Introduction

This section gives the overall architecture of the system and goes into details for each component.

4.1 Architecture of the System

Figure 3 presents the architecture of the whole system. The system is divided into four parts: Crawler, Parser, Human Editor and Classify Component. Crawler is the component using search engine to get the snippet for each query. Parser is the component used to extract potential opinion-element collocations. Human Editor is for the manual correction or modification of opinion-element collocations, one more notice is that Human Editor is added in this system for quality issue. In our previous studies, it doesn't exist because we approximately regard the potential opinion-element collocations extracted by Deparser is correct. The classifier adopts HowNet to classify collocations into domain-specific or domain-independent type for further use. The interchange files are all formatted in self-defined XML schema.

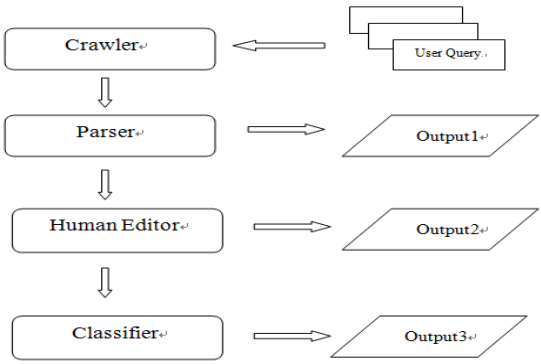


Fig. 3. System architecture

The system starts from user query, then it makes use of search engine to crawl and search snippet, after parsing, there are human editor and classifying processing, we get the final collocation dataset. After the Parser and Human Editor processing, there is also the output of collocations, compared to the final collocation, they are of different use.

4.2 Crawler

The crawler gets user input and sends the http request to the specified search engine using its public API, and then it processes the received data to get the snippet for each query. Also some data cleaning and preprocessing work should be completed in this component. Figure 4 shows a snapshot for this component. We choose Baidu as the default search engine because it is more professional in Chinese. The input sentences are selected directly from the corpus used for other advanced researches. We benefit from the snippet because it can return similar and related information except for the original query. As to data cleaning and preprocessing, two kinds of information are considered: replication and advertisement. Replica is caused mainly because the same article is reshipped by many sites and it is easy to kick off the replica by computing the similarity of the snippet. For ads, a list of common advertisement words, like "discount" and "for sale", is collected and we can discard the snippet if it hits too many obvious advertisements. How many snippets should we collect? By experience, we choose the first five search result pages as our sources for snippet because the quality for the search is not satisfied due to our observation on average. Of course this value can be modified in the configuration file.

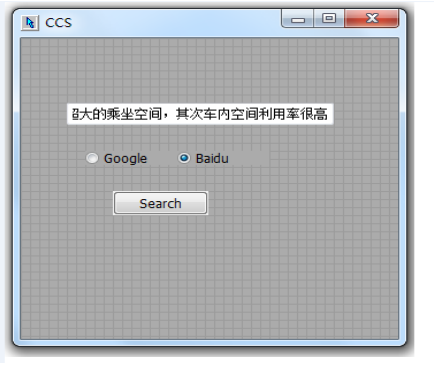


Fig. 4. Snapshot for the Crawler Component

4.3 Parser

The Parser component analyzes and extracts potential collocations from the output of the Crawler component. In the research work [8], we find three types of dependency relations that are important for opinion mining task, that is, SBV (subject-verb), VOB (verb-object) and ATT (attribute). For example, in the sentence "The boy is cute", the words "boy" and "cute" form an SBV dependency relation; in the sentence "I like swimming", the words "like" and "swimming" constitute a VOB dependency relation; in the phrase "beautiful pictures and delicious food", the words "beautiful" and "pictures" form an ATT relation. These three dependency types cover almost 90% of the opinion-element relations in the corpus we conduct experiments on, so it is reliable that we only consider these three types to extract opinion-element relations, which can be regarded as potential opinion-element collocations. Additionally, we examine

whether a collocation contains sentimental word in it and determine whether it is a potential opinion-element collocation. For efficiency issue, we can start multiple threads for the Parser component since it is the most time-consuming module in the whole system. Below is a simplified algorithm from the research [8] that they apply to extract potential collocations.

```
Input: Sentence S, Sentiment Dictionary SentDict
Output: Collocation Set ColSet
program ExtractPotentialOpinion Element Collocation:
  ColSet = {}
  DepRelationSet = Parse(S)
  Foreach DepRelation in DepRelationSet:
    If DepRelation in {SBV,ATT,VOB}:
        (word1, word2) = DepRelation
      // We assume the DepRelation is
      // composed by word1 and word2
        If word1 or word2 in SentDict:
        ColSet.Add(DepRelation)
```

4.4 Human Editor

This component is designed for human editing the result conducted by the Parser component because the opinion-element collocation extracted by Deparser may have errors especially when there are multiple topics and sentiments in a single sentence. So it is necessary to conquer this issue with a human editing process. Figure 5 gives a screenshot for this component.

Fig. 5. Snapshot for the Human Editor Component

In the figure, the top line is the extracted result from the Parser component. The words in red, that is, “空间(kongjian space)”, “乘坐(chengzuo ride)”, “表现(biaoxian representation)”, “材质(caizhi material quality)”, and “手感(shougan feel)”, represent topic and the words in yellow, namely, “大(da large)”, “舒适(shushi comfortable)”, “不错(bucuo well)”, “单薄硬朗(danboyinglang thin and tough)”, and “一般(yiban general)”, represent sentiment. The following table is the extracted potential opinion-element collocations, you can “edit”, “delete” or “save” a potential collocation, even you can “new” a collocation if it is missed.

4.5 Classifier

This component classifies extracted collocations into domain-specific or domain-independent type for further use. Intuitively speaking, the word "good" can almost modify any object, but some words are domain-specific, like "fuel-consuming", which are most probably specified to an engine. HowNet offers rich information for each lexicon it includes. For each collocation, we focus on both sentiment word and topic word, by examining the values given in the DEF property and the hierarchy information it offers, we can classify them to specified domains. As you could image, HowNet can't promise every word having its definition, so the ones that don't exist in HowNet are labeled as unknown. Due to our subsequence work, there are four domains: car, digital camera, PC and MP3, plus the domain-independent and unknown class, there are six classes in all. Randomly picking items in each domain-specific class, we find the classification effect is to our satisfaction.

5 Experiments

We conduct experiments from two perspectives, one is to test the system itself to prove that adopted approach is feasible and adaptive; the other is to use the collocations extracted from the system for further research to prove the collocation dataset is useful and valuable.

5.1 Corpus

The testing corpus is from COAE2008's public testing corpus[9], it is proper as a base to get more corpora from the Internet. The corpus has four domains, that is, car, mobile, PC and MP3. Each domain contains about 100 articles and each article contains 1 to 7 sentences. The article is considered to be of sentimental polarity, but the sentences it contains may be declarative sentences. The following is an example from the corpus: "哈弗(hafo Harvard)M2的(de)座椅(zuoyi seat)为(wei with)双色织布(shuangsezhibu two-yarn-dyed)面料(mianliao fabric)，与(yu with)内饰(neishi interior)整体(zhengti overall)色调(sediao tone)比较(bijiao more)协调(xietiao coordinated)，侧向支撑力(cexiangzhichengli lateral support force)不错(bucuo good)。驾驶员(jiashiyuan driver)座椅(zuoyi seat)为(wei can be)手动(shoudong manually)四向(sixiang four directions)调节(tiaojie adjust)，可(ke may)满足(manzu suitable)各种(gezhong different)身材(shencai stature)、体形(tixing bodily form)的(de)驾驶者(jiashizhe driver)，提供(tigong offer)最佳(zuijia best)驾驶(jiashi driving)姿态(zitai posture)。(Harvard M2 has seats with two-yarn-dyed fabric, which are more coordinated with the overall tone of the interior; the lateral support force is good. The driver's seat can be manually adjusted in four directions, which is suitable for the drivers of different stature and bodily form. It can offer the best driving posture)". The words in italics mean the sentiment words, you may notice that some sentences don't contain

[9] http://ir-china.org.cn/coae2008.html

any sentiment word at all, so most probably it is just a declaration. We will select the sentences that contain sentiment word to be the input queries for our system.

5.2 Experiments on System

To our knowledge, there is no similar work on the construction of collocation dataset before, so we try to evaluate and test from the following aspects:

Output/Input Rate. Here the "O/I Rate" is computed according to the following equation.

$$\text{O/I-Rate} = \frac{\#(\text{number of extracted collocations})}{\#(\text{number of input query})} \quad (1)$$

Table 1 shows the O/I Rate depending on different input query. The input query sentences are randomly selected from the COAE2008's public corpus and the extracted collocations are directly extracted after the Parsing component, in which no human intervention is involved. The rate decreases as the number of input query increases. It is because that the input sentences may come from the same article in the corpus, so the search result returned may be similar, therefore, there is duplication. But as you can image, this simple but novel approach could indeed get the intended collocations, 200 input query sentences could get as many as nearly 7000 collocation pairs, which is to our delights.

Table 1. O/I rate experiments results

#(input query)	#(extracted collocation)	O/I Rate
100	3567	35.67
200	6720	33.60
300	8340	27.80
400	10024	25.60
500	10943	21.87
Ave.		28.91

Human Involvement Effect. This criterion is used to evaluate how the human involvement affects, or in other words, we want to know how the Parser component works. Due to this is a time-consuming work, we just select 20 sentences to compare the results before and after human modification. We make statistic on #(edit number), #(delete number) and #(new number). Table 2 gives the result. As you could notice, the overall human effort affects 12% of the collocations extracted directly from the Parser component, so we think the results extracted by our system is reliable and the quality is ensured.

Table 2. Human validation results based on 20 queries

#(query)	#(collocation)	#(Edit)	#(New)	#(Delete)
20	711	42	17	27
Percentage		5.9%	2.3%	3.7%

Classification Effect. This result depends on the coverage fraction of HowNet. We extract collocations from 200 sentences in this section and classify the collocations into six classes (Section 4.5 gives the details). Altogether 6720 collocations and there are nearly 4000 collocations belong to the unlabeled class, the four domains(car, mobile phone, PC and MP3) counts about 1400. Randomly picking collocations from the four domains, we feel that the classification result is to our satisfaction. This is just a try, we can complete and mine further on this component, more semantic information should be taken into consideration.

5.3 Experiments on Collocations

We just list the result from one of our previous studies [11] to show that the collocations indeed could help in further research. Table 3 gives the result.

Table 3. Opinion-element relation extraction results

	P	R	F
Baseline1(Closest-pair)	51.60%	73.85%	79.60%
Baseline2(Parsing)	72.53%	85.99%	78.63%
Our method(Collocation)	73.92%	93.37%	82.47%

The experiment is conducted to extract the opinion-element relations that exist in a given opinioned sentence. The testing corpus is also the COAE2008's corpus. An opinion-element relation is a relation between a topic word and a sentiment word, the sentiment word modifies the sentiment word. In the Parser component, the extracted dependency relations are such relations if they are correct. From a more general level, an opinion-element relation is an opinion-element collocation, the former is specified to a sentence and the latter is oriented to the statistic of language usage. The collocation plays an important role in this experiment and you could see that the recall of our method obviously improves to the existing common method. The collocation dataset in this experiment could help seek the opinion-element relations whose composition words may apart from each other, perhaps exist in different sub-sentences, which is beyond the parser's ability to extract. It has proved the success application of the collocation dataset.

6 Conclusion

The motivation for this work depends on our studies for opinion mining: in the work [11], we extract opinion-element relations for Chinese and find it necessary to build such a collocation dataset to help get a better result, so we did. As a result, we have proved that our method and the collocation dataset play an import role in that work. Further, we think it can be extended to other uses, just list some: Chinese spelling check, for some domain-specific sentences, some words are rare and users who use Pinyin IME may get the wrong words printed on screen (the misspelled words have the same pronunciation with the correct one, but with different characters). For example, "蓝牙(lanya Bluetooth)" is often misspelled as "篮牙(lanya basket tooth)". If we get the context around the misspelled word and the content that can also be found in the domain-specified collocations, then we can correct such kinds of misspelling via some strategies. Hope we can benefit from this system in future research and some extensions can be made on this system to make it more powerful.

Acknowledgments. This research work is financially supported by the National Science Foundation of China (No. 60773087) and the UDS-SJTU Joint Research Lab for Language Technology. Besides, Information Retrieval Lab, Harbin Institute of Technology provides the Chinese syntactic analyzer LTP Deparser, and KEENAGE offers the HowNet platform. We sincerely thank them for their help.

References

1. Kim, S.M., Hovy, E.: Determining the Sentiment of Opinions. In: 20th International Conference on Computational Linguistics (COLING 2004), pp. 1367–1373. ACL, Stroudsburg (2004)
2. Matsumoto, S., Takamura, H., Okumura, M.: Sentiment Classification Using Word Subsequences and Dependency Sub-trees. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 301–311. Springer, Heidelberg (2005)
3. Popescu, A.M., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp. 339–346. ACL, Stroudsburg (2005)
4. Kim, S.M., Hovy, E.: Identifying and Analyzing Judgment Opinions. In: Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2006), pp. 200–207. ACL, Stroudsburg (2006)
5. Liu, B., Hu, M.Q., Cheng, J.S.: Opinion Observer: Analyzing and Comparing Opinions on the Web. In: 14th International Conference on World Wide Web (WWW 2005), pp. 342–351. ACM, New York (2005)
6. Pang, B., Lee, L.L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
7. Kobayashi, N., Inui, K., Matsumoto, Y.: Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining. In: 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), pp. 1065–1074. ACL, Stroudsburg (2007)

8. Lou, D.C., Yao, T.F.: Semantic Polarity Analysis and Opinion Mining on Chinese Review Sentences. Journal of Computer Application 26, 2622–2625 (2006) (in Chinese)
9. Zhang, J.F., Zhang, Q., Wu, L.D., Huang, X.J.: Subjective Relation Extraction in Chinese Opinion Mining. Journal of Chinese Information Processing 22, 55–59 (2008) (in Chinese)
10. Chen, Q.Z., Liu, Q.S., Yao, T.F.: Topic and Sentiment Relation Extraction on Chinese Opinioned Texts. In: 5th China National Conference on Information Retrieval (CCIR 2009), pp. 505–512. CIPSC, Beijing (2009) (in Chinese)
11. Chen, M.S., Yao, T.F.: Combining Dependency Parsing with Shallow Semantic Analysis for Chinese Opinion-element Relation Extraction. In: 4th International Universal Communication Symposium, pp. 299–305. IEEE Press, New York (2010)
12. Zengin, B.: Benefit of Google Search Engine in Learning and Teaching Collocations. Journal of Educational Research 34, 151–166 (2009)
13. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open Information Extraction from the Web. Communication of the ACM 51, 68–74 (2008)

Automatic Tagging of Interchangeable Characters in Pre-Qin Literature

Minxuan Feng[1,2], Liu Liu[1], and Ning Xi[3]

[1] School of Chinese Language and Literature, Nanjing Normal University, Nanjing, 210097, China
[2] Center of Language and Informatics, Jiangsu Higher Education Key Research Base of Social Studies, Nanjing, 210097, China
[3] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210046, China
fennel_2006@163.com, liuliu1989@gmail.com, xin@nlp.nju.edu.cn

Abstract. Interchangeable Characters (ICs) is an important issue in semantic analysis of Pre-Qin literature. This paper employs three knowledge databases resources for IC tagging: 1) IC frequency table built from 25 Pre-Qin literatures and *Commentary on the Thirteen Confucian Classics* based on *Chinese Dictionary*; 2) book-specific IC database based on philological and exegetic studies; 3) Academia Sinica IC bank based on the tagged corpus of ancient Chinese. Experiments are conducted to tag ICs in *Mo-tse*, *The Book of Filial Piety* and *The Songs of Chu* respectively and show that the second knowledge database, though of a small scale, is very reliable, that the third database can be a useful supplementary to it and that the first database alone can also provide useful information for the purpose. The research makes it clear that the construction of the IC knowledge base is of great significance in improving the performance of automatic tagging of IC.

Keywords: Interchangeable Characters, Pre-Qin Literatures, Automatic Tagging, Language Information Processing.

1 Introduction

The relationships among ancient-modern characters, Interchangeable Characters ("IC" for short), cognate characters, phonetic loan characters and homological characters have long been the focus of experts in exegetics and philology. Meanwhile, a large amount of relevant resources can also be found in many classic works, dictionaries, and commentaries.

This study employs three knowledge databases resources for IC tagging. There are IC frequency table, book-specific IC database, and Academia Sinica IC bank. Experiments are conducted to tag ICs in *Mo-tse*, *The Book of Filial Piety* and *The Songs of Chu*, using different resources and methods. The results show that the construction of the IC knowledge base is of great significance in improving the performance of automatic tagging of IC.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 344–355, 2013.

2 Current Research on Interchangeable Characters

Wang formally defines IC of ancient Chinese phonology as group of words which have the same or similar pronunciation and can be used temporarily or unconditionally in interchangeable ways in ancient written Chinese [1]. Zhou further explains that two words which have the same or similar pronunciation and meanings and can be used interchangeably with each other in writing are called Unconditional Interchangeable Characters ("UIC" for short). "反 fan3" and [1] "返 fan3", "知 zhi1"and "智 zhi4" are two examples of UIC. Two words with the same or similar pronunciation but different meanings of which one can sometimes be interchanged with the other are called temporary interchangeable characters, a.k.a. phonetic loan characters ("PLC" for short)[2]. Kang argues that UIC and PLC is similar in that the two words co-exist with exactly the same or similar pronunciation, but differ in that the original word and the loan word are not distinguished in UIC, but are in PLC. For instance, the first personal pronoun "yu2" does not have its own written form, and is represented by "予 yu2"or "余 yu2" with the same pronunciation. The modal word "wei2" does not have its own written form either, and is represented by "唯 wei2", "惟 wei2", "维 wei2" with the same pronunciation [3]. The set of PLC makes up the larger part of the set of IC. Therefore, this research shall focus mainly on PLC tagging, without explicitly distinguishing it from the set of UIC.

The chapter "Compilation Principles, Semantic Analysis, Handling interchangeability" in *Chinese Dictionary* states that IC is the interchangeable usage of words with the same or similar pronunciations but different original meaning in the ancient written Chinese, and the character A can be interchanged with B only if: 1) A and B co-exist in the same period; 2) A and B have the same or similar pronunciations; and 3) A and B do not have the same original meaning. As can be seen, the essence behind the IC is to borrow the pronunciation, but not the meaning. The relationship in pronunciation is highlighted. For example, in the sentence 授事以能，则人上功 (*ShouShiYiNeng, ZeRenShangGong)*[2] in *Asking in Guan-Tse*, *尚 shang4* is interchanged with *上 shang4*, the pronunciation of 上 is used to represent the meaning of 尚 but 上 does not convey its original meaning. However, the senses of A and B are not necessarily different. Meaning cannot be regarded as one of the standards in identifying IC.

IC presupposes the co-existence of the original word and the loan word, which is a outstanding features differing from the ancient-modern characters (AMC). By AMC, it meant a group of two characters, one is the ancient character, and the other is the modern character. In AMC, the ancient character and the modern character have explicit temporal sequential relation. When writing the IC, the ancients were not cautious about the meanings conveyed from the character forms and structures. They simply treated the characters as phonetic symbols. However, in AMC, the modern

[1] "反 and 返" is IC means "返" is interchanged by "反", or specifically, "反" is the loan character and "返" is the original character. It's similar in the following examples.

[2] The sentence means "Things can award, the people work."

character is reshaped by adding or modifying some ideograph symbols of the ancient character based on the ideographic functions of Chinese characters. In this way, the burden of preserving too many meanings on the ancient characters can be relieved to some extent, and the ancient characters can be visually distinguished from the modern characters. In other words, the modern characters preserve all or parts of the meanings of the ancient character. In addition, IC has the same or similar pronunciation, which does not lead to the increase in the number of Chinese characters. However, in AMC, the ancient character is created while the ancient character is retained, hence the number of Chinese characters will be increased. The fundamental difference of IC and AMC is that the former is synchronic and while the latter is diachronic. However, the properties of them are not absolutely opposite. As long as the modern character is created and the ancient character remained in use, they cannot be treated as AMC. In practice, however, it's not easy to tell exactly the period of the creation of each character. It's unreasonable to say that one character is created on some period on the basis of its usage in that period based on some literatures. The creation of the character could be in an earlier period if confirmed by the corresponding literatures.. Therefore, our works are only based on materials at hand. Even then, we still have a few findings. In an example in Chinese Dictionary, 悦yue4 is interchanged by 说 yue4 in three senses of the character 说 yue4 with an exception in one sense, 阅 yue4 is interchanged by 说yue4: 1) used later as 悦yue4, means happy or pleasure; 2) used as 悦yue4, meaning be fond of; 3) used as 悦yue4, meaning please sb.; and 4) used as 悦yue4, means respect. In the dictionary, 说yue4 and 悦yue4 are treated as AMC. However, based on the statistics in the 25 Pre-Qin literatures in CHANT [4], however, 说yue4 appears 1380 times in 24 literatures, and 悦yue4 appears 146 times in 14 literatures. Based on the statistics in the Academia Sinica tagged corpus of the ancient Chinese [5] on 23 literatures (The Songs of Chu,Wu-Tse are excluded because they have not been labeled), 说yue4 appears 863 times in 21 literatures, and 悦yue4 appears 87 times in 8 literatures. Specifically in Mencius, the number of the occurrences of 悦yue4 is one and a half times that of 说yue4. In Book of Filial, 说yue4 never appears, but 悦yue4 appears 5 times. These observations have strongly demonstrated that the two words are synchronic as early as Qin period, and should be treated as ICs. In AS tagged corpus Zuo's Commentary, It's worth noting that two instances of 说 yue4 have been tagged as PLC, where the skepticism surely lies about their being AMCs.

Besides, many scholars have studied the book-specific IC. Ge analyzes the loan characters and their original characters which have undergone great changes or need complicate phonological analysis to account for the change, points out the foundations on ancient Chinese phonology for these IC, and clarifies the basic clues on the variation of the pronunciations [6]. Luo finds four categories of misinterpretations of the IC from six prevailing versions of *The Book of Songs*: mistaking the loan characters for the original character, mistaking variant characters (VC) for IC, mistaking AMC for IC, and mistaking semantic extended characters for IC. She argued that IC,

Table 1. The frequcies of 说yue4 and 悦yue4 in 25 Pre-Qin literatures

Literature	CHANT		AS	
	说yue4	悦yue4	说yue4	悦yue4
Han Fei-Tse	223	23	39	2
Mencius	21	53	21	53
Lu's Commentaries of History	210	1	207	2
Mo-Tse	213	3	14	0
Book of History	21	3	5	0
Hsun-Tse	129	2	23	0
Yan-Tse	54	19	68	2
Chuang-Tse	64	24	75	14
Gongyang's Commentary of the Spring and Autumn Annals	9	0	8	0
Guliang's Commentary of the Spring and Autumn Annals	7	0	7	0
Zuo's Commentary	93	0	95	0
Guan-Tse	57	7	57	6
Discourse on the States	44	3	44	3
The Book of Changes	43	0	41	0
Lao-Tse	1	0	0	0
The Book of Rites	49	0	48	0
Book of Etiquette and Ceremonial	57	0	39	0
Rites of the zhou dynasty	12	0	12	0
The Analects of Confucius	21	0	21	0
The Book of Lord Shang	26	0	27	0
The Book of Songs	13	0	11	0
Sun-Tse	0	1	1	0
The Book of Filial Piety	0	5	0	5
The Songs of Chu	10	1	-	-
Wu-Tse	3	1	-	-
Total	**1380**	**146**	**863**	**87**

VC, AMC, and semantic extended characters should be strictly distinguished [7]. Cui investigates the ICs in *Hsun-Tse*, and concludes as follows: 1) the number of ICs are large; 2) the relationship between the loan and the original character in IC is complicate; 3) the ICs where the loan and the original character share common character shape make up the majority; 4) the loan and the original characters co-exist; 5) not applying the meaning of original character, but the meanings of IC. The characters in *Lao-Tse* of Chu Bamboo Slips belong to Chu Characters during the Warring States period, where a large amount of ICs exist [8-9]. Nie examines and supplements 44 groups of ICs which are not collected in *Chinese Dictionary* but used by *Lao-Tse* of Chu Bamboo Slips, based on the *Mawangdui Silk Texts* [10]. Wu sorts out 140 groups of ICs from *Mo-Tse* in the dissertation, of which 49 groups are not collected by *Gu-JinTongJiaHuiDian* [11]. The dissertation also studies the collection of ICs based on

Dictionary of Cognate and *Dictionary of Ancient Interchangeable Characters*, and finds that the ICs with dual sound rhyming make up a big part. The more similar the two characters are, the more likely they are ICs [12]. In addition, as requested by the research of the ancient Chinese, Academia Sinica built the ancient Chinese corpus in 1990, which was originated by Huang Juren, Tan Pusen, Chen Kejian, and Wei Peiquan. In the year of 1995, they started to build the AS tagged corpus of ancient Chinese. The corpus contains 36 literatures, including *The Book of History, The Book of Songs,The Book of Changes, Book of Etiquette and Ceremonial, Rites of the Zhou Dynasty, The Book of Rites, Gongyang's Commentary of the Spring and Autumn Annals, Guliang's Commentary of the Spring and Autumn Annals, Zuo's Commentary, Discourse on the States, Stratagems of the Warring States, The Analects of Confucius, Mencius, Mo-Tse, Chuang-Tse, Hsun-Tse, Han Fei-Tse, Lu's Commentaries of History, Lao-Tse, The Book of Lord Shang, Guan-Tse, Yan-Tse, Sun-Tse, The Book of Rites edited by Dai De, External Commentary to the Han Odes, Wei Liao-Tse, The Six Arts of War, Wen-Tse, Book of Filial, The School Sayings of Confucius, Records of the* Grand Historian, Xin Yu, Xin Xu, Chun Qiu Fan Lu, Huai Nan-Tse, ShuiHuDi Bamboo Slips of Qin. The corpus is manually tagged, aided with machine. It's of great value because the POS tags, some special usages, as well as UICs (tagged by jj) are tagged. However, there are still some omissions, for example, the tagged sentence 杀 (VC1O) 而 (C) 埋 (VC2) 之 (NH) 马 矢 (NA2)[+attr] 之 (T) 中 (NG) 。 (ShaErMaiZhiMaShiZhi Zhong.Killed him then buried him in the horse dung.) of Eighteenth years of Duke Wen in Zuo's Commentary, 矢 shi3 is not treated as an IC with 屎 shi3 because 矢 shi3 appears in the word 马矢 MaShi。

From the above, we see that the phenomenon of IC dates back to Pre-Qin, and it was becoming more and more abundant in the following periods. Our work built the database of ICs of Pre-Qin for purpose of automatic tagging. We hope that the database can benefit the automatic tagging of ICs in various periods.

3 Building Knowledge Bases of ICs

3.1 IC Frequency Table

We collected groups of ICs based on information in *Chinese Dictionary* [13]. We treat the pattern *a同b* or *a通b* in the explanations (*a* and *b* are constraint to single character word),such as *丐同≦考≧, 亂通≦率≧,* as an informative indicator of a group of IC. Having removed the duplicate ICs due to multiple explanations and two-character words, we obtained a list of 4577 groups of ICs.

To further investigate the distribution of IC of Pre-Qin, we count these ICs in the corpus of 25 Pre-Qin literatures in CHANT. In addition, the corpus of *Commentary on the Thirteen Confucian Classics* (CTCC) [14] is used.

We divide the 4577 groups of ICs into 6 categories based on their statistics in 25 Pre-Qin literatures, as follows:

1) 250 groups with both original and loan characters absent. For example, 仛（奼cha4）,伴（丰feng1）,伮（努nu3）,侳（酩ming3）,傌（罵ma4）.

These groups occur rarely in CTCC. Therefore, it's difficult for us to analyze the interchangeability for them.

2) 629 groups with original and loan characters which have similar number of occurrences. 318 of the groups are those of which the loan characters appear slightly more frequently than the original character, such as 亶（殫dan4）,倉（蒼cang1）,傳（專zhuan1）,傴（嫗yu4）,僉（憸xian1）, and 311 groups are those where the original characters appear slightly more frequently than the loan characters, such as 俠（狹xia2）,傒（蹊xi1）,傜（徭yao2）,倗（瞷jian4）,儉（險xian3）. As can be seen, these two sub-groups are two classics of ICs.
3) 1473 groups where the loan characters appear obviously more frequently than the original characters, such as 亂（率shuai4）,亶（癉dan4）,來（的de）, 倉（滄cang1）,倫（掄lun2）. In these groups, the loan character mainly serves as the interchangeability usage.
4) 1061 groups where the original characters appear obviously more frequently than the loan characters, such as 佁（似si4）,佽（次ci4）,來（不bu4）, 倅（粹cui4）,傳（士shi4）. In these groups, the loan character occasionally serves as the interchangeability usage.
5) 515 groups where the original characters appear but the loan characters never appear, such as 伓（伾pi1）,倪（鴑nu2）,侶（似si4）,侶（昭zhao1）, 佺（悛quan1）. 316 groups of this categories are counted in the CTCC , such as（傍ben1, 奔ben1，641，2376）3,（晴qing2,情qing2，646，1487），（肰ran2, 然ran2，3343，11934），（蔀jie2,節jie2，757，4264），（鏸hui4, 惠hui4，667，1542）.
6) 649 groups where the loan characters appear but the original characters never appear, such as 亞（椏ya4）,倉（艙cang1）,備（賠pei2）,傰（繃beng1）, 僇（勠lu4）. 372 groups of this categories are counted in the CTCC, such as（達da2, 韃da2,365，1681），（會kuai4, 儈kuai4,1303，6224），（伐fa2, 戛fa2,1903， 5328），（服fu2, 腷fu2,1919， 12880），（侯hou2, 堠hou2,4521，19770）.

In the last two categories of IC, the loan character and the original character do not co-occur with each other. This indicates that further research is needed to decide whether they belong to IC or AMC.

As discussed in section 1, the loan character preserve the meaning of the original word, therefore, the loan character should have the same POS tag with the original character. Considering this, we further annotated the POS tags to original characters in the list based on the human-annotated Chinese corpus of Pre-Qin literatures [15].

[3] The statistics is hereafter represented by 4-tuples such as (傍ben1, 奔ben1, 641, 2376) where" 奔ben1"is the original character, "傍ben1"is the loan character,"奔ben1"appears 641 times in the 25 pre-Qin literatures, and appears 2376 times in CTCC.

3.2 Book-Specific IC Database

Research works on philology and exegetics contain abundant information on IC. Especially, the book-specific works which are based on reliable and informative data are important resources in studying IC. We build a database of book-specific IC on five literatures, including *Mo-Tse, Hsun-Tse, Mencius, The Book of Songs* and *Lao-Tse*.

As illustrated by Table 2, each record in the database consists of 8 fields (from left to right): loan character, original character, pinyin, sense, example sentence, source, frequency, and number of books which contains the character. Currently, we have only collected 225 ICs (with 473 occurrences) into the database.

Table 2. Samples in database of book-specific IC

Loan Character	Original Character	PinYin	Sense	Example Sentence	Source	Frequency	Number of Books
僇	戮	lu4	killing	……	Mo-Tse	10	1
論	掄	lun2	choose	……	Mo-Tse	1	1
當	嘗	chang2	try to	……	Mo-Tse	7	1
離	罹	li2	suffer	……	Mo-Tse Hsun-Tse	3	2
蚤	早	zao3	morning	……	Mo-Tse Mencius	13	2
鄉	向	xiang4	formerly	……	Mo-Tse	12	1
葆	保	bao3	protect	……	Mo-Tse	17	1
從	纵	zong4	indulge	……	Mo-Tse	1	1
進	峻	jun4	high	……	Hsun-Tse	1	1

3.3 The AS-IC Bank

For comparison with the studies on the 25 literatures of Pre-Qin in CHANT, we extract the information of the IC from 23 literatures in the AS tagged corpus, which are the same with those in CHANT. The collected 507 IC types (1070 occurrences) are used to build the AS-IC bank. Note that no IC is collected from *Gongyang's Commentary of the Spring and Autumn Annals*, *Sun-Tse*, and *Book of Filial*. The distribution of IC over the 23 literatures is listed in Table 3.

Among the 507 ICs, there are 12 ICs occurring more than 10 times: *正zheng4* (31 times),*从cong2* (13 times),*止zhi3* (13 times),*共gong4* (12 times), *不bu4* (11times),*惟wei2*(11 times),*政zheng4*(11 times),*以yi3*(10 times),*通tong1*(10 times), *宅zhai2*(10 times), *气qi4*(10 times),*静jing4*(10 times); 4 ICs occurring 9 times: *光guang1, 为wei2,假jia3,臧zhang1;* 4 ICs occurring 8 times:*已yi3,至zhi4,训kun4,官guan1*; 5 ICs occurring 7 times: *而er2,是shi4, 曰yue1, 亡wang2, 重zhong4* ; 9 ICs occurring 6 times : *周zhou1, 当dang1, 取qu3, 之zhi1, 植zhi2, 难nan2, 臣chen2, 子zi3, 言yan2, 出chi1;* 14 ICs occurring 5 times : *宾bin1, 平ping2, 厘li3, 义yi4, 辟pi4, 及ji2, 德de2, 物wu4, 鬲ge2, 皇huang2, 极ji2, 图tu2, 趣qu4, 讫qi4*; 18 ICs occurring

Table 3. Distribution of IC over the 23 literatures

Literature	Occurrence
Han Fei-Tse	61
Mencius	12
Lu's Commentaries of History	148
Mo-Tse	48
Book of History	150
Hsun-Tse	8
Yan-Tse	55
Chuang-Tse	12
Gongyang's Commentary of the Spring and Autumn Annals	0
Guliang's Commentary of the Spring and Autumn Annals	4
Zuo's Commentary	59
Guan-Tse	247
Discourse on the States	13
The Book of Changes	39
Lao-Tse	3
The Book of Rites	55
Book of Etiquette and Ceremonial	10
Rites of the zhou dynasty	19
The Analects of Confucius	6
The Book of Lord Shang	21
The Book of Songs	100
Sun-Tse	0
The Book of Filial Piety	0
Total	**1070**

4 times: *或huo4, 中zhong1, 闻wei2, 人ren2, 耐nai4, 服fu2, 事shi4, 藏cang2, 乱luan4, 能neng2, 畏wei4, 绥sui2, 故gu4, 上shang4, 似si4, 纯chun2, 财cai2, 秋qiu1*; 40 ICs occurring 3 times: *秩zhi4, 宝bao3, 税shui4, 将jiang1, 盖gai4* and so on ; 98 ICs occurring twice, such as *保bao3, 食shi2, 谷gu3, 今jin1, 归gui1* and so on ; and 303 IC occurring only once, such as *妾qie4, 荷he2, 主zhi3, 多duo1, 目mu4* and so on. The AS corpus is tagged without giving the original character. Accordingly, no information of the original character is included in the AS-IC bank.

4 Experiment

To verify the values of the above knowledge bases, this research tries to tag the IC in three Pre-Qin corpora: *Mo-Tse, The Book of Filial Piety* and *The Songs of Chu*, using different resources and methods.

4.1 Experiment on *Mo-Tse*

We first used the book-specific IC database to tag *Mo-Tse* in CHANT. Compared with the AS results where 48 occurrences of IC are tagged. our results finds 140 occurrences of IC where 11 occurrences are identical with that in As results, which are 4 occurrences of *正zheng4*, 3 occurrences of *当dang1*, and 1 occurrence of *次ci4*, *以yi3*, *放fang4,* and *谨jin3*.

We then looked up the 140 collected groups of IC in *Chinese Dictionary*. All groups appear in it except that *次ci4 and 态i4* is not found. Of the 37 occurrences of IC which are tagged in the AS but not in our results, 18 occurrence of IC can be found in the IC frequency table. Again, we looked up the 18 occurrences in *Chinese Dictionary*, and found that the 13 occurrences are ICs, including 7 occurrences of *政zheng4*, and 1 occurrence of *灵ling2*, *贲ben1*, *操cao1*, *秖di1*, *譁huan1*, and *攸you1* respectively, where *譁huan1* and *攸you1* are considered interchangeable with *欢huan1* and *悠you1* but not with *权quan2* and *彼bi3* respectively by *Chinese Dictionary*, which indicated that *譁huan1* and *权quan2*, *攸you1* and *彼bi3* are rare cases of ICs. The other 5 occurrences are not ICs, including 1 occurrence of *辜gu1*, *转zhuan3* and *政zheng4*, 2 occurrences of *荐jian4*. There is not clear evidence that the remaining 19 occurrences are also ICs.

From the above analysis, we see that the resources of the conventional exegetics are of high quality. The use of book-specific IC database can achieve a very high quality and a moderate recall in tagging the ICs in *Mo-tse*. The book-specific IC database can be enhanced with machine-aided tagged corpus, such as the as *Mo-Tse* corpus which provides 7 additional groups of IC beyond the 140 groups, to further improve the recall of tagging.

4.2 Experiment on *Book of Filial*

We look up the IC where the two characters co-occur in the same commentary line (labeled by "○") of *Commentary on Book of Filial*, which also satisfying the following two conditions: 1) Both frequencies are higher than 1000, or both frequencies higher than 100 and the frequency of any character should not be more than twice that of the other character; 2) The original character appears after the loan character with a distance less than 2 characters.

Three occurrences of ICs are tagged in *The Book of Filial Piety* in CHANT, but only one occurrence (*脩xiu1* and *修xiu1*) is correctly tagged, and other occurrences of *是shi4* and *以yi3*, which are actually not ICs. This reveals that the above thresholds need to be adjusted.

We also found that in the sentence *德教加于百姓, 刑于四海*[4] in *Tian-Tse* Chapter of *Book of Filial*, *Chinese Dictionary* takes *刑xing2* and *形* xing2as IC, whereas

[4] The sentence means that "conduct will educate and transform (dejiao) the common people, serving as exemplary in all corners of the world".

the corresponding commentary says that *刑xing2* means penalty and does not consider them to be a group of IC.

The commentary of *The Book of Filial Piety* adds some explanations to the original text. For example, in the following text: *事亲者居上不骄，当庄敬以临下也，为下不乱，当恭谨以奉上也。在丑不争。丑，众也。争，竞也。当和顺以从众也。居上而骄则亡，为下而乱则刑，在丑而争则兵。谓以兵刃相加三者不除，虽日用三牲之养，犹为不孝也。三牲，太牢也，孝以不毁为先。言上三事皆可亡身，而不除之，虽日致太牢之养，固非孝也*[5].We find in the explanation (text in italic) *固gu4* and *故gu4* forms a group of IC, which shows the effectiveness of information in commentary. However, these texts are not found in CHANT.

As can be seen, it's suitable to use the commentary information to tag the ICs of the literatures such as *The Book of Filial Piety* and *Gongyang's Commentary of the Spring and Autumn Annals*, where the corresponding studies on exegetics are rare and no IC is found in the AS-IC corpus. We believe information beyond the relative positions, distance, and explanation style of can be explored to improve our results.

4.3 Experiment on *The Songs of Chu*

In this experiment, we use the frequency table as well as the POS-tagged corpus to tag the ICs in *The Songs of Chu.* Due to the lack of relevant resources, , we extract 70 example sentences (including 67 groups of IC) of *The Songs of Chu* from the explanations in *Chinese Dictionary*. Each group of IC is found in only one sentence, except that the group of *脩 xiu1* and *修 xiu1* is found in 3 sentences and the group of *游 you2* and *遊 you2* is found in only one sentence. These example sentences are used as reference to evaluate our tagged results.

First, we extract the groups of ICs from the IC frequency table with the frequencies of both the original and loan character which satisfy the following conditions: 1) lower than 100; or 2) lower than 700 wherein the frequency of any character should not be more than twice that of the other character. In this constraint, 1652 candidate groups of IC were extracted, and 35 groups (52% of the 67 groups) were also found in *Chinese Dictionary*. Second, we tag 735 groups of IC based on the POS-tagged corpus of "The Songs of Chu", with only 104 correct, of which 22 groups were found in the reference sentences. From the above, we can see that 13 groups were lost after we add the POS-tag information, possibly due to three reasons: 1)The corpus of *The Songs of Chu* in CHANT omit some characters, or uses spurious font, such that the ICs are not shown, they are *僞wei3* and *讹e2*, *憚dan4* and *殫dan1*, *戲xi4* and *巇xi1*, *秉bin3* and *禀bin3*, *內nei4* and *讷ne4*, *爰yuan2* and *喧xuan1*, *祗zhi1* and *振zhen4*,

[5] The sentence means that "Those who are truly able to serve their parents are not arrogant in high station, are not rebellious in a subordinate position, and are not contentious when only one among many. To be arrogant in high station leads to ruin; to be rebellious in low position incurs punishment; to be contentious among the many leads to violence. *Until these three attitudes—arrogance, defiance, and contentiousness—are set aside, even though someone were to fete their parents on beef, mutton, and pork, they still could not be deemed filial."*

鰥guan1 and *鰥guan1*, and *齎qi2* and *齎qi2*; 2) There are errors in the POS-tagged corpus. Some characters have flexible usages which are not collected in the list of the statistic usages. For example，*邈/a miao3* is interchanged by "*藐miao3*"，but in the POS-tagged corpus, there is a sentence *藐/n 蔓蔓/v 之/u 不/d 可/d 量/v 兮/y*, *遐/a xia2* is interchanged by *霞xia2*,but there is a sentence *載/v 營/a 魄/n 而/c 登/v 霞/n 兮/y* in the corpus.

This inconsistency reveals that the POS tag lists of characters need further refinement, and the flexible usages of IC should be taken into account. Even so, we made two discoveries: 1) 78 additional cases of IC are found beyond *Chi*nese Dictionary, which are 6 cases of *鯈shu1* and *倏shu1*, 2 cases of *圖tu2* and *度du4*, 2 cases of *晞xi1* and *睎xi1*, 2 cases of *斑ban1* and *班ban1*，4 cases of *班ban1* and *斑ban1*，11 cases of *流liu2* and *游you2*，32 cases of *脩xiu1* and *修xiu1*，10 cases of *游you2* and *遊you2*, 2 cases of *鄣zhang1* and*障zhang4*, 6 cases of *圜yuan2* and *圓yuan2*, and 1 case of *杼zhu4* and *抒shu1*; 2) An additional group of IC was found, that is *脅xie2* and *肤fu1* in the sentence *平/a 脅/n 曼/a 膚/n ，/w 何/r 以/p 肥/v 之/r ?/w*.

As we can see, in the case of absence of the electric versions of the commentaries for literatures such as *The Songs of Chu*, *Sun-Tse*, *Wu-Tse*, and the absence of the related IC-tagged corpus and the relevant search findings on exegetics, the POS-tagged corpus can be a useful knowledge resource in tagging the ICs.

5 Conclusions

Due to the fact that Pre-Qin corpus cannot meet the requirements for machine learning, and that no required statistics model is available for interchangeable character tagging, we build three knowledge resources:, an interchangeable character frequency table, an interchangeable character database and an Academia Sinica interchangeable character database, and conducts automatic interchangeable character tagging based on a large number of sources such as *Chinese Dictionary*, 25 Pre-Qin literatures, *CTCC*, Chinese exegetics, and Academia Sinica tagged corpus of old Chinese. For experiments, this research tags interchangeable characters in Pre-Qin corpus *Mo-Tse, The Book of Filial Piety* and *The Songs of Chu* respectively and achieves satisfactory result. It also discusses the usage of the preceding knowledge libraries for various corpuses. This paper performs the diachronic studies on interchangeable characters, which lays a sound foundation for more comprehensive studies of Pre-Qin interchangeable characters. In addition, the experience obtained in the process of constructing the knowledge databases shall benefit future research studies. In future studies, we shall make use of the phenomenon that certain interchangeable characters have same or similar pronunciation with their origins and incorporate phonetic information to refine the IC tagging in the corpus.

Acknowledgements. We thank anonymous reviewers for their constructive suggestions. This work is the staged achievement of the project "Study of the Construction

of Ancient Chinese Corpus" supported by National Social Foundation of China(10&ZD117),and "Word Knowledge Mining from Pre-Qin Literatures" supported by the research base of Philosophy and Social Sciences for Universities in Jiangsu(2010JDXM023).This work is financed by Philosophy and Social Sciences Foundation in Jiangsu Province(10YYB007),A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions(164320H107),and An Research Program for Postgraduate of Universities in Jiangsu(CXLX12_0357).

References

1. Wang, L.: Classical Chinese, p. 541. Zhonghua Book Company, Beijing (1962) (in Chinese)
2. Zhou, B.J.: Outline of Ancient Chinese Fine, p. 263. Hunan People's Publishing House, Changsha (1981) (in Chinese)
3. Kang, X.L.: Discussions on Interchangeable Characters. Master degree thesis of Shanxi University (2005) (in Chinese)
4. Pre-Qin Literature. CHANT (2010) (in Chinese), http://www.chant.org/
5. Academia Sinica in Taiwan: Academia Sinica Ancient Chinese Corpus (2010) (in Chinese), http://old_chinese.ling.sinica.edu.tw/
6. Ge, S.K.: Study on Interchangeable Characters in Book of Filial. Journal of Lianyungang Teachers College 2, 66–69 (2007) (in Chinese)
7. Luo, X.: Discrimination about Interchangeable Characters in Book of Filial. Journal of Huizhou University (Social Science Edition) 27, 73–76,128 (2007) (in Chinese)
8. Cui, Z.Z., Zh, X.B.: Study on Interchangeable Characters in Hsun tzu (1) –Statistics on Interchangeable Characters in Hsun tzu. Journal of Shijiazhuang Vocational Technology Institute 14, 33–35 (2002) (in Chinese)
9. Cui, Z.Z., Zh, X.B.: Study on Interchangeable Characters in Hsun tzu (2) –Characteristics on Interchangeable Characters in Hsun tzu in Hsun tzu. Journal of Shijiazhuang Vocational Technology Institute 15, 28–31 (2003) (in Chinese)
10. Nie, Z.Q., Li, D.: Guo Dian Chu Jian. A Study on the Phonetic Loan Characters Used in Lao-zi. Linguistics Study 25, 103–106 (2005) (in Chinese)
11. Gao, H., Dong, Z.A.: Ancient-Modern Interchangeable Characters. Qilu Press, Jinan (1989) (in Chinese)
12. Wu, D.D.: The Study on the Relation of Sound and Meaning about Tongjia in Mozi. Lanzhou. Master degree thesis of Lanzhou University (2008) (in Chinese)
13. Chinese Dictionary 2.0 (CD-ROM). The Publishing House of the Chinese Dictionary. Commercial Press (H.K.), Hong Kong (2000) (in Chinese)
14. Commentary on the Thirteen Confucian Classics. China Ancient Book. (2011), http://guji.artx.cn/list_2_1_1.html (in Chinese)
15. Chen, X.H.: Pre-Qin Chinese Character Tagging Corpus. Language Information Technology Research Center in the Base of Jiangsu University Philosophy and Social Science (2011) (in Chinese)

A Management Structures of Concepts Based on Ontology

Dexun Li[1] and Jinglian Gao[2,3]

[1] Fuyang Teachers College, Fuyang, China 236037
leeadesir@163.com
[2] Wuhan University, Wuhan, China 430072
[3] Guangdong Guobi Technology Stock Co., LTD, China 510620
gao@guobi.com

Abstract. Based on the comparison between management thesaurus and management domain ontology, this article proves the necessities and practicalities of ontology-based management knowledge organization, and the significance of ontological approaches, which should be adopted to management information organization. In addition, it also summarizes the basic situations of management structures of concepts, and designs the principles and procedures for its establishing, on the basis of which a management concept structure model is established.

Keywords: Retrieval, Management, Information Resources, structures of concepts.

1 Introduction

With the emergence of computer and the gradual popularization of internet, copy and free expression of digital, electronic, or network-based information resources and its digital text has become possible, which making management information resources augment daily. The rapid growth and effective use of management information has become a contradiction. How to effectively develop and control management information resources, and let it meet our needs of management information, and provide high-rank management information services, has become the hot issues.

Ontology-based knowledge organization provides a new space for management information organizations. It can make the management information resources have more accurate and complete semantics, and make computer a better understanding of information resources, as a result, the semantization and intelligentizing of information services will come true.

On the basis of other domain structures of concepts which have already been built [1-5], and from the perspective of servicing management practices, this article attempts to define the nature of management domain ontology, and discusses the principles, procedures and methods of establishing it, finally an ontology-based management knowledge organization framework has been built.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 356–363, 2013.

2 Management Thesaurus

The management information retrieval has become a problem ever since there is storage and processing activities to produce management information. In order to solve the contradictions between human information needs and information retrieval, researchers have been using Management Thesaurus. Though it has simple structure, in fact it is the most primitive ontology.

2.1 Management Thesaurus

Although management professional thesaurus can provide an expression of term structure, its aim is only involved in the relationships between terms in particular natural language. It is unhelpful to define concepts, and whether the expression and meanings of concepts is standard has been disregarded The relationship between concepts has not been shown either. In the face of the of various users' needs in the internet environment, it often fails to provide effective management information accurately. Its problems can be outlined as following.

The first problem is the reliability of expression. In many cases, it is difficult for users to faithfully express what he really need to retrieve only with keyword or keywords.

The second problem is the expression diversity. In language expression, the same concept can be expressed in different linguistic forms and may be retrieved with different keywords by different users.

The third problem is the expression island. In human brains, there is always varieties of relations among concepts. In information retrieval, users hope to get not only the document which contains the concepts, but also other information related to them. However, with the traditional information retrieval techniques, keywords are handled as isolated words, can not be extended, and thus form 'the expression island'.

2.2 Management Domain Ontology

For the purpose of finding ideal and exact management information, the management domain ontology should proceed from the perspective of knowledge, provide the approaches of organizing information resources, and progress towards the integrated, automated and intelligent management knowledge organization. The key lies in improving management information retrieval from traditional keyword-based level to knowledge-based or concepts-based level. The central role of management ontology is organizing information resources in the perspective of semantics. It defines the area of management, a series of concepts among management areas and relationships between management concepts, and provides a knowledge expressing language, which is more standardized, detailed, and can express the semantics more comprehensively. With abstract concepts and terminology of ontology, the knowledge structure of management domain can clarify the analysis [6].

It is generally acknowledged that ontology is derived from philosophy domain, and used to be a sub-subject of philosophy in a very long period. Heretofore, it covers

many fields, such as philosophy, knowledge engineering, artificial intelligence, etc. [7-9]. Management domain ontology includes many kinds of standard terms in management domain, the identification of relationships between these terms, and the precise definition to these terms [10]. It have a lot of advantages in knowledge expressing, reusing, sharing, and so on. As for management information resources, ontology can really meet users' information needs and habits of expression, overcome the drawbacks of traditional information retrieval techniques, provide an intelligent retrieval mechanism, and enhance the processing capabilities of retrieval systems. It improves management information retrieval from traditional keyword-based level to knowledge-based or concepts-based level, and reduces the terminal retrieval users' cognitive burden in the mean time. By this way, it can perfect the keyword-based retrieval, greatly promote the efficiency of management information retrieval, and eventually achieve the intelligent information retrieval.

3 How to Establish

Some question such as which concepts belong to management, and unique to management, what is the properties of concepts and the relations between concepts and properties, must be considered during management domain ontology establishing. The establishment principles and procedures of management structures of concepts will be interpreted respectively in this section.

3.1 Establishing Principles

The management domain ontology establishing should act on the following principles.

The first is the economy. During management domain ontology establishing, the advantage over the multilevel and inheritance mechanism of concepts classification should be taken to reduce duplicative definition, and ensure the economy.

The second is the Consistency of concepts Definition. The concepts definition in ontology should be consistent, logical, and the contradiction between definitions or instances must not be allowed.

The third is the standardization of terms. The term s in ontology should be standard, definite and objective. Term selection should refer to or use the international and domestic existing standard terms, which have clear semantics, standardized forms, and shareability. It should be objective and independent from a specific language environment, and give the natural language describing definition when possible. For example, ‘business management’ is not a definite term, its specific category should be defined clearly.

The fourth is the isometric of same semantic level. Semantic distance should be able to reflect the difference between concepts of different levels, and thus the concepts of same semantic level should keep isometric.

The fifth is the integrity and extensibility. Ontology is integral, and must contain all concepts of the field, but it is too difficult to achieve. Therefore, it should be

extensible, can keep pace with the times, can be added with some new concepts, and maintain its integrity at the same time of continuous development and improvement.

3.2 The Establishing Procedures

The establishing of management domain ontology is in a certain order, and follows the procedures of scoping, analysis, associate, accurate expression, inspection, evaluation and maintenance, etc.

The first step is scoping. It is must be made sure that the ontology belongs to management domain, the object of development is management domain ontology. Especially as management is a huge science system, which contains a number of sub-domains, such as manager, role of manager, management approaches, management functions, managed object, and so on.

The second step is analysis. All meanings of the term s in ontology and its relationship between them should be defined, the concepts should be defined by properties, the relationships between concepts should be defined by relationships, the constraints to properties of concepts should be represent by rules. The details is as follows:

The first is making clear the objects of management domain ontology, searching the data related to establishing management domain ontology, making full use of the established other related domain ontology or sub-ontology, getting rid of repetitive work, and drawing lessons on the methods and experiences from them.

The second is conceptualizing and modeling the basic entities or concepts during management domain ontology establishing. As for management, the classical management should be modeled, i.e., some knowledge which are inherited by successive management, identified in the application, and helpful to solve and deduce specific problems should be summarized, this is the basis of management domain ontology establishing. And then the approaches suitable for organization instance data to the knowledge base should be provided.

The third is determining the term s of management domain. As for management domain ontology, it must contain all management entities in a certain hierarchy, including plan, control and distribution, etc.

The terms of management domain must be used in the definition to properties of class or subclass in the hierarchy. Similarly, the differences between the members of class or subclass must also be defined by the term s of management domain.

The third step is associate. Though management domain ontology is limited to specific areas, the correlations between sub-ontologies and the related extensible contents should be considered during its establishing. This extension can be from the horizontal to the vertical, from the macro to the micro, and from the whole to the part, and so on. The relationship between levels seem more complicated, the low-level is an extension as far as the high-level is concerned, that is to say, they are the extensions of the upper concepts.

For example, if the ontology we are developing is for macro-management, the ontology can be extended to other areas of management, such as public administration, management science and engineering, business administration,

agriculture and forestry management, information and records management, library management, and so on. The class hierarchy model can be represented graphically as in Fig. 1.

The fourth step is accurate expression. As for the system of concepts, the hierarchical structure is just a frame, which cannot reflect the varied relations between concepts. Therefore, the properties of classes need to be clarified. Define slots are such descriptors used for describing the properties of classes and instances. It can give the formalization description to the natures of concepts, properties and relationships. Here are some examples designed for the classes.

Eg.1 Financial Management

Class Name: {financial management}

Inherit properties: {all attributes of its upper concept 'business management', 'Business Administration' and 'management'}

Adjacent concepts: {marketing, human resource management, etc.}

Sub-classes: {macro financial management, sector financial management, financial management of companies, financial management of nonprofit organizations, family financial management, etc.}

Agents: {managers, including managers of all strata and all fields, etc.}

Roles: {act as entrepreneur, leader, information master, resource distributor, etc.}

Approaches: {use various methods of economic management, taking interests and class nature into account}

Functions: {belongs to corporate finance activities and the handling of financial relations, including capital budgeting, corporate finance, personnel management, financial analysis, tax management and human capital management, etc.}

Objects: {various information of funds, including funds application and reimbursement, and the information of collection and payment, such as back section of a single payment order, borrowing orders, expense claims and reimbursement claims, etc.}

Eg.2 Marketing

Class Name: {marketing }

Inherit properties: {all attributes of its upper concept business management, business Administration and management}

Adjacent concepts: { financial management , human resource management, etc.}

Sub-classes: {maket investigation and study, select the target market , product development, product promotions, etc.}

Agents: {managers, including managers of all strata and all fields, etc.}

Roles: {act as entrepreneur, leader, information master, resource distributor, etc.}

Approaches: {including the analysis of market opportunities, the target market selection, the marketing strategy determining and marketing campaign management}

Functions: {create products and value, and exchanged them with others , to get the things required}

Objects: {ideas, products and services etc. }

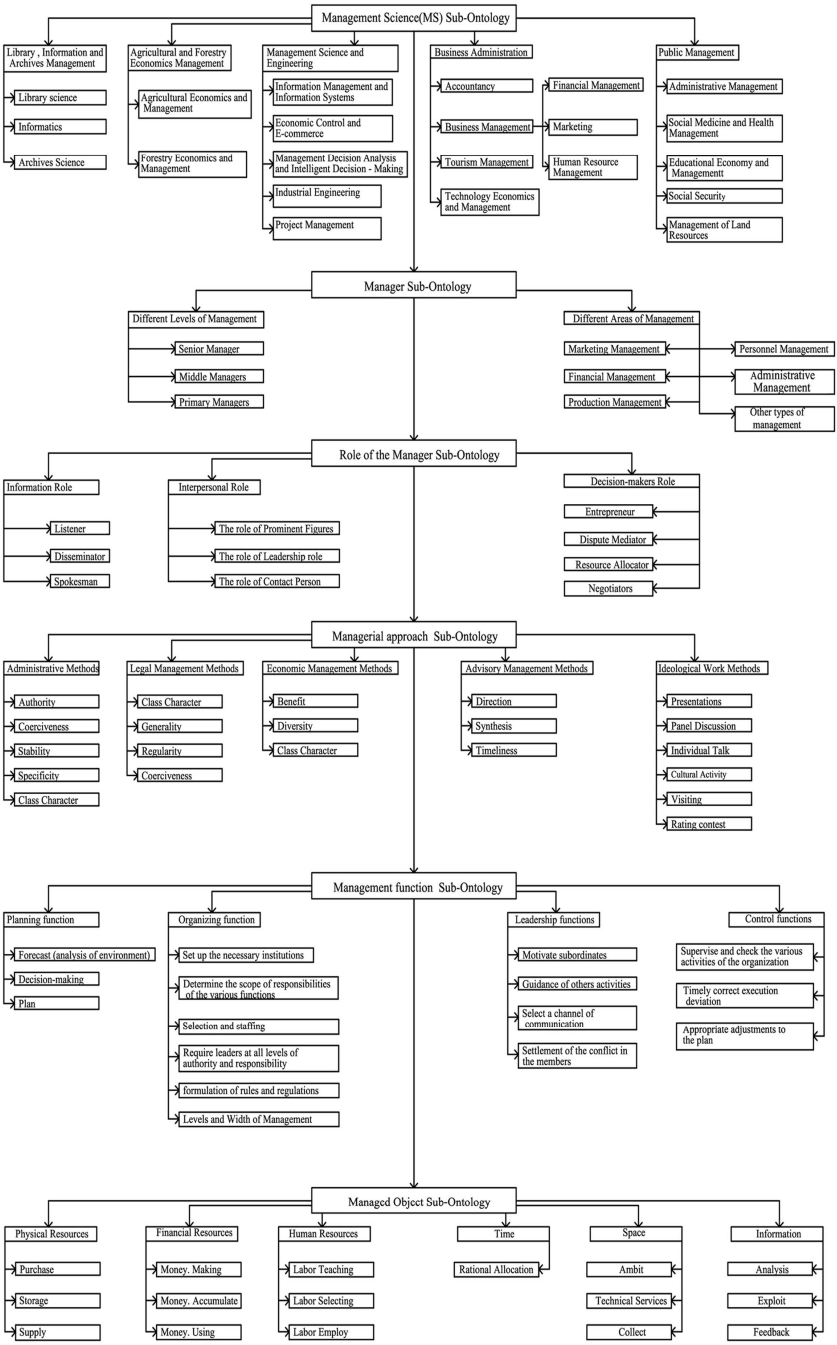


Fig. 1. The class hierarchy model

The fifth step is inspection, evaluation and maintenance. Guided by the establishing principles, whether there are contradictions and conflicts among the definitions of ontology, and whether it can reflect the semantics of concepts objectively should be checked. By making use of its extensibility, the established ontology should keep pace with the times, the ontology being implemented should be updated, and some new concepts should be added in its continuous development.

3.3 Summary

Intelligent information retrieval of management domain ontology can represent information of term's meanings. For example, the term *financial management* is no longer processed as a string, that is to say, there is no longer just a calculation of the frequency of string in a literature, the retrieval tools can develop the term relationship and obtain information about financial management directly. When the term *financial management* is input, system will transfer it to the nodes of concepts, just like the diagram shows above. All the information directly related to this concept can be accessed directly through this relations among terms.

By use of these relationships, the similar or related concepts can be speculated. In this way, users can obtain a particular concept and all information related to it. Users can simply use natural language, such as *list all financial management information*, to search query, the most direct answer can be presented automatically, including upper concepts, adjacent concepts and sub-concepts ,and so on.

4 Concluding Remarks

By combining ontological approaches with management information resources, this article tries to solve retrieval problems of some users who are in need of and interested in management information resources, and has established a management domain ontology by defining the core class, class hierarchy, and the slots. Through semantic-level natural language processing of users' query and web document information, the model has built a knowledge base of management domain ontology, matched the real user query needs with management domain ontology and its mapping source document, and finally presented the retrieval results to users after sort processing.

The application of domain theory and the definite expressing of data semantics will push up the management information services to a higher level, provide a high-quality services to users, lay foundation and provide instructive precedent, which is helpful to make the intelligent management information retrieval and the semantic Web establishment from the possibility to the reality.

Acknowledgments. This work has been supported by the CEEUSRO Project between the Anhui Kejian Education Investment and Development Co., Ltd. and Fuyang Teachers College, named *A Study on the Incentive and Restraint Mechanism in the Human Resource Management of Private Educational Institutions*, and the Fundamental Research Fund for the Central Universities (20081110202000116), named *The Embedded numeric keypad and handwriting Tibetan Input Method.*

References

1. Borst, W.N.: Construction of Engineering Ontologies for Knowledge Sharing and Reuse. Doctoral dissertation. University of Twente, Enschede (1997)
2. Qi, X., Xiong, Q.X., Li, Y.Q.: An e-Learning System based on Domain Ontology. In: Proceedings of 2007 International Symposium on Distributed Computing and Applications to Business. Engineering and Science, Vol. 2 (2007)
3. Zhang, X.K.: Logistics Domain Ontology Model and Its Application. In: Progress in Measurement and Testing—Proceedings of 2010 International Conference on Advanced Measurement and Test, IEEE CPS (2010)
4. Liu, Z.Y.: Research on Construction and Semantic Retrieval of Multiple Majors Domain Ontology. Doctoral dissertation. Beijing Jiaotong University (2010)
5. Jiang, X., Bian, Y.J., Wu, M.F.: Research on Domain Ontology Based Knowledge Retrieval Model. Library and Information Service 18, 116-119+144 (2010)
6. Uschold, M., Gruninger, M.: Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 11, 93–136 (1996)
7. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering, principles and methods. Data and Knowledge Engineering 25, 161–197 (1998)
8. Williams, A.T.: Ontologies. IEEE Intelligent Systems 1(2), 18–19 (1999)
9. Xiao, G.Z., Ji, D.H., Xiao, S.: The Types of Ontology and the Ontology Structure of Chinese Semantic Web. Yangtze River Academic 2, 111–117 (2011)
10. Chang, C.: Construction and Conversion of Ontology in Agricultural Information Management. Doctoral dissertation, Chinese Academy of Agricultural Sciences, Peking (2004)

A Tentative Study on the Annotation of Evidentiality

Qi Su[1,2] and Pengyuan Liu[3]

[1] School of Foreign Languages, Peking University, Beijing, China
[2] Key Laboratory of Computational Linguistics, Ministry of Education, China
[3] Applied Linguistics Research Institute, Beijing Language and Culture University, Beijing, China
{sukia,liupengyuan}@pku.edu.cn

Abstract. This article aims at presenting our ongoing work on the construction of a Chinese corpus in which the credibility of textual information is annotated. The linguistic markers of evidentiality in each sentence are identified as cues of credibility. We annotated both the scale and scope of the evidential markers. The annotated corpus can serve as a data basis for the research of information credibility. In this article, we analyze the theoretical underpinnings of our preliminary annotation guideline and the considerations on the choice of texts. Also, we discuss the possible hierarchy of evidentiality annotation.

Keywords: evidentiality, credibility, annotation, corpus.

1 Introduction

The trustworthiness of information is a big concern for the people who access the information. As more and more information are posted on the internet, the matter becomes more serious. Some natural language processing (NLP) researchers have noticed the issue and adopted the identification of information credibility as an interesting research topic. For incredible information, it may either be willful deception or just include some characters which weaken people's perception of its credibility. For the former, there have been some pilot works on how to spot deceptive language [1-2]. In CoNLL-2010, a shared task was proposed to detect hedges and their scope in natural languages, which is related with the latter aspect of information credibility [3].

Automatic analysis of texts cannot succeed without the support of annotated corpus. For the research of information credibility, the linguistic components encoded with the information of certainty/uncertainty contribute salient features for the automatic predication of credibility. For the above purpose, we construct a Chinese corpus in which the linguistic cues of credibility, *evidentiality,* are marked. Evidentiality is a linguistic category encoding how the speaker expresses the source of information and his commitment to the reliability of information. Thus, it can serve as an explicit linguistic cue for credibility.

In the following sections, we describe the undergoing project beginning with a survey of the evidentiality theory, and then discuss the principle and scheme of our annotation.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 364–372, 2013.

2 Overview of Evidentiality

Although recent years have witnessed an increasing interest in the research of evidentiality, the linguistic phenomenon began to attract linguists' attention very early in the beginning of the 20th century. During Franz Boas's 1911 study of the American Indian languages, he found that some verb suffixes are attached to express the sources and certainty of information in utterances [4]. Then the term "evidentiality" was adopted by Roman Jakobson in 1957 to describe the general linguistic phenomenon [5]. After that, the term of evidentiality has come into common usage. In 1980s, the milestone conference held in Berkley pioneered the research on evidentiality. Even since then the interest in the topic has been growing rapidly. The linguistic phenomenon has been dealt with from a wide variety of perspectives, e.g. typology, grammaticalization, syntax, pragmatics, and cognitive linguistics [6].

Evidentiality as a common linguistic behavior is pervasive in almost all languages. However, the research on evidentiality is mainly about inflectional languages, although there are already some researches for other languages like Chinese. The linguistic forms of evidentiality are termed as evidentials or evidential markers, which may be presented grammatically or/and lexically. With the usage of evidentials, languages provide a repertoire of devices for specifying speakers' assessment of the epistemic unsureness of their information.

2.1 Definitions of Evidentiality

There has been no consensus on the formal definition of evidentiality. In a narrow sense, evidentiality refers to a grammatical category which indicates the sources of information. As Bussmann stated in *Routledge Dictionary of Language and Linguistics*, evidentiality is "structural dimension of grammar that codifies the source of information transmitted by a speaker with the aid of various types of constructions" [7]. Aikenvald also takes the narrow viewpoint of evidentiality. She believes that the central meaning of evidentiality is merely the source of information [8].

However, the researchers who hold the broad sense insist that evidentiality is about the speaker's commitment towards the factuality of the information in addition to the expression of information sources. For instance,

a) She was rich.
b) I saw that she was rich.
c) I'm told she was rich.
d) Apparently she was rich.
e) It's possible that she was rich.
f) She must be rich.

In terms of the different expression of the proposition "*she was rich*", the speaker sometimes may provide the information source explicitly (e.g. b and c), or sometimes his epistemic judgment of the trustworthiness of the information (e.g. d, e and f).

Mushin defines evidentials as "…markers which qualify the reliability of information communicated in four primary ways. They specify the source of evidence

on which statements are based, their degree of precision, their probability, and expectations concerning their probability" [9]. According to Mushin's definition, in a broad view evidentiality can specify not only the information sources, but also the speaker's epistemic attitude towards the information. Evidential markers can be encoded in the utterance to convey the speaker's certainty or doubts about the information. With the development of evidentiality theory, most linguists now hold the broad view of evidentiality, in which evidentials have two main functions: (1) to indicate the source of information and (2) to express the speaker's degree of certainty about the information.

2.2 Hierarchies and Scales of Evidentiality

Under the narrow definition, Willett classified evidentiality into two types, direct and indirect, which may be further divided into subtypes [10]. The direct evidence involves the visual sense, the auditory sense, and other sensory. Indirect evidence means that the knowledge is acquired through either reported evidence (which may be specifically marked as hearsay or as part of the oral literature) or inferring evidence (which may be marked as involving either observable evidence or a mental construct only).

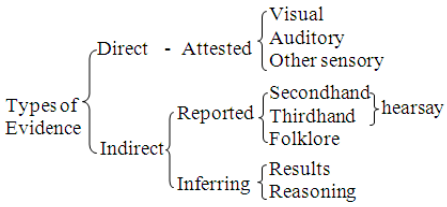


Fig. 1. Willet[10]'s classification of evidentiality

Aikehenvald also focuses on the narrow aspect of evidentiality. Based on the study of the grammars of over 500 languages, she summarized six evidential types: visual, non-visual sensory, inference, assumption, hearsay and quotative [8].

The evidentials belonging to different types and subtypes semantically signals the strength of the speaker's commitment to the validity of their utterance. Barners suggested a priority hierarchy of evidentials as [11]:

Visual > *Non-visual* > *Apparent* > *Second-hand* > *Assumed*

Whereas Oswalt proposed another hierarchy [12]:

Performative > *Factual* > *Visual* > *Auditory* > *Inferential* > *Quotative*

He believes that when the speaker is talking about the act he himself is performing, the information provided should be the most reliable.

As for the broad definition of evidentiality, Chafe's classification considered both the speaker's epistemic attitudes of the reliability of knowledge and information sources. In Chafe's classification model, knowledge may be regarded by a speaker as more or less reliable as indicated in figure 2 with the suggestion of a continuum from

the most reliable knowledge, at the top, to the least reliable, at the bottom [6]. He also pointed out that the reliability of the four modes of knowing is not fixed. Each mode of knowing can move up and down the scale of reliability.

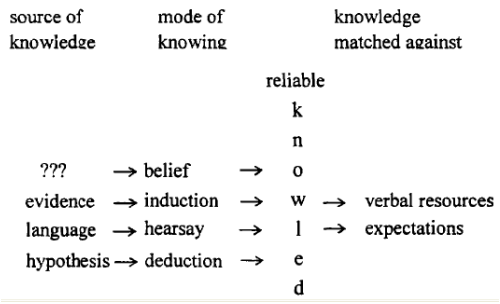


Fig. 2. Chafe's model of evidentiality

Hu further discussed and modified Chafe's model as follows [13]:

Table 1. Hu's model of evidentiality

Source of knowledge	Induction
Culture	Hearsay
Sense	Deduction
Language	Knowledge matched against
Hypothesis	Verbal Resources
Mode of knowing	Expectation
Belief	

Based on this model, Hu proposed seven types of evidentials in his comparative study of evidentials in news reports and debating discourse, which are belief (culture evidence), induction (sensory evidence), heresay (language evidence), deduction (hypothetic evidence), reliability, verbal resources, and expectation. He was also aware of the zero-marked evidentials, which means there is no explicit occurrence of any evidentials. Actually, when proposition is cast in the zero-marked evidenial, such proposition is often regarded as factual [13].

The scale of reliability imprinted in evidentials is actually context-determined, as many evidentials are polysemous and multi-functional [14]. Therefore, evidentiality should be studied in specific context. For example,

If the lights were on, they must have been at home.
If you must leave, do it quietly.

In sentence a), *must* is used to indicate logical probability or presumptive certainty. While in b), *must* means "to be determined to", which didn't show any certainty or uncertainty at all. So, in some languages, evidentiality is strongly related with lexico-semantics, which can be determined by the context.

3 Related Notions and Research

3.1 Biber's Investigation to the Marking of Stance

Biber and Finegan [15] used the term *stance* to mean the lexical and grammatical expression of attitude, feelings, judgments, or commitment concerning the propositional content of a message. Based on 500 texts drawn principally from the LOB and London-Lund corpora, they analyzed and identified adverbial, adjective, verbal and modal marker of stance markers. The stance markers were divided into 12 categories based on semantic and grammatical criteria, as shown in Table 2.

Table 2. Biber's stance categories

Affect markers	Certainty adverbs
Certainty verbs	Certainty adjectives
Doubt adverbs	Doubt verbs
Doubt adjectives	Hedges
Emphatics	Possibility modals
Necessity modals	Predictive modals

In the table we can find that Biber's stance can actually be divided into two parts: affect and evidentiality. For the evidentials which show certainty, they can be adjectives, verbs, adverbs, emphatics, and predictive modals. For those showing uncertainty, they can be adjectives, verbs, adverbs, hedges, possibility modals and necessity modals.

Table 3. Major stance categories investigated (only for evidentiality) in [15]

Certainty	Doubt
Adjectives (e.g., *impossible; obvious; true*)	Adjectives (e.g., *alleged; dubious; uncertain*)
Verbs (e.g., I *conclude*; This *demonstrates* that…)	Verbs (e.g., I *assume*; This *indicates* that…)
Adverbs (e.g., *assuredly; indeed; without doubt*)	Adverbs (e.g., *allegedly; perhaps; supposedly*)
Emphatics (e.g., *for sure; really; so* + ADJ)	Hedges (e.g., *at about; maybe; sort of*)
Predictive modals (e.g., *will; shall*)	Possibility modals (e.g., *might; could*)
	Necessity modals (e.g., *ought; should*)

Biber's investigation provides us the possible word categories of being evidentials as well as some tangible word cases.

3.2 Evidentiality and Epistemic Modality

Modality expresses the attitude of the speaker. Traditionally, it can be divided into two subcategories: deonitic modality and epistemic modality. Deontic modality refers to the speaker's attitude to social factors of obligation, responsibility and permission,

while epistemic modality involves the speaker's attitude to the status of the proposition, namely the possibilities of the factuality of the proposition [16]. Here are three different views on the relations between the notions of evidentiality and modality in linguistic studies. Some linguists tend to favor an overlapping relationship. Some tend to consider that they are conceptually distinguished from each other. Some linguists who hold the broad definition of evidentiality tend to consider epistemic modality is within the scope of evidentiality.

Although different views are held, we can still get to know that there are close relationship between modality and evidentiality. Epistemic modality can serve as a valued hint of the credibility of information.

3.3 Other Corpus-Based Research on Information Credibility

The research surveyed above mainly focus on the theoretical discussion of the linguistic devices of certainty or uncertainty. To date, there are still short of large scale corpus-based research on information credibility. Besides Biber and Finegan's work, which analyzed 500 texts extracted from LOB and London-Lund corpora, Rubin also conducted corpus based research. He proposed a model for uncertainty-certainty continuum and annotated a *The New York Times* dataset accordingly [17]. In his model, Rubin divided credibility into five dimensions, as follows.

Table 4. Rubin's model for uncertainty-certainty continuum

Absolute Certainty	Unambiguous or undisputable conviction, reassurance
High Certainty	High probability or firm knowledge
Moderate Certainty	Average likelihood or reasonable chances
Low Certainty	Distant possibility
Uncertainty	Hesitancy, lack of knowledge or lack of clarity

According to the model, a dataset of 80 articles (40 editorials and 40 news reports) was annotated, and the agreement between independent annotators was calculated. However, we found that the division on different dimensions is difficult to decide. The annotation may be quite subjective given the fuzzy boundary lines between the dimensions.

4 The Choice of Texts: Where Do We Perform Our Annotation Task on?

4.1 Studies on Evidentiality on Different Types of Texts

There have been several researches which analyzed the use of evidentials on specific styles of discourses. Chafe made a contrastive study on evidentiality between English conversation and academic writing [6]. Hu compared evidentiality between news reports and debate discourse [13]. He found that, to make news reports objective, the evidentials concerning reliability and belief are seldom used. However, some groups

of evidentials, like the reporting verbs (such as *said*, *argued*, *added*, etc.) and preposition (such as *according to*), are still commonly used. Those evidential markers could further improve the reliability and objectivity of the information.

4.2 Our Consideration on the Choice of Texts

Based on Hu[13] and Biber & Finegan[15]'s investigation, we find that although the collection of news reports may provide sufficient examples for information source, it is short of linguistic cases to show epistemic attitude. For a comprehensive research on information credibility, the data collection should better be more subjective. Under the consideration, a Chinese corpus which comprises of user-generated content (e.g. reviews and answers) is constructed in our project. A wordlist based filtering process is then conducted to dig out the texts which contain evidential markers. Based on the new text collection, we conduct our evidential annotation.

5 The Annotation Taxonomy

5.1 Construction of an Evidential Dictionary

Through Biber and Finegan [15]'s investigation as well as our previous research, we're aware of that only words belonging to some specific grammatical groups can act as evidentials, which mainly include attibutive/modal adverb, lexical verb, auxiliary verb and epistemic adjective. Actually, nouns can also be used as evidentials. However, the usage of nouns as evidential is extremely complex. That explains why we didn't deal with nouns and other word categories except for the above mentioned four in the current research effort.

Given the four possible grammatical categories, the corresponding items in a Chinese word sense dictionary was extracted and checked by the annotators. The task of the annotators here is to identify the possible evidentials, and compile a wordlist of evidential markers, which can be used to narrow down our pre-collected text set. The form of the word list is like:

huo xu (*maybe*)	epistemic attitude/uncertainty
ting jian(*hear*)	information source
yi ding (*must*)	epistemic attitude/certainty

In the list, an evidential and its role in expressing information credibility was specified.

5.2 Construction of an Evidentiality Annotated Corpus

Given the fact that there is still lack of large scale annotation on evidentiality for Chinese texts, we aim to construct such a corpus in our ongoing project. For the pre-collected user-generated corpus, a wordlist based filtering was conducted with the

purpose of getting a data collection which contains abundant evidentials. Then, trained annotators are requested to identify the credibility of each sentence as well as the linguistic components in the sentence on which to base their judgment for the credibility. The annotations are attached to texts with XML formed tags, which resemble the annotation used in CoNLL-2010 Shared Task "Learning to detect hedges and their scope in natural language text"[3].

<sentence id="S3.4" certainty="uncertain"> Denial by <ccue>others</ccue> of child sexual abuse is <ccue>common</ccue> and its reality is not easily accepted.</sentence>

Fig. 3. Example of a hedge cue annotated sentence in CoNLL 2010 shared task

The difference between our annotation and CoNLL 2010's annotation lies in that we have more detailed tags instead of the merely <ccue> tag used in CoNLL. Now in our annotation, the two types of evidentiality, i.e. epistemic attitude (tagged with "<e>") and information source (tagged with "<s>") were added. Also, for epistemic attitude, we labeled the two scale level of evidentiality degree, i.e. certainty (tagged with "<c>") and uncertainty (tagged with "<u>"). Although Rubin [17]'s five dimensions of credibility maybe useful, we found that it is difficult to reach consensus on what is "high" certainty or "moderate" certainty. So, we decide to adopt a coarse-grained dimension of credibility annotation in the current study, and leave the refinements to further research. Table 6 shows some annotated sentences in our corpus.

a. <eu>*Wo huaiyi*</eu> *zhe bushi zhende.*
<eu>*I doubt*</eu>*this is true.*
b. <s>*Zhongsuozhouzhi*</s>…
<s>*Everyone knows* </s> ,…

Fig. 4. Example of the annotated sentences

6 Conclusion

The credibility of information is an important issue in the information era. In this paper, we propose to explore the issue by the linguistic device of evidentiality. To achieve the aim, an annotated corpus is essential to provide significant research basis for automatic processing. In the ongoing project, we collect user-generated Chinese texts, and compile a Chinese evidentiality dictionary. Based on the corpus and the dictionary, the scale and scope of each evidential in sentences is annotated. We analyze the theoretical base of our annotation guideline and the choices of texts. Also, we discuss the possible category of evidential scales.

Acknowledgements. This work was supported by the project of National Natural Science Foundation of China (No.60903063).

References

1. Mihalcea, R., Strapparava, C.: The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 309–312 (2009)
2. Ott, M., Cardie, C., Hancock, J.: Estimating the Prevalence of Deception in Online Review Communities. In: Proceedings of WWW 2012 (2012)
3. Farkas, R., Vincze, V., Móra, G., Csirik, J., Szarvas, G.: The CoNLL 2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: Processing of CoNLL 2010 Shared Task (2010)
4. Boas, F.: Kwakiutl grammar, with a glossary of the suffixes. Transactions of the American Philosophical Society 37(3), 201–377 (1947)
5. Jakobson, R.S.: Verbal categories, and the Russian verb. Harvard University, Department of Slavic Languages and Literatures (1957)
6. Chafe, W.: Evidentiality: The Linguistic Coding of Epistemology, Evidentiality in English Conversation and Academic Writing. In: Chafe, Nichols (eds.) Evidentiality: The Linguistic Coding of Epistemology. Ablex, Norwood (1986)
7. Bussmann, H.: Routledge dictionary of language and linguistics. Routledge, London (1996)
8. Aikhenvald, A., Dixon, R. (eds.): Studies in evidentiality. John Benjamins Publishing Company, Amsterdam (2003)
9. Mushin, I.: Evidentiality and epistemological stance. John Benjamins, Amsterdam (2010)
10. Willett, T.: A crosslinguistic survey of the grammatication of evidentiality. Studies in Language 12, 51–97 (1988)
11. Barnes, J.: Problems of Dostoevsky's poetics (ed. And trans.) by C. Emerson. University of Minnesota Press, Minneapolis (1984)
12. Oswalt, R.: The evidential system of Kashaya. In: Evidentiality: The Linguistic Coding of Epistemology, Ablex, Norwood (1986)
13. Hu, Z.L.: Avidentiality in language. Foreign Language Teaching and Research 1, 9–15 (1994) (in Chinese)
14. Lazard, G.: On the grammaticalization of evidentiality. Journal of Pragmatics 33(3), 359–367 (2001)
15. Biber, D., Finegan, E.: Styles of Stance in English: Lexical and Grammatical Marking of Evidentiality and Affect. Text 9(1), 93–124 (1989)
16. Saeed, J.: Semantics. Blackwell, Oxford (1997)
17. Rubin, V.: Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. Information Processing and Management 46, 533–540 (2010)

The Semantic Category Restriction on Agent and Patient Syntactic Realization

Shiyong Kang[1,2]

[1] School of Journalism and Communication of Shandong University
[2] Institute of Chinese Information Processing, Ludong University, Yantai 264025, China
kangsy64@163.com

Abstract. The semantic category information is an important resource for semantic research in natural language processing. The semantic category information of noun composition serving as semantic component has a certain restriction on the syntactic position where this semantic component occurs. Based on the labeled corpus, this paper investigates the semantic category distribution on agent and patient syntactic realization, and summarizes the rules and characteristics of the restriction of semantic category on syntactic realization of semantic components.

Keywords: corpus, the semantic category, agent, patient, syntactic realization.

1 Introduction

The well-known linguist, Mr. Lin Xing-guang, pointed out that the research on semantic category is necessary and significant in terms of the information processing [1]. With the preliminary research on the restriction on mapping semantic components into syntactic components based on the labeled corpus, we found that it is useful for explaining the semantic relationship between predicate and noun composition to lucubrate the lexical-semantic features of semantic component [2]. Therefore, the semantic category information is an important resource for semantic research in natural language processing. This paper tries to summarize the rules and characteristics of the restriction of semantic category on syntactic realization of semantic components by describing the distribution of semantic category on agent and patient syntactic realization.

2 Construction of Corpus

Supported by the National Philosophy Social Sciences Foundation, firstly, we constructed "The corpus of syntax and semantics information in Modern Chinese", where both the syntactic structure and semantic structure information of each sentence were labeled. Secondly, we extracted example sentence database of syntactic realization of agent and patient from the corpus, where both agent and patient can

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 373–380, 2013.

project onto subject, object and adverbial, the subject is conventional coordination of agent, the object is conventional coordination of patient, and the other positions are unconventional coordination of agent and patient. Thirdly, we extracted agent subject, agent object, agent adverbial, patient subject, patient object, patient adverbial and their respective core verb from the example sentence database, put them into a Excel table, and use program to label the semantic category for core noun composition according to "Chinese Thesaurus"[3]. Finally, we generate "The corpus of syntax , semantics and semantic category information of agent and patient" which included the following six sub-tables, each of which describes a syntax semantics position, and these syntax semantics positions are agent subject (SSPV), agent object (PVOS), agent adverbial (DSPV), patient subject (SOPV), patient object (PVOO), patient adverbial (DOPV).

3 The Distribution of Semantic Category on Syntax Semantics Position

3.1 The Distribution of the Large-Class Semantic Category on Syntax Semantics Position

The "Chinese Thesaurus" has 12 large-class semantic categories and 94 middle-class semantic categories. In the following table, we exhibited the different semantic category distribution on agent and patient syntactic realization.

Table 1. The different semantic category distribution on agent and patient syntactic realization

component / semantic category	SSPV	PVOS	DSPV	PVOO	SOPV	DOPV
A (human)	10348	45	215	934	60	33
B (substance)	1957	113	24	2984	203	110
C (time and space)	160	/	5	320	10	5
D (abstract thing)	941	8	27	2702	94	42
E (character)	82	3	/	177	3	6
F (action)	/	/	/	15	/	1
G (mental activity)	20	1	/	92	1	3
H (activity)	80	/	1	273	/	/
I (phenomenon and state)	40	/	2	95	/	2
J (relevancy)	7	/	/	14		/
K (assistant)	/	/	/	/	/	/
L (honorific)	/	/	/	/	/	/

According to the occurrence frequency of various semantic categories in the six syntax semantics positions, some inequalities are given as follows:

a1. Agent Subject: A>B>D>C>E>H>I>G>J

a2. Agent Adverbial: A>B>D>C>I>H

a3. Agent Object: B>A>D>E>G

b1. Patient Object: B>D>A>C>H>E>I>G>J>F

b2. Patient Adverbial: B>D>A>E>C>G>I>F

b3. Patient Subject: B>D>A>C>E>G

According to the Table 1 and above inequalities, some features can be found as follows,

(1) The number of the types of semantic category which occurs in the conventional coordination of agent and patient is more than that occurs in the unconventional coordination. And the number of the types of semantic category which occurs in the conventional coordination of patient is more than that occurs in the conventional coordination of agent.

(2) A [human], B [substance], and D [abstract things] are the most common semantic categories serving as agent and the patient, accounting for 97.22% and 87.71% of the total agent and patient, respectively. They can also occur in all syntactic positions to which these two components can map. K [assistant] and L [honorific] generally do not serve as agent and patient. C [time and space], E [feature], G [mental activity], H [activity], and I [phenomenon and state] can serve as agent and patient, but only a small number of them serve as agent and patient, as well as they only occur in the conventional coordination of agent and patient, and do not occur or only sporadically occur in the unconventional coordination. F [Action] only serves as patient.

(3) As can be seen from above inequalities, A [human] prefers to serving as agent, and B [substance] prefers to serving as patient, both of which accord with our cognitive experience, but B [substance] appears in the front-end position of the inequality a3.

In general, in term of the large-class semantic category, the noun compositions of different semantic categories exhibit obvious imbalance in the above six semantic syntactic positions. So, in order to deeply analyze the reasons for forming above features and clearly understand the restriction of semantic category on semantic component syntactic realization, it is necessary to investigate the middle-class semantic category.

3.2 The Distribution of the Middle-Class Semantic Category on Syntax Semantics Position

By investigating the middle-class semantic category distribution on six language chunks generated by agent and patient syntactic realization, the middle-class semantic category can be roughly divided into the following three types, as shown in Table 2.

In the Table 2, the null projection indicates that the semantic category does not project onto any language chunk, and involves 15 middle-class semantic categories; the finite project indicates that the semantic category can only project onto one semantic component, and involves 7 middle-class semantic categories which only can serve as patient; and the focus projection includes two cases, denoted by I and II. The case I indicates the semantic category only project onto conventional coordination of semantic component, and involves 23 middle-class semantic categories; and the case II indicates the semantic category can project onto many positions, but the most of positions are conventional coordination, and few positions are unconventional coordination, the case II involves 49 middle-class semantic categories. According to the distribution of semantic category on the different syntax semantics positions, the semantic categories included in the focus projection are divided into three kinds of the tendency of projection.

a) Null projection: Ea, Fb, Fc, Fd, Gc, Hk, Hl, Ja, Jb, Ka, Kb, Kc, Kd, Ke, Kf

b) Finite projection: Bc, Fa, Hi, Hm, Hn, Ic, Jc

c) Focus projection:

I
- Tend to project onto Agent Subject: Ac, Ak, Bj, Dh, Id
- Tend to project onto Patient Object: Dg, Ee, Ga, Ha, Hb, Hd, Hf, Hh, Hc, Hg, Ia, Ih, Ie, Dl,
- The tendency of projection is not obvious: Ig, Je, Jd, He

II
- Tend to project onto Agent Subject: Aa, Ae, Ag, Ai, Am, Ab, Ad, Af, Ah, Aj, Al, An, Be, Bi, Bd, Bf, Dm
- Tend to project onto Patient Object: Bg, Bk, Bo, Bq, Bb, Bh, Bm, Bn, Bl, Bp, Br, Ca, Cb, Da, Dc, De, Di, Dk, Db, Dd, Df,, Dn, Dj, Ed, Ef, Gb, Ib
- The tendency of projection is not obvious: Ba, Ec, Eb, Hj, If

In the first one, the semantic categories which tend to project onto agent subject have the features of [+Dynamic], [+Control] and [+Specificity]. In these semantic categories, in addition to A [human] and Bi [animals] which have the feature of [+Vitality], Bf [weather], Bd [celestial bodies], Be [physiognomy], and so on, are independent of people's opinions, and rarely compelled by the external force. So these semantic categories rarely serve as patient. For example:

牵牛星[Bd类]在移动、漳河水[Be类]淹了他们的村庄
qianniuxing zai yidong, zhangheshui yan le tamen de cunzhuang
Altair was moving, Zhanghe water flooded their village

In the second one, the semantic categories which tend to project onto patient object have the features of [+ Controlled], [± Specificity]. The words which belong to these semantic categories have weak vitality, such as Bn[building], Bq[clothings], Bp[articles], Br[foodstuff, medicine], and so when they serve as agent, there exist many restrictions on them, such as, some semantic categories need the methods of personate, and some need the help of external force. For example:

小桥[Bn]走进我的心里
xiaoqiao zoujin wo de xinli
Bridges went into my heart

风筝[Bp]随风飘荡
fengzheng suifeng piaodang
The kite gones with the wind

In the third one, the number of semantic categories which project onto agent subject almost equal to the number of semantic categories which project onto patient object, and in these semantic categories, the feature of [+ Abstract] is obvious, while the feature of [±Controlled] is not obvious. These semantic categories include Ig[beginning and ending], Je[influence], If[circumstance] and Ec[color and taste].

4 The Feature of Restriction of Semantic Categories on Semantic Role Syntax Realization

There are 73 middle-class semantic categories which can serve as the agent mapping into subject, while there are 80 middle-class semantic categories which can serve as the patient mapping into object. What's more, the number of types of semantic category which occur in conventional coordination is more than in unconventional coordination. The typical features of agent are [+Control] and [+Dominate], the typical features of patient are [+ Controlled] and [+Dominated]. The human with vitality is the prototype of agent. As the main part of society, the human has powerful productivity and creativity, and control and dominate a large number of objects where even some objects are bestowed a certain controlling force fixedly or provisionally by human, and become sub-prototypes and non-prototypes of agent. The semantic category can choose its own semantic component and syntax position.

According to frequency at which the middle-class semantic category occurs in six positions, the inequalities are generated as follows, where the first ten semantic categories and their proportions in the total of six inequalities are given in Table 2.

Table 2. The first ten semantic categories and their proportions in the total of six inequalities

agent	syntax position	semantic categorysequence	proportion
	subject	Aa>Ah>Ab>Bi>Ba>Ae>Af>Bk>Dk>Dm	83.94%
	adverbial	Aa>Dm>Ae>Ab>Aj>Di>Bf>Ba>Da>Ag	87.23%
	object	Bf>Bg>Aa>Ab>Bi>Bh>Bl>Ae>Bk>Bd	81.74%
patient	syntax position	semantic categorysequence	proportion
	adverbial	Da>Aa>Dk>Bk>Ba>Bp>Br>Di>Bg>Dd	55.67%
	adverbial	Ba>Aa>Bk>Bp>Dj>Di>Bi>Da>Bq>Bm	68.84%
	subject	Aa>Dk>Da>Bn>Bh>Ba>Bp>Bo>Bi>Br	72.24%

In the three inequalities which are generated by agent mapping into three syntax position (subject, object and adverbial), the sum of the number of the first ten semantic categories accounts for 80% of the total. And the number of type of semantic categories which serve as patient is great. In the three inequalities which are generated by patient mapping into three syntax position, the sum of the number of first ten semantic categories accounts for 56% of the total. The proportion of other semantic categories (such as, Df[consciousness] and Dj[economic],) is also great..

In general, the semantic categories in the front of the six inequalities directly contact with current production and daily life, such as people (including the independent and collective), animals, plants, natural objects, and artificiality. Because of the different semantic features, the conventional coordinations of the semantic categories are different. The Aa[general term] occurs in the front of the six inequalities, which shows that the human has obvious feature of [+Control] and [+Controlled]. The human, as the main part of the society, interact with each other, and this complex relationship also accurately is reflected to the product of social life—the language. In addition, the Dk [cultural], Bk [body], and Di [social, politics and law] often serve as both patient and agent, because they are closely linked to people, bestowed with the ability of control, and have positive impact on the life of human society.

Be affected by the semantic feature of sentence pattern, Bd [celestial bodies], Bf [weather] and Bg [natural things] occur in the front of inequalities of agent mapping into object and occur in the back of inequalities of agent mapping into subject and adverbial.

In the sentences where agent maps into subject and agent led by preposition serves as adverbial, their structures emphasize the meaning of triggering, while in the sentences where agent maps into object, their structures emphasize the meaning of existence. Relatively, in the agent adverbial, agent subject, and agent object, the first one has the highest requirement on the Vitality of agent, and followed by the second one and the third one sequentially.

By exploring the restriction of semantic category (including large-class and middle-class) on semantic component' syntax realization, we found the following projection characteristics:

(1) polymerization

From the Table 2, we observe among the three types, the focus projection accounts for the biggest percentage. 72 middle-class semantic categories intensively project onto conventional coordinations of agent and patient, where 22 middle-class semantic categories preferentially choose agent subject, and 22 middle-class semantic categories preferentially choose patient object. In addition, there are rules in semantic categories which occur in unconventional coordinations, more than 70% of semantic categories correspondingly occur in the two unconventional coordinations. In a word, semantic categories with common semantic features gather together, and syntax semantics positions which they occur have uniformity and analogy.

(2) naturalization

The semantic categories which serve as prototypes of agent have the feature of [+Vitality] and [+Specificity], while some semantic categories with the feature of [-Vitality] and [-Specificity] serve as agent usually. The later has close contact with A

[human], such as, Dm [organization], Di [social, political and law], Dk [cultural], and Df [consciousness] are formulated and disseminated by human, and are bestowed some kind of ability to control by human. For example,

政府[Dm]采取了一切有效措施
zhengfu caiqu le yiqie youxiao cuoshi
Government has took all effective measures

科学[Dk]战胜了 愚昧
kexue zhansheng le yumei
science has overcame ignorance

Some semantic categories exhibit obvious features of [+Controlled] and [-Vitality], and sometimes serve as agent, such as Bo [equipment], and Bm [stuff], because they usually are dominated and controlled by human. For example,

马车[Bo]冲过来了
mache chong guolai le
The carriage is rushing

无情的泥土[Bm]吞噬了世上最可爱的人儿
wuqing de nitu tunshi le shishang zui ke'ai de ren
Merciless mud engulfed the most lovable person in the world

(3) inheritance

Some semantic categories have the relationship between the whole and the part. For example, Bk [body] and Bl [secretions and excretions] can be regarded as the parts of the human body, and inherit the ability of control from the human, so they have the same coordination capacity as A [human]. For example,

她的眼睛[Bk]告诉我……
ta de yanjing gaosu wo
He eyes told me

泪[Bl]顺着脸庞滑落下来
lei shunzhe lianpang hualuo xialai
Tears fall down along the face

Bh [Plant], Bp [supplies] and Br [Food and Drug] are dominated by human usually, but the Bh [plants] has the same feature of vitality as A [human], and inherits some capacity and emotion from human, so it often serves as agent. However, because Bp and Br do not have the features of vitality, their ability of serving as agent is far weaker than Bh class. Such as, "person (Aa01)" usually associates with "flower (Bh02)" through a verb (such as " cultivate," "buy," "pour" etc.), "flowers" provisionally acquires semantic features of human. In terms of the pragmatics, it is a

rhetorical phenomenon. And from the semantic point of view, the commonality between the two classes leads to inheritance of semantic categories' features. For example,

有些花[Bh]还选择昆虫
youxie hua hai xuanze kunchong
Some flowers selected insects

(4) variability

The co-taxonymy means some semantic categories locate in same class. Roughly, the distribution of co-taxonymy which serves as semantic component in syntactic coordination is consistent basically, but there exists still some "disharmony" sometimes.

The conventional coordination which some co-taxonymies project onto is consistent, but unconventional coordination is inconsistent, such as, the syntactic projection priority degree for Aa [general term], Ag [status], Ah [kinfolks], and Aj [relation] of A class serving as agent is "subject>adverbial>object", the syntactic projection priority degree for Ab [men and women] serving as agent is "subject>object>adverbial", and Ac [posture] and Ak [trait] generate vacancy in the object and adverbial. The formation of syntactic vacancy may be limited by the scale of corpus, but this imbalance and variability also reflect certain regularity when using the language, which remains to be further studied.

5 The Prospect

Needless to say, it is entirely possible for a large-scale labeled corpus including the syntactic structural information, the semantic structural information, and the semantic category information to establish the mapping and linking relationship among the sentence pattern, semantic relationship and lexical semantic features. The study on the rules and features of the restriction of semantic category on semantic components' syntactic realization has yet to be proceeded with. In the future, we will examine the other semantic components except agent and patient, lucubrate the small-class semantic category, and exploit the lexical semantic features of predicate verbs and the features of sentence pattern to reveal the overall corresponding mechanism between the semantic components with syntactic components.

References

1. Lin, X.G.: Glossary semantic and computational linguistics. Chinese Publishing House, Beijing (1999)
2. Xu, X.X., Kang, S.Y.: Correspondence between Syntactic and Semantic Components in Modern Chinese Based on Labeled Corpus. In: Recent Advance of Chinese Computing Technologies, Singapore (2007)
3. Mei, J.J., et al.: Chinese Thesaurus. Shanghai Lexicographic Publishing House, Shanghai (1983)

Towards an Event-Based Classification System for Non-natural Kind Nouns

Shan Wang and Chu-Ren Huang

Dept. of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
{wangshanstar,churenhuang}@gmail.com

Abstract. Nouns are usually divided syntactically or semantically. These classifications, however, ignore the fact that some nouns can represent events. This behavior is similar to verbs. This study examines three types of nouns (nominals, pure event nouns and entity nouns) based on eventive parameters (classifiers, argument structure and event structure). It establishes an event-based noun classification system for non-natural kind nouns, which not only enriches the research on noun classifications, but also facilitates event detection in natural texts.

Keywords: an event-based classification system, non-natural kind noun, classifier, argument structure, event structure.

1 Introduction

It is commonly accepted that nouns represent entities and verbs stand for actions. Traditionally nouns are usually classified semantically or syntactically using classifiers [1-4]. However, such classifications ignore the fact that some nouns can express events, in a way that is similar to verbs. In recent years, there has been growing interest in these event-representing nouns. Previous research in Mandarin Chinese includes the constructions in which they usually appear [5-7], their classifiers [4], [6], [8], internal and external temporal attributes [9], their properties from a Generative Lexicon perspective [10-16].

However, a noun classification system based on eventive features has not yet been established. Such a system will facilitate nominal event detection in natural language processing. In turn, it will help us to make event-based temporal inferences, which will support information extraction, question answering and text summarization. This paper will re-examine the characteristics of different types of nouns and establish an event-based classification system for non-natural kind nouns.

The data in this research are collected from four sources: (a) Corpus of Center for Chinese Linguistics (CCL)[1], (b) internet data accessed online via Google and Baidu,

[1] http://ccl.pku.edu.cn:8080/ccl_corpus/

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 381–395, 2013.

(c) a balanced Modern Chinese corpus *Sinica Corpus*[2], accessed through *Chinese Word Sketch Engine*[3], and (d) a few examples from native speakers.

2 Literature Review

[17] illustrates the differences among process nominals, result nominals and simple event nouns as shown in Table 1.

Table 1. Differences between Process Nominals, Result Nominals and Simple Event Nouns

	Determiner System		Argument Structure		Event Structure		
	Determiner	Plural	Subject-Oriented Adj.	Argument Taking	Frequency Adj.	Time Expression	Rational Clause
Process	the/Ø	−	+	+	+	+	+
Result	the, a, that	+	−	−	−	−	−
Simple Event Noun	the, a, that	+	−	−	−	−	−

[18] claims that in Mandarin Chinese process nominals, result nominals and simple event nouns have the same behavior as those in English.

Their analysis ignores the difference between natural and non-natural kind noun. [19-21] separate the domain of individuals into three distinct levels: (a) natural types, which direct at the formal and constitutive qualia roles; (b) artifactual types, which refer to telic or agentive roles; (c) complex types, which make references to the relation between types. [22] further discusses three linguistic diagnostics which motivate a fundamental distinction between natural and unnatural kinds. These diagnostics are: (a) Nominal Predication: How the common noun behaves predicatively; (b) Adjectival Predication: How adjectives modifying the common noun can be interpreted; (c) Interpretation in Coercive Contexts: How NPs with the common noun are interpreted in coercive environments. Since natural and non-natural kinds have significant differences, our study will only consider non-natural kind nouns.

3 Data Analysis

This paper examines nominals, pure event nouns and entity nouns. Their definitions are as follows.

Nominal: it refers to the noun that has a verbal form. Nominals include three types: (a) Process Nominal: it refers to the noun that has a process reading, which can last for a period; (b) Result Nominal: it refers to the noun that expresses a result, which is similar to an entity. For example, 调查*diàochá* 'investigation' is both a process nominal and a result nominal; and (c) Instant Nominal: it refers to the noun that has an instant event reading, such as 叛变*pànbiàn* 'renegading' and 奖励*jiǎnglì* 'rewarding'. Both process nominals and instant nominals are called event nominals.

[2] http://db1x.sinica.edu.tw/kiwi/mkiwi/
[3] http://wordsketch.ling.sinica.edu.tw/

Pure Event Noun: it refers to the noun that has a process reading but does not have a verbal form, such as 会议*huìyì* 'conference' and 婚礼*hūnlǐ* 'wedding'. They are called simple event nouns in [17-18].

Entity Noun: it refers to the noun that does not have an event reading and a verbal form. This type of nouns comprises Concrete Entity Nouns like 黑板*hēibǎn* 'black-board' and Abstract Entity Nouns like 政策*zhèngcè* 'policy'.

Table 2 illustrates these different types of nouns.

Table 2. Different Types of Nouns

Nominal	Even Nominal	Process Nominal
		Instant Nominal
	Result Nominal	
Pure Event Noun	Process Reading	
Entity Noun	Concrete Entity Noun	
	Abstract Entity Noun	

[17-18] did not examine instant nominals. To make corresponding comparison with their research, this paper does not take these nominals into consideration. Therefore, this study only examines process nominals, result nominals, pure event nouns and entity nouns that are non-natural kinds.

3.1 Classifier

[18] points out that of [17]'s tests (Table 1), the determiner and plural marking on nouns do not apply to Mandarin. This is because there is no overt contrast between indefinite and definite nouns, nor is there plural marking on nouns. Classifiers can be viewed as something in the determiner system that differentiate process and result readings: (a) process classifiers [e.g., 次*cì* 'occurrence', 回 *huí* 'occurrence (for come and go)', 遍*biàn* 'time (from beginning to the end)'] select process nouns; and (b) non-process classifiers [e.g. 个*gè* '(for most objects)', 条*tiáo* (for river, stick), 张*zhāng* '(for paper, table) '] select result nouns.

Most derived nominals can take both process and non-process classifiers [18]. For example:

一场报告 / 一篇报告
yī chǎng bàogào / yī piān bàogào
a CL reporting / a CL reporing

Some of the derived nominals only have the process reading, and thus only allow process classifiers [18].

一次休息/*一个休息
yī cì xiūxi / * yī gè xiūxi
a CL resting / * a CL resting

We are in favor of her analysis of using classifiers to identify process and result reading nominals. But there are two points we need to note: (a) 次*cì* 'once (re. frequency of event)' is only used to count the frequency of events. It cannot show whether an event

has a process or instant reading; (b) as much research has pointed out, 个*gè* 'CL' is a neutral classifier. Thus it is improper to treat 个*gè* 'CL' as a non-process classifier.

Mandarin Chinese Classifier and Noun-Classifier Collocation Dictionary ([23] and [24]) lists 35 event classifiers in Mandarin Chinese: 波*bō* 'of staggered event', 班*bān* 'of shift, scheduled flight/bus etc', 筆*bǐ*'of transaction', 步*bù* 'step (event procedures)', 泡*pào* 'a brewing (of tea etc.)', 盤*pán* 'a serving round (of a dish)', 幕*mù* 'cut (of a play)', 番*fān* 'times (of a repeated event)', 道*dào* 'of dishes of procedures', 檔*dàng* 'duration of run (of play, movie etc.)', 段*duàn* 'section (of play, etc.)', 頓*dùn* 'the process of a meal', 台*tái* 'a run of a traveling troupe', 堂*táng* 'a class', 趟*tàng* 'a journey', 通*tōng* 'a phone call', 輪*lún* 'a round', 回*huí* 'a roundtrip', 節*jié* 'a class, a session', 屆*jiè* 'an annual event', 件*jiàn* 'event', 局*jú* 'game', 期*qī* 'term', 起*qǐ* 'event (especially a happening, an accident)', 圈*quān* 'round (of majong)', 席*xí* 'lecture', 折*zhé* 'an act (in a Chinese play)', 陣*zhè* 'one of a sporadic event(s)', 椿*zhuāng* 'event', 場*chǎng* 'a (scheduled) event (with beginning and ending)', 齣*chū* 'a play', 任*rèn* 'term (of a termed position)', 宗*zōng* 'trade/transaction', 餐*cān* 'a meal', 次*cì* 'once (re. frequency of event)'. The mutual selection between 波*bō* 'of staggered event' and nouns has been conducted in [8]. These classifiers can assist in determining whether a noun has an eventive reading.

3.2 Argument Structure

3.2.1 Argument Taking

[18] uses对*duì* 'to'-PP and关于*guānyú* 'about'-PP to test whether a noun needs an argument as shown in (1).

(1) a. 他??(对灾情)的报道进行了三个小时。
Tā??(duì zāiqíng) de bàodào jìnxíng le sān gè xiǎoshí.
He to disaster DE reporting proceed ASP three CL hour
'His reporting to the disaster lasted three hours.'

b. 他({??对/关于}灾情)的报道发表了。
Tā ({?? duì/guānyú} zāiqíng) de bàodào fābiǎo le.
he to/about disaster DE report publish ASP
'His report (of the disaster) was published.'

c. 他({*对/关于}灾情)的文章发表了。
Tā ({*duì/guānyú} zāiqíng) de wénzhāng fābiǎo le.
he to/about disaster DE article publish ASP
'His article (about the disaster) was published.'

[18] claims that in (1a) the process nominal 报道b*àodào* 'reporting' requires an obligatory argument expressed by对*duì* 'to'-PP. In (1b) result nominal报道*bàodào* 'report' and in (1c) the entity noun 文章*wénzhāng* 'article' admit the adjunct 关于*guānyú* 'about'-PP, not the argument 对*duì* 'to'-PP. The difference between (1a) and (1b, 1c) indicates that对*duì* 'to'-PP introduces arguments, and thus process nominals take obligatory arguments; 关于*guānyú* 'about'-PP introduces adjuncts, and thus result nominals do not take obligatory arguments.

We agree that process nominals take arguments, while result nominals and entity nouns take adjuncts. We further find that pure event nouns also take arguments. Since both process nominals and pure event nouns express events, they both have compulsory event participants. That is to say, they both take arguments, as demonstrated in (2) and (3).

Process Nominal

(2) 他们对政策的**说明**进行了三个小时。
Tāmen duì zhèngcè de **shuōmíng** jìnxíng le sān gè xiǎoshí.
they to policy de explanation proceed ASP three CL hour
'Their explanation of the policy lasted three hours. '

Pure Event Noun

(3) 这次关于改善民生的**会议**进行了三天。
Zhè cì guānyú gǎishàn mínshēng de **huìyì** jinxing le sān tiān.
This CL about improve people's livelihood DE conference proceed ASP three day
'The meeting on improving people's livelihood lasted three days.'

In (2), 政策*zhèngcè* 'policy' is the event participant of the process nominal 说明 *shuōmíng* 'explanation'. In (3), 改善民生*gǎishàn mínshēng* 'improve people's livelihood' is the event participant of the pure event noun 会议*huìyì* 'conference'.

Neither result nominals nor entity nouns express events, so neither take arguments, as illustrated in (4) and (5). In (4), 造纸术*zàozhǐshù* 'papermaking' is a specification of the result nominal 发明*fǎmíng* 'invention', not an argument. In (5), 电器*diànqì* 'electrical appliance' explains what the concrete entity noun 商店*shāngdiàn* 'store' sells. 电器*diànqì* 'electrical appliance' is not an event participant, because 商店 *shāngdiàn* 'store' is an entity not an event. In (6), 团结当地高山族人民 introduces what the 政策*zhèngcè* 'policy' is. Since 政策*zhèngcè* 'policy' is an abstract entity noun, 团结当地高山族人民*tuánjié dāngdì gāoshān zú rénmín* 'unite the people of the local Gaoshan ethic group' is not an event participant.

Result Nominal

(4) 造纸术的发明促进了人类文明的传播。
Zàozhǐshù de **fǎmíng** cùjìn le rénlèi wénmíng de chuánbò.
Papermaking DE invention promote ASP mankind civilization DE spread
'The invention of papermaking promotes the spread of human civilization.'

Entity Noun:

(i) Concrete Entity Noun:

(5) 一些电器**商店**的冷气机销量升幅多达两成。
Yīxiē diànqì **shāngdiàn** de lěngqìjī xiāoliàng shēngfú
some electrical appliance store DE cold air conditioners sale increase
duōdá liǎngchéng.
up to 20%
'The sale of cold air conditioners in some electrical appliance stores increased up to 20%.'

(ii) Abstract Entity Noun:

(6) 他还采取了团结当地高山族人民的**政策**, 密切了台湾地区的民族关系。
Tā hái cǎiqǔ le tuánjié dāngdì gāoshān zú rénmín de **zhèngcè**,
He also take ASP unite local Gaoshan ethic group people DE policy,
mìqiè le táiwān dìqū de mínzú guānxì.
intimate ASP Taiwan area DE ethic group relation
'He has also taken a policy of uniting the people of the local Gaoshan ethic group, which intimated the ethnic relations in Taiwan.'

However, we do not agree that对*duì* 'to'-PP introduces arguments and关于*guānyú* 'about'-PP introduces adjuncts. First, process nominals admit both对*duì* 'to'-PP and 关于*guānyú* 'about'-PP. For example, in (7a) and (7b), the process nominal说明 *shuōmíng* 'explanation' admits 对政策*duì zhèngcè* 'to the policy' and 关于政策 *guānyú zhèngcè* 'about the policy' respectively.

(7) a. 他们对政策的说明进行了三个小时。(process nominal)
Tāmen duì zhèngcè de **shuōmíng** jìnxíng le sān gè xiǎoshí.
they to policy de explanation proceed ASP three CL hour
'Their explanation to the policy lasted three hours. '
b. 他们关于政策的说明进行了三个小时。(process nominal)
Tāmen guānyú zhèngcè de **shuōmíng** jìnxíng le sān gè xiǎoshí.
they about policy de explanation proceed ASP three CL hour
'Their explanation about the policy lasted three hours.'

Secondly, pure event nouns also take arguments, which can be introduced by关于 *guānyú* 'about'-PP, not 对*duì* 'to'-PP, as shown in (8).

(8) 关于全球贸易的**会议**进行了两天。
Guānyú quánqiú màoyì de **huìyì** jìnxíng le liǎng tiān.
About global trade DE conference last ASP two day
'The conference on global trade lasted two days.'

[25] divides arguments into true arguments, default arguments, and shadow arguments. This paper follows this distinction.

True Arguments: parameters of the lexical item that are syntactically realized. *Tom* is the true argument of *jump* in (9).

(9) Tom jumped.

Default Arguments: parameters of the lexical item that are not syntactically expressed, but participate in the qualia. *War* takes two default arguments which are optional in the syntax. However, they are logically obligatory as shown in (10).

(10) the war between the U.S. and Vietnam

Shadow Arguments: parameters which are semantically incorporated into the lexical item. *Kick* incorporate *leg* in its meaning, so *with his right leg* is a shadow argument in (11).

(11) He kicked the door with his right leg.

Process nominals and pure event nouns often have default arguments. For instance, 会议*huìyì* 'conference' is a pure event noun, it takes default arguments. In (12a), no argument of会议*huìyì* 'conference' is syntactically expressed, but logically it has two default arguments. (12b) two default arguments are expressed: topic (developing missile technology) and interlocutors (Zhōu'ēnlái, Qian Xuesen, etc.).

(12) a. 对于这次会, 毛泽东在闭幕时的讲话中说："这个会议开得很好。"
Duìyú zhè cì huì, máozédōng zài bìmù shí de jiǎnghuà
About this CL meeting, Mao Zedong in closing session time DE speech
zhōng shuō: "Zhè ge huìyì kāi de hěn hǎo."
in say: 'this CL meeting hold DE very good'
"For this meeting, Mao Zedong in his closing speech said: 'This meeting went very well.'"

b.周恩来主持中央军委会议, 听取钱学森关于在中国发展导弹技术的规划设想。
Zhōu'ēnlái zhǔchí zhōngyāng jūnwěi huìyì, tīngqǔ Qián xuésēn
Zhou Enlai preside the Central Military Commission meeting, hear Qian Xuesen
guānyú zài zhōngguó fāzhǎn dǎodàn jìshù de guīhuà shèxiǎng.
about in China develop missile technology DE planning assumption
'Zhou Enlai presided over the Central Military Commission meeting to hear Qian Xuesen's planning assumption on developing missile technology in China.'

3.2.2 Subject-Oriented Adjective

[18] tests which types of nouns admits subject-oriented adjectives by using不怀好意的*bùhuáihǎoyìde* 'malicious' as shown in (13).

(13) a. [他不怀好意的 ??(对灾情)的报道]进行了三个小时。
[Tā bùhuáihǎoyìde??(duì zāiqíng) de **bàodào**] jìnxíng le sān gè xiǎoshí.
he malicious to disaster DE reporting proceed ASP three CL hours
'His malicious reporting of the disaster lasted three hours.'

b.??他不怀好意的(关于灾情的)报道发表了。
?? Tā bùhuáihǎoyìde (guānyú zāiqíng de) **bàodào** fābiǎo le.
he malicious about disaster DE reporting publish ASP

c.??他不怀好意的(关于灾情的)文章发表了。
?? Tā bùhuáihǎoyìde (guānyú zāiqíng de) **wénzhāng** fābiǎo le.
he malicious about disaster DE article publish ASP

In (13a) the process nominal 报道*bàodào* 'reporting' admits the subject-oriented adjective不怀好意的*bùhuáihǎoyìde* 'malicious'. Neither does the result nominal 报道*bàodào* 'report' nor the entity noun 文章*wénzhāng* 'article' admits this adjective.

In our analysis, not only process nominals, but also pure event nouns, result nominals and abstract entity nouns can admit subject-oriented adjectives, as shown in (14)-(17).

Process Nominal

(14) 克林顿本人都支持他的努力, 但却明显地遭到克林顿许多高级顾问们蓄意的、强烈的**抵制**。

Kèlíndùn běnrén dōu zhīchí tāde nǔlì, dànquè míngxiǎnde zāodào kèlíndùn
Clinton oneself even support his effort, but clearly suffer Clinton
xǔduō gāojí gùwènmen xùyìde, qiángliède **dǐzhì**.
many senior adviser deliberate, strong resistance
'Clinton himself supported his efforts, but clearly many of Clinton's senior advisers displayed deliberate and strong resistance.'

Pure Event Noun

(15) a. 这是一场他精心策划的婚礼。

Zhè shì yī chǎng tā jīngxīn cèhuà de **hūnlǐ**.
This is a CL he thoughtfully planned wedding
'This is a wedding that he thoughtfully planned.'

b. 这是敌人处心积虑的阴谋。

Zhè shì dírén chǔxīnjīlǜde **yīnmóu**.
This is enemy deliberate conspiracy
'This is the enemy's deliberate conspiracy.'

Result Nominal

(16) 他的那条不怀好意的**建议**

tāde nà tiáo bùhuáihǎoyìde **jiànyì**
his that CL malicious suggestion
'his that malicious suggestion'

Entity Noun:

(i) Abstract Entity Noun

(17) 敌人不怀好意的政策 (admit)

dírén bùhuáihǎoyìde **zhèngcè**
enemy malicious policy
'enemy's malicious policy'

(ii) Concrete Entity Noun

(18) *不怀好意的**桌子** (doesn't admit)

bùhuáihǎoyì de zhuōzi
malicious DE table

In (14)-(17), 蓄意的*xùyìde* 'deliberate', 精心策划的*jīngxīncèhuàde* 'thoughtfully planned', 处心积虑的*chǔxīnjīlǜde* 'deliberate', 不怀好意的*bùhuáihǎoyìde* 'malicious' are all subject-oriented adjectives. In (14) the process nominal 抵制*dǐzhì* 'resistance' can be modified by 蓄意的*xùyìde* 'deliberate'. In (15), the pure event noun 婚礼*hūnlǐ* 'wedding' can be modified by 精心策划的*jīngxīncèhuàde* 'thoughtfully planned'; another pure event noun 阴谋*yīnmóu* **'**conspiracy**'** can be modified by 处心积虑的*chǔxīnjīlǜde* 'deliberate'. In (16), the result nominal 建议*jiànyì* 'suggestion' can be modified by 不怀好意的*bùhuáihǎoyìde* 'malicious'. In (17), 不怀好意

的*bùhuáihǎoyìde* 'malicious' can modify the abstract entity noun 政策*zhèngcè* 'policy'.

The only type that does not admit the subject-oriented adjectives is the concrete entity noun, as shown in (18). 桌子*zhuōzi* 'table' cannot be modified by不怀好意的*bùhuáihǎoyìde* 'malicious'.

In summary, this section has illustrated that process nominals and pure event nouns admit subject-oriented adjectives and take arguments. Result nominals and abstract entity nouns admit subject-oriented adjectives, but do not take arguments. Concrete entity nouns neither admit subject-oriented adjectives nor take arguments.

3.3 Event Structure

3.3.1 Frequency Adjectives

[18] uses 经常不断的*jīngchángbùduànde* 'frequent' to test which type of nouns admits frequency adjectives. She shows that only process nominals admit it as shown in (19).

(19) a. 他经常不断的 ??(对灾情)的报道十分有用。
Tā jīngchángbùduàn-de??(Duì zāiqíng) de bàodào shífēn yǒuyòng.
he frequently to disaster DE reporting very useful
'His frequent reporting of the disaster is very useful.'
b. *他经常不断的(关于灾情的)报道发表了。
*Tā jīngchángbùduàn-de (guānyú zāiqíng de) bàodào fābiǎo le.
he frequently about disaster DE report publish ASP
c. *他经常不断的(关于灾情的)文章发表了。
*Tā jīngchángbùduàn-de (guānyú zāiqíng de) wénzhāng fābiǎo le.
he frequently about disaster DE article publish ASP

In (19a), the process nominal 报道*bàodào* 'reporting' can be modified by经常不断的*jīngchángbùduànde* 'frequent'. In (19b) and (19c), this adjective cannot modify the result nominal 报道*bàodào* 'report' and the entity noun 文章*wénzhāng* 'article'.

Our analysis shows the same result as [18]. That is, process nominals allow frequency adjectives. Result nominals and entity nouns (concrete entity nouns and abstract entity nouns) do not allow them. Furthermore, we test one more type: pure event nouns. The result shows that this type of nouns also admits frequency adjectives, as indicated in (20).

(20) ……帶來了頻繁的化學災害……
…… dàilái le pínfánde huàxué **zāihài**……
bring ASP frequent chemical disaster
'(It) brought frequent chemical disasters……'

In (20), the pure event noun 災害*zāihài* 'disaster' can be modified by the frequency adjective 頻繁的*pínfánde* 'frequent'.

3.3.2 Durative Time Expressions

[18] tests which type of nouns can be modified by durative time expressions. She finds that process nominals allow such modification, while result nominals and entity nouns do not, as shown in (21).

(21) a. 他 ??(对灾情的)三个小时的报道十分有用。
Tā??(duì zāiqíng de) sān gè xiǎoshí de bàodào shífēn yǒuyòng.
he to disaster DE three GL hour DE reporting very useful
'His reporting of the disaster for three hours is very useful.'
b. *他关于灾情的三个小时的报道发表了。
*Tā guānyú zāiqíng de sān gè xiǎoshí de bàodào fābiǎo le.
he about disaster DE three CL hour DE report publish ASP
c. *他关于灾情的三个小时的文章发表了。
*Tā guānyú zāiqíng de sān gè xiǎoshí de wénzhāng fābiǎo le.
He about disaster DE three CL hour DE article publish ASP

In (21a) the durative time expression 三个小时 *sān gè xiǎoshí* 'three hours' can modify the process nominal 报道*bàodào* 'reporting'; while in (21b) and (21c) it cannot modify either the result nominal 报道*bàodào* 'report' or the entity noun 文章 *wénzhāng* 'article'.

We agree with her analysis. That is, process nominals allow while result nominals and entity nouns do not allow durative modification. Moreover, we find that pure event nouns also admit modification of durative time expressions, as shown in (22).

(22) 他们召开了三个小时的会议。
Tāmen zhàokāi le sān gè xiǎoshí de **huìyì**.
They hold ASP three CL hour DE meeting
'They held a three-hour meeting.'

In (22), the durative time expression 三个小时 *sān gè xiǎoshí* 'three hours' modifies the pure event noun 会议*huìyì* 'meeting'.

3.3.3 Rationale Clauses

[18] tests whether rationale clauses can modify different nouns as shown in (23).

(23) a. 他为了出风头的 ??(对灾情)的报道进行了三个小时。
Tā wèile chūfēngtóu de??(Duì zāiqíng) de **bàodào** jìnxíng le sān gè
he to show off DE to disaster DE reporting proceed ASP three CL
xiǎoshí.
hour
'His reporting of the disaster in order to show off lasted three hours.'
b. ??他为了出风头的(关于灾情的)报道发表了。
?? Tā wèile chūfēngtóu de (guānyú zāiqíng de) **bàodào** fābiǎo le.
he to show off DE about disaster DE report publish ASP
c. ??他为了出风头的(关于灾情的)文章发表了。
?? Tā wèile chūfēngtóu de (guānyú zāiqíng de) **wénzhāng** fābiǎo le.
he to show off DE about disaster DE article publish ASP

[18] finds that only process nominals admit rationale clauses as shown in (23a). The rationale clause 为了出风头的 *wèile chūfēngtóu de* 'to show off' modifies the process nominal 报道*bàodào* 'reporting'. In (23b) and (23c), neither the result nominal 报道*bàodào* 'report' nor the entity noun 文章*wénzhāng* 'article' can be modified by the rationale clause.

We agree that process nominals can be modified by rationale clauses as shown in (24). The rationale clause 为了丰富群众文化生活*wèile fēngfù qúnzhòng wénhuà shēnghuó* 'in order to enrich the cultural life of the masses' modifies the process nominal 演出*yǎnchū* 'performance'. Furthermore, we note that pure event nouns, result nominals and entity nouns all admit rationale clauses as depicted from (25) to (28). In (22), the rationale clause 为了筹款*wèile chóukuǎn* 'to raise fund' modifies the pure event noun 音乐会*yīnyuèhuì* 'concert'. In (26), the result nominal 发明*fāmíng* 'invention' is modified by the rationale clause 为了解馋wèile jiěchán 'to satisfy a craving for delicious food'. In (27), the rationale clause 为了方便交流 *wèile fāngbiàn jiāoliú* 'to facilitate communication' modifies the concrete entity noun 手机 shǒujī **'**mobile phone'. In (28), the abstract entity noun 做法*zuòfǎ* 'way' is modified by the rationale clause 为了降火*wèile jiànghuǒ* 'to decrease internal heat'.

Process Nominal

(24)艺术团的这次为了丰富群众文化生活的**演出**进行了三个小时。

Yìshùtuán de zhè cì wèile fēngfù qúnzhòng wénhuà shēnghuó de **yǎnchū**
troupe DE this CL for enrich the masses cultural life DE performance
jìnxíng le sān gè xiǎoshí.
proceed ASP three CL hour

'The troupe's performance, in order to enrich the cultural life of the masses, lasted three hours.'

Pure Event Noun

(25)他们这场为了筹款的**音乐会**圆满结束了。

Tāmen zhè chǎng wèile chóukuǎn de **yīnyuèhuì** yuánmǎn jiéshù le.
They this CL for fund-raising DE concert successfully end ASP

'Their fund-raising concert successfully ended.'

Result Nominal

(26)由此推断素鸡应该就是和尚们为了解馋的**发明**。

Yóucǐ tuīduàn sùjī yīnggāi jiùshì héshàngmen wèile
From this infer vegetarian chicken should be monk in order to
jiěchán de **fāmíng**.
satisfy a craving for delicious food DE invention

'From this (we can) infer that vegetarian chicken was the invention of monks to satisfy their craving for delicious food.'

Entity Noun

(i) Concrete Entity Noun

(27)人们发明了为了方便交流的**手机**。

Rénmen fāmíng le wèile fāngbiàn jiāoliú de **shǒujī**.
People invent ASP in order to facilitate communication DE mobile phone
'People invented the mobile phone to facilitate communication.'

手机*shǒujī* 'mobile phone' as an artifactual-type noun has a telic role, according to Generative Lexicon Theory [21] and [25-27].

(ii) Abstract Entity Noun

(28) 她们喝凉茶是一种为了降火的**做法**。
Tāmen hē liángchá shì yī zhǒng wèile jiànghuǒ de zuòfǎ.
they drink herbal tea is one CL to decrease internal heat DE way
'They drink herbal tea as a way to decrease internal heat.'

In summary, [18] has tested whether the three types of eventive expressions, *frequency adjectives, durative time expressions and rationale clauses*, can modify different nouns. She finds that only process nominals allow all of them, while result nominals and entity nouns allow neither. Thus she claims that Chinese process nominals and English process nominals are similar in event structure.

In our analysis, however, we find that both process nominals and pure event nouns allow all of the three types of eventive expressions. Entity nouns (Abstract Entity Nouns and Concrete Entity Nouns) do not admit frequency adjectives and durative time expressions. All these non-natural kind nouns allow rationale clauses, so such clauses don't help in determining event structure.

[17-18] claim that complex process nominals have a verbal form and license argument structure and event structure. Simple event nouns do not have a verbal form and thus do not license argument structure and event structure. However, we find that in Mandarin Chinese, process nominals and pure event nouns behave similarly in taking arguments and licensing event structure, so the *complex* and *simple* contrast in English does not exist in Chinese. Thus, it is necessary to re-define what is *complex* and what is *simple* for Chinese. Following [25] we treat complex process nominals as nouns with more than one subevent, such as accomplishments. The new term for this type of nouns is process nominals. Simple event nouns are nouns with only one subevent or many similar subevents, such as activities. The new term for this type of nouns is pure event nouns or pure nouns that represent events.

The most common and basic event type of event-representing nouns is a process (or called an activity) [28], as show in (29).

(29) 他对问题的解释进行了三个小时。
Tā duì wèntí de jiěshì jinxing le sān gè xiǎoshí.
he to problem DE explanation last ASP three CL hour
'His explanation to the problem lasted three hours.'

In (29), the process nominal 解释*jiěshì* 'explanation' represent an event that lasts some time, so its event type is process.

Activities can be shifted to be accomplishments when there is a natural endpoint. [28] explains three ways that lead to the shift: demonstratives, localizers and quantifiers. We further find that some lexical words can also trigger the aspectual shift, as demonstrated in (30).

(30) 经济的发展持续了三个月就停滞了。
Jīngjì de fǎzhǎn chíxù le sān gè yuè jiù tíngzhì le.
economy DE development last ASP three CL month at once stagnate ASP
'The economy development lasted three months and then stagnated.'

In (30), 停滞*tíngzhì* 'stagnate' gives the 发展*fǎzhǎn* 'development' event an end-point, which shifts it from an activity to an accomplishment.

4 Conclusions and Future Work

To sum up, this research has examined the following parameters of an event: (i) process classifier vs. individual classifier (classifier), (ii) argument-taking (argument structure), (iii) subject-oriented adjective (argument structure), (iv) frequency adjective (event structure), (v) durative time expression (event structure), and (vi) rationale clause (event structure).

Based on these parameters, we propose establishing an event-based classification system for non-natural kind nouns, as demonstrated in Table 3.

This table indicates that process nominals and pure event nouns have the same behavior; result nominals and abstract entity nouns have the same behavior. Concrete entity nouns parallel largely with result nominals and abstract entity nouns, except that they do not take subject-oriented adjectives. These eventive parameters, except rationale clauses, behave well in distinguishing nouns with a process reading (viz. process nominals and pure event nouns) from those without (viz. result nominals and entity nouns). Based on these parameters, we have established an event-based noun classification system for non-natural kind nouns. These parameters can facilitate the detection of nominal events in natural texts. In future work, we would use them to detect such events.

Table 3. An Event-Based Classification System for Non-Natural Kind Nouns

		Classifier		Argument Structure		Event Structure		
		Process	Individual	Argument Taking	Subject-oriented Adj.	Frequency Adj.	Time Expression	Rationale Clause
Nominal	Process Nominal	+	-	+	+	+	+	+
	Result Nominal	-	+	-	+	-	-	+
Pure Event Noun	Process Reading Only	+	-	+	+	+	+	+
Entity Noun	Concrete Entity Noun	-	+	-	-	-	-	+
	Abstract Entity Noun	-	+	-	+	-	-	+

Acknowledgments. We would like to thank Prof. James Pustejovsky, Prof. Nianwen Xue, Prof. Haihua Pan and the anonymous reviewers for their comments. The remaining errors are ours. This work is supported by a General Research Fund (GRF) sponsored by the Research Grants Council (Project No. 543810, 544011), and the studentship of The Hong Kong Polytechnic University.

References

1. Zhu, D.X.: Grammar Handouts. The Commercial Press, Beijing (1982)
2. Chao, Y.-R.: A Grammar of Spoken Chinese. University of California Press, Berkeley (1968)
3. Lǔ, S.X., Problems on Chinese Grammar Analysis. The Commercial Press, Beijing (1979)
4. Wang, H., Zhu, X.: Subcategorization and Quantitative Research on Modern Chinese Nouns. In: Modern Chinese Grammar Studies that Face the Challenges of New Century: The International Conference on Modern Chinese Grammar:1998, Shandong Education Press, Jinan (2000)
5. Chu, Z.X.: An Investigation on Temporal Adaption of Nouns. In: Lu, J. (ed.) Modern Chinese Grammar Studies that Face the Challenges of New Century: The International Conference on Modern Chinese Grammar:1998, Shandong Education Press, Jinan (2000)
6. Ma, Q.Z.: Verbs with denotational Meaning and Nouns with Predicative Meaning. In: Research and Exploration of the Grammar. The Commercial Press, Beijing (1995)
7. Han, L.: Analysing the Word Class Status of Event Nouns. Journal of Ningxia University (Humanities & Social Sciences Edition) (1), 6–10 (2010)
8. Wang, S., Huang, C.-R.: Event Classifiers and Their Selected Nouns. In: The 19th Annual Conference of the International Association of Chinese Linguistics (IACL-19), vol. 71. Nankai University, Tianjin (2011)
9. Liu, S.: A Study of Temporality of Common Nouns. Language Teaching and Linguistic Studies (4), 25–35 (2004)
10. Wang, S., Huang, C.-R.: Domain Relevance of Event Coercion in Compound Nouns. In: The 6th International Conference on Contemporary Chinese Grammar (ICCCG-6). I-Shou University, Kaohsiung (2011)
11. Wang, S., Huang, C.-R.: Compound Event Nouns of the 'Modifier-head' Type in Mandarin Chinese. In: Gao, H.H., Dong, M. (eds.) Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC-25), pp. 511–518. Nanyang Technological University, Singapore (2011)
12. Wang, S., Huang, C.-R.: Temporal Properties of Event Nouns in Mandarin Chinese. In: The 57th Annual International Linguistic Association Conference (ILA-57), New York, USA (2012)
13. Wang, S., Huang, C.-R.: Qualia Structure of Event Nouns in Mandarin Chinese. In: The 2nd International Symposium on Chinese Language and Discourse, Nanyang Technological University, Singapore (2012)
14. Wang, S., Huang, C.-R.: A Constraint-based Linguistic Model for Event Nouns, Forum on 'Y. R. Chao and Linguistics'. In: Workshop of the 20th Annual Conference of the International Association of Chinese Linguistics (IACL-20), The Hong Kong Institute of Education, Hong Kong (2012)

15. Wang, S., Huang, C.-R.: Type Construction of Event Nouns in Mandarin Chinese. In: The First Workshop on Generative Lexicon for Asian Languages (GLAL-1) Workshop of The 26th Pacific Asia Conference on Language, Information and Computation (PACLIC-26), Bali, Indonesia (2012)
16. Wang, S., Huang, C.-R.: Compositionality of NN Compounds: A Case Study on [N_1+Artifactual-Type Event Nouns]. In: The 26th Pacific Asia Conference on Language, Information and Computation (PACLIC-26), Bali, Indonesia (2012)
17. Grimshaw, J.B.: Argument structure. MIT Press, Cambridge (1990)
18. Fu, J.: On Deriving Chinese Derived Nominals: Evidence for V-to-N Raising. University of Massachusetts Amherst, Amherst (1994)
19. Pustejovsky, J.: Construction and the Logic of Concepts. In: Bouillon, P., Busa, F. (eds.) The Language of Word Meaning, pp. 91–123. Cambridge University Press (2001)
20. Pustejovsky, J.: Type theory and lexical decomposition. Journal of Cognitive Science 6, 39–76 (2006)
21. Pustejovsky, J., Jezek, E.: Semantic Coercion in Language: Beyond Distributional Analysis. Distributional Models of the Lexicon in Linguistics and Cognitive Science, Special Issue of Italian Journal of Linguistics/Rivista di Linguistica (2008)
22. Pustejovsky, J.: Type Theory and Lexical Decomposition. Journal of Cognitive Science 7(1), 39–76 (2006)
23. Huang, C.-R., Chen, K.-J., Lai, Q.-X.: Mandarin Chinese Classifier and Noun-Classifier Collocation Dictionary. Mandarin Daily Press, Taipei (1997)
24. Huang, C.-R., Ahrens, K.: Individuals, Kinds and Events: Classifier Coercion of Nouns. Language Sciences 25(4), 353–373 (2003)
25. Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge (1995)
26. Pustejovsky, J.: Type Construction and the Logic of Concepts. In: Bouillon, P., Busa, F. (eds.) The Language of Word Meaning, pp. 91–123. Cambridge University Press (2001)
27. Asher, N., Pustejovsky, J.: A Type Composition Logic for Generative Lexicon. Journal of Cognitive Science 7(1), 1–38 (2006)
28. Wang, S., Lee, S., Huang, C.-R.: A Corpus-based Analysis of Semantic Type System of Event Nouns: A Case Study on huìyì. In: Jing-Schmidt, Z. (ed.) Proceedings of the 23rd North American Conference on Chinese Lingusitics (NACCL-23), Eugene, Oregon, USA, pp. 18–34 (2011)

A Form of Verb + Object Displaced Separable Slots: "N's+B+Ax"

Chunling Li[1] and Xiaoxiao Wang[2]

[1] College of Liberal Arts, Shenyang Normal University, Shenyang, China
lchl4019@Sina.com
[2] College of International Education, Shenyang Normal University, Shenyang, China
yuanxiao8211@hotmail.com

Abstract. This paper combines semantics and syntax to study the displaced separable form of verb+object separable words AB in modern Chinese in a comprehensive perspective, namely the "N's+B+Ax" form. The paper discusses and analyzes the characteristics of such a form and selection and restriction within the quantitative verb+object separable words AB. It further analyzes the internal semantic features and rules of use for separable words falling into this category. The outcome of the study can serve as a reference for computer linguistics and foreign language teaching.

Keywords: V+O form, separable words, semantic feature.

1 Introduction

V+O separable words are unique grammatical phenomena in modern Chinese language. The features of them lie in the fact that their meanings are generated in the same way as composite words, but their forms are similar to phrases. Thus they are in the middle between words and phrases. They are not only difficult for ontological study of language, but also key to Chinese language study and computer linguistics.

In certain context, when a V+O separable word AB is used together with another component X, morpheme B is translocated before A in order to highlight the importance of morpheme B in this separable word, thus AB becomes a displaced separable form "N's+B+Ax" [1]. Such a form has the communication effect that the two morphemes "A" and "B" are displaced and separated before using the separable word in sentence in order to make it a concrete event. Specifically, it is to make morpheme "B" and the component element of "B" in the separable word "AB" become topic, and morpheme "A" and the component element of "A" in the separable word "AB" become comment. Thus, a concrete and bounded event could be expressed. This kind of pragmatic displacement further makes the boundless and indiscrete separable word AB become bounded and discrete and better served for concrete event[2].

Of all the current study on separable words, researchers have already noticed the translocation in separable words and reached consensus on this point[3-4]. For example, morpheme B must have qualifying component before it. Morpheme A must be accompanied with adverbial or complement, and morpheme B must be put ahead

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 396–405, 2013.

when modal complement is used. These research outcomes have undoubtedly provided us with some reference and assistance. However, they only describe the features of displaced separable words in general terms, the actual contexts are quite different. Besides, they didn't analyze the quantified separable words, nor did they study the different situations which may occur when various separable words fall into this form. This paper just fills up this vacancy in studying and analyzing the separable words in this form under different contexts, so as to discuss the selection and restriction for quantified separable words. It also analyzes the internal semantic features and rules of use for separable words in this form.

2 Analysis of Separable Form "N's+B+Ax"

The separable form "N's+B+Ax" refers to a sentence which contains privately agentive attribute, privately affected attribute and the common form of the privately agentive attribute and the privately affected attribute. The definition and viewpoint on "privately agentive attribute" has taken reference from Prof. Guozheng Xiao's dissertation "On Privately Agentive Attribute" (1986).a. When N denotes subject, then N is the privately agentive subject. Here the subject is a general term, including dative component sometimes; b. When N denotes the common form of subject and object, N can be interpreted as either privately agentive subject or privately affected subject; c. When N denotes object, N is the privately affected subject. For example:

1) Tā de wǔ tiào de hěn hǎo。
   He dances well.
2) Wǒ de kè shàng wán le。
   I finished my class.
3) Tā de xīn shāng tòu le。
   He is heart-broken.

In example 1, "he" is the privately agentive subject, that is "he dances, and he dances well"-"he" is the agentive subject; in example 2, "I" can be interpreted as either privately agentive subject or privately affected subject, that is " I have class and I finished my class -"I" is the agentive subject. At the same time, the sentence can be also interpreted as "Other person finished teaching me", hereby "I" is privately affected subject[5]; in example 3, "he" is the privately affected subject, that is to say "he is hurt, and he is heart broken"- "he" is the affected subject. The following are the comprehensive study and analysis of the separable words for the above 3 examples from four angles of semantic feature, structure mark, syntax formation and sense and cognition.

2.1 Semantic Feature

When N denotes the subject or the common form of subject and object, N is equivalent to agentive subject or can be interpreted as either privately agentive subject or privately affected subject. In this case, the form is interpreted as privately denoting the subject of an action or event, or privately denoting both subject and object of an

action or event. As long as an action or event has its subject, the separable words in the form "N's+B+Ax" should have the semantic feature of [+autonomous], for example:

4）Tā de niú chuī de gòu dà de le。
He exaggerated too much.
5）Tā de fā lǐ de hěn hǎo。
He is quite a barber.
6）※Tā de jìng yǐjīng chū le。
※Because separable words are unique phenomena of Chinese language, the wrong examples are meaningless to translate. Thus they did not appear in this paper.

Example 4 and 5 are feasible sentences, so N and B have formed possessive relation. In example 4, "he" is the privately agentive subject, it has the semantic features of [+autonomous][+sustained][-properties][-directed/coordinated]. Separable words with such features fall into this form, in which N denotes privately agentive subject. There are 325 separable words with these semantic features, including "boast(*chuīniú*), take lead(*dàitóu*), wear the pant(*dāngjiā*), turnover(*fānshēn*), cheer(*gānbēi*), hook on(*guàgōu*), spend the New Year(*guònián*), look back(*huítóu*), work extra(*jiābān*), carry on(*jiēbān*), open mouth(*kāikǒu*), serve the meal(*kāifàn*), skip class(*kuàngkè*), evade taxation(*lòushuì*), study(xuéxí), queue in(*páiduì*), step up onto the floor(*shàngtái*), be on duty(*shàngbān*), stretch hands(*shēnshǒu*), avoid tax(*tōushuì*), go to countryside(*xiàxiāng*), put off(*yánqī*), make sentence(*zàojù*), stay in hospital(*zhùyuàn*), walk(*zǒudào*), commit an offence(*zuòàn*), stay on duty(*zuòbān*), let go(*fàngshǒu*), make a sigh(*tànqì*), take a travel(*chūchāi*), be pregnant(*huáiyùn*), go to school(*shàngxué*), go hunting(*dǎliè*), go sleeping(*shuìjiào*), go swimming(*yóuyǒng*), make a turn(*guǎiwān*), turn round(*zhuǎnwān*), make a rebellion(*zàofǎn*), etc.. And here the separable words with semantic feature of [-directed/coordinated] refer to those with [-directed] or [-coordinated] feature.

There are some exceptions, some separable words having these features cannot be the privately agentive subject, such as "price cut (*jiàngjià*), ventilation (*tōngfēng*), price hike (*zhǎngjià*), rectification (*zhěngfēng*), turn round (*zhuǎnxiàng*)" etc. For these words, morpheme B cannot form possessive relation with N. In example 5, N-"he" can be interpreted as either privately agentive subject or privately affected subject, the separable words here have the semantic features of [+autonomous] [+sustained][+directed/coordinated]. Example 6 is impossible, N and B can never form possessive relation. The separable words in this case have the semantic features of [+sustained][+fast changing], and separable words with such features cannot be included in this form. However, study shows that some individual separable words with such features can be accepted in this form. Under this situation, N is comparable to the agentive subject or the common form of the agentive subject and affected subject. Such separable words are "engagement(*dìnghūn*), get married(*jiéhūn*), begin(*kāitóu*), divorce(*líhūn*), ask for leave(*qǐngjià*), make order(*xiàlìng*) etc.. Study also shows that as long agentive subject is available in a sentence, then most separable words in the form "N's+B+Ax" have the common semantic features of [+autonomous] [+sustained].

When N denotes object, N becomes privately affected attribute. When a sentence has an object, the separable words will either have the semantic feature of [+autonomous] or [-autonomous]. But in oral communication, there are mostly separable words with [-autonomous], for example:

7）Wǒ de xíng biàn le。
I changed my figure.
8）Tā de kuī chī gòu le。
He had suffered a huge loss.
9）Nǐ de fú yǐjīng xiǎng dào tóu le。
You have had enough of your luck.
10）※Tā de fǎ fàn le。

Example 7 and 8 are feasible sentences. In example 7, N is the affected subject. In this case, the separable words having the semantic features of [-autonomous] [-directed/coordinated] can be used. In the displaced form, most separable words with the [-autonomous] semantic feature can become the privately affected subject in the form "N's+B+Ax"; in example 8, N can be interpreted as either agentive subject or affected subject, the separable words in this example have the semantic features of [-autonomous] [+directed/coordinated]. Separable words with such features include "cheated(shàngdàng), lose face(diūrén), worry about(cāoxīn), concern about(dānxīn), suffer loss(chīkuī), etc.; in example 9, N-"you" is a privately affected subject. The separable words in this example have the semantic features of [+autonomous] [+sustained][+properties][-directed/coordinated]. Separable words with such features include "enjoy happiness(xiǎngfú), terrible(yàomìng), cruel-hearted(hěnxīn), try hard(jìnlì), obedient(tīnghuà) etc.; example 10 is impossible, as morpheme B (law) and N (he) don't form up possessive relation. Separable words with such features include "reach maturity(dàoqī), break law(fànfǎ)".

2.2 Structure Mark

"N's+B+Ax" form must have "'s" as the structure mark. "'s" is the prerequisite for N to become the privately agentive/affected attribute, otherwise there will never be the definition of privately agentive/affected attribute. For example:

11）Tā dú xī de hěn duō。
He takes quite a lot of drugs.
12）Tā xīn huī tòu le。
He is very desolate.

In example 11, "he" is the agentive subject; while in example 12, "he" is the affected subject. The agentive/affected subject is formed without "'s". Privately agentive/affected attribute is formed only when personal pronoun is followed by "'s", for example:

11）′Tā de dú xī de hěn duō。
He takes too many drugs.
12）′Tā de xīn huī tòu le。
His heart is very much hurt.

In the two above examples, when "'s" is added, the sentence of privately agentive/affected attribute is formed. On the surface, "N's+B"-"his drug" and "his heart" are modifier-head construction. But in fact, N-"he" is the privately agentive/affected attributes, namely it is equivalent to privately agentive/affected subject. Therefore, the above two examples can be changed into the following structures, with identifiable privately agentive/affected subjects:

11）"Tā xīdú, tā dú xī de hěn duō。
He takes drugs, and he takes quite a lot.

12）"Tā huīxīn, tā xīn huī tòu le。
He is hurt, and his heart is broken.

The "'s" can only be omitted when the context (mostly questioning context) is clear enough without affecting the semantic expression, for example:

13）A: Shuí de qǔ pǔ wán le?
Who has finished composing the song?
B: Wǒ qǔ pǔ wán le ＝ Wǒ de qǔ pǔ wán le。
I finished composing the song.＝ My song has been composed.

14）A: Shuí de piào tóu wán le?
Who has finished voting?
B: Tā piào tóu wán le ＝ Tā de piào tóu wán le。
He finished voting.= His vote is completed.

2.3 Syntax Formation

First, the sentences of privately agentive/affected attributes are different from other sentences of dominant agentive/affected attributes.

The privately agentive/affected attributes are not dominant, and are expressed in the modifier-head construction "N's B". They are quite different from the dominant agentive/affected attributes. The dominant agentive/affected attributes follow non-attributive-noun structure, "'s" of "N's V" is omitted to get the subject-predicate structure "N+V" with agentive/affected subject. While the structure with privately agentive/affected attribute is attributive-noun structure, it can never assume the "N+V" structure by omitting "'s" [5], for example: "That's all for my performance." "my performance" follows "N's V" structure, "I" is the dominant agentive attribute, the sentence will be "I perform" without "'s'. In another example, "my class is finished.", "my class" is the sentence of privately agentive/affected attribute-"N's B", "I class" is impossible if "'s" is omitted.

Second, the "N's+B+Ax" form of separable words can sometimes lead to different meanings.

Study shows that some separable words with the semantic features of [+autonomous] [+sustained][+directed/coordinated] or [-autonomous][+directed/coordinated] will have different meanings when they are used in the "N's+B+Ax" form. The ambiguity is endowed by the form itself, for example:

15）Wǒ de zhēn dǎ wán le。
I was given an injection.

16）Tā de xīn cāo gòu le。
His worry is too much for him.

In example 15, the separable word has the semantic features of [+autonomous] [+sustained][+directed/coordinated], in example 16, the separable word has the semantic features of[-autonomous][+directed/coordinated]. Both examples have different meanings. The attribute N-"I" and "he" can be interpreted as either privately agentive subject or privately affected subject.

The sentence of example 15 leads two meanings, "Wǒ gěi biérén dǎzhēn dǎ wán le。(I finished giving injection to others.)"- "I" is the agentive subject or "Biérén gěi wǒ dǎzhēn dǎ wán le。(Other people gave injection to me.) " -"me" is the affected subject。

The sentence of example 16 can be interpreted as "Tā cāo biérén de xīn cāo gòu le。(He had worried enough about others.)"-"he" is the agentive subject or "Biérén wèi tā cāoxīn cāo gòu le。(Others had worried enough about him.)"-"him" is the affected subject.

Third, in the "N's+B+Ax" form, morpheme B of the separable words can be preceded with momentous indicative components, for example:

17）Tā de zhè huí zǎo xǐ de kě chèdǐ。
This time he took a good bath.

18）Nǐ de zhè cì chāi kě méi bái chū。
This time your travel is worthwhile.

In these two examples, "*zhèhuí* (this time)" and "*zhècì*(this time)" are indicative components.

Forth, "N's+B+Ax" form can be divided into two propositions.

When the separable words with semantic features of [+autonomous] [+sustained][+directed/coordinated] are put into the "N's+B+Ax" separable form, they can be transformed into "ABA de…", thus the separable words must have the semantic features of [+autonomous] [+sustained]. The morpheme B should have the features of action verbs, and the structure can be divided into two propositions, namely "NAB" and "BAx". That is, "N's+B+Ax"→"NA BAx"→"NAB+BAx". In reverse, "NAB+BAx" can be changed into "NA BAx", for example:

19）Tā de huǎng sā de kě gòu dà de。
His lie is too wild.
→ Tā sāhuǎng sā de kě gòu dà de。
He has told a wild lie.
→ Tā sāhuǎng, huǎng sā de kě gòu dà de。
He told a lie, and the lie is too wild.

19）′Tā sāhuǎng, huǎng sā de kě gòu dà de。
He told a lie, and the lie is too wild.
→Tā sāhuǎng sā de kě gòu dà de。
His lie is too wild.

2.4 Sense and Cognition

Privately agentive attribute (or the privately affected attribute) “from the angle of sense and cognition, an agentive attribute (or an affected attribute) is equivalent to an agentive subject (or an affected subject).” The usage of privately agentive attribute N of the slot [“‘N’s+B+Ax”] is similar to the usage of word which is both privately agentive attribute and privately affected attribute. In order to save space, we explain them together in this paper and use the same address of "privately agentive attribute". And we will separately explain if there is exception. By using the following methods, this kind of agentive attribute (or affected attribute) can be converted into the agentive subject (or the affected subject). a. The modifier marker “’s” can be deleted to make N become subject of the sentence. “N+B+Ax” forms subject-predicate structure and is used as predicate in the sentence. b. After deleting the modifier marker "的", we can repeat the verb morpheme "A" of the separable word “AB”. Thus N becomes subject of the sentence. Then “N+B+Ax” forms the serial verb construction and can be used as predicate in the sentence. Most of separable words which have the semantic feature of [-autonomous] can not be transformed like this. c. “NAx” can be used to answer the question about N and N becomes subject of “AX” in this context. E.g. the examples above can be converted into:

1) ′Tā de wǔ tiào de hěn hǎo。
He dances well.
→Tā， wǔ tiào de hěn hǎo。
He dances well.
→Tā tiàowǔ tiào de hěn hǎo。
He dances well.
Shuí de wǔ tiào de hěn hǎo?
Who dances well?
→Tā tiào de hěn hǎo。（＝Tā de wǔ tiào de hěn hǎo。）
He dances well. (＝He dances well.)

2) ′Wǒ de kè shàng wán le。
I finished my class.
→Wǒ， kè shàng wán le。
I finished my class.
→ Wǒ shàngkè shàng wán le。
I finished my class.
Shuí shàngkè shàng wán le?
Who finished class?
→Wǒ shàng wán le。（= Wǒ de kè shàng wán le。）
I finished my class. (＝I finished my class.)

3) ′Tā de xīn shāng tòu le。
He is heart-broken.
→Tā， xīn shāng tòu le。
He is heart-broken.
→Tā shāngxīn shāng tòu le。

He is heart-broken.
Shuí de xīn shāng tòu le?
Who is heart-broken?
→Tā xīn shāng tòu le。(＝Tā de xīn shāng tòu le。)
He is heart-broken. (＝He is heart-broken.)

The privately agentive attribute N in example 1)′and 2)′ is used as the subject and the separable word "AB" which is represented by "A" of slot "Ax" should have the feature of the action verb and the semantic feature of [+autonomous]. No matter what the concrete forms of "AX" are, so long as it contains the separable word "AB" which has the feature of action, N often denotes subject. It mainly manifested in the following four ways.

- "x" of "AX" is often the modal particle"*le*" at the end of the sentence, and the sentences without "*le*" are often unfeasible. E.g:

20）Tā de kuǎn cún le。(Bǐ jiào ：※Tā de kuǎn cún。)
He saved up some money.
21）Tā de hūn lí le。(Bǐ jiào ：※ Tā de hūn lí。)
He divorced.

- The adverbial modifier is used before "Ax".E.g:

22）Tā de kè gāng bǔ。(Bǐ jiào ：※ Tā de kè bǔ。)
He just made a miss lesson.
23）Nǐ de kuǎn yǐjīng huì le。(Bǐ jiào ：※ Nǐ de kuǎn huì。)
You already remitted money.

- "Ax" is verb-compliment structure.

24）Tā de wǔ tiào de hěn hǎo。
He dances well.
25）Tā de zhuāng huà de hěn nóng。
She uses too much makeup.

- "Ax" is the "*shì……de*"form and it mainly emphasizes the content in the middle of "*shì……de*". E.g:

26）Tā de hūn shì qùnián jié de。
He got married last year.
27）Wǒ de shū shì zài wǔhàn niàn de。
I was educated in Wuhan.

When the privately affected attribute N in example 3)′is used as the subject of the sentence, the separable words "AB" which is represented by "A" of slot "Ax" should be separable words of non-action and most of them should have the semantic feature of [-autonomous]. Some separable words which have the semantic features of [+autonomous][+properties][-directed/coordinated] can also be used in this kind of slot. And the concrete forms of "Ax" are mainly showed in the following two ways.

- "Ax" is verb-compliment structure.

  28）Tā de xīn shāng tòu le。
  He is heart-broken.
  29）Tā de huà tīng bù jìnqù。
  His word was unaccepted.

- The adverbial such as"*méi shǎo*"of "*kě méi shǎo*"often can be used before "Ax" to emphasize a large number. E.g.:

  30）Tā de lì méi shǎo fèi。
  He made a great effort.
  31）Nǐ de shì kě méi shǎo chū。
  You made a lot of troubles.

3 Conclusion

Study shows that "N's+B+Ax" separable form and the separable words which can be used in this form have the following features and rules of use:

3.1 Features of "N's+B+Ax" Separable Form

- The structure mark of "'s" in the form is a prerequisite of the privately agentive/affected attributes.
- N is the privately subject (or object), instead of dominant subject (or object). It serves as the privately agentive/affected subject in a sentence.
- For separable words in this form, the morpheme B and N can form up possessive relation, which can be expressed as "N's B".

3.2 The Characteristics and Applying Rules of the Separable Words Which Can be Used in "N's+B+Ax" Separable Form

- Morpheme B of separable word AB is often object morpheme which is often countable, concrete and used with modifiers. When it is used with A, B often shows its original meaning instead of transferred meaning.
- Morpheme A of separable word AB often has the semantic feature of [+autonomous] or it is the independent morpheme. When the verb morpheme A has the semantic feature of [-autonomous], most of morpheme B can be retouched by adjective. If B can not be retouched by adjective, when it is used with A, the separable word often has the semantic feature of [-autonomous].
- There must be some other elements which are used before or after the verb morpheme A in this displaced separable form. They can be adverbial, complement, or modal particle at the end. However, sometimes morpheme A can be used alone in a symmetrical sentence.

Acknowledgements. This work has been supported by the Major Projects of Chinese National Social Science Foundation (11&ZD189).

References

1. Li, C.L.: A Study on the "V+O" Separable Words & Its Separable Slots in Modern Chinese. Doctoral dissertation of Wuhan University (2008)
2. Li, C.L.: A Study on the Construction of Theoretical System of Separable Slots in Modern Chinese. Social Science Journal 1, 196–199 (2009)
3. Zhao, J.M.: Discussion on Expandable V+O Construction. Language Teaching and Studies. 2, 5 (1984)
4. Huang, X.Q.: A Semantic Research on Separable Character group. Doctoral Dissertation (2003)
5. Xiao, G.Z.: Privately Agentive Attribute Language Study 4, 4–10 (1986)

Innovative Use of *Xiā* in Modern Taiwan Mandarin: A Witness to Pragmaticalization

Yu-Chih Lin

Graduate Institute of Linguistics, National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan (R.O.C.)
r99142009@ntu.edu.tw

Abstract. The lexeme 瞎 *xiā* (literally meaning 'blind') exhibits distinct related senses in corresponding syntactic environment in modern Taiwan Mandarin. Through a synchronic analysis, the present study examines the semantics, the sense evolution, and the corresponding syntactic distribution of the three senses of *xiā* in modern Taiwan Mandarin Chinese. It is shown that the original physical sense ($xiā_1$) has been metaphorically extended to the mental domain ($xiā_2$) to convey 'conceptual blindness'. This abstracted sense has furthered undergone a process called "pragmaticalization" in the recent decade, denoting an evaluative-deontic attitude 'lousy' towards the context or the hearer ($xiā_3$). The evolution is supported by cross-linguistic general trends of semantic extensions as well as historical evidence drawn from comparison between the two corpora.

Keywords: semantic extension, pragmaticalization, evaluative predicate.

1 Introduction

The polysemous lexeme 瞎 *xiā* shows multiple related senses in various syntactic environment in modern Taiwan Mandarin, as the following examples illustrate:

(1) a. xiā zi (blind-suffix)
b. kū xiā (cry-blind)
c. xiā cāi (blind-guess)
d. xiā huà (blind-speech)
e. hǎo xiā (very-blind)

In (1a), the attributive *xiā* forms part of a compound noun; in (1b), *xiā* denotes the resulted state in a resultative construction; in (1c), *xiā* functions as an adverbial modifying a verb; in (1d), *xiā* is an adjective modifying a deverbal noun; in (1e), *xiā* conveys a new sense as an evaluative predicate. Some of these senses are more tangible (1a-1b), while others are more elusive (1c-1e). The meaning conveyed by the evaluative predicate use is especially vague, and theoretically and historically speaking, this use is thought to be a later development of the lexeme. The goal of the

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 406–415, 2013.

present research then is to (i) employ a synchronic analysis to delineate the evolution of the major senses of the polysemous lexeme *xiā*, including their semantics and syntax (with evidence of the emergence of the evaluative predicate use from corpora); and (ii) establish the evaluative predicate use of *xiā* as a result of pragmaticalization, with semantic and structural evidence.

2 Methodology

2.1 Data Collection

The data collected for the present study come from two sources: (i) Academia Sinica Balanced Corpus of Modern Chinese (selecting only the genre of journalism), intended to be representative of the 1990's, and (ii) Udn (a journalism database) [udndata.com], limited within a three-year time span from 2008/11/29 to 2011/11/29, intended to be representative of the recent decade.

2.2 Semantic Change

According to Hilpert [1], body part terms have been identified cross-linguistically as a productive source of extensions in the mental lexicon, into both lexical and grammatical senses. There is a general trend in the process of semantic change: a physical/tangible sense is developed into an abstract/intangible sense through metonymic and metaphorical extensions, for example:

(2) Semantic extension of *eye* in Basque, Bokobaru, and Busa (from Hilpert, 2007)

eye → vision → {attention, beauty}

The semantic extension observed in (2) is a result of chained metonymies: the derivation of the sense 'vision' hinges on the metonymy PERCEPTION OF ORGAN FOR PERCEPTION; the senses 'attention, beauty' are further achieved through PERCEPTION FOR ATTENTION metonymy and PERCEPTION FOR QUALITY PERCEIVED metonymy. Since *xiā* is associated with the bodily function, its semantic development is likely to conform to this trend. In other words, the more physical sense would be the earlier primary sense, while the more emotional and mental meanings are lexical extensions of the original sense.

2.3 Pragmaticalization

The evaluative predicate use of *xiā* is a pragmatic use, and I suspect that it has undergone a process called pragmaticalization, which enables propositional elements such as adverbials to develop more abstract and pragmatic meanings. Typically this process involves a broadening of scope, often correlating with a change in sentence

position, on a syntactic level [2]. This process results in "pragmaticalized elements" that express the speaker's attitude towards the hearer and the context and, in terms of distribution, they show syntactic detachability and mobility [3].

This correspondence between formal/distributional distinction and semantic/pragmatic meaning is highlighted by Tyler and Evans [4]: The formal aspects of language have conceptual significance. Given this correspondence, I hypothesize that the innovative use of *xiā* is pragmaticalized and highly interactional and thus forms a constituent unit that resides at a structural level higher than the other *xiā* senses. Figure 1 schematizes the constituent structure, which consists of three levels: the discoursal level, clausal level, and phrasal level. A pragmaticalized element (S2), and presumably the innovative *xiā*, is likely to reside at the discoursal level, with a predicating scope equal to or wider than the clause (S1), while other senses of *xiā* resides below the clausal level.

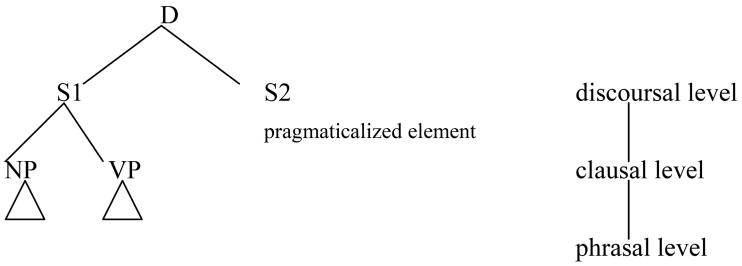


Fig. 1. Constituent structure of clauses with adjoined pragmaticalized elements (adapted from Aijmer, 1996: 4)

In Section 4.2, this structural tree will be employed as a yardstick for the quantitative analysis of the distribution of the three *xiā* senses.

3 Three Major Senses of *xiā*

Three distinct senses of *xiā* have been identified from the data collected. These senses are only "major" senses, because there may be idiosyncratic extended uses of each major sense that, having not stabilized, do not form senses themselves, but are dependent on the major ones. Seven constructions have also been observed to be the milieu where *xiā* occurs; each of them can be allocated to one of the three senses. This allocation shows that the sense classification does have structural evidence.

3.1 Physical *xiā* (*xiā*$_1$): 'Visual Malfunction'

This sense of *xiā* largely means a malfunction of vision, corresponding to the English equivalent *blind*. It is assumed to be the earliest sense, since it is physical and exists in the objective world. It does not involve any subjective judgment, and therefore does not collocate with any forms of degree words such as *hěn* 'very' and *chāo* 'super'. The following are three common constructions where *xiā*$_1$ occurs:

(3) adj. + N: inherent attribute
他和這位女伴還沒有見過面，正在擔心對方會不會是個<瞎>子或什麼的。
tā hàn zhè wèi nǚ bàn hái méi yǒu jiàn guò miàn, zhèng zài dān xīn duì fāng huì bú huì shì ge xiā zi huò shén me de
'He hasn't met his date yet and is worrying if she is *blind* or something.'

(4) V + N (eye): anticausative
那種文章也能上報的話，一定是編輯先生看稿看得迷糊了，或是<瞎>了眼睛了。
nà zhǒng wén zhāng yě néng shàng bào de huà, yí dìng shì biān jí xiān shēng kàn gǎo kàn de mí hú le, huò shì xiā le yǎn jīng le
'If that kind of article could be published, the editor must have been dazed by reading drafts, or *blinded*.'

(5) V + adj.: resulted state
你媽哭<瞎>了眼睛。
nǐ mā kū xiā le yǎn jīng
'Your mom cried her eyes *blind*.'

(3) is an Adjective-N compound, where *xiā* signifies an inherent quality of N. (4) is an anticausative construction, with *xiā* as verb, meaning 'to be caused to lose vision', followed by an object usually related to eyes, such as *yǎn jīng* 'eyes', *shuāng mù* 'double eyes', and *yòu yǎn* 'right eye'. (5) is a resultative construction, with V denoting a resulting event and *xiā* denoting the resulted state. Common instances of this construction are *kū xiā* 'cry-blind' and *dǎ xiā* 'hit-blind'.

It is thought that this sense itself ordinarily conveys a neutral attitude of the speaker, in a plain description of factual attribute of an entity, but this sense can also be intended to carry judgmental attitude of the speaker. The judgmental use is context-induced, because the extra subjective attitude can be canceled out by the stronger attitudinal indicators, such as conditionals *yí dìng shì* 'must be', as in (4), or *hǎo xiàng* 'as if'.

3.2 Abstracted *xiā* (*xiā*$_2$) 'Lack of Consideration'

This second sense of *xiā* is an abstracted version of *xiā*$_1$. It does not refer to the tangible bodily function anymore, but enters the conceptual domain where the subjective attitude of the speaker is involved. The following are two constructions where *xiā*$_2$ mostly occurs:

(6) adv. + V: manner adverbial; predicate modifier
a. 妳別<瞎>猜亂冤枉好人好不好?
nǐ bié xiā cāi luàn yuān wǎng hǎo rén hǎo bù hǎo
'Will you stop guessing *wildly* and falsely accusing the good man?'
b. 陳希煌顯得相當氣憤，直說這篇不實的報導「莫名其妙」、「<瞎>掰」。

chén xī huáng xiǎn de xiāng dāng qì fèn, zhí shuō zhè piān bù shí de bào dǎo mò míng qí miào, xiā bāi
'Chén Xī-Huáng seemed rather mad, saying this unreal report was "non-sense" and *so made-up*.'

(7) adj. + N: attributive modifier; 'fake; groundless'
…是睜眼說<瞎>話，是「說謊大王」。
…shì zhēng yǎn shuō xiā huà, shì shuō huǎng dà wáng
'…[They] are saying *blind* words with seeing eyes; they are liar kings.'

(6) is a compound verb construction, where *xiā* functions as an adverbial that modifies the manner of the verb. From the examples (6a-b), it is inferred that $xiā_2$ basically means 'lack of consideration', 'aimlessly', 'at random', or 'groundlessly'. As these English equivalents suggest, the use of $xiā_2$ evokes a semantic frame of "goal achievement": *xiā* denotes an action not implemented in the direction of the intended goal or expectation due to the malfunction of the conceptual ability. Note that this semantic frame of "goal achievement" also applies to $xiā_1$: Because the visual ability is defected, a physical entity cannot be perceived, and thus the expected goal is not achieved. The shared semantic frame supports the idea that $xiā_2$ originated from $xiā_1$, and the mapping from "visual ability" to "mental ability" involves a conceptual metaphor VISION IS THOUGHT.

Note that in (6b), the main verb in the compound carries negative attitude: *bāi* 'make up speech'. In addition, the adverbial use of *xiā* often co-occurs with emotion-loaded negatively polarized phrases: *luàn yuān wǎng* 'falsely accuse' (6a), and *mò míng qí miào* 'nonsense' (6b). Both these point to the hypothesis that the judgmental and emotional sense arises when $xiā_1$ is developed into $xiā_2$.

In (7), *xiā* functions as an attributive modifier. What differs from (3) is that $xiā_1$ only refers to an objectively inherent property, while $xiā_2$ expresses the subjective evaluation of the speaker. In this construction, *xiā* means 'fake' and 'groundless', usually used to modify a deverbal noun, like *huà* 'speech'. On the surface, this evaluation is intended at the following action-denoting noun, but propositionally this attitude is intended at the person who performs the action. In (7), *xiā huà* 'blind speech' means words without consideration, and the functional shift of *xiā* exploits the metonymy ACTION FOR ACTOR: by "blind speech" one actually means the speech giver is "conceptually" blind.

3.3 Pragmatic *xiā* ($xiā_3$): Evaluative-Deontic Attitude

The new sense of *xiā* is elusive. Consider the following examples:

(8) (N) + degree adverbial + adj.: 'lousy; lame'
a. 英文翻譯讓網友直呼「好<瞎>」。
yīng wén fān yì ràng wǎng yǒu zhí hū hǎo xiā
'The English translation made the Internet users exclaim "so *lame*(?)".'
b. 所有追她的人，都要先過她媽這關，也因此當媒體大篇幅報導她的緋聞時，連林媽媽都說，「這些傳聞好<瞎>。」

suǒ yǒu zhuī tā de rén, dōu yào xiān guò tā mā zhè guan, yě yīn cǐ dāng méi tǐ dà piān fú bào dǎo tā de fēi wén shí, lián lín mā ma dōu shuō, zhè xiē chuán wén hǎo xiā

'All her suitors have to pass her mother's test, so when the media reported on her affairs, even Ms. Lin said, "These rumors are so *groundless*(?)".'

(9) V + V: main predicate + intensifier verb

a. 問起她和馮德倫、徐若瑄的三角緋聞，她說：「<瞎>爆了！而且我和徐若瑄一直有聯絡。」

wèn qǐ tā hàn féng dé lún, xú ruò xuān de sān jiǎo fēi wén, tā shuō, xiā bào le, ér qiě wǒ hàn xú ruò xuān yì zhí yǒu lián luò

'Asked about the rumor of love triangle with Féng Dé-Lún and Xú Ruò-Xuān, she said, "So *groundless*(?)! And Xú Ruò-Xuān and I have always been in touch."'

(8) is a copular construction where the main predicate consists of *xiā* and a degree word such as *hǎo* 'good', *chāo* 'super', *fēi cháng* 'abnormally', *hěn* 'very', *zhēn* 'really', *tài* 'too', *zuì* 'most', *gòu* 'enough'. In (9), *xiā* alongside with an intensifier verb such as *tòu* 'through' and *bào* 'explode', constitutes a verbal predicate. When subjectless as in (8a) and (9), the two collocates, roughly meaning 'lousy' or 'lame', form interjection-like elements which predicate previous unspecified propositions, unlike $xiā_2$, whose semantic scope ranges over only the verb (6) or a deverbal noun (7). Because $xiā_3$ exhibits a wider semantic-pragmatic scope, it requires grounding of the context, and the constructions it co-occurs with—interjection and copular construction—exactly provide such possibility.

The previous two senses $xiā_1$ and $xiā_2$ do not match here in (8) and (9), and it is even difficult to see any common attributes between the modified NPs *fān yì* 'translations' (8a) and *chuán wén* 'rumors' (8b). More often than not, one can hardly provide an exact explanation of this use without referencing to a context. This vagueness is associated with the fact that syntactically less specified environments may provide the potential for semantic elusiveness: the innovative uses occur in copular (8b) and interjection (8a, 9) constructions, which are lenient in allowing types of predicates to occur in the syntactic slots. More importantly, much of the meaning of the evaluative predicate *xiā* hinges on the attitudinal evaluation of the context and the hearer by the speaker: for the evaluated party, there is an average desired goal which he is obliged to achieve but fails. Here we see the evocation of the same semantic frame "goal achievement" as in $xiā_2$ at work. What distinguishes $xiā_2$ and $xiā_3$ though lies in the degree of involvement of the evaluating party: the intent of judgment is decreased, and because the speaker is less emotionally involved, the judgment sounds more objective. The self-distancing function in $xiā_3$ may arise from the higher structural position of the expression. For $xiā_3$, the predicated is an accumulation of the context, while for $xiā_2$, the predicated is structurally specified. The propositional meaning of $xiā_3$ varies every time it is used in an actual instance. This property shows that this sense possesses little remnant semantic content; it mainly functions pragmatically/interactionally.

The meaning extension of the lexeme *xiā* is thus shaped. From $xiā_1$ to $xiā_2$ and to $xiā_3$, the semantic contents of the original sense are seen to be reduced, and at later stages of the semantic development, the meaning becomes less truth-conditional as well as less referential. The sense evolution is schematized as the following figure:

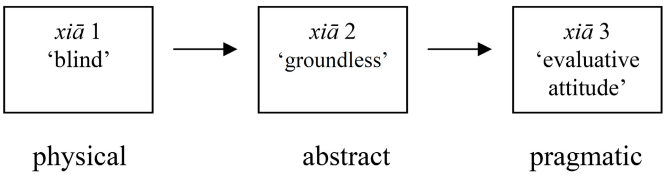


Fig. 2. Sense evolution of *xiā*

3.4 *Xiā* $_3$ as an Innovation: Evidence from Corpus

Based on the assumption that in the process of semantic change, the physical/tangible sense is developed into the abstract/intangible sense through metonymic and metaphorical extensions, it is suspected that $xiā_3$ is an innovation: among the three sense in (2), $xiā_1$ is physical; the semantic content is reduced and resulting in $xiā_2$; $xiā_3$ is further on its way to retain only the interactional sense. Therefore, I make the following hypothesis:

(10) **Hypothesis 1:** $xiā_3$ is a new sense arising in the recent decade.

To test this hypothesis, I compare the occurrence of these three senses in two corpora: the Sinica corpus and the Udn corpus. The following is the contingency table showing the relation between the three *xiā* senses and the two corpora:

Table 1. Contingency between corpora and senses

	xiā 1		*xiā 2*		*xiā 3*		**Marginal**
Sinica	29	(27)	38	(30)	0	(10)	67
Udn	409	(411)	439	(447)	168	(158)	1016
Marginal	438		477		168		1083

The result of the chi-square test is significant ($\chi^2_{obt} = 13.08 > 9.21$; $df = 2$; $\alpha = .01$). It is inferred that $xiā_3$ is an innovative use that did not arise until in the 2000's because no occurrence of $xiā_3$ is found in the more recent corpus. Examining Table 1, it is found that there is little gap between obtained and expected frequencies for $xiā_1$, while there is a considerable difference between the obtained and expected frequencies for $xiā_2$ and especially $xiā_3$. This means that xi_1 is relatively stabilized in use over time independent of the other two senses, while $xiā_2$ and $xiā_3$, due to their semantic proximity, co-vary with each other.

4 Meaning/Function Change Mechanisms

4.1 Metaphorical Extension

As mentioned in Section 3 above, *xiā*$_1$ 'blind' refers to the bodily function, while *xiā*$_2$ 'aimlessly' has something to do with the malfunction of thinking, and thus can be roughly translated as 'mentally blind'. In other words, the original core meaning of *xiā*$_1$ is placed in another sphere and becomes *xiā*$_2$. More specifically, the "vision frame" which *xiā*$_1$ evokes is transformed into the "goal achievement frame" of *xiā*$_2$ when mapped onto the mental domain. The schema of vision consists of a figure with a vantage point, a fictive trajectory, and a landmark. *xiā*$_1$ is related to the perceptual ability, with a perceptual organ as the figure, a trajectory of sight, and a perceived physical entity as landmark. *xiā*$_2$ and *xiā*$_3$ are related to the conceptual ability, with a conceptual mental eye, the process of achievement as the fictive trajectory, the goal as the landmark. It can be observed that the shared figure-trajectory-landmark configuration makes the mapping of ICM's possible, and the development of the "unattained goal" sense seems natural.

4.2 Pragmaticalization

We have seen in Section 3 that *xiā*$_2$ describes the manner of an action as evaluated by the speaker while *xiā*$_3$ denotes an attitudinal evaluation of the context by the speaker. *xiā*$_2$ and *xiā*$_3$ may appear in similar contexts, only that they exhibit different semantic scopes and attitudinal intensity:

(6b) 陳希煌顯得相當氣憤，直說這篇不實的報導「莫名其妙」、「<瞎>掰」。
chén xī huáng xiǎn de xiāng dāng qì fèn, zhí shuō zhè piān bù shí de bào dǎo mò míng qí miào, xiā bāi
'Chén Xī-Huáng seemed rather mad, saying this unreal report was "non-sense" and *so made-up*.'

(8b) 所有追她的人，都要先過她媽這關，也因此當媒體大篇幅報導她的緋聞時，連林媽媽都說，「這些傳聞好<瞎>。」
suǒ yǒu zhuī tā de rén, dōu yào xiān guò tā mā zhè guan, yě yīn cǐ dāng méi tǐ dà piān fú bào dǎo tā de fēi wén shí, lián lín mā ma dōu shuō, zhè xiē chuán wén hǎo xiā
'All her suitors have to pass her mother's test, so when the media reported on her affairs, even Ms. Lin said, "These rumors are so *groundless*(?)".'

Xiā in (6b) and (8b) modifies the "rumors" or "news reports", but with different attitudes from the speaker towards them: (6b) is emotional and criticizing, while (8b) sounds more objective and aloof. One of the driving forces for development from *xiā*$_2$ to *xiā*$_3$ may be intersubjectification: the broadened functional scope can avoid direct evaluation of the specifics and thus avoid direct confrontation with the evaluated party. The function of evaluation of the whole context then is the source of semantic

elusiveness but at the same time takes into consideration the feeling of the hearer. In this sense, the semantic scope is correlated with different degrees of self-involvement and emotion.

The semantic scope usually parallels the position of the expression on the constituent structure: The wider the scope, the higher on the structure and the more peripheral to the clause. Since *xiā*$_3$ is a pragmatic use, often occurring turn-initially and subjectless, it should be assumed that:

(11) **Hypothesis 2:** *xiā*$_3$ is a result of pragmaticalization and discursivization, which broaden the scope of *xiā* and move the *xiā* expression towards a higher structural level in discourse.

To test this hypothesis, I compare the distribution of these three senses in discourse in the Udn corpus in terms of their position at the constituent structural tree to see whether they are situated at the phrasal level, clausal level, or discoursal level. A contingency table that shows the relation between the three senses and the three structural levels is made, as shown in Table 2.

Table 2. Contingency between senses and structural levels

	discoursal		*clausal*		*phrasal*		**Marginal**
xiā 1	0	(16)	100	(80)	309	(313)	409
xiā 2	0	(17)	0	(86)	439	(336)	439
xiā 3	39	(6)	99	(33)	30	(129)	168
Marginal	39		199		778		1016

The result of the chi-square test is significant (χ^2_{obt} = 545.10 > 13.28; *df* = 4; α = .01). It is inferred that *xiā*$_3$ is a linguistic unit at a higher structural level, while *xiā*$_1$ and *xiā*$_2$ are linguistic units at a lower structural level. Examining the contingency table, it is discovered that the gap between obtained and expected values is most obvious in the column of the discoursal level: *xiā*$_1$ and *xiā*$_2$ never occur at this level, while roughly one fourth of the occurrence of *xiā*$_3$ resides at this level. In addition, as seen in Test 1, *xiā*$_2$ and *xiā*$_3$ co-vary with each other. This co-variance can also be attributed to the semantic proximity between the two senses: *xiā*$_2$ and *xiā*$_3$ similarly expresses an evaluative attitude (but differing in emotional involvement), while functioning at, or occupy, different levels of the constituent structure.

5 Conclusion

This study has examined through a synchronic analysis the semantics, the sense evolution, and the corresponding syntactic distribution of three major senses of *xiā* in modern Taiwan Mandarin Chinese.

It is claimed that there are three major senses of *xiā*. The original physical sense ($xi\bar{a}_1$) has been metaphorically extended to the mental domain ($xi\bar{a}_2$), which has furthered been pragmaticalized in the recent decade to denote an evaluative-deontic attitude towards a context ($xi\bar{a}_3$). The development is not only supported theoretically, but also empirically through comparison of corpora.

Examining the syntactic environment of the three senses, the present study supports that $xi\bar{a}_3$, mainly functioning as an interjection-like subjectless predicate or verbal predicate in a copular construction, is a product of pragmaticalization with a widened semantic scope and thus residing at higher levels in discourse.

References

1. Hilpert, M.: Chained metonymies in lexicon and grammar. In: Radden, G., Köpcke, K.-M., Berg, T., Siemund, P. (eds.) Aspects of Meaning Construction, pp. 77–98. John Benjamins, Amsterdam (2007)
2. Defour, T., et al.: Degrees of pragmaticalization: The divergent histories of *actually* and *actuellement*. In: Lauwers, P., Vanderbauwhede, G., Verleyen, S. (eds.) Pragmatic Markers and Pragmaticalization: Lessons from False Friends: Special Issue of Languages in Contrast, vol. 10(2), pp. 166–193 (2010)
3. Aijmer, K.: *I think*—an English modal particle. In: Swan, T., Westvik, O.J. (eds.) Modality in Germanic Languages: Historical and Comparative Perspectives, pp. 1–47. Mouton de Gruyter, Berlin (1997)
4. Tyler, A., Evans, V.: Reconsidering prepositional polysemy networks: The case of *over*. Language 77(4), 724–765 (1996)

Study on "*shi*" as a Demonstrative Pronoun in Modern Chinese

Shuhao Qu[1] and Yi Yu[2]

[1] College of International Cultural Exchange, Huazhong Normal University, Wuhan, China
qushuhao@gmail.com
[2] College of Foreign Languages, Huazhong Normal University, Wuhan, China
yyhs@163.com

Abstract. In modern Chinese, a considerable number of *shi* still possess the grammatical attributes of a demonstrative pronoun. This usage originates from that of *shi* as a predicate in ancient Chinese. As a demonstrative pronoun, *shi* functions as a very flexible anaphor, whose characteristics may help us to distinguish it as a demonstrative pronoun from *shi* as a verb.

Keywords: *shi*, demonstrative pronoun, anaphora.

1 Introduction

As is well known, "*shi*" was originally an adjective. In The Analysis of Chinese Characters, Xu [1] said, *shi* means "straight". It was also used as a demonstrative pronoun, which meant "this". In modern Chinese, *shi* is mainly used as two types of verbs, which include the linking verb that appears before a noun (or noun phrase) to indicate judgment, and the auxiliary verb before a verb (or verb phrase), an adjective (or adjective phrase) or a clause to indicate the focus of certain information. As a linking verb, *shi* connects a Subject to additional information about the Subject. In most cases, it can not be omitted except for the few sentences in which a noun (or noun phrase) acts as a predicate. As an auxiliary verb, *shi* plays the role of affirmation or emphasis, which is often used with "*de*", a modal particle that exists at the end of a sentence, and can be omitted.[1] Lü [1] divides *shi* into nine subclasses according to its usages, and regards all of them as verbs, which "mainly perform the function of affirmation and connection".

However, the development and evolution of language may not always be uniform. Zhang [1] indicates that *shi* has ten kinds of "additional usages" besides of verb, adverb and conjunction, whose grammatical attributes "are difficult to determine" and it doesn't mention the problem of *shi* as adjectives and pronouns.[2]

[1] For more information, please refer to Xing, F.Y. [1].
[2] The *shi* used as an adverb in his work is regarded as an auxiliary verb mentioned above, and *shi* used as a conjunction is now generally regarded as a verb.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 416–426, 2013.

Furthermore, Fan [1] classifies homonyms *shi* into six types, namely main verb, auxiliary verb, adjective, interjection, pronoun and non-predictive adjective.[3]

However, at the same time, he also considers that the adjective *shi* which exists mostly in dialect and literary works, is used less and less in modern Chinese; generally speaking, the demonstrative pronoun *shi* is seldom used in oral Chinese, but only in written Chinese and Chinese idioms, which originate from the ancient Chinese.

In this article, we mainly focus on the following questions:

Firstly, do there still exist *shi* used as one-syllable adjectives in modern Chinese? If so, are they used less and less?

Secondly, are there any *shi* used as demonstrative pronouns in the language domain of spoken Chinese? If so, what are the specific circumstances?

Thirdly, what are the grammatical characteristics of *shi* used as pronouns and demonstrative pronouns in modern Chinese?

By the way, in view of the poor reliability of self-made texts, all the examples in this paper are chosen from authoritative texts.

2 The Adjective Usages and Pronoun Usages of *shi* in Modern Chinese

One typical characteristic of *shi* used as verbs and auxiliary verbs in modern Chinese is that they can work with objects, which include substantive objects and predicative objects. In contrast, as adjectives or pronouns, *shi* don't work with objects, which is also one of the highly characteristic features that differentiate adjectives and pronouns from verbs in modern Chinese.

After carefully retrieving modern Chinese corpus at the Language Education and Research Center in Huazhong Normal University, we come to find many sentences containing *shi* that should not be regarded as verbs. Some good examples are provided as follows:

Group I

(1) After Wei Yi spoke, everyone nodded repeatedly and said, "This is a good place of ambush." Liu Yuan-xing, the commander of Battalion I followed Wei Yi's words and said, "*Ke bu shi*? (*ke* not right)(Isn't it?) This is Qinghuabian....."

(*Defend Yan'an* by Du Peng-cheng)

(2) From the blue veins on the forehead of the major Gumi, Lap Dog knew Gumi really lost his temper. He shivered with fear and stammered, "Tycoon, no, no, no. *Ni shuo de shi*, (you say *de* right)(What you said is right), is, is, is I..... I, I, I can't....."

(*Behind Enemy Lines* by Feng Zhi)

[3] In his paper, *shi* is considered as "an interjection" which serves as a response in the sentence and should be regarded as an independent clause. In fact, it can also be seen as an adjective. It is not so acceptable to define it as "an analog form of sound" because there exist the corresponding negative forms of *shi* such as "*bushi* (not), *budui* (not right)", etc. In addition, When used in a sentence such as "*Zhe ren haochilanzuo, shi huor dou bu gan.* (The man is so lazy that he won't do any job.)", it should be regarded as a verb.

(3) Yuntao said, "Since Uncle Zhong moved to the new house, every time he saw me plowing in the good land, he would silently bring a jar of boiled rice to me; otherwise, Aunt Zhong would ask me to have dinner at her home in a loud voice. You see, it is a daily thing for me to plow the field, how should it always make them spend money in treating me!" Jiangtao thought, "*Zhe ye shi*. (this also right)(This is right.)"

(*Keep the Red Flag* by Liang Bin)

(4) His sword was not slow and the young man had never thought that he would make such a sneak attack ---- He killed White Snake. *Zhuge Lei ben gai ganji ta cai shi*, (Zhuge Lei originally should greatful him indeed right)(Zhuge Lei should have been grateful to him for it). Why did Zhuge want to kill him?

(*Top One Knife of the Storm* by Gu Long)

(5) Yang Guo said, "I daren't accept the title such as Young Hero. *Ni jiao wo Yang Guo bian shi*. (you call me Yang Guo just right)(It is right that you call me Yang Guo.)"

(*The Return of the Condor Heroes* by Jin Yong)

(6) "You needn't guess. In the voucher of donation for government posts you had brought, the family name Yuan is not wrong. *Mingzi bu shi*. (first name not right)(But the first name is not right)； of course, it was borrowed."

(*Hu Xue-yan, a Businessman with a Red-Topped Hat* by Gao Yang)

Group II

(7) In many cases men are more loyal than women. *Wo zhe ge pengyou jiu shi*. (my this *ge* friend just is)(This friend of mine is a case in point.)

(*The Absolute Privacy* by An Dun)

(8) My idea that I want a child is very strange; it is completely separated from marriage. The normal way of thinking is that it is natural to have a baby when two persons are in love, *danshi wo bu shi*. (but I not is)(but I don't think in such a way.) I do not want his child; neither do I know whose child I should have. So I felt very sad.

(*The Absolute Privacy* by An Dun)

(9) At first, my father treated them reverently. As an old man, he kept on expressing his repentance, hoping to offer an apology to the public through reporters and make some necessary explanations for me to the public. Those reporters were very pious and showed that they could set the record straight. Therefore, my father really trusted them. *Wo muqin ye shi*. (my mother also is)(So did my mother.)

(*Ran's Father* by Liang Xiao-sheng)

(10) The girl curled her lip and said, "Why would you have dinner with them? Please eat it. Take off the clothes for you have sweated a lot."

Qingmiao made a touch with hands. *Zhen shi*. (really is)(It's right!) Since an unknown moment, his two pieces of clothes had been soaked with sweat. With a smirk, he quickly took them off.

(*Xia Qing-miao and His Master* by Hao Ran)

(11) Hearing that, Liu Bu-cai burst into a laughter and then made a sigh silently which implied his sneer at others' not being sensible. Aunt Seven got angry and answered immediately, "Uncle Liu, what's wrong with my words?"

"What you said is not wrong and your have a warm heart. But, even so, you are just getting yourself into trouble. As a proverb says, "a wise judge can't make accurate

judgments on household conflicts"; Sister Seven! Even if you were Bao Zheng and could make the correct judgments, there would still be a trouble!"

"What! I can't understand it."

"Sister Seven, you are always clever, but ignorant this time. As to a lawsuit, either the plaintiff wins, or the defendant wins. Why should we take the trouble to help one side and harm the other?"

Aunt Seven suddenly realized that if in the future she helped Mrs. Hu, she would surely offend Hu Xueyan; Isn't it "to help one side and harm the other"? "Well, well, Uncle Liu! *Ni ye shi*. (you also is)(You are the same.) Why should you beat about the bush instead of telling me the truth directly? Luckily, I had become more careful than before and inquired patiently; otherwise, I would be led astray?"

(*Hu Xue-yan, a Businessman with a Red-Topped Hat* by Gao Yang)

(12) Going out of the restaurant, I had been a little drunk. Leaning on Wu Di, I asked, "Do you think I am bad?"

She helped me and walked carefully with her head bent down. She didn't answer.

"Bad! Bad! Really bad!"

I laughed at Wu Di, "*Ni ye shi*. (you also is)(You are the same.) You are aware that I am bad, but you would approach me."

(*Half is Flame, Half is Sea Water* by Wang Shuo)

(13) The other person blushed, whose face is as red as that of Guan Yu, and said, "I am now on vacation."

(I:) "Oh! On vacation! Hurrah! These days we were so busy that everybody desires the opportunity of vacation. I had just told Xiao Ge yesterday that it would be the greatest bliss if I could get a leave of eight or ten days. If so, I would go shopping and drink tea with ease in the streets. I would do everything that women can do outside the office! Mr Guo! You see, as to vacation, I am also happy! *Zhen shi*, (really is)(Just like what you say,) how long you have already been on vacation? "

(*The Storm of Love and Hatred* by Liang Feng-yi)

(14) That night, my mom and I stacked book leaves. In doing it, my mind became absent and I thought of one way of chess.

My mother said with a sigh, "*Ni ye shi*. (you also is)(Just like you.) You neither go to the cinema movies nor go to the park. You just play chess in such a way....."

(*The Chess King* by A Cheng)

In the above two groups of cases, the *shi* we translated twice obviously don't play the role of "connection". In terms of form, some *shi* only have the preceding items instead of being connected with the following items; others even have none of these two items. The preceding items or following items should not be regarded as being omitted or implied, for it is impossible to retrieve exactly the content of what has been "omitted" from the context. It is difficult to classify these *shi* as anyone of the nine usages described in the work of Lü [1].

2.1 Usage of *shi* as Adjectives

From the semantic perspective, the shi in Group I have the meaning close to "*zhengque*, *dui* (correct or right)". Syntactically, they have the following features:

Being Used without Objects. In the above cases (1) to (6), *shi* appear without objects. In addition, it is impossible to find the omitted object, as shown especially in cases (2) to (6).

Serving Alone as the Predicate. As shown in cases (1) to (6), all the *shi* function as the predicate or part of the predicate in each sentence. Especially in case (2), *shi* alone serves as the predicate.

Being Modified by Adverbs. In case (1) and cases (3) to (6) of Group I, *shi* is preceded by adverbs such as "*bu* (not), *cai*(indeed) and *jiu* (just)". In cases (2), *shi* is modified by *hen*(much), an adverb of degree.

From the above analysis, we may find such *shi* of modern Chinese have the grammatical characteristics more close to those of adjectives and thus should be regarded as adjectives. As an adjective, *shi* is used with relatively great frequency in different registers, so we should pay more attention to it.

2.2 Usage of *shi* as Pronouns

The *shi* of Group II have some kind of close relationship with those in Group I. In both groups, *shi* have very similar grammatical characteristics. They don't go with objects, but can independently act as predicates, and can be modified by adverbs. The only difference is that *shi* of Group II are semantically closer to "*zheyang*, *nayang*, *ruci* (like this, like that)". They have the feature of "indefinite reference" [4], and could be considered as "pronouns that indicate properties or situations" [5] or "predicative demonstrative pronouns" [6] in modern Chinese.

Usages of *shi* as pronouns are a bit complicated. In cases (7) to (14), they can be subdivided into the following three types.

Basic Usage of *shi* as Pronouns. The *shi* of cases (7) to (10) in Group II can be substituted by "zheyang, nayang (like this, like that)". They refer back to the situation mentioned above in each text and act as the predicate or main part the predicate in each sentence.

Additional Usage One of *shi* as Pronouns. The *shi* of case (11) and case (12) in Group II cannot be substituted by "*zheyang*, *nayang* (like this, like that)". It is somewhat difficult to interpret the anaphora mainly because such *shi* contain the implicit comparisons. When the content of each comparison is detected, *shi* will be correctly interpreted.

Let's take case (11) for instance. Aunt Seven firstly realized her fault of not thinking carefully. She quickly changed the topic to blame Uncle Liu for his implicit utterance, so she said, "*Ni ye shi*!" Case (12) is even more implicit, what "I" mean is

[4] For more information, please refer to Xing F.Y. [2].
[5] Lü S.X. [2].
[6] Zhu D.X. [1].

that "I" am bad. Although Wu Di was aware that "I" am bad, but still would approach me. That means, Wu Di was also some unreasonable. So "I" said to her, "*Ni ye shi*".

Actually, this extended usage of *shi* is a variation in the course of its linguistic development. It often occurs in colloquial Chinese and has a unique pragmatic effect.

Additional Usage Two of *shi* as Pronouns. We can compare cases (11) and (12) with cases (13) and (14). Comparatively speaking, in the latter cases, it is even harder to interpret the anaphora of *shi* for it requires much more background information. Such *shi* are more like modal particles or vocative expressions, whose connotations are quite obscure.

It seems that the monosyllabic demonstrative pronoun *shi* is not used scarcely in modern Chinese. Actually, it often exists in oral Chinese.

3 Features of *shi* as Demonstrative Pronoun

According to Li [1], "Subject + *shi*" is a structure that contains omission, but the omission involving *shi* differs from other kinds of omissions in that it shows distinctive characteristics. More exactly, in the structure of "Subject + *shi*", *shi* is related with IP (inflectional phrase); in the structure of "auxiliary verb + main verb", the auxiliary verb is connected with a verb phrase; in the structure of "verb + Object", the verb goes with the Object. This theory can explain the difference of the pairing sentences in cases (15) to (17) as follows: [7]

(15) a. *Tamen dagai bu hui lai le; women ye shi. (Women ye dagai bu hui lai le.)*
they probably not will come *le*; we also is (we also probably not will come le)
They probably will not come; neither will we. (We probably will not come, either.)

b. *Wo renshi ta hen jiu le; wo baba ye shi. (Wo baba ye renshi ta hen jiu le.)*
I know him very long *le*; my father also is (my father also know him very long le)
I've known him for a long time; so has my father. (My father has known him for a long time, too.)

(16) a. *Wo yao renzhen de zuo gongke; ta ye yao. (Ta ye yao renzhen de zuo gongke.)*
I want careful *de* do homework, he also want (he also want careful de do homework)
I will do my homework carefully, so will he. (He will do his homework carefully, too.)

b. *Wo yao tanwang ta san ci; tamen ye yao. (Tamen ye yao tanwang ta san ci.)*
I want see him three time; they also want. (they also want see him three time)

[7] Examples (15) to (17) and the following related analysis come from Li [1]. Actually, the "adverbials" mentioned by Li refer to VP-adverbials, which include predicate adverbials and complements in traditional Chinese grammar studies. In addition, the interpretive information in brackets is provided by the authors of this paper. The symbol "≠" means "not equal to".

I will visit him three times; so will they. (They will visit him three times, too.)

(17) a. *Wo renzhen de zuo gongke; tamen ye zuo le. (≠Tamen ye renzhen de zuo gongke le.)*

I careful *de* do homework; they also do *le*. (≠they careful *de* do homework le)

I finished my homework carefully; he finished it, too. (≠ He finished my homework carefully, too.)

b. *Wo renshi ta hen jiu le; wo baba ye renshi. (≠ Wo baba ye renshi ta hen jiu le)*

I know him very long *le*; my father also know. (≠ my father also know him very long *le*)

I've known him for a long time; my father also knows him. (≠ My father has known him for a long time, too.)

The *shi* of case (15) replaces "all the ingredients after the subject, including the adverbial, negative word and auxiliary verb"; "In analyzing the elliptical sentences that contain auxiliary verbs (such as '*yao*'(want), the author) or linking verbs (such as '*shi*', the author), the adverbial phrase should be taken into consideration"; "in analyzing the elliptical structures that involve verbs, the adverbial phrase must be excluded and focus be put onto the omitted object".

Li's opinion is very instructive. She tells us, *shi* has the features that differentiate it from other common auxiliary verbs and main verbs. However, Li's research does not deal with why *shi* has such characteristics and what information such features convey to us.

According to our observations, *shi* that has referential meaning in the end of the sentence can not be regarded as a verb, auxiliary or verb. Instead, it should be regarded as a demonstrative pronoun because like other demonstrative pronouns, it has the characteristics of anaphora, "indefinite reference". In terms of content, what *shi* refers back to is sometimes equal to, but sometimes greater than IP, or sometimes part of IP, and even has no direct relationship with IP proceeding in a text. Accordingly, the interpretation of anaphora might go on in a very liberal state, and this free anaphora exactly provides the most powerful evidence that proves *shi* is a demonstrative pronoun.

3.1 Direct Anaphora

The direct anaphora (explicit anaphora) means to refer to an explicit antecedent in the text, the antecedent and the anaphor share the same meaning. The explicit anaphora of shi that act as demonstrative pronouns can be divided into the following conditions.

The Anaphor Is Equal to IP. We believe that it is a coincidence that *shi* replaces the IP in the preceding sentence, inflection phrase IP is just a coincidence, which can be illustrated with a few examples as follows:

(18) Who can live peacefully and at the same time make others live peacefully? Me ----*Xianzai de wo jiu shi*.(now *de* I just is)(It's me.)

(*The Retirement* by Bai Hua)

(19) So she grabbed the handbag to go out of the office, and said to Ying Jia-cheng with a smile, "Goodbye! Have fun!"

Ying Jia-cheng pulls the door open quite gracefully for Yue Qiu-xin, and responded to her, "Good-bye! *Ni ye shi*. (you also is)(The same to you.)"

(*Three Hundred Days of Passion* by Liang Fengyi)

(20) They parted, with the same belief that they knew each other well. *Er shiji shang que bu shi*. (but fact on but not is)(But actually they did not.)

(*The Family* by Ba Jin)

The Anaphor Is Greater Than IP. In some conditions, the "+ *shi*" phrase in the short sentence refers back to all the parts after the subject in the preceding sentence. Since Chinese is a topic-prominent language, it can also refer to the CP (Clause Phrase) after the topic. What the "+ *shi*" phrase refers to is greater in content than IP. Consider the following examples.

(21)"You're studying! Hello, Great Scholar! But as to interpersonal relationship, you still behave like a baby. Maybe you are the youngest in your family ....."

"*Ni bu ye shi ma*? (you not also is *ma*)(You are the same, aren't you?)"

(Honey Girl by Cen Kai-lun)

(22) Husband: I really don't want to mention that thing. My wife and I, now also avoid mentioning it. Once it's mentioned, one could not sleep well for several nights. It applies to her, *wo ye shi*. (I also is.)(and me too.)

(*The Decade of 100 People* by Feng Ji-cai)

(23) Ran said, "My father is such a person that cannot commit an error in a lifetime. *Wo muqin ye shi*. (my mother also is.)(So is my mother.)"

(*Ran's Father* by Liang Xiao-sheng)

In case (21), "*Ni bu ye shi ma*?" is equal to "Aren't you the youngest too?" In case (22), "*wo ye shi*" is equal to "I could not sleep well for several nights." In case (23) "*Wo muqin ye shi*." is equal to "My mother is such a person that cannot commit an error in a lifetime." In each case, the "+ *shi*" phrase refers to what the Subject + Predicate structure expresses in the preceding sentence. Thus it is greater in content than the IP in the discourse.

The Anaphor Is Part of IP. On some occasions, *shi* functions so flexibly that it can refer back to any grammatical part in the preceding sentence. After careful analysis, we may find such usages can be classified into several types as follows.

shi Refers Back the Subject. Examples as:

(24) Suddenly, Wilma uttered a cry, and hurriedly fled into the car. Because numerous soldier ants had come, *bian di dou shi*. (every ground all is)(and covered every inch of the land.)

(*Tony! Tony!* by Zhu Bang-fu)

(25) "I don't believe it! Some females being single, that is what I had seen before. *Wo de gumu jiu shi*. (my *de* aunt just is)(My aunt is just one.)"

(*The Red Sun* by Wu Qiang)

shi Refers Back to the Verb Phrase. Sometimes, *shi* refers back to the verb phrase. But what the anaphor substitutes does not include the auxiliary verb or adverbial. Take the following examples for instance:

(26) Juehui seemed to understand, and then said gently, "Brother! Do you feel lonely here? ... *Wo xiaode ni yiding hui gandao jimo*. (I know you surely will feel lonely)(I know you will surely feel lonely.) *Wo ye shi*. (I also is)(So do I.) At home, no one understands me ..."

(*The Family* by Ba Jin)

(27) "Don't you think it's very hard?"

"Hard? Not, I am able to adapt to and even enjoy it."

"*Wo ke shou bu liao*. (I indeed suffer not *liao*)(I really can't stand it.)"

"*Wo ye shi*. (I also is)(Me too.) So ill-at-ease! You are destined to be the hostess of a wealthy family."

(*To Return Your Last Life* by Cen Kai-lun)

In case (26), what shi refers back to does not include the adverbial "*yiding* (surely)" and the auxiliary verb "*hui*(will)". In case (27), it does not include the adverbial "*ke* (really)".

shi Refers Back to the Object. The objects mentioned here aren't the objects of the *shi* that indicates judgment in each preceding sentence. Accordingly, such *shi* should not be considered as a linking verb. Here are some examples.

(28) My boss also believes that the girl isn't a decent person. Because nobody can excel in virtues over me. If you want to look for someone well-educated, *zhouwei de dou shi*. (around *de* all is)(all of who around you are well-educated.) Why is he so interested in such a girl?

(*The Absolute Privacy* by An Dun)

(29) The women pointed to the channel by the village and said, "Vegetables can be washed here", "In the next section, we wash clothes", and "My family needn't fetch water to home any longer; *yi chu men jiu shi*. (once out door just is)(you can find water just outside of your door.)" ...

(*Three Mile Bay* by Zhao Shu-li)

(30) "Well! Huaqian once said that you would not take a girl to the disco. To be your girlfriend is really not easy. *Xing'er wo bu shi*. (fortunately I not is)(But fortunately, I'm not your girl.) Whether happy or bored, I will come to play here."

(*To Return Your Last Life* by Cen Kai-lun)

shi Refers Back to the Attributive. One case is provided here:

(31) "People's Liberation Army are the forces of the proletariat!" Shuishan said seriously.

"*Wo ye shi*! (I also is)(Me too!)" Renbao blurted, "I have enough qualities. I would sell my house, my land, and all my properties and take my wife and children to join ....."

(*The Flowers of Spring* by Feng De-ying)

***shi* Refers Back to the Non-grammatical Units.** In the text where *shi* refers back to the non-grammatical units, we need to use the ability of association and sufficient

background information to determine what is referred back to by *shi*. In this sense, what *shi* refers back to almost has no direct relationship with IP.

(32) He replied, "Today I was almost cheated and nearly lost my face to humiliate the Ondor Grassland! Brother! The hearts of the dukes are vicious!" "*Wendu'er Wang ye shi ma*?(Ondor king also is *ma*)(Ondor Duke is such a case?)" I asked.

(*The Snow Horse* by Feng Ling-zhi)

(33) In the "mutual markets", there were some "mutual men", namely, horse broker. They had a powerful say on the qualities and quantities of goods. Thus, they could manipulate the trade and make a fortune in a short time. However, they were firstly required to master the foreign language; secondly, they should be able to build relationship with the market monitor. Therefore, many barbarians served us the "mutual men"; *An Lu-shan jiu shi*.(An Lu-shan just is)(An Lu-shan was just one of them.)

(*Hu Xue-yan, a Businessman with a Red-Topped Hat* by Gao Yang)

In case (32), "*Wendu'er Wang ye shi ma*?" means "The heart of Ondor Duke is vicious, too?" In case (33), "*An Lu-shan jiu shi*." actually means "Lu-shan was a barbarian and once acted as a mutual man and he had all of the above characteristics". In both cases, what shi refers back to appears in the antecedent sentence, but its referential item is accurately not a grammatical unit and cannot be obtained from the antecedent sentence. We need to process those antecedent sentences in a complex way. In fact, cases (7) - (10) have the same conditions.

3.2 Indirect Anaphora

Indirect anaphora (implicit anaphora) means the anaphor refers back to an implicit antecedent. What *shi* refers to will indirectly become part of the hearer's knowledge, which can not be obtained by direct speech, but be inferred from what has been talked about.

In fact, Additional Usage One and Two of shi as demonstrative pronouns mentioned above are indirect anaphora, which might be illustrated by cases (11) to (14). Consider a few more examples:

(34) Fang Da-sheng: (haltingly) I naturally can not say there is. (He bowed his head). You should remember you really loved me, *wo ye shi*. (I also is)(and me too.)

(*The Sunrise* by Cao Yu)

(35) (She :) "You are reviewing English?"

(I :) "I'm reading an extra-curricular book. I just read it for fun. What about you?"

(She :) "*Wo ye shi*. (I also is)(Me too.) Whether to revise or not doesn't matter."

(*When the Sunset Disappeared* by Li Ping)

In case (34), "*wo ye shi*." means "I really love you, too". The information that shi refers to can not be obtained directly from the antecedent sentence. In case (35), "*Wo ye shi*." seemingly should be interpreted as "I am reading an extracurricular book, too.

I just read it for fun, too." which might be regarded as a good example of omission. However, that's not really the case. From the context, we can see that "she" came unexpectedly when "I" was reading a book. "*Wo ye shi*." should be interpreted as "just like you, I would not review English". In this case, what *shi* refers back to can only be inferred from the context by using the background information.

The indirect anaphora of *shi* demonstrates that it possesses the grammatical features different from those of verbs (main verb and auxiliary verb), and it is not reasonable to insist that something is "omitted" or "implied" after *shi*.

4 Conclusion

In the above, we have discussed the grammatical attributes of *shi* at the end of the sentence in modern Chinese. Some of them should be seen as adjectives, and some should be viewed as demonstrative pronouns. The usage of *shi* as the adjective and demonstrative pronoun originates from the usage of *shi* as the predicate in ancient Chinese. As a demonstrative pronoun, *shi* functions as a very flexible anaphor, which is a typical feature shared by all demonstrative pronouns.

References

1. Fan, X.: The Classification of Homonymous "*shi*". Lexicographical Studies 2, 22–33 (1996)
2. Li, Y.H.: Ellipsis and Missing Objects. Linguistic Sciences 2, 3–19 (2005)
3. Lü, S.X.: Eight Hundred Words in Modern Chinese (New Edition). The Commercial Press, Beijing (1980)
4. Lü, S.X.: Demonstrative Pronouns in Modern Chinese (New Edition). Xuelin Publishing House, Shanghai (1985)
5. Wang D.Y.: A Study of the Referential Functions of "*zhe*" and "*na*". Xuelin Publishing House, Shanghai (2005)
6. Xing F.Y.: Modern Chinese. Higher Education Press, Beijing (1991)
7. Xing F.Y.: Three Hundred Questions in Chinese Grammar. The Commercial Press, Beijing (2002)
8. Xu, S.: The Analysis of Chinese Characters. China Bookstore, Beijing (1989)
9. Zhang J.: A Comprehensive Study of "*shi*". Henan People's Press, Zhengzhou (1960)
10. Zhu D.X.: The Chinese Grammar Handout. The Commercial Press, Beijing (1982)

On the Transferred Designation of "Subject 1 + Subject 2 + Predicate" Structures in Modern Chinese

Tai Pan[1] and Yi Yu[2]

[1] College of Foreign Students' Education, Wuhan University, Wuhan, China
pantai@whu.edu.cn
[2] College of Foreign Languages, Huazhong Normal University, Wuhan, China
yyhs@163.com

Abstract. The reason of the transferred designation (TD) of "VP + de" structures which have realized every valence of their verbs in the surface is that such structures may play as frames of "Subject 2 + Predicate" in "Subject 1 + Subject 2 + Predicate"(SSP) structures. What these frames are designating are their beginning Subjects. If there is no Subject 1, then the "VP + de" structure could not be used as TD. According to different syntactic and semantic characteristics of Subject 1, these transferred designations of SSP structures show much difference respectively.

Keywords: SSP structures, transferred designation, valence.

1 SSP and the TD of "VP + de"

1.1 TD and SD of "VP + de" Structures

Similar to some structures like the attributive clauses led by "that" or "which" in English, there exist some syntax methods of nominalization in modern Chinese, one of which is to add a "*de*" at the end of a verb or a predicative phrase. Zhu [3] pointed out systematically the problem of self designation (SD) and transferred designation (TD) of "VP + *de*" in modern Chinese. He said that if a "VP + *de*" has absent valence, then this structure should be a TD. If there is no absent valence, then "VP + *de*" must be a SD. For example, "chi1 de" (eat *de*) (food) acts as a TD, because it does not mean the action of "eat" but something to be eaten. While "wo3 chi1 shui3guo3 de" (I eat fruit *de*) (when/where…I ate fruits) can only be a SD, just because this phrase refers to the exact action of "eat". TD and SD, these two methods of nominalization can form nominal phrases in modern Chinese, but the first can usually be realized at the grammatical positions of Subject, Object and Attributive while the second can only be found at the place of Attributive. This paper mainly focuses on the TD of "VP + *de*" in modern Chinese.

Zhu [1] worked out a famous formula for SD and TD of "VP + *de*": $p = n - m$. Here, n means the amount of valence of the verb in "VP + *de*", m means the amount of valence which already appears in the surface of this structure, and p, the difference of n and m, can be called as the Index of Ambiguity of "VP + *de*". If $p = 0$, this "VP + *de*" must be a SD. If $p = 1$, this structure should be a TD of the absent valence. If $p \geq 2$, this

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 427–437, 2013.

structure may have two or more than two transferred designations of the absent valences, and this "VP + *de*" is ambiguous. For example:

(1) wo3 gei3 ta1 yi1 ben3 shu1 de (yuanyin) （p = n − m = 3 − 3 = 0）
I give him one *ben* book *de* (reason)
(a reason) I gave him a book
(2) wo3 gei3 ta1 de (yi1 ben3 shu1) （p = n − m = 3 − 2 = 1）
I give him *de* (one *ben* book)
(a book) I gave him
(3) wo3 gei3 le yi1 ben3 shu1 de (ren2) （p = n − m = 3 − 2 = 1）
I give *le* one *ben* book *de* (person)
(to whom) I gave a book
(4) gei3 ta1 yi1 ben3 shu1 de (ren2) （p = n − m = 3 − 2 = 1）
give him one *ben* book *de* (person)
(who) gave him a book
(5) wo3 gei3 de (ren2 $_{\text{Dative}}$ / shu1) （p = n − m = 3 − 1 = 2）
I give *de* (person $_{\text{Dative}}$ / book)
(to whom $_{\text{Dative}}$)I gave / (some books) I gave
(6) gei3 ta1 de (ren2 $_{\text{Agentive}}$ / shu1) （p = n − m = 3 − 1 = 2）
give him *de* (person $_{\text{Agentive}}$ / book)
(who $_{\text{Agentive}}$) gave him / (some books) gave him
(7) gei3 le yi1 ben3 shu1 de (ren2 $_{\text{Agentive}}$ / ren2 $_{\text{Dative}}$) （p = n − m = 3 − 1 = 2）
give *le* one *ben* book *de* (person $_{\text{Agentive}}$ / person $_{\text{Dative}}$)
(who $_{\text{Agentive}}$ / to whom $_{\text{Dative}}$) gave a book

In these examples, "gei3"(to give) is a verb with three valences, and the structures headed by "gei3" show very strict regularity of SD or TD according to their Index of Ambiguity.

1.2 TD of SSP

After Zhu, Lu J.M. (in Shen [1] and Yuan [1]) mentioned such problems a lot of times in his works. In his opinion, there are some structures of "VP + *de*" that may still contain transferred designations even though their Index of Ambiguity is zero. Here are his examples:

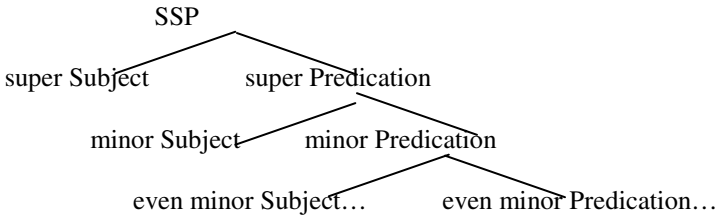
(8) chong2zi zhu4 le xinr1 de (tao2zi)（p = n − m = 2 − 2 = 0）
worm eat *le* core *de* (peach)
(some peaches with their) cores eaten by worms
(9) hai2zi kao3shang le Bei3jing1 Da4xue2 de (jia1zhang3)（p = n − m = 2 − 2 = 0）
child succeed in tests *le* Beijing University *de* (parents)
(some parents whose) child succeeded in tests and entered Beijing University
(10) wo3 si1 le feng1mian4 de (shu1)（p = n − m = 2 − 2 = 0）
I tear *le* cover *de* (book)
(some books with their) covers torn by me

(11) ger4 gao1 de (yun4dong4yuan2) (p = n − m = 1 − 1 = 0)
stature high *de* (athlete)
(some athletes being) high in stature

(12) shair3 hong2 de (yue4ji4hua1) (p = n − m = 1 − 1 = 0)
color red *de* (rose)
(some roses being) red-colored

(13) chuan1zhuo2 jiang3jiu de (gu1niang) (p = n − m = 1 − 1 = 0)
dress nice *de* (girl)
(some girls being) nice-dressed

This grammar phenomenon can be well explained by the theory of Subject 1 + Subject 2 + Predication (SSP). In modern Chinese, a Subject + Predication structure may have another Subject. In the above examples, all the verbal phrases can play as frames of "Subject2 + Predication" which form "super Predications" of other Subjects, the "super Subjects".



Therefore, example (8) to (13) can be revised to examples as following:

(14) zhe4xie1 tao2zi chong2zi zhu4 le xinr1
these peach worm eat *le* core
these peaches have their cores been eaten by worms

(15) zhe4 ge4 jia1zhang3 hai2zi kao3shang le Bei3jing1 Da4xue2
this *ge* parent child succeed in tests *le* Beijing University
this father/mother has his/her child succeeded in tests and entered Beijing University

(16) zhe4 ben3 shu1 wo3 si1 le feng1mian4
this *ben* book I tear *le* cover
this book has its cover torn by me

(17)na4 ge4 yun4dong4yuan2 ger4 gao1
that *ge* athlete stature high
that athlete is high in stature

(18) zhe4 duo3 yue4ji4hua1 shair3 hong2
this *duo* rose color red
this rose is red-colored

(19) na4 ge4 gu1niang chuan1zhuo2 jiang3jiu
that *ge* girl dress nice
that girl is nicely dressed

The reason of forming transferred designations of these "VP + *de*" structures, which all of their valences are already realized in form, is that these verbal phrases can serve as super Predicates of SSPs. What these "VP + *de*" structures may transferredly designate

are their super Subjects. If there is no super Subject, it will be impossible to form a TD. The Phrases in following examples are incorrect because they lack the super Subjects.

(20) * zhe4xie1 chong2zi tao2zi xinr1 zhu4 le →
* tao2zi xinr1 zhu4 le de (chong2zi)
(21)* zhe4 ge4 jia1zhang3 Bei3jing1 Da4xue2 hai2zi kao3shang4 le →
* Bei3jing1 Da4xue2 hai2zi kao3 shang4 le de (jia1zhang3)
(22)* wo3 zhe4 ben3 shu1 feng1mian4 si1 le →
* zhe4 ben3 shu1 feng1mian4 si1 le de (ren2)
(23)* ger4 na4 ge4 yun4dong4yuan2 hen3 gao1 →
* na4 ge4 yun4dong4yuan2 hen3 gao1 de (ger4)
(24)* shair3 zhe4 duo3 yue4ji4hua1 hen3 hong2 →
* zhe4 duo3 yue4ji4hua1 hen3 hong2 de (shair3)
(25)* chuan1zhuo2 na4 ge4 gu1niang hen3 jiang3jiu →
* na4 ge4 gu1niang hen3 jiang3jiu de (chuang1zhuo2) [1]

Maybe a VP is already an SSP, but it can still form a TD if it has another even super Subject before it. Here are some examples:

(26) xinr1 chong2zi zhu4 le → zhe4xie1 tao2zi xinr1 chong2zi zhu4 le → xinr1 chong2zi zhu4 le de (tao2zi)
(27) hai2zi da4xue2 kao3shang4 le → na4 ge4 jia1zhang3 hai3zi da4xue2 kao3shang4 le →hai2zi da4xue2 kao3shang4 le de (jia1zhang3)
(28) feng1mian4 wo3 si1 le → zhe4 ben3 shu1 feng1mian4 wo3 si1 le → feng1mian4 wo3 si1 le de (shu1)

Of course, the above analysis of the TD of SSP is not an end. More questions will be dealt with in this paper and the following ones will be the focus of our research.

What is the principle of the TD of SSP in Chinese?

How can this principle guide us in teaching Chinese as a foreign language and in Chinese information processing?

2 Principles of TD of SSP

Before our discussion, there are some questions that should be clarified.

First of all, SSP is not a sentence, but a phrase. As the basic units for communication between people, sentences have certain moods and rising or falling tones. The existence of sentences is restricted by some grammatical and pragmatical rules.[2] SSP is just a part of a sentence; it is an abstract grammatical element. Therefore, it is obvious that we can not judge whether a certain SSP is correct only according to the rules of sentences. For example:

(29) Zhe4 jian4 yi1fu ni3 hai2 mei2you3 ding4 kou4zi ne.
this *jian* dress you still not sew button *ne*

[1] "jiang3jiu" can be an adjective which means "nice, well", and be a transitive verb as well which means "pay attention to". If it is a transitive verb, the phrase "na4 ge4 gu1niang hen3 jiang3jiu de" is correct, and its meaning is "what that girl paid attention to".

[2] For more information, please refer to Xing F.Y. [1].

You have still not sewed on buttons for this dress
→ * ni3 hai2 mei2you3 ding4 kou4zi ne de (yi1fu)
→ ni3 mei2you3 ding4 kou4zi de (yi1fu)
(some dresses that) you did not sew on buttons

(30)Zai4 da4 de kun4nan wo3men ye3 bu4 pa4.
even big *de* difficulty we still not afraid
We are not afraid of even bigger difficulties.
→ ? wo3men ye bu4 pa4 de (kun4nan)
→ wo3men bu4 pa4 de (kun4nan)
(what) we are not afraid of

Secondly, the rules of Chinese sentences include some indirect principles besides some direct principles we already know. For instance, Subject inclines to be definite while Object inclines to be indefinite. The SSP we talk about here does not concern such problems. Consider the following examples.

(31) ta1 zhe4 ge4 fang2zi mai3 le
he this *ge* house buy *le*
he bought this house
→ * zhe4 ge4 fang2zi mai3 le de (ren2)
→ fang2zi mai3 le de (ren2)
(someone who) bought a house

Thirdly, a few verbs may be very special in that they can only form certain kind of TD anyway. For example:

(32) A jia1 B deng3yu2 C
A plus B equal C
A plus B is C
→ * A jia1 B deng3yu2 de
→ deng3yu2 C de (yun4suan4shi4)
(some expressions that) equal C

(33) Xiao3wang2 shi Shang4hai3 ren2
Xiaowang is Shanghai person
Xiaowang comes from Shanghai.
→ * Xiao3wang2 shi4 de
→ shi4 Shang4hai3 ren2 de (ren2)
(somebody who) comes from Shanghai

2.1 Typical TD of SSP

"VP + de" with Obligatory Case(s) Missing. According to our investigation, if some obligatory Cases do not appear in a "VP + *de*" structure, this structure can certainly form a TD. It is easy to understand. Obligatory Cases refer to some Cases which are semantically required by the verb. Agentive, Objective and Dative are naturally obligatory while other Cases, like Instrumental, Location, Time, may be or may not be obligatory. Whether a Case is obligatory or not depends on the semantic characteristic of the verb.

"VP + de" with the Subject 1 and Subject 2 in a Possessive Relationship or a Whole-Part Relationship. For example:

(34) zhe4 ge4 ren2 xin1yanr3 hao3
this *ge* people heart good
this man is good-minded
→ xin1yanr3 hao3 de (ren2)
(somebody who is) good-minded

(35) wo3men2 ban1 yi1 ban4 shi4 nan2fang1 ren2
our class one half is south people
half of our class comes from the south
→ yi1ban4 shi4 nan2fang1 ren2 de (ban1ji2)
(a class) half of which come from the south

(36) can1jia1 zhe4 xiang4 ke1yan2 gong1zuo4 de ren2, nian2ji qing1 de zhan4 duo1shu4
participate in this *xiang* research work *de* people, age young *de* take up most
Youths take up the most part of this research team.
→ nian2ji qing1 de zhan4 duo1shu4 de (team)
(a team with) youths taking up the most part.

Actually, some nouns also could be Predicates in Chinese and have the same function to form transferred designations as verbs. For Example:

(37) zhe4 zhang1 zhuo1zi san1 tiao2 tui3
this *zhang* table three *tiao* leg
this table has three legs
→ san1 tiao2 tui3 de (zhuo1zi)
(a table with) three legs

(38) zhe4li de shan1 lu4 shi2ba1 wan1
here *de* mountain road eighteen turn
the road of this mountain has a lot of turns
→ shi2ba1 wan1 de (shan1 lu4)
(a mountain road with) a lot of turns

In the above two examples, "zhe4 ge4 ren2"(this *ge* people)(this person) and "xin1yanr3"(mind) have a possessive relationship while "zhe4 zhang1 zhuo1zi"(this *zhang* table)(this table) and "san1 tiao2 tui3"(three *tiao* leg)(three legs) have a whole-part relationship, therefore the formers can be Subject 1 while the latters can be Subject 2. The SSP formed by them has the possibility to form TD.

2.2 Conditional TD of SSP

"VP + *de*" with Subject 1 and Subject 2 in an Indirect Relationship. Sometimes the relationship between Subject 1 and Subject 2 is not very clear. However, Subject 1 should always have a close relationship with a part of super Predicate. If the Subject 1 has a possessive relationship or whole-part relationship with the Object of the super Predicate, then this SSP can form a TD. For example:

(39) zhe4 kuai4 dan4gao1 wo3 chi1 le yi1 ban4
this *kuai* cake I eat *le* one half

I've eaten a half of this cake
→ wo3 chi1 le yi1 ban4 de (dan4gao1)
(a cake with) a half of it being eaten by me

(40) jia1 li de shir4, ni3 ying4gai1 duo1 zuo4 yi1xie1
home inside *de* work, you should more do some
you should do more housework
→ ni3 ying4gai1 duo1 zuo4 yi1xie1 de (shir4)
(some works that) you should do more

If the Subject 1 only has a relationship with the verb, generally speaking, it is impossible to form a TD. Such Subject 1 is similar to a section grammatically independent to Subject 2 + Predicate. We can add some prepositions, like "guan1yu2"(about), "dui4yu2"(as for), before this Subject 1. For example:

(41) lan2qiu2 sai4 ta1men shi1bai4 le
basketball game they lose *le*
they lost the basketball game
→ * ta1men shi1bai4 le de (bi3sai4)

(42)zhe4 jian4 shir4 za2men shang4dang4 le
this *jian* thing we be cheated *le*
we've been cheated on this case
→ * zan2men shang4dang4 le de (shir4)

(43)zhe4 jian4 shir4 wo3 bu4 guai4 ni3
this *jian* thing I not blame you
I don not blame you for this case
→ * wo3 bu4 guai4 ni3 de (shir4)

"VP + *de*" with Subject 1 or Subject 2 Being Indefinite. Here, the word "indefinite" means some nominal structures in Chinese which have the meaning of everyone, anywhere, everything etc.[3] If Subject 1 or Subject 2 is indefinite, an SSP can form a TD only when this SSP has missing valence(s). For example:

(44) shen2me huor2 wo3men dou1 gan4
whatever job we all do
we shall do any kind of job
→ wo3men gan4 de (huor2)
(some jobs that) we shall do

(45) wo3men shen2me huor2 dou1 gan4
we whatever job all do
we shall do any kind of job
→ shen2me huor2 dou1 gan4 de (ren2)
(someone who) will do any kind of job

(46) qi2yu2 de ren2 jiu4 ni3 wang4-wang wo3, wo3 wang4-wang ni3
rest *de* people then you look me, I look you
then, the rest people are looking at each other
→ * ni3 wang4-wang wo3, wo3 wang4-wang ni3 de (ren2)

[3] Refer to Lu J.M. [1].

(47) zan2men lia3 shui2 ye3 bie2 guan3 shui2
we two who also not control who
we two do not control each other
→ * shui2 ye3 bie2 guan3 shui2 de (ren2)

"VP + *de*" with Subject 1 Being Locative or Instrumental of the Predicate. If the Subject 1 is the Case of Locative or Instrumental of the verb in "VP + *de*", and it is semantically required by the verb, then this "VP + *de*" can form a TD. If the Subject 1 is not semantically required by the verb, this "VP + *de*" can only form a SD. Please compare:

(48) zhe4 jian1 wu1zi wo3men fang4 dong1xi
this *jian* room we store thing
we stored goods in the room
→ wo3men fang1 dong1xi de (wu1zi)
(a room where) we stored goods
→ zhe4 jian1 wu1zi shi4 wo3men fang4 dong1xi de
this room is where we stored goods

(49) zhe4 ba3 dao1 wo3 qie1 rou4
this *ba* knife I cut meat
I cut the meat with this knife
→ Wo3 qie1 rou4 de (dao1)
(a knife that) I cut meat
→ zhe4 ba3 dao1 shi4 wo3 qie1 rou4 de
this knife is what I used to cut meat

(50) xia4wu3 wo3men kai1 hui4
afternoon we have meeting
this afternoon we shall have a meeting
→ wo3men kai1 hui4 de (shi2hou4)
(a time when) we have a meeting
→ * xia4wu3 liang3 dian3 shi4 wo3men kai1 hui4 de

(51) nan2fang1 zhe4xie1 tian1 zheng4 xia4 yu3
south these day *zheng* fall rain
it is raining these days in the south
→ zhe4xie tian1 zheng4 xia4 yu3 de (di4fang)
(somewhere) is raining in these days
→ * nan2fang1 shi4 zhe4xie1 tian1 zheng4 xia4 yu3 de

Unlike to "kai1(hui4)" and "xia4(yu3)", "fang4" and "qie1" always requires a place and an instrument to appear with them in the same clause. We can not image how to do the action of "fang4" without a place, neither to suggest an action of "qie1" without an instrument. In this condition, Locative and Instrumental are obligatory. Just because of this reason, if these obligatory Cases are missing, the TD may be ambiguous. For example:

(52) fang4 dong1xi de(ren2/wu1zi)
store thing *de*(people/room)

(someone/ somewhere) stored goods
→ ta1 jiu4shi4 zuo2tian1 fang4 dong4xi de
he is yesterday store thing *de*
he is who stored goods yesterday
→ zhe4 ge4 wu1zi shi4 fang4 dong1xi de
this *ge* room store thing *de*
this room is where stored goods

(53)qie1 rou4 de(ren2/dao1)
cut meat *de*(people/knife)
(someone who/ something used to) cut meat
→ ta1 shi4 qie1 rou4 de
he is cut meat *de*
he is who cut meat
→ zhe4 ba3 dao1 shi4 qie1 rou4 de
this *ba* knife is cut meat *de*
this knife is used to cut meat

2.3 Some SSP Structures That Can Not Form TD

If Subject 1 and Subject 2 are synonymous or totally indicate each other, this SSP can not form a TD. For example:

(54) wo3 de xing4ming4 shi4 xiao3
my *de* life thing small
the matter of my life is unimportant
→ * shi4 xiao3 de (shi4)

(55)ta1 zhe4 ge4 ren2 lao3 kai1xiao3chai1
he this *ge* people always absent-minded
he is always absent-minded
→ * zhe4 ge4 ren2 lao3 kai1xiao3chai1 de (ren2)

(56)ta1men lia3 yi1 ge4 ban4 jin1, yi1 ge4 ba1 liang3
they two one *ge* half *jin*, one *ge* eight *liang*[4]
they two are Tweedledum and Tweedledee
→ * yi1 ge4 ban4 jin1, yi1 ge4 ba1 liang3 de (ren2)

(57)Xiao3zhang1 he2 Xiao3wang2 na3ge4 dou1 hao3
Xiaozhang and Xiaowang who both good
both of Xiaozhang and Xiaowang are good
→ * na3ge4 dou1 hao3 de (ren2)

(58)tiao4wu3, chang4ge1 yang4yang4 jing1tong1
dancing, singing everything be expert at
(someone is) expert at everything like dancing and singing
→ * yang4yang4 jing1tong1 de (shi4)

[4] "*jin*" and "*liang*" are measuring units used both in the ancient and modern China. One "*jin*" which contains 10 *liang* equals 500 grams in modern time. But in ancient time, one *jin* sometimes equaled 16 *liang*. That means, a half of *jin* was 8 *liang*.

3 Conclusion and Others

What we have discussed above can be concluded as follows:

Firstly, if a "VP + *de*" structure has obligatory Case(s) missing, it can undoubtedly form a TD. What it transferredly designates is/are the missing Case(s).

Secondly, if Subject 1 has such a close relationship (possessive or whole-part) with Subject 2 or Object, then this SSP can easily form a TD.

Thirdly, if Subject 1 is the obligatory Case of Locative, or Instrumental which is semantically required by the verb, then the structure of "Subject 2 + VP + *de*" can form a TD and refers to the Subject 1.

Fourthly, an SSP with its Subject 1 and Subject 2 being synonymous can not form a TD.

These four principles are meaningful when we do the Chinese language processing. They can guide us to infer what a complex "VP + *de*" designates. We may create a model to "calculate" its meaning with the help of computer.

Furthermore, in our opinion, the theory of TD of SSP can resolve some difficult problems in teaching Chinese as a foreign language. For example, what is the difference between these following two phrases?

(59) ping2guo3 san1 kuai4 qian2 yi1 jin1
apple three *kuai* money one *jin*
the apples are three *kuai* per *jin*
(60)ping2guo3 yi1 jin1 san1 kuai4 qian2
apple one *jin* three *kuai* money
one *jin* of apples are three *kuai*

It is obvious that these two sentences have different focuses. Their deep structures are syntactically different to each other according to the theory of TD of SSP. The former structure "ping2guo3 // san1 kuai4 qian2 yi1 jin1" is an SSP with a noun phrase as its minor Predicate while the latter "ping2guo3 yi1 jin1 // san1 kuai4 qian2" is not. If we do not analyze these two phrases in this way, we probably can not explain the following phenomena:

(61) ping2guo3 san1 kuai4 qian2 yi1 jin1
→ san1 kuai4 qian2 yi1 jin1 de (ping2guo3)
(some apples with the price of) three *kuai* per *jin*
(62) ping2guo3 yi1 jin1 san1 kuai4 qian2
→ * yi1 jin1 san1 kuai4 qian2 de (ping2guo3)
→ san1 kuai4 qian2 de (yi1xie1 ping2guo3)
(some apples costing) three *kuai*
(63) niu2zai3ku4 ba1shi2 kuai4 qian2 liang3 tiao2
jeans eighty *kuai* two *tiao*
jeans are eighty *kuai* for two
→ bai1shi2 kuai4 liang3 tiao2 de(niu2zai3ku4)
(some jeans with the price of) eighty *kuai* for two
(64) niu2zai3ku4 liang3 tiao2 bai1shi2 kuai4 qian2
jeans eighty *kuai* two *tiao*
two jeans are eighty *kuai*

≠ liang3 tiao2 bai1shi2 kuai4 qian2 de(niu2zai3ku4)
two jeans with the price of eighty *kuai* for each
→ bai1shi2 kuai4 qian2 de (liang3 tiao2 niu2zai3ku4)
(two jeans costing) eighty *kuai*

References

1. Chen, P.: Double NP Constructions and Topic-Comment Articulation in Chinese. Chinese Language 6, 493–507 (2004)
2. Fillmore, C.J.: The Case for Case. The Commercial Press, Peking (2002)
3. Hu, Y.S.: On the Nominal Constituent at the Beginning of Sentences in Chinese. Language Teaching and Linguistic Studies 4, 13–20 (1982)
4. Lu, J.M.: Indefinite Subjects and Others. Chinese Language 3, 161–167 (1986)
5. Shen, Y.: Zheng Ding'ou: Studies on Valent Grammar in Modern Chinese. Peking University Press, Peking (1995)
6. Xing, F.Y.: A Hypothesis of Clausal Pivot in Chinese. Chinese Language 6, 420–428 (1995)
7. Yuan, Y.L.: A Cognitive Research on Monovalent Nouns. Chinese Language 4, 241–253 (1994)
8. Zhu, D.X.: "de" Structures and Judgment Sentences. Chinese Language 1 & 2, 23–26, 104–110 (1978)
9. Zhu, D.X.: The Chinese Grammar Handout. The Commercial Press, Peking (1982)
10. Zhu, D.X.: Self Designation and Transferred Designation. Dialects 1, 16–31 (1983)

The Research on Sequential Meaning Extension: A Case Study on the Polysemy of 看(kan)

Xiaofang Ouyang

School of Chinese Linguistics & Literature/Center for Study of Language & Information,
Wuhan University, Luojia Mountain,
430072, Wuhan, P.R. China
bbrirao@126.com

Abstract. Meaning extension patterns can help to construct the large-scale semantic knowledge base. After analyzing the meanings of 看(kan) in modern Chinese, we found a new extension pattern——sequential extension. That pattern is also reflected in the meaning extensions of some other high frequency verbs. The meanings extended with the sequential pattern are characterized by the following: (1) overall continuity, (2) activating the before and after, (3) the distance determining the degree of activation. On the view of cognition, the sequential extension pattern roots in the mode of ACTING-RESULT and metonymic mechanism.

Keywords: Sequential extension, meaning extension pattern, 看(kan), meaning continuum.

1 Background

Semantic processing is the key problem of natural language processing now. As the basic foundation of semantic processing, the construction of large-scale semantic knowledge base (SKB) is becoming one of the most important developing orientations. With the success of some SKBs as WordNet, the paradigmatic semantic network with hierarchies, connections and nodes has been accepted universally as the form of knowledge representation. That needs to reveal the relations between word meanings. Meaning extension plays an important role in the meaning relations. Research shows that extension is the basic form of the motion of word meanings and meaning extension is regular. The laws of meaning extensions can help to clear the complicated relations between word meanings, and to provide the foundation for constructing the SKB.

The study on meaning extension has a long history. It is generally acknowledged that the description about extended meaning by Kai Xu in Southern Tang is the beginning. Since then meaning extension has increasingly become an important issue of language studies. Especially in the recent years, scholars have worked on meaning extensions from multiple angles and made lots of achievements that provided a reference for us. Some representative achievements offered some reference to our research [1-3]. After analyzing the meanings of 看(kan) in modern Chinese, we found a new extension

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 438–447, 2013.

pattern—sequential extension. This paper documents the study process, analyzes the characters and tries to prove the existence of the sequential extension pattern.

2 The Meaning Analysis and Organization of 看(kan)

看(kan) is the core of visual verbs in modern Chinese and one of the most frequently used words in daily communication. According to the statistics in 3000 Common Words in Modern Chinese, 看(kan, 4th tone) is ranked 44. In addition, 看(kan, 1st tone) is included in our research field of 看(kan) for the following reasons: (1) the two 看 with different tones share the same historical origin, (2) there is a close semantic relation between them, (3) the phonetic difference doesn't play the distinctive role for the machine.

The traditional dictionary definitions can provide some reference for our meaning analysis. As the representative achievement of traditional dictionary, *the Contemporary Chinese Dictionary* (5th edition) [4] records eight senses of 看(kan, 4th tone) and two senses of 看(kan, 1st tone) as follows:

看(kan, 4th tone): (P762)

① 动(dong, v.), 使视线接触人或物(shi shixian jiechu ren huo wu, direct one's gaze towards somebody or something): 看书(kanshu, read a book)｜看电影(kan dianying, see a movie)｜看了他一眼(kan le ta yiyan, give him a look)

② 动(dong, v.), 观察并加以判断(guancha bing jiayi panduan, observe and make a judgment): 看他是个可靠的人(kan ta shi ge haoren, think he is a reliable person)｜看你这个办法好不好(kan ni zhege banfa haobuhao, find out whether your way is good or not)

③ 动(dong, v.), 访问(fangwen, visit): 看望(kanwang, visit)｜看朋友(kan pengyou, visit a friend)

④ 动(dong, v.), 对待(duidai, treat): 看待(kandai, treat)｜另眼相看(lingyan xiangkan, pay special regard to)

⑤ 动(dong, v.), 诊治(zhenzhi, make a diagnosis and give treatment): 王大夫把我的病看好了(Wang daifu ba wode bing kanhao le, Doctor Wang cured me of my illness)

⑥ 动(dong, v.), 照料(zhaoliao, take care of): 照看(zhaokan, take care of)｜衣帽自看(yimao zikan, take care of your own hats and coats)

⑦ 动(dong, v.), 用在表示动作或变化的词或词组前面 (yongzai biaoshi dongzuo huo bianhua de ci huo cizu qianmian, used in front of a verb or phrase that means action or change), 表示预见到某种变化趋势 (biaoxian yujiandao mouzhong bianhua qushi, to express anticipating some variation tendency), 或者提醒对方注意可能发生或将要发生的某种不好的事情或情况 (huozhe tixing duifang zhuyi keneng fasheng huo jiangyao fasheng de mouzhong buhao de shiqing huo qingkuang, or to draw the other's attention to some bad thing that may or will happen): 行情看涨(hangqing kanzhang, market will be strong)｜别跑！看摔着！(Bie pao! Kan shuaizhe! Don't run. Mind you don't fall.)

⑧ 助(zhu, aux.), 用在动词或动词结构后面(yongzai dongci huo dongci jiegou houmian, used after a verb or a verb structure), 表示试一试(biaoshi shiyishi, to express making an attempt), 前面的动词常用重叠式(qianmian de dongci changyong chongdieshi, the verb in front is usually reduplicated): 想想看(xiangxiang kan, try to think it over)｜找找看(zhaozhao kan, just try to find)｜等一等看(dengyideng kan, wait and see)

看(kan, 1st tone): 动(dong, v.) (P761)

① 守护照料(kanhu zhaoliao, look after)

② 看押监视(kanya jianshi, detain and watch)

The intention of *the Contemporary Chinese Dictionary* is to popularize Putonghua and promote the standardization of modern Chinese. So it is written for the native Chinese speakers. The separation and description of word meanings in *the Contemporary Chinese Dictionary* are not suitable for constructing the machine-readable semantic network. For example, the first semantic item of 看(kan, 4th tone) in the Dictionary (5th edition) is 使视线接触人或物 (shi shixian jiechu ren huo wu, direct one's gaze towards somebody or something): 看书 (kanshu, read a book)｜看电影 (kan dianying, see a movie)｜看了他一眼(kan le ta yiyan, give him a look). But the typical example 看书 means reading a book usually. The 看(kan) here focuses on interpreting with mind rather than eye contact. The 看(kan) in 看电影(kan dianying) is aimed at enjoying or deriving pleasure by sight and hearing. It is not just eye contact, either. So we need to newly conclude the glossemes of 看(kan). To summarize, the existing problems of traditional definitions mainly display in three aspects: (1) different super-glossemes are combined into one semantic item; (2) the hierarchies between the glossemes are not clear; (3) the separation within one glosseme is not clear [5]. So we should reanalyze the meanings of 看(kan).

From the large-scale real corpus[1] we found the meanings of 看(kan) are very complicated. At least 13 glossemes that express concrete behaviors can be concluded as follows: [2]

$看_1$-视线接触(shixian jiechu, look): 我在看那边(wo zai kan nabian, I am looking there);

$看_2$-看见(kanjian, see): 我看他手里拎了两个袋子(wo kan ta shouli lin le liangge daizi, I saw he took several bags);

$看_3$-读取(duqu, read): 她喜欢读书看报(ta xihuan dushu kanbao, she likes reading books and newspapers);

$看_4$-评阅(pingyue, read and appraise): 学校从今年开始施行电脑看卷(xuexiao cong jinnian kaishi shixing diannao kanjuan, from this year the school begins to use the computer to evaluate the test papers);

[1] The main source of the corpus is `http://ccl.pku.edu.cn:8080/`

[2] Only the word meanings of 看(kan) that express concrete behavior categories are discussed in this paper. The abstract 看(kan) used in the special syntactic structures, such as 看来(kanlai, it seems), 尝尝看(changchang kan, try to taste) and 看摔着(kan shuaizhe, mind you don't fall), will be studied in another paper.

看$_5$-赏看(shangkan, watch and enjoy): 看风景(kan fengjing, enjoy the scenery)/我们打算明天去看球赛(women dasuan mingtian qu kan qiusai, we are going to watch the ball game tomorrow);

看$_6$-察看(chakan, observe): 看地形(kan dixing, detect the landform);

看$_7$-诊察(zhencha, diagnose): 中医靠把脉看病(zhongyi kao bamai kanbing, Chinese medicine doctors take one's pulse to diagnose his illness);

看$_8$-诊治(zhenzhi, diagnose and treat): 王医生把我的病看好了(Wang yisheng ba wode bing kanhao le, Doctor Wang cured my illness);

看$_9$-看望(kanwang, visit): 母亲出去看朋友了(muqin chuqu kan pengyou le, my mother went to visit her friend);

看$_{10}$-照看(zhaokan, take care of): 衣帽自看(yimao zikan, take care of your own hats and coats)/你留在家里看孩子(ni liuzai jiali kan haizi, you stay home to look after the kids)

看$_{11}$-看管(kanguan, detain and watch): 看犯人(kan fanren, guard prisoners);

看$_{12}$-(观察之后)判断或预测(guanchazhihou panduan huo yuce, judge the present situation or predict the future situation after observation): 我看你有烦心事儿(wok an ni you fanxin shier, I see you've got a trouble)/我看马上就要下雨了(wo kan mashang jiuyao xiayu le, I think it will rain soon);

看$_{13}$-看待(kandai, treat or regard…as…): 没人把你当才子看(meiren ba ni dang caizi kan, nobody regards him as a gifted scholar).

3 The Sequential Extension Existing in the 看(kan)'s Meaning Development

The polysemy in modern Chinese is the result of meaning development and evolution. Because visual behavior is very important for human, visual verbs were generated very early. They have been used very frequently and widely for a long time. So their meanings are primordial and basic. The development and evolution of visual verbs' meanings are not easy to be broken. The whole process is very complete and tightly linked to the actual human life. So it is possible to extract some representative law of meaning extension from the core visual verb-看(kan). 看(kan) has developed at least 13 glossemes that express concrete behaviors in modern Chinese, but the relations between those meanings are clear, as follows:

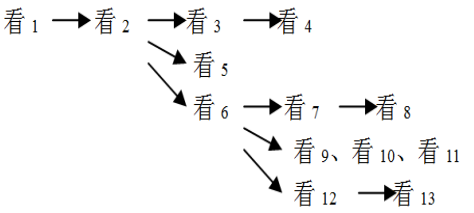


Fig. 1. The relations between 13 meanings of kan

According to Fig1, we can see some clear extension paths. For example, 看$_1$(look)→看$_2$(get visual information)→看$_6$(analyze the information)→看$_{12}$(make the judgment)→看$_{13}$(treat). From 看$_1$(visual act) to 看$_{13}$(treat), the five senses of 看 form a whole event chain. They happen in sequence and the former is always the cause of the latter. It shows a new meaning extension pattern-sequential extension. Let's see the following example sentences.

A. 我看了他一眼(Wo kan le ta yiyan).
I gave him a look. 看$_1$(visual act)

B. 我看他在吃饭, 就没打扰他(Wo kan ta zai chifan, jiu mei darao ta).
I saw that he was eating. So I didn't bother him.
看$_2$(the result of 看$_1$, get the visual information）

C. 看他吃饭的样子, 很着急(Kan ta chifan de yangzi, hen zhaoji).
From the way he ate, he was in a hurry.
看$_6$(analyze the visual information)

D. 我看他是要赶时间(Wo kan ta shi yao gan shijian).
I thought he had something urgent.
看$_{12}$(make a judgment after analysis)

E. 不该把他当闲人看了(Bugai ba ta dang xianren kan le).
I shouldn't treat him as an idler.
看$_{13}$(treat sb in some way according to the judgment)

The 看 in Example A is the concrete visual action. The 了 expresses that the action happened and completed. The 一眼 is the quantity of the motion. Those implies that the agent-我 (I) will get the visual information by visual action. But the focus of the sentence is still the action, not the result. The 看 in Example B draws forth the content of the visual information-他在吃饭 (ta zai shifan, he was eating). The 看(kan) here focuses on the result of the visual action. It is the continuation of 看$_1$. 看$_1$ and 看$_2$ are cause and effect. The 看(kan) in Example C emphasizes on observing and analyzing which are based on the visual information. It is the continuation of 看$_2$ and implies some decision will be made. Then a judgment is made in Example D. The agent takes up a corresponding attitude to treat the patient in Example E. From Example A to E, the five senses of 看(kan) are used one by one. Five sentences describe a whole event chain. With the concrete visual action as the start point, a series of senses were extended in time order and by cause-and-effect relationship. They form a complete continuity.

Word meanings are generated in the use of language. The same word will develop different meanings when it is used in different situations. Those meanings are in the extension sequence in some order. In the real communication, only a part of the continuity will be focused on. After this happens many times, a new sense is created. When the sense is used in the context, the focus point will be activated first. Meanwhile the before or after in the continuity may be activated, too. That is to say, the sense will be used as a lever to activate more information in the language communication. Only the focus point should be chosen as the lexical definition of dictionaries.

But sometimes the before and after part activated are also included into the scope of the definition. For example, 重视 (zhongshi, attach importance to) focuses on treating with some attitude. But the definition in *the Contemporary Chinese Dictionary* (5th edition) is 认为人的德才优良或事物的作用重要而认真对待 (renwei ren de decai youliang huo shiwu de zuoyong zhongyao er renzhen duidai, think somebody possess intelligence and virtue or something important, so attach importance to). In that definition, the part before 认真对待 (renzhen duidai, attach importance to) is just the activated neighboring part-making judgment. The process of observing and analyzing before judgment is implied. The following two extension sequences of 看 also reflect the character.

看$_1$(look)→看$_2$(get visual information)→看$_3$(read) →看$_4$(appraise)
看$_1$(look)→看$_2$(get visual information)→看$_5$(enjoy)

The 看 in 我在看书 (wo zai kanshu, I am reading the book) focuses on acquiring the content of the book and activates the before parts-the visual acting and getting the visual information. The 看 in 我在看风景 (wo zai kan fengjing, I am enjoying the scenery) focuses on enjoying the scenery, meanwhile activates the behavior of directing the gaze toward the scenery and acquiring the visual information. The frequency of 看$_1$ being activated with 看$_3$ or 看$_5$ is very high. So the meanings of 看$_3$ and 看$_5$ are merged into the sense: 使视线接触 (shi shixian jiechu, direct one's gaze towards). And their real focuses are ignored. But it can't be denied that 看书 (kanshu, read a book) and 看风景 (kan fengjing, enjoy the scenery) can be replaced with 阅读书 (yuedu shu, read a book) and 欣赏风景 (xinshang fengjing, enjoy the scenery). So 看$_3$ and 看$_5$ should be distinct senses and shouldn't be simply merged into 看$_1$. Besides, 看$_3$ may be extended to a sequential meaning in some special situations. For example, the 看 in 看卷子 (kan juanzi, read and appraise the test papers) is the sequential meaning of 看$_3$. The process of reading is extended to the next step-analyzing and judging. That is 看$_4$. So we can also say 评卷子 (ping juanzi, grade paper) or 判卷子 (pan juanzi, mark paper). 看(kan, look) →阅(yue, read) →评 (ping, appraise) is a semantic continuity. 评(ping, appraising) must be based on 看 (kan, look) and 阅(yue, read).

In addition, we found that the back meanings in the semantic continuity can always activate the neighboring before part. And the larger the distance is, the less likely the before part will be activated. The latter sense in the sequence is relatively far away from the start point-看$_1$, so the visual acting is rarely activated in the content. For example, the 看 in 我把他当弟弟看 (wo ba ta dang didi kan, I regard him as my younger brother) is 看$_{13}$ that is located at the end of the sequence. So 看$_{13}$ here may activate the neighboring before-看$_{12}$ (making the judgment), but won't imply any visual acting and information. 我(wo, I) may get the information that lets 我(I) think 他(ta, him) is like 弟弟(didi, younger brother) by eyes, ears or others. Especially when the patient is abstract, the possibility of activating 看$_1$ is very low. For example, the patient in 我看这个问题很严重 (wo kan zhege wenti hen yanzhong, I think this

problem is very serious) is 这个问题 (zhege wenti, this problem) which is abstract. So the 看$_{12}$ (make a judgment) here is not related to visual acting. 看$_1$ isn't activated, but the neighboring 看$_6$(observe and analyze) is. So the second sense of 看 in *the Contemporary Chinese Dictionary* (5th edition) is 观察并加以判断 (guancha bing jiayi panduan, observe and make a judgment) that merged the two neighboring parts-看$_6$ and 看$_{12}$ in the sequence into one definition item.

From above, we can see that only one point in the semantic continuity can be the focus in one special context. The before and after are just the activated items. If there are two possible focuses in one sentence, it will lead to ambiguity. For example, the 看 in 我在看房子(wo zai kan fangzi) may focus on 看$_1$, 看$_6$ or others. So there are at least two ways to understand the sentence-I am looking at the house or I am inspecting the house. It is an ambiguous sentence. If it is changed to 我在看房子怎么样 (wo kan fangzi zenmeyang, I am inspecting how the house is), the focus is stable on 看$_6$. It is disambiguated.

4 The Sequential Extension Existing in Other Verbs' Meaning Development

After analyzing the meanings of 看(kan) in modern Chinese corpora, we found the sequential extension pattern. But it is just a hypothesis and still need to be proved by studying the diachronic corpora. Another way to prove it is to find the same extension pattern existing in other verbs' meaning development. We did it.

听(ting, listen):

F. 我在听他们说话(Wo zai ting tamen shuohua).
I was listening to them. 听$_1$: listen, the hearing action

G. 我听他们说要让我去乡下(Wo ting tamen shuo yao rang wo qu xiangxia).
I heard them say that let me go to the country.
听$_2$: hear, the result of listening

H. 听他们的语气，很强硬(Ting tamen de yuqi, hen qiangying).
From their tone, they were very tough. 听$_3$: analyze

I. 你听出来了吗(Ni ting chulai le ma)?
Did you figure out? 听$_4$: make a judgment

The 听(ting) in Example F. focuses on the hearing motion process and activates the corresponding result. The focus of the 听(ting) in Example G is the consequence of the listening. The before part, listening, is activated and the following mental process may be implied. In Example H, the 听 focuses on analyzing the auditory information and then the agent must make some judgment. The judgment is made by the 听 in Example I. From Example F to I, the content expressed by all those 听 forms a complete and interlocking event chain: listening →hearing →analyzing →judging. With the hearing action as the start point, the meaning development of 听(ting) also followed the sequential pattern.

打(da, hit):

J. 有人在打门(You ren zai damen).
Somebody is hitting the door. 打 $_1$: hit

K. 他家的门给人打了(Tajia de men gei ren da le).
His door was broken. 打 $_2$: break

The 打(da) in Example J refers to the motion process of hitting by hand. The 打(da) in Example K expresses the result of hitting—broken that implies the motion happened before.

学(xue, study):

L. 我在学知识(Wo zai xue zhishi).
I am studying some knowledge. 学 $_1$: study

M. 跟着他我学了很多知识(Gen zhe ta wo xue le henduo zhishi).
I mastered much knowledge from him. 学 $_2$: master

The 学(xue) in Example L and M also shows the meaning development from the process to the result.

想(xiang, think):

N. 我在想那天吃饭的情景(Wo zai xiang natian chifan de qingjing).
The scene of dinner calls back to my mind. 想 $_1$: call back to mind

O. 我想吃饭了(Wo xiang chifan le).
I want to have dinner. 想 $_2$: want

P. 我得想办法解决吃饭问题(Wo dei xiang banfa jiejue chifan wenti).
I have to think about some way to get to eat. 想 $_3$: use one's mind

Q. 我想在附近找家快餐店随便吃点儿(Wo xiang zai fujin zhaojia kuaicandian suibian chi dianr).
I decided to eat something in the fast food restaurant nearby.
想 $_4$: decide, plan

The meaning of 想(xiang) in Example N should be the original meaning. Something reappearing in one's mind is 想$_1$. The reappearing thing may arouse the agent's desire of possession or experiencing that again. That is the 想$_2$ in Example O. Then the agent needs to use his head to realize his desire such as the 想$_3$ in Example P. After thought, some decision or plan comes out. The 想$_4$ is the end of the semantic sequence.

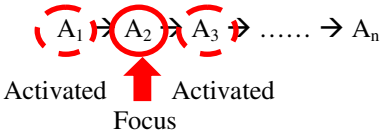
5 Conclusion

Through the above content, some evidence is offered to improve the existing of the sequential extension in some verbs' meaning development. And the meanings extended with the sequential pattern are characterized by the following:

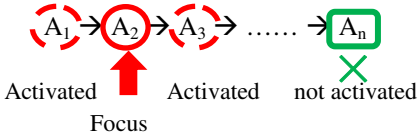
(1) Overall continuity. With some concrete action meaning as the start point, a series of senses extended in time order and by cause-and-effect relationship. They form a complete continuity. The content expressed by them can form a whole and interlocking event chain.

$$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow \ldots\ldots \rightarrow A_n$$

(2) Activating the before and after. When some sense in the semantic continuity is focused on in the specific context, the neighboring before and after parts may be activated, too. So the semantic information in the real language is more than the sense itself. The sense is used as a lever to activate more information in the language communication.



(3) The distance determining the degree of activation. When some meaning point is focused on in the content, the neighboring before and after parts, especially the before, are the most likely to be activated. And the further the other parts are from the focus meaning, the less likely they are activated.



Considering from many aspects, we think that the sequential extension pattern roots in the following three:

(1) The mode of ACTING-RESULT existing in the objective world. So-called SEQUENTIAL is the reflection of ACTING-RESULT in the objective world. As a kind of acting force, visual action must lead to the corresponding result. It will cause another new acting force and then some new result will be produced. That is a fundamental law in the world. The law must be reflected in people's cognitive world and language level.

(2) The cognition of people about the mode of ACTING-RESULT. The mode of ACTING-RESULT exists in the objective world. Only through recognition and abstraction of the complicated experience, the mode can be acquired by human and reflected in the language.

(3) The effect of the metonymic mechanism of thinking. According to the cognitive linguistics, the metonymic mechanism is based on proximity principle and prominence principle. In the semantic continuity produced by the sequential extension, each part is on the neighboring timeline and logically independent one another. The relationship between them is just as proximity. As the determinant of each sense, the focus is similar to the prominence. So on the view of cognition, the sequential extension is the production of the metonymic mechanism.

Acknowledgments. This research is supported by the Major Projects of National Social Science Foundation of China (11&ZD189) and the Post-70s Scholars Academic Development Program of Wuhan University: Interdisciplinary Research Team of Applied Linguistics.

References

1. Bai, Z.L.: Meaning Extension and Deducing Word's Meaning by Extension. Research in Ancient Chinese Language 4, 29 (1991)
2. Lu, Z.D., Wang, N.: Gloss and Chinese Exegetics, pp. 113–124. Shanxi Education Press, Taiyuan (1994)
3. Xu, G.Q.: The System Theory of Modern Chinese Vocabulary, pp. 230–233. Peking University Press, Beijing (1999)
4. Chinese Academy of Social Sciences Research Institutes of Language Department of Dictionary: Contemporary Chinese Dictionary, 5th edn., pp. 761–762. Commercial Press, Beijing (2005)
5. Ouyang, X.F.: The Cognitive Study of the Paradigmatic Semantic Network Related to Kan, pp. 20–24. Doctoral Dissertation of Wuhan University (2009)

The Semantic Feature Analysis and Formalization of the "Appearance" Attribute Nouns

Yanping Xu and Jincheng Zhang

School of Literature and Journalism, Hubei Engineering University, Xiaogan, China, 432000
{me19751219,zhang_jin_20}@163.com

Abstract. The "appearance" attribute nouns in Mandarin Chinese are the nouns describing the appearance of things. There is a necessary connection between the semantic features of the "appearance" attribute nouns and their syntax functions. Analyzing the semantic features of the "appearance" attribute nouns could improve the research on syntax-semantics interface directly. Formalizing the semantic features By Parse-Annotate Translate (PATR) and Copenhagen Tree Tracer (CTT) system could help computer to understand the attribute nouns automatically to some extent.

Keywords: "appearance" attribute nouns, semantic feature, formalization.

1 Introduction

With the rising of the Strict Lexicalism in the late 20th century, the focus of language description has shifted to the lexical level from syntactic level. More and more scholars have realized that semantic feature of words will play a key role in the research and paid more attention to this kind of semantic feature in this condition. So far, they have made a lot of achievements in this research area. But there are many problems needed to be investigated, such as how to extract semantic features and how to formalize the semantic features. These directly restrict the development of language information processing. Therefore, analyzing and formalizing the semantic feature is becoming more and more important. Based on the theory of Frame Semantics, we will analyze the semantic features of the "appearance" attribute nouns and test them in the Copenhagen Tree Tracer (CTT) system.

2 The Object and the Theory Basis of the Study

2.1 The Object of the Study

We intend to study attribute nouns in Mandarin Chinese. The attribute refers to some inherent aspects of things and phenomena, such as "color", "taste", "length", etc. At the same time, it could be some relationships between things and phenomena too, such as "proportional", "distance", "impression" and so on. In a word, the attribute nouns in Mandarin Chinese are the nouns referring to attributes.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 448–458, 2013.

The attribute nouns have many subcategories, such as "thing attribute nouns ","event attribute nouns ", "space attribute nouns ", etc. And the "appearance" attribute nouns belong to "thing attribute nouns ". In this paper, we will analyze ten "appearance" attribute nouns. They are "外形(wàixíng, external form)", "外貌(wàimào, aspect)", "外表(wàibiǎo, likeness)", "外观(wàiguān, facade)", "相貌(xiàngmào, personal appearance)", "容貌(róngmào, facial features)", "长相(zhǎngxiàng, appearance)", "模样(múyàng, looks)", "式样(shìyàng, style)" and "形状(xíngzhuàng, shape)".

2.2 The Corpus of the Study

The linguistic data come from the Modern Chinese Corpus of Peking University (CCL corpus) which is developed by the Center for Chinese Linguistics of Peking University in China.

2.3 The Theory Basis of the Study

The Frame Semantics was proposed by American linguist Charles J. Fillmore in the 1970s. Fillmore considered that words represented the classification of experience (categorization) and each member of the categorization depended on the activation of knowledge and experience background. In other words, the semantic frame is the situation which is activated by words [1]. As a kind of schematic scenarios, the semantic frame could be used to describe the lexical semantic. Here we take the commodity exchange as an example. The schematic scenario constituted by "buyer", "seller", "goods" and "money" provide the semantic frames by which we can describe the lexical semantic of "buy" and "sell". Therefore, we will analyze the semantic features of attribute nouns based on the Frame Semantics.

3 The Semantic Feature Analysis

We obtained 16684 records about the ten "appearance" attribute nouns from CCL corpus firstly. According to life experience, Modern Chinese Dictionary explanations (2005 edition) and the linguistic data coming from CCL corpus, we extracted the semantic frame of the "appearance" attribute nouns which consisted of entity and attribute value. The entity is the thing or the phenomenon on which the attribute depends. The attribute value indicates the quantity of attribute or how the attribute is [2].

(1) 年轻时的戴维相貌英俊 (Niánqīng shí de Dài Wéi xiàngmào yīngjùn, When Dai Wei was young, he was handsome) 。

In sentence (1), the entity of attribute noun "相貌 (xiàngmào, personal appearance)" is "戴维（Dài Wéi, Dai Wei）", and the attribute value of attribute noun "相貌 (xiàngmào, personal appearance)" is "英俊 (yīngjùn, handsome)".

Lin Xingguang pointed out that the meanings of verbs were decided on the semantic feature set of collocation constituents [3]. The elements of the semantic frame

could be described by the collocation constituents of the attribute nouns in the same sentence. Based on the semantic frame, analyzing semantic features of the attribute nouns is to analyze the semantic features of their collocation constituents.

3.1 The Semantic Feature Analysis of Entities

According to the analysis based on CCL corpus, we could conclude that there are three types of entities among the ten "appearance" attribute nouns.

The first type of entity expresses animate which means living things. "相貌 (xiàngmào, personal appearance)", "容貌 (róngmào, facial features)", "长相 (zhǎngxiàng, appearance)" and "模样 (múyàng, looks)" have this type of entity[1].

The second type of entity expresses animate partly. "外形 (wàixíng, external form)", "外貌 (wàimào, aspect)" and "外表 (wàibiǎo, likeness)" have this type of entity.

The third type of entity expresses inanimate which means non-living things. "外观 (wàiguān, facade)", "式样 (shìyàng, style)" and "形状 (xíngzhuàng, shape)" have this type of entity.

The difference between the living things and the non-living things is whether they have life. The categories of "animate" and "inanimate" are the reflection of this kind of difference in a language system. Then we need formal standards to decide whether a category belong to the "animate" or the "inanimate" in a language.

Long Tao considered that the [+animate] nouns could add individual measure words ahead, but could not add metrical measure words ahead [4]. When adding volitional verb ahead, the [+animate] nouns could collocate with modal verbs and mental verbs which could not be negative by "不（bù, no）". The [-animate] nouns may not have these three features above.

According to Long Tao's standard, we will identify the syntactic characteristics of the three types of entities above. Through collocating with mental verb "希望 (xīwàng, hope), we discover that there are many difference between these three types of entity nouns .The entity nouns of "相貌(xiàngmào, personal appearance)", "容貌 (róngmào, facial features)", "长相(zhǎngxiàng, appearance)", "模样(múyàng, looks)" could add "希望(xīwàng, hope)" behind. But at the same time, the entity nouns of "外观 (wàiguān, facade)", "式样 (shìyàng, style)" and "形状 (xíngzhuàng, shape)" could not add "希望 (xīwàng, hope)" behind. Meanwhile, there were two possibilities for the entity nouns of "外形 (wàixíng, external form)", "外表 (wàibiǎo, likeness)" and "外貌 (wàimào, aspect)".

---

[1] "容貌(róngmào, facial features)", "长相(zhǎngxiàng, appearance)", "模样(múyàng, looks)" could collocate with their entities to construct the personification in very fewer sentences, such as "我读大学时，家里的月饼已经很有模样了(Wǒ dú dàxué shí, jiā li shāolái de yuèbǐng yǐjīng hěn yǒu múyàng le, When I was at college, the moon cakes coming from home had looked very nice)".

(2) 她长相美丽 (Tā zhǎngxiàng měilì, She looks beautiful) 。

→ 她希望长相美丽 (Tā xīwàng zhǎngxiàng měilì, She hopes that her appearance is beautiful) 。

(3) 这件衣服式样简单 (Zhè jiàn yīfú shìyàng jiǎndān, This dress has a simple style) 。

→这件衣服希望式样简单 (Zhè jiàn yīfú xīwàng shìyàng jiǎndān, This dress hopes that it has a simple style) 。 ※

(4) 刘翔有俊朗的外形 (Liú Xiáng yǒu jùnlǎng de wàixíng, Liu Xiang is handsome) 。

→ 刘翔希望有俊朗的外形 (Liú Xiáng xīwàng yǒu jùnlǎng de wàixíng, Liu Xiang hopes that he is handsome) 。

(5) 这辆汽车外形美观 (Zhè liàng qìchē wàixíng měiguān, The shape of the car is beautiful) 。

→这辆汽车希望外形美观 (Zhè liàng qìchē xīwàng wàixíng měiguān, The car hopes that its shape is beautiful) 。※

"希望 (xīwàng, hope)" could enter into sentence (2) and sentence (4), while it could not enter into sentence (3) and sentence (5). These show that the entity nouns of "外形 (wàixíng, external form)", "外表 (wàibiǎo, likeness)" and "外貌 (wàimào, aspect)" could collocate with "希望 (xīwàng, hope)" partly. The syntactic characteristics are closely related to the semantic features of the “appearance “attribute nouns. According to Tao Long's standards, we obtained three semantic features. "相貌 (xiàngmào, personal appearance)", "容貌 (róngmào, facial features)", "长相 (zhǎngxiàng, appearance)" and "模样 (múyàng, looks)" have the [+animate] feature. "外观 (wàiguān, facade)", "式样 (shìyàng, style)" and "形状 (xíngzhuàng, shape)" have the [-animate] feature. "外形 (wàixíng, external form)", "外表 (wàibiǎo, likeness)" and “外貌 (wàimào, aspect)" have the [±animate] feature.

3.2 The Semantic Feature Analysis of Attribute Values

Based on analysis of the CCL corpus, we could conclude that there are three groups of attribute values in the ten "appearance" attribute nouns. From these three groups, we could obtain three semantic features.

The first semantic feature is [+trait]. This value describes the nature and the state of attribute. They are mainly expressed by some adjectives and nouns.

The typical adjectives are "美丽 (měilì, beautiful)", "笨拙(bènzhuō, clumsy)", "新颖(xīnyǐng, novel)", "豪华(háohuá, luxurious)" and "可爱(kě'ài, lovely)"[2] which could collocate with "不(bù, no)", "有点儿(yǒudiǎnr, a bit of)", "很(hěn, very)" and "

[2] "不同 (bùtóng, different)" and "相同 (xiāngtóng, same)"can be used to compare the quantity of the attribute nouns. But we don’t intend to analyze them, because they can only indicate the different semantic of attribute nouns and can not select the attribute nouns.

最(zuì, most)".Through the adverbs "有点儿 (yǒudiǎnr, a bit of)", "很 (hěn, very)" and "最 (zuì, most), these adjectives could describe different degree of the trait. The nouns could also describe the attribute. They mainly enter into the structures such as "弯弓形状 (wāngōng xíngzhuàng, the shape likes a curved bow)", "大男孩模样 (dà nánhái múyàng, a big boy's looks)" and act as the modifier of the attribute nouns. By doing so, these nouns could describe the trait with no degree. From this kind of value we could extract the [+trait] feature.

The second semantic feature is [+quantity]. People always describe the appearance with some phrases such as "一种 (yīzhǒng, a type of)", "一副 (yīfú, a pair of)", "一类 (yīlèi, a kind of)", "很多 (hěnduō, a lot of)", "繁多 (fánduō, various)", and so on. Then we could extract the feature [+quantity] from these constitutes which modify attribute nouns.

The third semantic feature is [+content]. The appearance of things is describable. For examples, in the sentence "树叶有圆形、条形等形状 (Shùyè yǒu yuánxíng, tiáoxíng děng xíngzhuàng, the leaves have many shapes, such as round, bar, etc) ", "形状"(xíngzhuàng, shape) can be defined by "圆形 (yuánxíng, round) " and "条形 (tiáoxíng, bar)". Therefore, "圆形 (yuánxíng, round) " and "条形 (tiáoxíng, bar)" are the content of "形状"(xíngzhuàng, shape). There are many special structures such as "A 的属性是 B (A de shǔxìng shì B, The attribute of A is B)" and "A 有 C、D 等属性 (A yǒu C, D děng shǔxìng, A has some attributes such as C, D, etc)" often selected by this semantic feature [5]. So we could extract the [+content] feature from the constituents of B, C and D.

According to the analysis above, we could construct a feature system of these "appearance" attribute nouns.

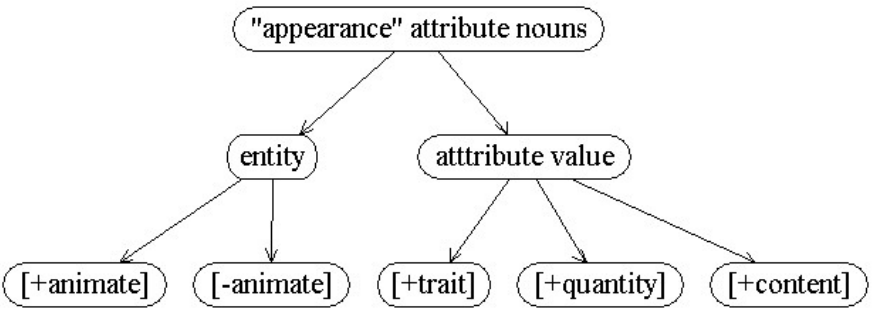


Fig. 1. The Semantic Feature System

At the same time, we could construct a semantic feature set of the "appearance" attribute nouns.

Table 1. The Semantic Feature Set

Attribute noun	The semantic feature of entity	The semantic feature of attribute value
相貌(xiàngmào, personal appearance)	[+animate]	[+trait +quantity +content]
容貌(róngmào, facial features) 长相(zhǎngxiàng, appearance) 模样(múyàng, looks)	[+animate]	[+trait +quantity +content]
外观(wàiguān, facade) 式样(shìyàng, style) 形状(xíngzhuàng, shape)	[±animate]	[+trait +quantity +content]
外形(wàixíng, external form) 外貌(wàimào, aspect) 外表(wàibiǎo, likeness)	[-animate]	[+trait +quantity +content]

4 The Semantic Feature Formalization

Many scholars have discussed the question of how to formal a language. And there are some formal grammar systems, such as Head-Driven Phrase Structure Grammar, Lexical Function Grammar, and so on. In this paper we intend to formalize the semantic features by Parse-Annotate Translate (PATR)[3]. Parse-Annotate Translate is well suitable for describing the semantic feature of words by feature structure. Based on the formalization, we will test the semantic features and the syntactic structures in Copenhagen Tree Tracer (CTT) which is developed by Matthias Trautner Kromann of Copenhagen Business School[4].

4.1 The Grammar System of Formalization

The PATR consists of the lexical rules and the syntactic rules. The lexical rules describe complex categorizations of the words and the syntactic rules describe the construction rules of the categorizations. During the formalization, we summarized the lexical rules from the semantic features, formal features, frame elements and

[3] http://www.sil.org/pcpatr/
[4] http://www.buch-kromann.dk/matthias/ctt/

syntactic constituents of the attribute nouns acted firstly. Then we summarized the syntactic structures of the attribute noun and organized the different complex categorizations based on the lexical rules. Finally, we constructed the syntactic rules.

In this paper, we propose a set of tags to express the semantic features and syntactic structures.

Table 2. The Tag Set

Tag	Meaning	Tag	Meaning
s	sentence	sx	attribute
n	noun	sxz	attribute value
v	verb	cat	formal feature
a	adjective	frame	frame element
np	noun phrase	sem	semantic feature
vp	verb phrase	zt	entity
ap	adjective phrase	xz	trait
mp	quantified phrase	nr	content
ys	animate		

Take sentence (4) as an example. We could express the semantic features of the attribute noun "外形" and construct the lexical rules and the syntactic rules.

```
% Lexical Rules
lex('刘翔(Liú Xiáng, Liu Xiang)',N):-
N>>cat===='n',
N>>frame===='zt',
N>>sem===='ys'.

lex('有(yǒu, has)',V):-
V>>cat===='v'.

lex('俊朗(jùnlǎng, handsome)',A):-
A>>cat===='a',
A>>frame===='sxz',
A>>sem===='xz'.

lex('的(de)',De):-
De>>cat===='de'.

lex('外形(wàixíng, external form)',N):-
N>>cat===='n',
N>>sem===='sx'.
```

```
% Syntactic Rules
S--->[N1,VP1]:-
S>>cat===s,
N1>>cat===n,
N1>>frame===zt,
N1>>sem===ys,
VP1>>cat===vp.

VP1--->[V,NP1]:-
VP1>>cat===vp,
V>>cat===v,
NP1>>cat===np.

NP1--->[A,De,N2]:-
NP1>>cat===np,
A>>cat===a,
A>>frame===sxz,
A>>sem===xz,
De>>cat===de,
N2>>cat===n,
N2>>sem===sx.
```

According to the lexical rules and the syntactic rules above, we could obtain a syntax tree by CTT.

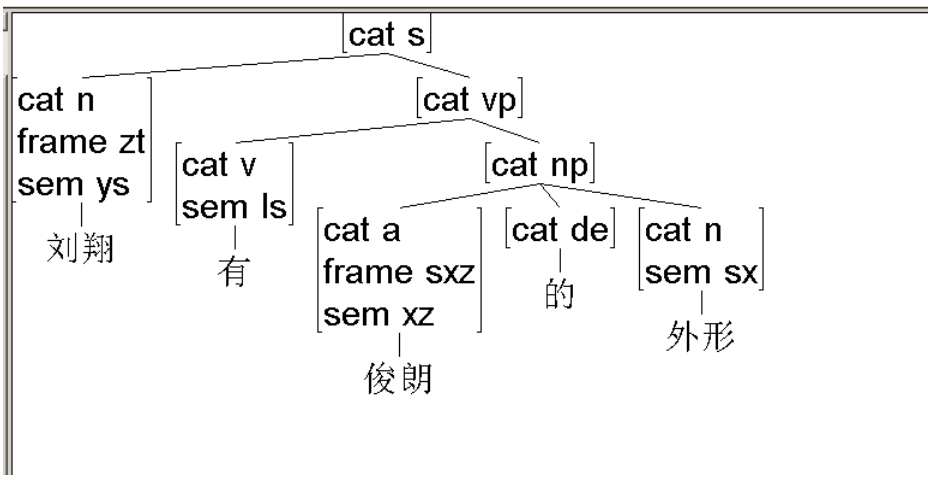


Fig. 2. The Syntax Tree of sentence (4)

"刘翔 (Liú Xiáng, Liu Xiang)" has the [+animate] feature which acts as the entity of "外形 (wàixíng, external form)". "俊朗 (jùnlǎng, handsome)" has [+trait] feature which acts as the value of "外形 (wàixíng, external form)". So "外形 (wàixíng, external form)"in sentence (4) has the [+animate] feature and the [+trait] feature.

4.2 The Meaning of Formalization

There are two meanings of formalizing the semantic features of the attribute nouns by PATR.

Firstly, the formalization could help the computer to recognize the attribute nouns.

Two important factors are involved in automatic recognizing of attribute nouns. One factor is necessary linguistic knowledge, such as syntactic knowledge and semantic knowledge, etc. The other factor is the recognizable form of this linguistic knowledge. In fact, the two factors are also involved in analyzing and formalizing the semantic features of attribute nouns. Based on the semantic frame, we could extract semantic features and necessary linguistic knowledge of attribute nouns. If the computers have the knowledge, they will understand attribute nouns automatically. In a word, what we have done above is providing some support for automatic recognition.

Secondly, the formalization could test the extracted semantic features whether reasonable or not.

All of the extracted semantic features could reflect syntax functions of attribute nouns, such as collocating with some language units, repelling some language units and the collocation patterns. Zhao Shiju considered that lexical semantic decide the property, the function, the collocation patterns and the expression forms of the language elements [6]. Xu Yanping regarded that the trend that analyzing semantic features combined with the grammar was inevitable [7]. In other words, the semantic features of the attribute nouns closely related to their syntactic characteristics. If the semantic features we extracted are unreasonable, the sentence comprehension will be affected somewhat. A good method testing the sentence comprehension is drawing a tree by CTT [8].

(6) 一个丫环模样的女孩出来了。(Yī gè yāhuán múyàng de nǚhái chūlái le, A girl liking the servant came out)

Which is the entity of the attribute noun "模样 (múyàng, looks)", "丫环 (yāhuán, servant girl)" or "女孩 (nǚhái, girl)"? It is difficult to answer. We could explain it by drawing the syntax tree of sentence (6).

The syntax tree shows that "女孩 (nǚhái, girl)" acts as the entity of "模样 (múyàng, looks)" and its attribute value is "丫环 (yāhuán, servant girl)". So the entity "女孩 (nǚhái, girl)"of "模样 (múyàng, looks)" should contain the [+animate] feature, and "丫环 (yāhuán, servant girl)" acting as the attribute value should contain the [+trait] feature. It is clear that the semantic features in the parsing tree are reasonable.

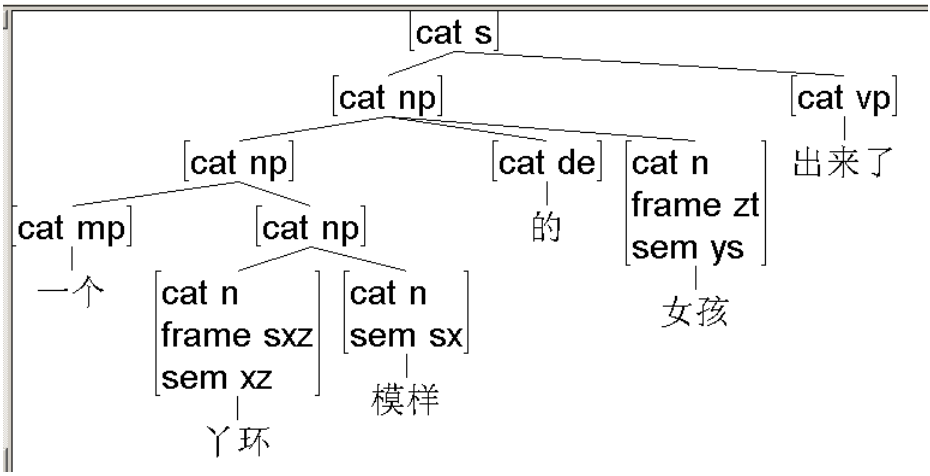


Fig. 3. The Syntax Tree of sentence (6)

5 Conclusion

According to the Frame Semantics proposed by American linguist Charles J. Fillmore, we analyzed the semantic features of the ten "appearance" attribute nouns. By Parse-Annotate Translate (PATR) and Copenhagen Tree Tracer (CTT) system, we formalized the semantic features and draw up the syntax tree of the sentences which contained attribute nouns. What we have done in this paper not only demonstrate how to extract the semantic features but also provide some support to understand the relationship between the semantic features of "appearance" attribute nouns and their syntactic functions. All these could improve the research on syntax-semantics interface directly. In addition, formalizing the semantic features by PATR could help computers to understand the attribute nouns automatically to some extent. Through analyzing and formalizing the semantic features, we could continually probe into the disambiguation of the attribute nouns and extract the information of “entity-attribute-value” from the natural language.

Acknowledgements. This work was supported by the Youth Social Science Foundation of Ministry of Education of China (Grant No. 11YJC740120), the planning project of serving enterprise of young teachers in higher school of Hubei province (Grant No. XD20100732), the Social Sciences in Hubei Province Youth Project (Grant No. 2006q117).

References

1. Charles, J.F.: Frame Semantics and the Nature of Language. Annals of the NY Academy of Sciences, 20–32 (1976)
2. Xu, Y.P., Zhang, J.C.: The Selection Differences on the Structure of “( ) + Verb + Quantifier” between “length (长度)” and “extent (长短)”. Studies in Language and Linguistics, 87–90 (2012)

3. Lin, X.G.: Lexical Semantics and Computational Linguistics. Language Publishing House, Beijing (1999)
4. Long, T.: Grammatical Form and Spatial Meaning of "Animate" and "Inanimate" Noun. Yangtze River Academic, 156–163 (2006)
5. Xu, Y.P., Zhang, J.C.: The Syntactic Structures Chosen by ［+Content］ of the Attribute Noun. Journal of Chongqing University of Posts and Telecommunications, 125–129 (2012)
6. Zhao, S.J.: Lexical Decisive Function of Meaning on Grammar. Wuhan University Journal, 173–179 (2008)
7. Xu, Y.P.: The Trend That Analyzing Semantic Features Combined with the Grammar Is Inevitable. Theory Monthly, 128–130 (2010)
8. Zhang, J.P., Feng, Z.W.: The Study on Digitized Chinese Grammar Teaching. In: Zhang, P. (ed.) Research and Application of Digitized Chinese Teaching and Learning, p. 278. Language Publishing House, Beijing (2006)

Semantic Derivation Patterns of the Chinese Character "SHENG"—A Perspective from Metaphor

Weidu Xiong[1,2] and Ling Zhao[3]

[1] School of Chinese Language and Literature, Wuhan University, Wuhan, China
[2] College of Foreign Languages, South-Central University for Nationalities, Wuhan, China
kumane@163.com
[3] School of Foreign Language and Literature, Wuhan University, Wuhan, Hubei 430072
lingzhao2006@126.com

Abstract. Polysemy is a common phenomenon in all languages. Polysemy is the capacity for a sign to have multiple meanings, but can be a "semantic primitive" and "derived meanings". The semantic primitive is not equal to core meaning, and it represents a concept. In this paper, we will analyze the Chinese character "SHENG" from the perspective of metaphor. The semantic derivation of the semantic primitive has "a Concept Derivation Layer" and "a Property Projection Layer". We propose that all the words have the same characteristic feature types with some explicit, and others implicit. Only those explicit characteristic feature types can be projected outwards, thereupon new meaning can be formed.

Keywords: metaphor, polysemy, derivation pattern, SHENG, prototype theory.

1 Introduction

Polysemy is a universal phenomenon in all languages. Researches on polysemy focus on the evolution and extension of several semantic meanings. According to Prototype Theory, one meaning occupies a prototypical core position among all the polysemous meanings [1]. Other meanings are nothing but derived meanings of the core meaning, associating with the core meaning in various dimensions. Polysemy, as a matter of fact, is the result of people's increasing demands on the richness of expression in line with the development of society [2-4]. On the other hand, the Principle of Economy decides that people tend to describe the new concepts with the existing lexicon. Therefore, driven by the mechanism of metaphor and metonymy, diverse cognitive domains project into the corresponding factors of one cognitive domain, hence forming the polysemous word [5-6].

If viewed in solitary, the senses of polysemous word are scattered and disordered individuals, which is difficult to discern the internal connections. However, taking the perspective of metaphor, we find that every polysemous word has a primitive meaning, which may not equal to certain sense of the polysemous word, together with several derived meanings. The various senses of polysemous word are evolved from

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 459–472, 2013.

the primitive meaning by means of the cognitive process of metaphor. In some sense, metaphor constitutes human's conceptual system [7-9].

This paper argues that polysemous word is composed of a "primitive meaning", which represents the core conception, and a multitude of derived meanings. Put it simply, it is "Polysemous[1] Meaning = Primitive Meaning[2]+ Derived Meanings". The concept of primitive meaning does not equal to core sense. The core sense is the most central, most frequently used, and the most eminent sense, while the primitive meaning stands for a concept that is not necessarily a sense.

The key to polysemous word is the internal connections, which is also where the fundamental reason of emergence of polysemy lies. Derivation is a cognitive process, from which we would extract the metaphorical cognitive paths. The metaphorical cognitive path is a kind of externalized and explicit idea flow with some features that are able to represent a country's socialized characteristics of thinking. Constructing the path means qualitative description of the language users' thinking patterns.

From the metaphorical point of view, this paper analyses the polysemy semantic primitive mode and derived meaning mode with the example of the polysemous Chinese character "SHENG".

2 The Analysis on Polysemous Word "SHENG" from the Metaphorical[3] Perspective

No matter what relationships among the meanings of polysemous word are, it is inevitable to bring in the concept of metaphor in the extension from primitive meaning to its derived meanings. The emergence of metaphor originates from the relevance or similarities between the source domain of primitive meaning and the target domain of the derived meanings. Lakoff stated that when the source domain and the target

---

[1] The concept of polysemous word usually confuse with the concept of homonymy. According to the definition of Hatch and Brown, homonymy refers to those variations who have the same spelling but share no common in meaning [10]. That is, homonymous words are several items that happen to have the same phonetic pronunciation or the same phonological form. Homonymy is not discussed in this paper. Instead, polysemy in strict sense is the only focus.

[2] The concept of primitive meaning stems from the author's supervisor Professor Xiao Guozheng's notion of "conceptual base", the major arguments are that several "base words" exist in the lexicon whereas that those "base words" are conceptual rather than in the form of senses. The term "primitive meaning" attempts to explore the microcosmic world of polysemy in depth, which holds that although polysemous word has a range of meanings, it also owns a part in a base position, i.e. the primitive meaning.

[3] Sheng Jiaxuan assumed that metonymy is the transition from one concept to another, which is a reanalysis; metaphor is the projection from one concept to another similar one, which is analogy [8]. No matter which process is, it is concerned with the projection process from one concept to another concept. In addition, metaphor generally embraces the metonymic factors and vice versa, which makes the two indispensible from each other. Therefore, this paper does not make strict distinction between metaphor and metonymy and metaphor is viewed as projection process in a broad sense.

domain belong to the same cognitive domain, they would show relevance by which people are able to refer to the object with one of its prominent features [6]. This process does not involve in the feature transference of objects. On the other hand, when the source domain and the target domain belong to two different cognitive domains, they would show relevance by which people are able to understand one thing through the other. This process usually involves in the projection and transference of features of objects. By whatever means, people could comprehend unfamiliar and complex concept in the aid of familiar and understood concept, thereby building the connections between the two concepts [11].

Take the polysemous word "SHENG" as an example. Modern Chinese Dictionary (the Fifth Edition) divides the senses of "SHENG" into four categories with the marks of "SHENG1" to "SHENG4" respectively. In this way, the homonymous words and polysemous words are distinguished: the four words, from "SHENG1" to "SHENG4" are in the relationship of homonymy. Except for "SHENG4", the other three words are polysemous words in the strict sense. Next, we will take "SHENG1" as an example to reveal the metaphorical categories and features of this example.

2.1 Establishment of Research Scope

The sense division of "SHENG1" in Modern Chinese Dictionary (the Fifth Edition) is as following in Table 1.

Table 1. Sense division of "SHENG1"

	POS	Definition	Example
①	V	Produce; bear; give birth to; be born	优~优育 / yōu shēng yōu yù / give a good birth and good care
②	V	Grow	~根 / shēng gēn / take root
③	N	Existence; live (in contrast with "die")	舍~忘死 / shě shēng wàng sǐ / disregard one's own life
④	N	livelihood;	谋~ / móu shēng / seek a livelihood
⑤	N	life;	丧~ / sang shēng / lose one's life
⑥	N	all one's life;	一~一世/ yì shēng yí shì / in all one's life
⑦	A	Energetic; alive; living	~物 / shēng wù / living things
⑧	V	Cause; bear; beget; create	~病 / shēng bìng / get ill ~效 / shēng xiào / go into effect 惹是~非/ rě shì shēng fēi / stir up trouble
⑨	V	Make something (firewood, coal, etc) start to burn	~火 / shēng huǒ / make a fire
⑩	N	Family name	

Among ten senses of "SHENG1", we should exclude"SHENG1⑩" because "SHENG" being used as a family name is a particular usage, which has no evident

relationship with the other senses, thus out of discussion. In the following section, the nine senses will be elaborated.

Further observation of the parts of speech of nine senses "SHENG1①"~"SHENG1⑨" showed that ①②⑧⑨ are verbs, each of which can be used as a discrete word, thus being marked as <Verb>. As for the five senses from ③~⑦, no parts of speech are identified. Among them, ③④⑤⑥ is nominal constituents that can not occur as separate words but serve as lexical morphemes. However, from the aspect of meaning, they are similar to nouns, only they can not occur independently in a sentence. For the convenience of discussion, this paper does not make strict distinction between nouns and nominal lexical morpheme but investigate from conceptual perspective. Therefore, the two kinds of constituents are both marked as <Noun>. Similarly, ⑦ is marked as <Adjective>. Next we will discuss the transition issue in respect of verb, noun, and adjectives using senses ①~⑨.

2.2 Sense Combination in View of Metaphor

2.2.1 Sense Combination in View of Structural Metaphors

"①" is most frequently used to describe the process of birth that things grow out of nothing (Example 1), "②" is mostly used to describe the course of development after birth (Example 2).

Such as:

Example 1: 一个 人 [生] 下来, 首先 受到 的 教育 来自 家庭。
yí gè rén shēng xià lái, shǒu xiān shòu dào de jiào yù lái zì jiā ting.
One people be born down come, first come in for education come from family.
When one was born, the education he or she received at the first time was from the family.

Example 2: 万 物 有 根 则 [生]。
wàn wù yǒu gēn zé shēng.
All things on earth have root will grow.
With roots, all things on earth grow.

Example 3: 我 [生] 于 一个 普通的 工人 家庭。
wǒ shēng yú yí gè pǔ tōng de gōng rén jiā ting.
I be born at one common worker family.
I was born in a family of ordinary workman.

Example 1 contains the message of "HUMAN + BIRTH", which is an apparent realization of the change "GROWING OUT OF NOTHING", and it stands for Sense①. Example 2 carries the message of "PLANT + GROWTH", which is not a process of "(0→1)" but a "(1→n)" procedure of reproduction and development, and it stands for Sense②.

However, Example 3 is slightly different. It both refers to the change of "BIRTH (0→1)" and the procedure of "GROWTH (1→n)".

Although there are some divergences between the cognitive domains of the two concepts, their conceptual structures share consistency to a high degree. They all indicate the process of "GROWTH"; only ① stands for the "GROWING OUT OF NOTHING" whereas ② stands for the "EXPANDING FROM NOTHING". Thus the compositions of the concepts nearly share one-to-one correspondence. Through verb-verb combination, we apply the concepts employed in one conceptual structure to another, hence the thought pattern of structural metaphor. Therefore, we suggest combining ① and ② and identified the concept as "a. GROWTH", and subcategorized as "a1: 0→1" and "a2: 1→n" as a continuum. This is the combination of verb and verb.

Similarly, "⑦" signifies the property of "SURVIVAL" (Example 4). "③" (Example 5), "④" (Example 6), "⑤" (Example 7) and "⑥" (Example 8) signify the results of the property in unlike dimensions. ③ is the "STATUS" of affairs of the feature, ④ the "STYLE" of the feature, ⑤ the "MATERIALIZATION" of the feature (the material carrier of "LIVING") and ⑥ the "QUANTIFICATION" of the feature (measurement of "LIVING").

Such as:

Example 4: [生] 机体的 发展 依靠 细胞 环境的 不同。

shēng jī tǐ de fā zhǎn yī kào xì bāo huán jìng de bù tong.

Life organism develop rely on cell environment different.

The development of living thing/organism depends on the conditions of cell conditions.

Example 5: 儒家 宣扬 "死 [生] 有 命，富 贵 在 天"的 唯心 史 观。

rú jiā xuān yang "sǐ shēng yǒu mìng, fù guì zài tiān" de wéi xīn shǐ guān.

Confucian publicize "death life have life, wealth rank at heaven" idealistic history viewpoint.

The Confucius School advocates the historical idealism that "Death and life have determined appointments, riches and honors depend upon heaven".

Example 6: 他一个公子哥儿，谋 [生] 的手段是星点儿也不会的。

tā yí gè gōng zi gē er, móu shēng de shǒu duàn shì xīng diǎn er yě bú huì de.

He a son of feudal prince or high official brother, seek a livelihood method be star point too not can.

He is the son of high official, knowing nothing of seeking livelihood.

Example 7: 所不同的，基督 因 要 救人 而 受苦 舍 [生]，释迦 则 因 要 济人 而 留 [生] 受苦。

suǒ bù tong de, jī dū yīn yào jiù rén ér shòu kǔ shě shēng, shì jiā zé yīn yào jì rén ér liú shēng shòu kǔ.

Difference, Christ because will save people suffer sacrifice, Shakya because will save people save one's life suffer.

What the difference is that the Christ sacrificed his life and suffered for saving people, and Sakyamuni stayed alive and suffered for relieving people.

Example 8: 在 她 脸上 可以 看出 她 对 人 [生] 充满 希望。

zài tā liǎn shàng kě yǐ kàn chù tā duì rén shēng chōng mǎn xī wàng.

At she face can see she to life be full of hope.

In her face showed that she was full of hope for the whole life.

In Example 4, “shēng jī tǐ (living thing/organism)” means “LIVING + BODY”, and “SHENG” is the property of "SURVIVAL" here, hence Sense⑦.

Different from Example 4, “SHENG” in Example 5, which is in contrast with the concept of “DEATH”, does not indicate “SURVIVAL” but the results of the property, hence Sense③.

Example 6 “móu shēng = SEEK + WAY + LIVE”, in which “móu = SEEK”, “shēng = A WAY OF SURVIVAL”, hence Sense④.

Although in Example 7 “shě shēng (sacrifice one’s life)” and “liú shēng (stay alive)” are in opposite, but they have the same morpheme “shēng”. “shě shēng = SACRIFICE + LIFE” while “liú shēng = SAVE + LIFE”, and "shēng (SHENG)" expresses the same meaning of “LIFE”, hence Sense⑤. “Life” is the emblematic materialization of “SURVIVAL” and in effect the material carrier.

“rén shēng” in Example 8 means “HUMAN + A WHOLE LIFE”. “shēng (SHENG)” has been a unit that could be used to measure life, hence Sense⑥. This kind of unit is used for quantified the feature of “SURVIVAL”.

This way, we suggest combining ③~⑦ and defined as “b. SURVIVAL”. This category concerns about the combination of adjective and noun, and can be subcategorized as “b1. PROPERTY”, “b2. RESULT”, while b1 is corresponding to ⑦, and b2 is corresponding to ③ - ⑥. Under b2, we divide it into the noun-noun combination of “b21. STATUS”, “b22. STYLE”, “b23. MATERIALIZATION” and “b24. QUANTIFICATION”.

2.2.2 Sense Combination in View of Ontological Metaphors

Ontological metaphors are used when expressing abstract and shapeless concepts, for instance thoughts, feelings, and psychological activities, people tend to employ the existed and understood substances or the concepts that describe substances to explain and elaborate the abstract ones.

Sense ⑨ is originally used to depict concrete things. But when it is used to collocate the abstract concepts, for instance, “~病 /shēng bìng / get ill fall sick (⑧)” , it make the abstract and obscure concept concrete for us to quantify and recognize.

For example:

Example 9: 徐 区长 急得 心 [生] 一 计。

xú qū zhǎng jí dé xīn shēng yí jì.

Xu district head worry heart happen one stratagem.

Governor Xu has quick wits in an emergency.

The key information in Example 9 is “have quick wits (xīn shēng yí jì)”, which means devise a strategy in mind (“HIT UPON + STRATEGY + IN HEART”). “STRATEGY” is an abstract concept, and “devising strategy or having quick wits” is the coming up of the abstract “STRATEGY” due to the outside stimulus.

Example 10: 他 从容地 搬 回 一 捆 木柴， [生] 上 火，默默地 坐 待 最后 的 时刻。

tā cóng róng de bān huí yì kǔn mù chái, shēng shàng huǒ, mò mò de zuò dài zuì hòu de shí kè.

He calm move back one bundle firewood, fire up fire, silently sit wait last moment.

He carried back a bunch of firewood calmly, lit the fire, and waited in quiet for the last moment.

Example 10 contains the information of “he light (shēng) the fire”, which means to make a fire, i.e. turning the states of “NO FIRE” to “HAVE A FIRE” by the operation of man.

The root of the two is the “HAPPEN” of incentive events only that the target of ⑧ is concrete substance and the target of ⑨ an abstract concept. We propose combining the two and defined as “c. INDUCE” and subcategorized as “c1. SPECIFIC” and “c2. ABSTRACT”. This is about the verb-verb combination.

Certainly, we also notice that some forms of expression have been fossilized, for example, “"~火 / shēng huǒ /light a fire ” and “生字/ shēng zì / vocabulary ”, which are dead metaphor that we are not aware of the contained metaphor. This, however, proves that metaphor what we live by. When we come across something unknown, people project the mastered structures of source domain to the structures of the new target domain.

2.3 Extraction of Primitive Meaning

With the above combined concepts a~c as object, we would extract the primitive meanings. The common characteristic of “a. GROWTH”, “b. SURVIVAL” and “c. INDUCE” is that they all represent “A CHANGE GROWING OUT OF NOTHING”. Specifically, “a” is the process of “CHANGE”, “b” is the results of “CHANGE”, while “c” has the meaning of “a” and “CAUSE”, a further step of “a”, and “c” can be seen as the derivation of “a”.

Thus we extract the primitive meaning ⊙ as “A CHANGE OUT OF NOTHING” and its derivations of “a” and “b”; “c” is not directly correlate with the primitive meaning but a result of further derivation of “a”.

2.4 Analysis of Derivation of Primitive Meaning

From the above process of extracting the primitive meaning, we found that the primitive meaning of polysemous word “SHENG1” is a concept but not a sense, and an exclusive one. The derived meaning from the primitive “a”, “b”, and “c” also appear as concept. They are not in the same level but form a hierarchy. In the bottom is “a” and “b", being the products of proceduring and conceptualization of primitive meaning respectively. Under a there is the second layer “c”, the inferior layer of “a”, and does not directly correlate with “b”. Moreover, “a”, “b” and “c” form together as the Concept Derivation Layer. Under “a”, “b” and “c”, there exists the Property Projection Layer. For example, according to the conceptual property, there are subcategories

of “a1”, “a2”, and “a1” and “a2” project into the terminal forming Sense① and Sense ②. By the same token, the two subcategories of “b”, “b1” project to the terminal, forming Sense⑦, while b2 further divides into “b21”, “b22”, “b23” and “b24”, corresponding with Senses ③~⑥. Concept “c” is further categorized as “c1” and “c2”, which project to the terminal as Sense⑧ and Sense⑨. The dotted lines in the Concept Derivation Layer stand for the first layer, while the full lines stand for the second layer. In the Property Projection Layer, dotted lines stand for the first layer and the full lines stand for the second layer. If there is the third layer or the fourth layer, the concentric dotted layer is added in the interior of the layer. Below is Figure 1:

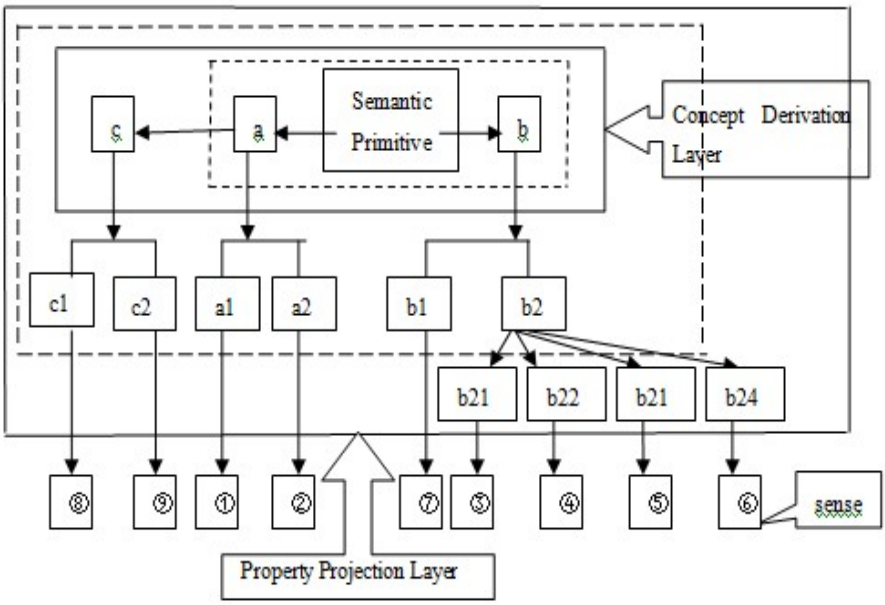


Fig. 1. Primitive meaning derivation process

3 Empirical Analysis Based on CCL Corpus

In this section we will verify the above conclusions of “SHENG1” by means of CCL Corpus (Center for Chinese Linguistics PKU). This paper makes queries in the newspaper data in CCL Corpus with the keyword of “SHENG”. Due to the limit of time and space, we select the first 331sentences for analysis.

In these 331 sentences, excluding those cases not belonging to “SHENG1”, like "（活）生生 / huó shēng shēng / actual", "学生 / xué shēng / students ", "先生 / xiān sheng / sir "and" 花生 / huā shēng / peanuts", there are 247 sentences left to be discussed. Among them, the statistics of the frequency of the terms are as following in Table 2.

In 11 cases labeled with question mark, “SHENG” has more than one meaning sense, and they all show the form of a monosyllabic verb "SHENG". Among them, there are 5 sentences meaning “BORN”, and 2 sentences meaning "TO GIVE BIRTH" (E.g. 3; [生]得漂亮可人/ shēng dé piào liàng kě rén / Born beautiful and pleasant "). They belong to the sense ①. There are also 4 sentences meaning "produce" (E.g. “钱能[生]钱/ qián néng shēng qián / money draws money”, “~现象宛如一阵暴雨中的冰雹或隐或现、或[生]或化/ xiàn xiàng wǎn rú yí zhèn bào yǔ zhōng de bīng báo huò yǐn huò xiàn, huò shēng huò huà / Some phenomenon like the hail in a burst of heavy rain, or hidden or appear”, “心[生]不满/ xīn shēng bù mǎn / Disgruntled”, “哪儿[生]出来的自费歌手/ nǎ er shēng chū lái de zì fèi gē shǒu / Where does the singer at their own expense come from.”). They belong the sense ⑧. Combing with other columns, the total of sense ① has 24 sentences, the total of sense ② has 2 sentences, sense ③ has 78 sentences, sense ④ has 25 sentences, the sense ⑤ has 10 sentences, sense ⑥ has 5 sentences, sense ⑦ has 2 sentences, sense ⑧ has 77 sentences, and sense ⑨ has zero.

Table 2. Examples drawn from CCL Corpus of “SHENG1”

Lexical item	生活 /shēnghuó /life	生产 /shēngchǎn /produce	发生 /fāshēng /happen	生意 /shēngyi /business	产生 /chǎnshēng /produce
Frequency	61	28	28	23	11
Meaning sense	③	⑧	⑧	④	⑧
Lexical item	人生 /rénshēng /life	生 /shēng /born	生存 /shēngcún /survive	生财 /shēngcái /making money	生命 /shēngming /life
Frequency	11	11	6	6	5
Meaning sense	③	?	③	⑧	⑤
Lexical item	卫生（间） /wèishēng(jiān) /bathroom	一/半生 /yì/bànshēng /lifetime/half a lifetime	新生 /xīnshēng /new born	诞生 /dànshēng /born	生日 /shēngrì /birthday
Frequency	5	5	4	4	3
Meaning sense	⑤	⑥	①	①	①
Lexical item	出生 /chūshēng /born	独生 /dúshēng /only child	生机 /shēngjī /vitality	~为生 /wéishēng /live for	生发精 /shēngfājīng /hair tonic
Frequency	3	3	2	2	2
Meaning sense	①	①	⑦	④	②
	Others			one for each (total: 24)	

(1) Category “a”.

Sense ① includes the examples of “新生/ xīn shēng / newborn, 诞生/ dàn shēng/ to give birth, 生日/ shēng rì / birthday, 出生/ chū shēng / born, 独生/ dú shēng / only child,

生/ shēng / born", describing the process of person growing out of nothing. According to Modern Chinese Dictionary (the Fifth Edition), their definitions are as follows:

[新生/ xīn shēng /newborn]: ① newly born; ② new life.

[诞生/ dàn shēng/ be born]: give birth (to a newborn)

[生日/ shēng rì / Birthday]: an anniversary of the day on which a person was born (or the celebration of it).

[出生/ chū shēng /be born]: (a person) come into existence through birth.

[独生/ dú shēng / Only Child]: the only son / daughter.

[生/ shēng]: there are 11 examples containing monosyllabic verb of "SHENG", with five sentences meaning "BE BORN", two sentences meaning "TO GIVE BIRTH". All these seven sentences are classified the sense ①.

All of the examples are related to the change of "GROWING OUT OF NOTHING" about "HUMAN".

Sense ② includes the example of "生发精/ shēng fà jīng / hair tonic", describing the process of non-biological growing out of nothing. It is worth mentioning that two sentences with sense ① can be considered either as "GROWTH (0→1)" or "GROWTH (1→ n)". This also proves that sense ① and sense ② being indivisibility.

These two categories contain 247 sentences accounting for 10.5% of all.

(2) Category "b".

Sense ③ includes three categories of "生活/ shēng huó / life, 生存/ shēng cún / survival, 人生/ rén shēng / life ", meaning the status of "SURVIVAL".

[生活/ shēng huó / life]: ① the course of existence of an individual; the actions and events that occur in living; ② the experience of living; the course of human events and activities; ③ an account of the series of events making up a person's life.

[生存/ shēng cún / survival]: a state of surviving.

[人生/ rén shēng /life]: the period from the present until death.

Through the definitions, we can see that "Life" being inseparable with "SURVIVAL", and all of them are a property of "SURVIVAL".

Sense ④ includes two categories of " 生意/ shēng yì / business, ~为生/ wéi shēng / ~live for ", describing the style of "SURVIVAL".

[生意/ shēng yì / Business]: a commercial or industrial enterprise and the people who operate it.

[~为生/ wéi shēng / ~live for]: (a way) to make a living.

"Business" can also be described as "WAY + COMMERCIAL + (MAKE A LIVING)". Both of them having the meaning of "MAKE A LIVING", can be classified to the sense ④, the way of "SURVIVAL".

Sense ⑤ includes two categories of "生命/ shēng mìng / life, 卫生（间）/ wèi shēng (jiān)/ bathroom ", indicating the material carriers of the "存活/ survival".

[生命/ shēng mìng / Life]: the organic phenomenon that distinguishes living organisms from nonliving ones.

[卫生/ wèi shēng / hygiene]: the science concerned with the prevention of illness and maintenance of health.

"卫生/hygiene" can be interpreted as "GUARD + Life", unified with the word "LIFE". Since "LIFE" is a form of existence, both of them being interpreted as the material carrier of “SURVIVAL”.

Sense ⑥ includes “一/半生/ yì/bàn shēng / lifetime/ half a lifetime”, and is a measurement of quantity of the "SURVIVAL".

[一生/ yì shēng / a lifetime]：the period between birth and the present time.

Sense ⑦ includes the "生机 /shēng jī /vitality ", meaning a quality of "SURVIVAL". They all belong to the group of "SURVIVAL". These categories account for 48.6%.

[生机/ shēng jī / vitality]：①a chance of survival; ②an energetic style.

(3) Category “c”.

Sense ⑧ includes five categories of "生产/ shēng chǎn / produce, 发生/ fā shēng / happen, 产生/ chǎn shēng / produce, 生/ shēng / born, 生财/ shēng cái / making money ", and can be used to describe "a process of something growing out of nothing by external force (induced) ".

[生产/ shēng chǎn / produce]：create or manufacture a man-made product.

[发生/ fā shēng / happen]：happen, occur, or be the case in the course of events or by chance.

[产生/ chǎn shēng / produce]：cause to occur or exist.

[生/ shēng] ：there are 11 examples containing monosyllabic verb of "SHENG", with four sentences meaning "PRODUCE". All of them are categorized as the sense ⑧.

[生财/ shēng cái / making money]：increase wealth.

"Produce" is an activity by using "Tool"; "making money" is to use some methods to "increase wealth"; and "happen" is not a natural transformation, but a result of external force.

These categories account for 31.2% of all valid examples.

Obviously, the first category "a. a process of something growing out of nothing” and the third category "c. a process of something growing out of nothing by external force (induced)" are associated closely, and the former is the root of the latter. They can be looked as a derivative relationship of “a → c”. The second category “b. the status of survival” is also related with “a”. No matter what kind of the relationship is, their cores contain "a change of growing out of nothing", which is the so-called "semantic primitive". Through the validation based on the above corpus, it provides an evidence for the derivation pattern of "SHENG1".

4 Semantic Primitive Mode and Derived Meaning Mode from Polysemous Word “SHENG1”

4.1 Semantic Primitive Mode and Derived Meaning Mode

From the analysis of the primitive meaning derivation of “SHENG1”, we summarize the Polysemy Semantic Primitive Mode and Derived Meaning Mode as follows:

(1) The primitive meaning is one and only one

There is one and only primitive meaning that is able to assemble the senses of polysemous words, and the senses are products of primitive meaning derivation.

(2) Primitive meaning derivation firstly occurs in the Concept Derivation Layer

The primitive meaning firstly derived in the Concept Derivation Layer, thus forming the first layer of derived meaning "a" and "b"; a further derived in the Concept Derivation Layer, thus forming the second layer of derived meaning "c". The supposed formula "Polysemous Meaning = Primitive Meaning + Derived Meanings" means "a", "b" and "c" in the Concept Derivation Layer. If necessary, more derivation will occur in this layer or in more layers.

(3) The primitive meaning derivation secondly occurs in the Property Projection Layer

After forming the derived meaning "a", "b" and "c", the polysemous word continued to derive in the Property Projection Layer. In the first layer of the Property Projection Layer, according to its property, "a" is further divided as "a1" and "a2"; "b" is further divided as "b1" and "b2"; "c" is further divided into "c1" and "c2". Later, "b2" proceeded to divide in the second layer of the Property Projection Layer, forming "b21", "b22", "b23" and "b24". If necessary, more derivation will occur in this layer or in more layers.

(4) Projection into the terminals

In the final stage, all the divided items projected to correspond to the senses in the dictionary text.

"SHENG2" "SHENG3" and "SHENG4" share the same primitive derivative structures. However, if "SHENG1" to "SHENG4" is put together, we come across the issue of primitive meaning transference[4], which will be discussed in later papers. Project outward to the terminals

4.2 The Incentive Mechanism of Semantic Primitive Mode and Derived Meaning Mode

Primitive meaning derivation does not occur randomly. Take "SHENG1" as an example. Why does the primitive meaning "the change from nothing to existence" can derive "a", "b" and "c"? What is the incentive mechanism?

The key to the questions is the property of primitive meaning. The primitive meaning should be polygon, each side of it representing a kind of property, some explicit but some implicit. The explicit property can be derived to form derived meaning through metaphor; the derived meaning further be derived and form the senses at the terminal. Instead the implicit property is hidden and obscured. Nevertheless, it will not remain implicit all the time. Under certain conditions, it will be transformed into

[4] Primitive meaning transference is the suggestion of the author's supervisor Professor Xiao Guozheng, hereby the author extending the thanks.

explicit property and be derived to form new derived meanings (as is shown in Figure 2). This is the root of sense increase of polysemous word.

The characteristic types of primitive meaning of all polysemous words are consistent. The difference lies in the properties of unlike polysemous words. Specially speaking, the characteristic types include time, space, substance, action, and property. Take "SHENG1" as an example. The derived meaning "a" is the realization of the feature of "ACTIONS"; the derived meaning is the realization of the feature of "CHARACTERS". The first layer of the Concept Derivation Layer is formed. The five features have other hypostasis attributes and their realization is the internal cause of the further derivation of the Concept Derivation Layer. Specifically, the meaning "c" derived from "a" has the feature of "CHANGE", which is a sub-feature of "ACTIONS".

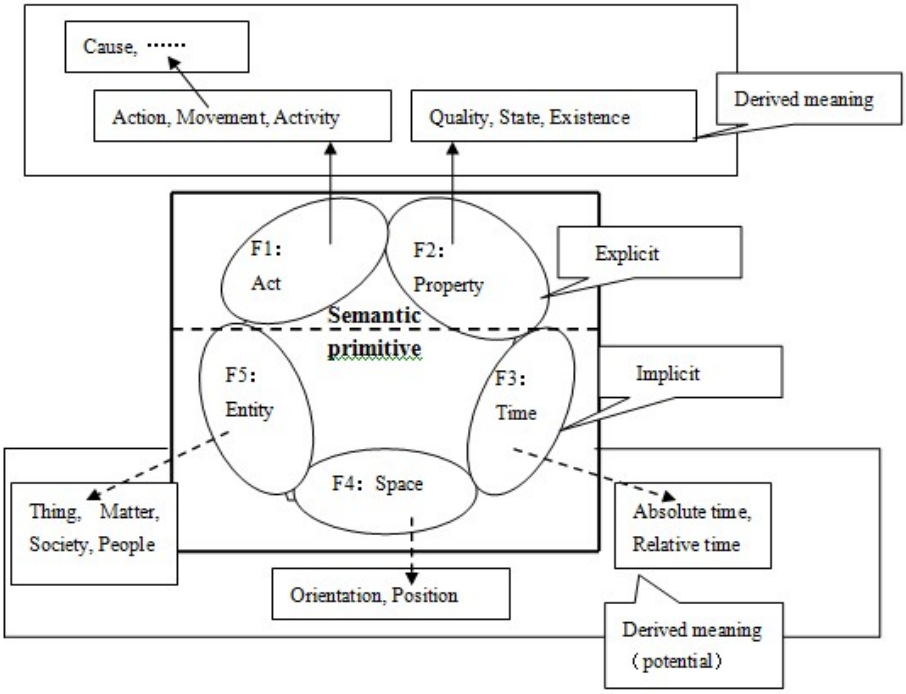


Fig. 2. The incentive mechanism of "SHENG1"

5 Conclusion

From the above discussion we have drawn the following conclusions:

1. Polysemous word has only one primitive meaning which does not equal to core meaning but is a concept;

2. The characteristic types of primitive meaning of all polysemous words are uniform. Some properties are explicit while some properties are implicit;

3. Derivation is aided by metaphorical thinking, which makes the explicit properties of the primitive meaning externalized.

Acknowledgements. This article is sponsored by "the Fundamental Research Funds for the Central Universities", South-Central University for Nationalities, "Semantic Derivation Patterns of the Chinese and Japanese Polysemous Word –A Perspective from Metaphor (CSQ12031)".

References

1. Fillmore, C.J.: Frame Semantics: Linguistics in the Morning Calm, pp. 112–137. Hanshifl, Seoul (1982)
2. Zhang, L.: On the Relationships of Polysemous Word Meanings in Modern Chinese. Journal of Hebei University (Philosophy and Social Science) 22, 62–67 (1997)
3. Ge, B.Y.: Modern Chinese lexicology, pp. 582–621. Shandong Priovence Renmin Press, Jinan (2001)
4. Zhao, K.Q.: Ancient Chinese lexicology, pp. 82–127. The Commercial Press, Beijing (2005)
5. Jiang, S.Y.: Outline of Ancient Chinese Vocabulary, pp. 56–126. Peking University Press, Beijing (2005)
6. Lakoff, G.: Women, Fire and Dangerous Things: What Categories Reveal about the Mind, pp. 58–64. The University of Chicago Press, Chicago (1987)
7. Wang, W.B.: The Generation and Evolution of Metaphorical Meaning. Foreign Languages and Their Teaching 4, 13–17 (2007)
8. Shen, J.X.: The Function of the Modern Chinese Grammar, Pragmatics, Cognitive Research, pp. 398–419. The Commercial Press, Beijing (2004)
9. Marina, R.: The Extent of the Literal Metaphor, Polysemy and Theories of Concepts. Beijing University Press, Beijing (2004)
10. Hatch, E., Brown, C.: Vocabulary, Semantics and Language Education, p. 49. Cambridge University Press, Cambridge (1995)
11. Taylor: Linguistic Categorization: Prototypes in Linguistic Theory. Foreign Language Teaching and Research Press, Beijing (2001)

Study of Semantic Features of Dimensional Adjective *Cu* 'Thick' in Mandarin Chinese

Ying Wu

School of Foreign Studies, Hunan University of Science and Technology, Xiangtan 411201, China
wuyingywu@foxmail.com

Abstract. *Cu* 'thick' is an adjective which is used to describe an object's spatial dimension of thickness. *Cu* 'thick' used to describe cylindrical objects shares the same sense in essence with *Cu* 'thick' used to describe granular objects, the latter is the special case of the former. *Cu* 'thick' used to describe linear objects both refers to the diameter of the cross-section of a cylindrical object and refers to one minimal dimension of a flat objet which corresponds to *hou* 'thick , therefore, it should be comprehended as the combination of this two kinds of senses. The semantic features of *Cu* 'thick' are as follows: the minimal dimension(s), dependency, implicating qualities of weight or strength.

Keywords: dimensional adjective, *Cu* 'thick', semantic features.

1 Introduction

The senses of *Cu* 'thick' explained in <Modern Chinese Dictionary> are as follows:

① (a cylindrical object) the cross-section is large: *cu sha* 'a thick yarn', *zhe ke shu hen cu* 'this tree is very thick.'

② (a linear object) the distance between two long sides is not nearer: *cu xian tiao* 'a thick line', *cu mei da yan* 'thick eyebrows and big eyes'

③ a granule is big: *cu sha* 'thick sands'

From the senses explained in dictionary, we find out that *Cu* 'thick' has three different meanings: Firstly, the adjective describes the cross-section of cylindrical objects and it refers to two dimensions, such as *cu qian bi* 'a thick pencil'. Secondly, the adjective describes the distance between two long sides of linear objects and it refers to a dimension, as is the case for *cu zi* 'thick characters'. Thirdly, the adjective describes the overall shape of granular objects and it refers to three or multiple dimensions, such as *cu sha* 'thick sands'. The dictionary gives a thorough and clear account of the usages of *Cu* 'thick', but a fuzzy and obscure explanation of its meanings, especially, the explanation of the meaning of *Cu* 'thick' in *cu xian tiao* 'a thick line', *cu mei mao* 'thick eyebrows', which makes people difficult to understand, because the meaning explained in the dictionary is not correspondent with the cognition that people make about *cu xian tiao* 'thick line', *cu mei mao* 'thick eyebrows' in daily life. Through the in-depth study of *Cu* 'thick' in Chinese, we believe that the important semantic features of *Cu* 'thick' are as follows: the minimal dimension(s), dependency, implicating qualities such as weight or strength.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 473–482, 2013.

2 Semantic Features

2.1 The Minimal Dimension(s)

According to earlier researches on *thick* in some languages in the West, German *dick* 'thick' can either refer to one minimal dimension, such as the minimal dimension of a flat board, (which corresponds to Chinese *hou* 'thick'), or it can refer to two minimal dimensions, such as the two smaller dimensions of a cigarette (which corresponds to Chinese *cu* 'thick'). Moreover, it can refer to three approximately equal dimensions, such as all three dimensions of a ball.[1] French *épais* 'thick' can refer to one minimal dimension, such as for flat, two-dimensional objects, *gros* 'thick' can refer to zero dimension, such as for dots, and it can either refer to one minimal dimension of a linear object, or can refer to two smaller dimensions of a cylindrical object.[2] Swedish *tjock* 'thick' either refers to the two minimal dimensions of cylindrical objects, or it refers to the minimal dimension of flat objects, in some cases, *tjock* 'thick' may refer to three approximately equal dimensions, too.[3]

In Chinese, *Cu* 'thick' can either refer to one minimal dimension and describe linear objects, such as *zi mu* 'letters', *xian tiao* 'lines', *mei mao* 'eyebrows', or refer to two minimal dimensions and describe cylindrical objects, such as *shu zhi* 'branch', *mu gun* 'stick', *gan zi* 'pole', Moreover, it can refer to three or multiple minimal dimensions and describe granular objects, such as *sha zi* 'sands', *yan* 'salts', *mian fen* 'flour'. The three or multiple minimal dimensions of granular objects are approximately equal, whose differences can be ignored. These three equal dimensions are smaller in comparison to those big objects with three equal dimensions, such as ball, apple. Granular objects, like sands, can be idealized to zero-dimensional dots that doesn't occupy any space, that is, they are zero-dimensional objects, but sphere-shaped objects, like ball, is three-dimensional objects. The dimension(s) referred to as *Cu* 'thick' are the smallest one(s) in all dimensions of an object, such as the cross-section of the cylindrical objects, the thickness of the geometrical linear objects and the overall shape of the granular objects.

The Cylindrical Objects. Usually, *Cu* 'thick' is used to describe the cross-section of cylindrical objects, we regard the cross-section of the cylindrical objects as a rounded surface which consists of two equal dimensions. Correspondingly, in mathematics, we can draw an incircle or a circumcircle with a square. However, some scholars claim that the cross-section of the cylindrical objects refers to one dimension. Lyons states that for unoriented three-dimensional entities, the maximal extension of the object is identified as its *length*, if the two other dimensions of the object are negligible in comparison to its *length*, these two dimensions will be collapse into the single dimension of thickness, such as *a long thick pole*.[4] In fact, whether the cross-section of cylindrical objects is considered as two dimensions or one dimension is the same matter in essence, the cross-section still refers to a rounded surface, Lyons does nothing but amalgamates two dimensions into one dimension. However, the openings of some objects are also rounded surfaces like the cross-section of the cylindrical objects, such as *wan kou* 'the opening of a bowl', *dong kou* 'the opening of a hole', *guan kou* 'the opening of a pipe', these objects are described as *kuan* 'wide', rather than

described as *Cu* 'thick', *kuan* 'wide' refers to the diameter of the opening, rather than to the area of the rounded surface of the opening. Try to compare the phrase of *wan kou kuan* 'as wide as the opening of a bowl' with the phrase of *wan kou cu* 'as thick as the opening of a bowl', the senses of this two phrases are different, the former refers to the diameter of *wan kou* 'the opening of a bowl', the latter refers to the area of *wan kou* 'the opening of a bowl'.

The use of *Cu* 'thick' to describe objects with a cylindrical shape is the most typical, which almost cover two-thirds instances in the corpus. As long as the shape of an object is cylindrical, it can be described as *Cu* 'thick', and no matter whether a cylindrical object is soft or stiff, hollow or solid, such as *shu gan* 'trunk', *qian bi* 'pencil', *teng* 'rattan', *tie si* 'iron wire', *tou fa* 'hair', *sheng zi* 'rope', *shui guan* 'water pipe', *xue guan* 'blood vessel', *mao kong* 'trichopore' etc. In (1)-(3), three such examples are shown:

(1) Ta shou zhu yi gen *cu cu de mu gun*, liang jian bu duan de song dong, hao xiang hu xi hen chi li.

'His hands hold a very thick wooden stick and two shoulders continuously stirred up, as if he breathed very difficultly.'

(2) Ta bu you de da liang zhe ta de ce mian, da liang zhe ta *cu* ying de *tou fa* he yan jing.

'She could not help looking at his side and looked at his thick, hard hairs and eyes.'

(3) Ta fa xian wei sheng jian li na xie *cu* de xi de *guan dao*, quan zai bian cheng shu gan he teng tiao.

'He found out that those thick or thin pipes in toilet are all being changed into trunks and rattans.'

Cu 'thick' can be used about the human's body parts and animals' body parts that can easily be idealized to cylinders. In (4)-(6), the 3 instances are given:

(4) Ta shen chu xiang wo xiao tui na yang *cu* de *ge bo*, xiang lai jiu wo de tou fa.

'He reached out an arm that is as thick as my leg, and wanted to catch my hairs.'

(5) Ya Li Shan Da gen ta de mei mei yi yang gao, kuan jian pang, *cu bo zi*, tu nao dai, yi zui jia ya.

'Alexander is as tall as his sister, with wide shoulders, a thick neck, a bald head and mouthed-dentures.'

(6) Zhe shi kong long de yi jie *yao zhui*, you ba jiu cun *cu* xi, yi chi wu gao.

'This is a section of the dinosaur's lumbar spine, with 8-9 inch thick, 1.5 feet high.'

In (4)-(5), *ge bo* 'arm' and *bo zi* 'neck' are the human's body parts with a cylindrical shape. In addition, Some other human's body parts can be described as *Cu* 'thick', such as *tui* 'legs', *shou zhi* 'fingers', *jiao zhi* 'toes', *yao* 'waist', *bi zi* 'nose', *jin mai* 'vein'. In Chinese, human's body parts and animals' body parts can be regarded as *Cu* 'thick', but human's body and animal's body can not be like this, for example, we can't speak of *hen cu de nan ren* 'a very fat man', *hen cu de zhu* 'a very fat pig', if we want to express the same concept, we describe human's body with *pang*

'fat', and describe animal's body with *fei* 'fat', such as *hen pang de nan ren* 'a very fat man', *hen fei de zhu* 'a very fat pig'. However, if a dimension of some animals stands out very remarkably, that is, these animals can be considered as one-dimensional objects, or typical cylindrical objects, it can be described as *Cu* 'thick', such as *she* 'snakes', *qiu yin* 'earthworms', *shan yu* 'eels', *chong* 'worms'. In (7), one such example is shown:

(7) Qian bi na me *cu*, huo chai gan na me chang de qing se da *rou chong*, yin cang zai bao gu miao de ye bei mian.

'A green worm that is as thick as a pencil and is as long as a match stem hides on the back of a maize seedlings' leaf.'

Some other objects are cylinders which are made up of some substance temporarily, such as liquid, light, smoke. In (8)-(10), 3 instances are given:

(8) Kong jian yi wei *cu* er mi de *yu tiao* zhan you le, tiao xi jian hai mi man zhe shui hua.

'Space is already occupied by thick and dense rain-strips and gaps between strips pervade spray.'

(9) Wu shu dao you *cu* you da de qi cai *guang zhu* tong guan qing tian, cheng xian chu yi ge shuo da wu peng, ban lan wu bi de shan xing.

'Countless streaks of thick and large colorful light-crosses go through the sky, presenting a huge and gorgeous sector.'

(10) Ta zai yi ge pu zi gen qian ting zhu, yong jiao ti le ti pu men, han le yi sheng shen me, zui li pen chu le *cu cu de* yi dao *bai qi*.

'He stopped at a shop, ticked the shop's door, cried something and spewed a streak of white gas from his mouth.'

Cu 'thick' is usually used to describe the cylindrical objects, but in some special cases, it can be used to describe some square strip-shaped objects that is some filiform food, such as *rou si* 'shredded meat', *huang gua si* 'cucumber stick', *luo bu si* 'radish strip'. Here, *Cu* 'thick' refers to a square cross-section and this cross-section must be very small, if the square cross-section of some objects is larger, they are not described as *Cu* 'thick'.

The Linear Objects. *Cu* 'thick' can describe some liner objects which are comprehended as one-dimensional objects and is approximated to a geometric line, such as *zi* 'characters', *bi hua* 'strokes', *jian tou* 'arrows'. These objects are symbols written with pen. In (11)-(12), 2 instances are shown:

(11) Zhang shun tui kai men jin lai, shou li na zhe ge bai zhi feng, shang mian hua zhe *ji cu* de *lan zi*.

'Zhang Shun pushed the door and came, holding a white paper seal painted the extremely thick blue characters on it.'

(12) "Zhao dai suo li de wan hui zhao pai shang de *jian tou* zhe me *cu*." Liu Shu You bi hua wan kou da xiao.

'"The arrows on the sign of the party in hostel are very thick like this", Liu Shuyou gestured as the size of the opening of a bowl.'

Mei mao 'eyebrows' can also be comprehended as one-dimensional lines. One instance is given in (13).

(13) San tiao cu zhou wen ke zai kai lang de qian e shang, *cu mei* xia de da yan jing ye you yu zhou wen de ya po er xian de xiao xie.

'Three thick wrinkles are graved on cheerful forehead and the big eyes under thick eyebrows appear a little small because of the oppression of the wrinkles.'

Vogel states that collocation like *thick eyebrows* is problematic special case. They are extended in shape. They are probably perceived as three-dimensional, either as cylinders or as cuboids. The meaning becomes more complicated because eyebrows consist of hair. This means that an element of density may be given priority in the interpretation process.[3] Vogel doesn't give us any explanation about why *thick eyebrows* indicate great extension and dense hair. We believe that the linear objects described as *Cu* 'thick' should be regarded as three-dimensional, both as cylindrical objects and as flat objects. *Cu* 'thick' refers to the diameter of the cross-section of a cylindrical object and at the same time it also refers to the minimal dimension of a flat objet which corresponds to *hou* 'thick, therefore, when *Cu* 'thick' is applied to linear objects, it should be comprehended as the combination of this two kinds of senses. First of all, let's talk about the situation that linear objects are considered as cylindrical ones. When a cylindrical line becomes thick, the diameter of this line is greater, that is, the distance between two long sides of the line becomes greater (we also can call the distance between two long sides of a line as width). On the whole, the overall shape of this line extended largely. Therefore, the sizes of linear objects added to the thickness are greater than the original ones. Secondly, let's discuss the situation that linear objects are considered as flat ones. Here, *Cu* 'thick' is synonymous to *hou* 'thick', they all are referred as the third dimension which converted the two-dimensional surface into a three-dimension volume. Weydt suggest that language user form images of objects in subsequent steps: In the first step, a two-dimensional description arises. In the second step, a third dimension is then added. *Thick* is the third dimension that is added. *Thick* may describe the minimal dimension of a cuboid, such as *thick board*.[5] We think that the same interpretation would be given to *Cu* 'thick' described lines. In real world, even if a line is very thin and very small, it still is a three-dimensional object. We explain it with two steps, too. In the first place, we perceive a line as a two-dimensional object which can be described as *long* and *wide*. Then, when this two-dimensional line becomes *Cu* 'thick', it is furnitured with a volume. *Cu* 'thick' is used to refer to the third dimension of the line which is added in the second place. It is the third dimension of *Cu* 'thick' that make the line have a volume. Because the three-dimensional cuboid objects described as *Cu* 'thick' are all solid, the cuboid lines with *thickness* are blacker in color and thicker in tone than the original ones. Now, we should unify the first step and the second step to understand the sense of *Cu* 'thick' described the linear objects. Thus, the linear objects regarded as *Cu* 'thick' are not only greater in size, but also are blacker and thicker in color than the original ones, such as *cu zi* 'thick characters', *cu jian tou* 'thick arrows', *cu jing tan hao* 'thick exclamation mark'. The pictures make them stand out from the surroundings around them and the degree of the salience is enhanced largely. Therefore, when people want to emphasize a word or a sentence, they often make the relevant words become thicker, that is, make them become bigger, blacker and strengthen the stereo feeling of the characters, so as to get the effect of emphasis and highlight.

Cu mei mao 'thick eyebrows' in example (13) refers to great size, dense hair and black color of eyebrows. *Cu* 'thick' in both *cu mei mao* 'thick eyebrows' and *cu zi* 'thick characters' shares the same meaning, it refers to the diameter of the cross-section of cylindrical eyebrows and the same time it also refers to the minimal dimension of flat eyebrows. In our daily life, if our eyebrows are sparse and thin, we need to pencil eyebrows with brush, which can get the effect of the beauty. The process of penciling the eyebrows is the one in which we make the eyebrows become greater in size, denser in hair and blacker in color, as well as enhance the degree of the stereo feeling, that is, strengthen its thickness. Here, *cu du* 'thickness', in fact, both refers to *kuan du* 'width' of eyebrows and refers to *hou du* 'thickness' of them, *hou* 'thick' is interconnected with *cu* 'thick' when they are related to the linear objects, because the two concepts expressed by two words *cu* 'thick' and *hou* 'thick' respectively in Chinese is expressed only by one word *thick* in English.

The Granular Objects. *Cu* 'thick' can also be used to describe all three dimensions or the overall shape of the granular objects and it refers to three, approximately equal, dimensions, even the sizes of this three dimensions have some differences, they can be ignored. The granular objects, such as *sha zi* 'sands', *yan* 'salts', *mian fen* 'flour', *mai fu zi* 'wheat bran', *liang shi* 'grain', are made up of countless small, disparate parts. Only these objects are very small, can they be referred to as *cu* 'thick', otherwise, they can not be described as *cu* 'thick', for example, *luan shi* 'graits' that is the same classes with *sha zi* 'sands' can not be referred to as *cu* 'thick', because the sizes of them are greater than that of *sha zi* 'sands'. *Cu* 'thick' that describes the overall shape of the granular objects has the same function with *da* 'big' which can also describe the overall shape of the objects, but the meanings of this two words are completely different, because the granular objects described as *Cu* 'thick' can't be described as *da* 'big', for example, we can not speak of *da sha zi* 'big sands', *da yan* 'big salts', *da mian fen* 'big flour', but we can say *cu sha* 'thick sands', *cu yan* 'thick salts', *cu mian fen* 'thick flour'. Ren yongjun have explained about this phenomenon. He writes that the objects described as *cu* 'thick' and *xi* 'thin' are ones in which one dimension referred to as *cu* 'thick' and *xi* 'thin' is not salient and significant, and the other dimension is salient and main, identified as *chang* 'long' or *duan* 'short'. Because the dimension described as *cu* 'thick' is not salient and it is a minimal one, so, whether an object is thick or thin, it is still regarded as *xiao* 'small', rather than regarded as *da* 'big'. As long as we can identify the dimension described as *cu* 'thick' and *xi* 'thin', the dimension referred to as *chang* 'long' and *duan* 'short' is contain in it, because the dimension of *thickness* depends on the dimension of *length*. Yongjun Ren claims that because the dimension of thickness is smaller, correspondingly, the dimension of length will also be smaller, so the objects described as *cu* 'thick' and *xi* 'thin' always are small.[6] According to his explanation, the reason that *cu* 'thick' and *xi* 'thin' can describe the overall shape of the objects and only can describe small granular objects is that the dimension of *thickness* depends on and contains the dimension of *length*. We think that his explanation isn't correct. The dimension of thickness refers to the minimal dimension of the objects, but we can't infer that as the dimension of thickness becomes smaller, the dimension of length will correspondingly become smaller. Though the dimension of thickness depends on the dimension of

length, their meanings are different. We believe that whether *cu* 'thick' and *xi* 'thin' can describes the overall shape of the objects or not depends on the shape of the objects. Because the granular objects are small and the differences of the three dimensions in size can be ignored, that is, their three dimensions are approximately equal, So the shape of these granular objects can be identified as a cube or a sphere. The typical usage of *Cu* 'thick' is that it refers to the cross-section of the cylindrical objects, when the differences in size between three dimensions of an object is so small that it can be considered a sphere, the all cross-sections of a sphere can be referred to as *cu* 'thick', thus, *cu* 'thick' can be considered to refer to the overall shape of the granular objects. *cu* 'thick' referred to the cross-section of the cylindrical objects is the special case of *cu* 'thick' referred to the overall shape of the granular objects. For a sphere, its three dimensions are equal, but for a cylinder, its two dimensions are equal and a dimension is very salient. We believe that the fact that *cu* 'thick' refers to the cross-section of the cylindrical objects doesn't change from beginning to the end, just because the object to describe has changed. In Chinese, *cu* 'thick' referred to three dimensions can only describe the small, inanimate granular objects. If they are animate, they can't be described as *cu* 'thick', for example, we can't say *cu shi zi* 'thick louses', *cu tiao zao* 'thick fleas'. In German, *thick* can describe some bigger sphere-shape objects, such as *thick ball*, *thick apple*, which is the same usage with *big*. This usage of *thick* in German demonstrates again that the reason that *cu* 'thick' can refer to the overall shape of the objects is not because the dimension of thickness contains the dimension of length like Ren yongjun's explanation, but because of the shape of an object, namely, the sphere-shape.

2.2 Dependency

Lyons separates unoriented entities and space from oriented ones. For unoriented, three-dimensional entities, the maximal extension of the object is identified as its *length*. For oriented, three-dimensional entities, the maximal extension of the object is referred to as *high*.[4] In fact, the dimension of *height* can be considered as the *length* with orientation and direction, that is, *high* is the special case of *long*. If an object has the dimension of thickness, this object must have the dimension of length or height, so, the dimension of thickness depends on the dimension of length, after the maximal dimension of an object is described as *long* or *high*, then, and the minimal dimension of an object is referred as *thick*. *Length* is the base on which *thickness* can exist, and *thickness* is the derivant of *length*, so, they are interrelated and interdependent. If you want to understand the concept of the *thickness* of an object, the first appeared in your brain is not an isolated cross-section, but a cylindrical mage schema with its *length*, which is the background schema to cognize the concept of *thickness*. The cross-section of an object only can be identified on the basis of the cylindrical mage schema, and then you can understand the concept of its *thickness*, for example, *cu gun zi* 'a thick stick', we first associate this object with a long cylinder, in other words, first of all, it should be *chang gui zi* 'a long stick', and then we can identify its cross-section and cognize the dimension of *thickness* of this stick. Moreover, we can comprehend the dependency relationship of this two dimensions through the use of *cu*/*xi*

'thick/thin' combining with *chang/duan* 'long/short', *gao/ai* 'high/short'. In corpus, the usage of *cu* 'thick' combining with *chang/duan* 'long/short' can often be seen, such as *you cu you chang de xue jia yan* 'a thick and long cigar', *you cu you chang de bian zi* 'a thick and long pigtail', *you cu you duan de zhi tiao* 'a thick and short branch', *you cu you duan de bo zi* 'a thick and short neck'. The phrases formed by combining *cu* 'thick' with *gao* 'high' often be used to describe humans or trees, such as *you cu you gao de shu mu* 'a thick and high tree', *you cu you gao de han zi* 'a fat and tall man'. The usage of *xi* 'thin' combining with *chang* 'long' can often be seen, such as *you xi you chang de xian* ' a thin and long thread', *you xi you chang de mei mao* 'thick and long eyebrows', *you xi you chang de shou zhi* 'a thin and long finger'.

2.3 Implicating Qualities of Strength or Weight

Cu 'thick' implicates qualities of strength or weight, that is, *cu* 'thick' has meaning of strength or weight, so, the objects referred to as *cu* 'thick' are strong or heavy. Because the dimension of thickness depends on the dimension of length, on the condition that the lengths of two objects are equal, the thicker is an object, the larger its volume is. Usually, thick objects are made of solid matter, the larger is the volume of an object, the heavier its weight is. The objects described as *cu* 'thick' often can be grasped by the user's hands and taken to move around. If an object is thicker, people need much more strength to grasp it or to move it around. Therefore, the thicker is an object, the much more strength people need. For example, a man with a thick waist is often very fat, the larger is his volume, the heavier he is, if someone wants to push him, he needs much more strength, in return, which demonstrate that the strength of the man himself pushed by others is also large. In corpus, the objects referred to as *cu* 'thick' all implicate the qualities of strength or weight. In (14)-(15), two instances are given.

(14) Wo bu gan he ta jiao liang, ta de *bi bang* zu you hai wan *cu*, zuan qi quan tou hun shen gu bao.

'I dare not fight with him, because his arms are as thick as a big bowl, the muscles on his whole body bulge when he carrys his fists.'

(15) Ta zai ye di li pao le liang ke wan kou *cu* de *bai yang shu*, yi jing ren de li qi tuo le hui lai, zai zai yuan zi de liang bian.

'He chopped down two poplar trees in wild lands which were as thick as the opening of a bowl , and dragged back with an amazing strength, planting both sides of the yard.'

In (14), the phrase *the muscles on his whole body bulge when he carry his fists* indicates that a person with thick arms has much more strength or is very strong. In (15), the phrase *dragged back with an amazing strength* indicates that the thick poplar trees are very heavy.

Cu 'thick' and *zhong* 'heavy' share with the same meaning, so the two words often combine with each other to be applied to a sentence, two example are shown in(16)-(17).

(16) Si Jia cong cong qiao le yi yan, kan dao na shi yi ge *you cu you zhong de jin jie zhi*.

'Scarlett looked at it in a hurry, and see a thick and heavy, gold ring.'

(17) Dang shi suo wei de 'cheng gan', qi shi zhi shi yi gen *you cu you zhong de song shu shu gan*.

'At that time, so-called 'pole', in fact, is a thick and heavy pine tree trunk.'

Because *Cu* 'thick' shares the same meaning with *zhong* 'heavy' and the use frequency of combining with each other to be applied to a sentence is very high, the two words gradually fuse into one word *cu zhong* 'thick and heavy' in the process of their application. The morpheme *zhong* 'heavy' not only means to be heavy, but also means to be strong, such as the morpheme *zhong* 'heavy' in *cu zhong de yuan mu* 'a thick and heavy round wood', *cu zhong de shou* 'a thick and strong hand'. The word of *cu zhong* 'thick and heavy' is a compound word with coordinating relation. Because only the two morphemes have same or similar meaning, can they combine with each other to form a compound word, so, the fact that this two words of *cu* 'thick' and *zhong* 'heavy' fuse into one word of *cu zhong* 'thick and heavy/strong' prove that *cu* 'thick' implicates the qualities of strength or weight.

3 Conclusion

We have studied the semantic features of *cu* 'thick' on the basis of the previous researches, and have gained some new understandings about the semantics of *cu* 'thick', they are as follows: the meaning of *cu* 'thick' applied to the cylindrical objects is same with the meaning of *cu* 'thick' used to describe the granular objects, and they all refers to the cross-section of an object, their only distinction is that the objects they describe are different, one is used about cylindrical-shaped objects, the other is used about sphere-shaped objects. *Cu* 'thick' referred to the cross-section of the cylindrical objects is the special case of *cu* 'thick' referred to the overall shape of the granular objects. *Cu* 'thick' described linear objects both refers to the diameter of a cylindrical object and refers to one minimal dimension of a flat objet which corresponds to *hou* 'thick, therefore, it should be comprehended as the combination of this two kinds of senses. The semantic features of *Cu* 'thick' are as follows: the minimal dimension(s), dependency, implicating the qualities of weight or strength. these studies not only make us understand the senses and the semantic features of *cu* 'thick' deeply and correctly, but also can supply with the theoretical basis for the dictionary compilation and the foreign Chinese teaching.

This article has only studied the spatial senses of *Cu* 'thick'. Many metaphorical senses of *Cu* 'thick' have derived from its spatial senses. Which metaphorical senses *Cu* 'thick' have? What is the relationship between the spatial senses and the metaphorical senses, as well as, between the metaphorical senses? What is the mechanism of their evolution? As for these questions, we will do more researches in further.

Acknowledgments. Financial Support from Project Supported by Humanities and Social Sciences Youth Research Fund of Education Department(12YJC740113), Philosophy and Social Sciences Research Fund of Hunan Province(Xiang Zhe She Ling[2011], NO: 12).

References

1. Bierwisch, M.: Some Semantic Universals of German Adjectival. Foundations of Language 3, 1–36 (1967)
2. Vandeloise, C.: The Role of Resistance in The Meanings of Thickness. Leuvense Bijdragen 82(1), 29–47 (1993)
3. Vogel, A.: Swedish Dimensional Adjectives. Doctor's Degree Thesis in Stockholm University, 170–202 (2004)
4. Lyons, J.: Semantics. Cambridge University Press, Cambridge, etc. (1977)
5. Weydt, H., Birgitte, S.-L.: The Meaning of Dimensional Adjectives. Discovering The Semantic Process. Lexicology 4(2), 199–236 (1998)
6. Ren, Y.-J.: Study of The Semantics of the Spatial Dimensional Words in Modern Chinese. Master's Degree Thesis in Yanbian University, pp. 23–27 (2000)

Paradigmatic Semantic Network Construction of Psychological Adjectives in Mandarin Chinese—With a Case of Semantic Metadata Network Denoting 聪明 Congming "Smart"

Yuan Tao[1] and Zhanhao Jiang[2]

[1] School of Foreign Languages, Shannxi Normal University, Xi'an, China, 710062
taoyuanhuahua@126.com
[2] School of English Studies, Xi'an International Studies University, Xi'an, China, 710128
jzh89@yahoo.com.cn

Abstract. Psychological adjective in mandarin Chinese is one of the semantic groups of Chinese lexicons. With the construction of semantic metadata network denoting 聪明 *congming* "smart" as an example, this article, on the basis of Chinese defining metalanguage and paradigmatic semantic theory, made a probe into the paradigmatic semantic network construction of psychological adjectives so as to provide some support for computers to understand lexical items in Chinese.

Keywords: psychological adjective, paradigmatic, semantic network.

1 Definitions of Psychological Adjective

Psychological adjective is a relatively closed set in semantics and morphology. Therefore, two factors are involved when we delimit what is a psychological adjective (or in short, psych-adjective): In semantics, it is based on men's cognition and mental activity in their everyday perception. In morphology, this set is based on the definition and classification of Chinese adjectives set up by Dexi Zhu (cited in [1]). As for the status of adjective in mandarin Chinese, there have been controversies: Zhu[2] regarded adjective as part of the predicate and adjectives in modern Chinese can be classified as attributive adjectives and state adjectives, thereby considering adjective as an independent class of word. Many other scholars in China also held that adjectives in Chinese should be thought of as an independent class of word[3-6]. However, Zhao[1]maintained that Chinese is an adjective-verb-oriented language and that adjective is only a subcategory of verb. In our opinion, adjective can be delimited from the semantic perspective for it is deployed to describe characteristics of and to depict features and status of different things[7-8].

Then what does "psychological" imply? It is closely related with the English word "mind" and subsumes a variety of abilities concerning people's thinking: to feel, to observe, to understand, to judge, to choose, to remember, to imagine, to hypothesize and to reason, which all come under the umbrella phrase "mental ability" to guide people's action and behavior. Among all the definitions of mental ability, I prefer to

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 483–491, 2013.

follow George Boeree's definition: people's mental ability consists of: 1) ability to acquire knowledge; 2) ability to apply knowledge; and 3) ability to reason[9]. Therefore, adjectives reflecting such three abilities in Chinese can be labeled as psychological adjectives.

Psychological adjectives are often used to describe mental capacities of both human beings and other animals. Most of them are thoroughly abstract, which will be a hinder for computer to understand natural language. Therefore, on the basis of analyzing and describing such adjectives, especially adjectives denoting "聪明" *congming* "smart" and the like, this article, with the help of Chinese defining metalanguage theory, aims at providing a model to construct semantic metadata network to provide support for information processing in natural language.

2 The Inception of Metalanguage and Its Three Forms

The term metalanguage was originally invented by the Polish logician Alfred Tarski in the 1930s. It "indicates a language that is about language, one lever 'up' from the language itself, the 'object language'. A metalanguage indicates, comments on, examines, criticizes etc. what happens on the level of the object language" [10]. Since its inception, it used to be deployed mainly to explain the logical paradox.

Tarski delineated the necessity to demarcate the natural language and formal language when he tried to distinguish, from the perspective of philosophy and logic, the truth of a statement itself from the existential probability of the sentence. Natural language is closely related to the object and can be regarded as the object language when it is consistent with the metalanguage. However, the language that is used to describe the object is metatlanguage in the sense of philosophy and logic.

Besides, there are two other metalanguages, namely, one in lexicography and the other in natural language processing. Although such two metalanguages are on the basis of the metalanguage in philosophy, they two are mainly used in the applied linguistics and serve for dictionary compilation and natural language processing.

As for the second metalanguage(one in lexicography), Wierzbicka once expressed her ideas:

1. The lexicon of any language can be divided into two parts: a small set of words (or morphemes) that can be regarded as indefinable, and a large set of words that can be regarded as definable and that in fact can be defined in terms of the words from the set of indefinables.
2. For any language, its indefinables can be listed and the other words of this language can be defined in terms of these language-specific indefinables.
3. Although the set of indefinables is in each case language specific, one can hypothesize that each such set realizes, in its own way, the same universal and innate "alphabet of human thoughts."… Consequently, the number of indefinables is probably the same in all languages, and the individual indefinables can be matched cross-linguistically. Of course the indefinables of different languages cannot be expected to be equivalent in all respects; they can, nonetheless, be regarded as SEMANTICALLY equivalent…. In this sense (and only in this sense), semantic primitives can be identified with lexical universals[11:209-210].

The third metalanguage, taking the form of primitives, aims at formalizing and regularizing the meaning of words so as to help computers recognize and understand the natural language.

Metalanguage has now become a common means to account for natural language and has thus received considerable attention from many linguists. Is it an artificial language or a natural language? How can we apply it to dictionary compilation and natural language processing?

3 Chinese Defining Metalanguage and Semantic Network

3.1 Chinese Defining Metalanguage

We have, up to now, discussed the inception, the function, and the forms of metalanguage. Then how about the metalanguage in Chinese?

As for its definition and classification(artificial language or natural language), many scholars voiced their opinions. Li[12] pointed out that there should be demarcation between "metalanguage in linguistics" and "metalanguage in logics". In his opinion, the former should be classified as natural language. An[13]divided metalanguage into three types according to its function:

- Explanatory metalanguage: an instrumental language to explain utterance itself
- Defining metalanguage: an instrumental language to interpret lexical items such as words and phrases
- Analytical metalanguage: a language to describe and analyze glosseme.

After a statistical analysis of the word frequency, comparison, and validation of lexical items from many corpuses, he extracted the primitive defining words that consist of the core words and the extended words. Meanwhile, he ranked those interpretive words and made a list of such words alphabetically.

Su[14] subcategorized words from *Modern Chinese Dictionary*(1996 edition) and made a statistical analysis of them as well as an investigation of relative word frequency. He initially outlined both semantic features and pragmatic features of the defining metalanguage and pointed out that defining metalanguage provides a feasible access to dealing with the problem of "repetitive occurrence" in entry explanation in a dictionary, viz., the similar words are used to explain a dozen of synonyms without delineating their nuanced differences. Here are some dictionary entries in Chinese(Sentences in parentheses are their approximate English equivalents from Collins CoBUILT English Dictionary)[15]:

机智：脑筋灵活，能够随机应变(机智*jizhi* "resourceful": Someone who is resourceful is good at finding ways of dealing with problems.)

机敏：机警灵敏(机敏*jimin* "quick-witted": Someone who is quick-witted is intelligent and good at thinking quickly)

机巧：灵活巧妙(机巧：*jiqiao* "dexterous": Someone who is dexterous has the ability to perform a difficult action quickly and skillfully with the hands, or the ability to think quickly and effectively)

机灵：聪明伶俐；机智(机灵 *jiling* “shrewd”: A shrewd person is able to understand and judge a situation quickly and to use this understanding to their own advantage.)

From the Chinese explanations, we found that those synonyms explain one another and we can’t have a clear idea of their subtle differences, which will be difficult for readers and Chinese learners from abroad in their writing in terms of diction. However, the problem of “repetitive occurrence” can be solved. Here we take “机智”, a primitive word, as an example.

From the examples, we may conclude that “机智”, as a primitive word, is a primary part of its synonyms such as “机敏”, “机巧”, “机灵” in terms of meaning. However, the typical features of certain word are also listed in the above explanation, for example, [+敏捷 *minjie* “quick-minded”] is unique to the word “机敏”; [+巧妙 *qiaomiao* “tactful”] and [+口语*kouyu* “colloquial”] to “机巧”; [+灵巧*lingqiao* “skillful”] [+口语] to “机灵”. By so doing, we can avoid the problem of “repetitive occurrence” in entry explanation in a dictionary, on the one hand and fill the gap between natural language itself and artificial language explanation on the other, which will definitely give much support to the natural language processing by computer. Therefore, the current article, by means of interpreting psychological adjectives, especially the adjectives denoting “聪明” and its like with defining primitive words, attempts to describe the semantic relationships between words so as to construct a semantic metadata network of “聪明” type words.

3.2 Primitive Words Choice and Their Shared Meanings Extraction

Two factors are involved in our primitive words choice: word frequency and word appearance times according to a semantic group in a corpus. For example, by retrieval of the semantic group-“聪明” from the Peking University Corpus, we learn the frequencies of the following words: “聪明”(8992 times), 智慧 *zhihui* “wise”(8810 times, half of them are nouns in terms of part of speech), 聪慧 *conghui* “intelligent” (463 times), 聪颖 *congying* “bright”(325 times). Other words such as 明慧 *minghui* “bright”, 颖悟 *yingwu* “[of teenage]clever”, 颖慧 *yinghui*, “formal]intelligent” appear fewer than 100 times in the corpus. Given the frequency, we may take “聪明” as the primitive word for its highest frequency.

According to the research purpose, we regard psychological adjectives as a big semantic group and extract their shared meanings that have five features as follows:

Feature 1(Semantic class): it is concerned with their semantic property that can be used to classify “psychological adjectives” in a broad sense. Such feature, for example, can be “positive”(“active”) or “negative”(“passive”) that are embodied in those adjectives.

Feature 2(Semantic orientation): it is also related to semantic property, but aims at describing the opposite aspects of the semantic meanings of those adjectives: natural aspects or nurtural aspects of the semantic meaning of those adjectives. Some adjectives can denote mental capacities that people are born with: 聪明, 颖慧, 慧黠 *huixia* “clever and artful” while some can only be used to describe people’s mental

capacities that are nurtured in their life such as: 渊博*yuanbo* "erudite", 饱学*baoxue* "learned", 神机妙算*shenjimiaosuan* "crafty".

Feature 3(Semantic primitive): it deals with the specific and shared semantic features of adjectives in certain semantic group. Such features can be partly extracted from the entry definitions in the dictionary and partly from the nuanced differences we experience in our reading.

Feature 4(Original quantity feature): it covers "the adjectives' intrinsic feature that exists by themselves without reference to other adverbs denoting degree or other means such as comparison denoting measurement or scale. It consists of extremeness and high degree."[16]

Feature 5(Social meaning): it is closely connected with adjectives' unique semantic property and their commentary features. For example, being colloquial is typical of "机灵" while being formal and in written text are typical of "颖慧".

3.3 Paradigmatic and Semantic Metadata Network Construction

What we are going to do is to establish the paradigmatic semantic network that reflects the paradigmatic relationships among words. Lexical meanings, the same as grammatical meanings, can not only be expressed in a static and a general manner from the perspective of linguistic system but also expressed in a dynamic and concrete way from the perspective of linguistic function. The former is the potential meaning in a paradigmatic way while the latter is closely related to the pragmatic meaning of and the functional aspects of lexical items[17]. We are going to give a static and generative description of the semantic network and pay our attention to the ontological meaning and the relational meaning of words under discussion. In the network, the nodes are established according to the positions of the lexical items. The nodes are connected by lines that demonstrate the semantic relations between the nodes where the lexical items are pinpointed.

Semantic network is a virtual and formalized net. Plereme, sememe, glosseme and semantic field provide us with the basis and the tools to help establish such net. The theory underlying such net is the semantic field theory. This net is different from WordNet which also focuses on the semantic relationships in language in the following two aspects:

1) WordNet is designed to construct such relationships among lexical items in a language as synomymy, antonymy, hyponymy, homonymy, and meronymy. In contrast, paradigmatic semantic network is aimed at depicting more various semantic relationships(including social semantic meaning) among lexical items.

2) WordNet focuses its attention on the syntagmatic relationships among nouns, verbs, adjectives and adverbs while our paradigmatic semantic network attempts to, first, extract one core word from each semantic group, and then to establish the relationships between the core word and other words from the same semantic group, thereby offering a network with the core word as the guide so as to make easier the recognition and retrieval of the words in the course of language processing.

Now, in the following section, we will construct a whole paradigmatic network of psychological adjectives, in which the semantic metadata network of "聪明" is pin-pointed, as shown in Figure. 1.

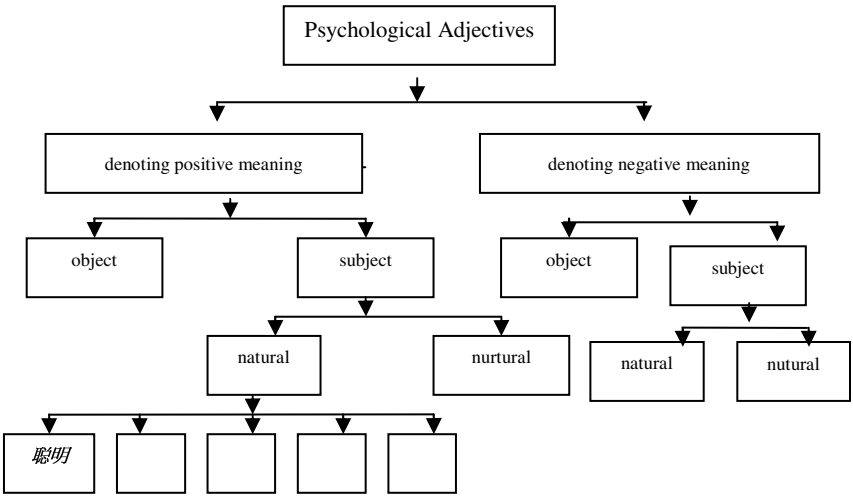


Fig. 1. Paradigmatic and Semantic Network of Psychological Adjectives

4 Semantic Metadata Network of Semantic Group Denoting 聪明

We have already positioned the semantic metadata network of "聪明" in the paradigmatic and semantic network of psychological adjectives, as shown in Figure 1. Still further, relationships between words belonging to the same semantic group are also complex. Then the following section, with the semantic metadata network of "聪明" as the research purpose, will discuss and illustrate the detailed relationships between words with such features as [+积极*jiji* "positive"], [+主体*zhuti* "subject"] and [+先天 *xiantian* "nurture"] in the network by means of metalanguage interpretation.

4.1 Defining Metalanguage Explanation of Words in the Semantic Group of 聪明

Through entry-to-entry comparison and semantic filtering, we chose the following words in the semantic group of "聪明" in our research from the following dictionaries: *Modern Language Dictionary*[18], *Synonym Thesaurus* [19], *Common Adjectives Classification Dictionary*[20], *Practical Chinese Adjectives Dictionary*[21]. Those adjectives can be defined as:

聪明：[+积极] [+主体] [+先天] [+聪明]
智慧：[+积极] [+主体] [+先天] [+聪明]
聪慧：[+积极] [+主体] [+先天] [+聪明] [+智慧]
明慧：[+积极] [+主体] [+先天] [+聪明] [+智慧] [+书面]
聪颖：[+积极] [+主体] [+先天] [+聪明] [+程度高]
颖悟：[+积极] [+主体] [+先天] [+聪明] [+程度高] [+书面]
灵性：[+积极] [+主体] [+先天] [+聪明] [+神通]
颖慧：[+积极] [+主体] [+先天] [+聪明] [+智慧] [+程度高] [+书面]
颖异：[+积极] [+主体] [+先天] [+聪明] [+突出] [+书面]
慧黠：[+积极] [+主体] [+先天] [+聪明] [+狡猾] [+书面]
岐嶷：[+积极] [+主体] [+先天] [+聪明] [+幼年] [+古语]
嶷嶷[1]：[+积极] [+主体] [+先天] [+聪明] [+幼小] [+古语][2]

From the interpretation, we may conclude that as is the case with “机智” being the primitive word and also the primary part of its synonyms such as “机敏”, “机巧”, “机灵”, so is the word “聪明” being the elementary shared component of its synonyms such as “智慧”, “聪慧”, “明慧”, “聪颖”, to name just a few.

4.2 Relationships of the Primitive Word to the Other Words in the Semantic Metadata Network

In the same semantic metadata network, there exists a primitive word and other related words that can be labeled as “external words”. Following the method mentioned in Section 3.1, we may also position those “external words” in the semantic metadata network by establishing the relationship of those words to the primitive word. By so doing we can provide a better access and a more feasible approach for computer to identifying those synonyms with nuanced and subtle differences.

Such relationships in the network can be tentatively summarized as:

Meaning Increase(labeled as A): There are more meanings of one “external word” than the primitive word. For instance, [+敏捷] is added to “机敏” on the basis of all the meanings of “机智”.

Meaning Strengthening(labeled as D): The external word is more stronger in meaning than the primitive word. For example, “聪颖” is in more strong sense when used to describe a person than “聪慧” is.

Word Register(labeled as Y): Some “external word” may have its own feature, i.e., such feature is typical of the word. Take “岐嶷” as an example. Both [+幼年] and [+古语] are unique to “岐嶷” in light of meaning interpretation.

---

[1] 灵性(*lingxing*, intelligence),颖异(*yingyi*, extraordinarily intelligent), 岐嶷(*qini*, particularly bright when young), 嶷嶷(*nini*, extremely intelligent when young).

[2] [+积极]=[+positive], [+主体]=[+subject], [+先天] =[+natural], [+聪明]=[+smart], [+智慧]=[+wisdom], [+书面] (*shumian*)=[+written], [+程度高](*chengdugao*)=[+high degree], [+神通](*shentong*)=[+mysterious],[+突出] (*tuchu*)=[+eminent], [+狡猾](*jiaohua*)= [+sly], [+幼年] (*younian*)=[+younger], [+幼小](*youxiao*)=[+young and small], [+古语](*guyu*)=[+archaic]).

Meaning Equivalence(labeled as T): The external word is nearly the same as the primitive word. “智慧”与“聪明” are a good case in point.

4.3 Metadata Network of the Semantic Group denoting 聪明

With the help of the primitive word definition and the relationships summarized between the primitive word and the “external words”, we attempt to construct the metadata network of the semantic group denoting “聪明” as shown in Figure. 2.

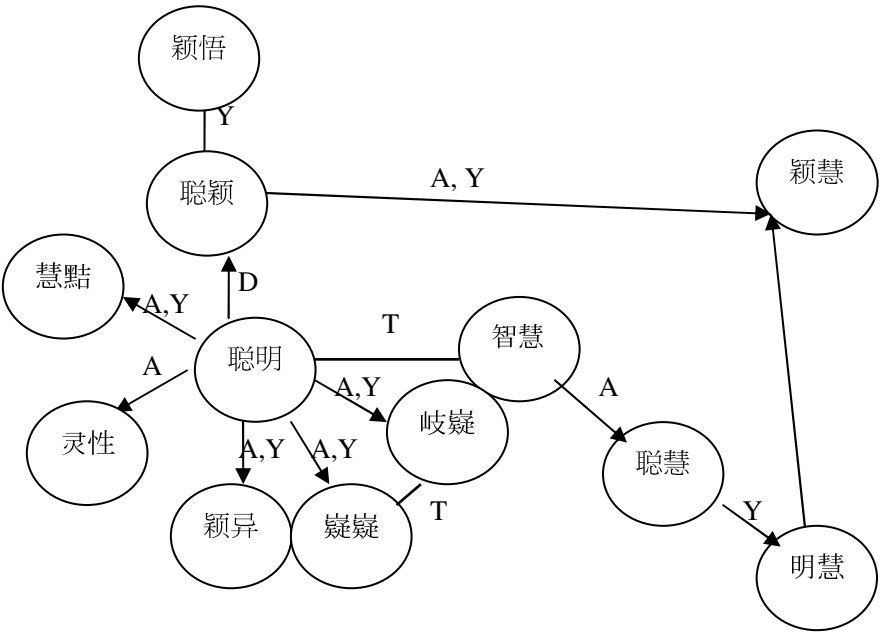


Fig. 2. Metadata Network of the Semantic Group Denoting ***congming***

5 Conclusion

With the help of Chinese defining metalanguage, this article made an attempt at describing and interpreting psychological adjectives by means of the primitive word. Due to the limited space, this article confined the description in certain semantic group, viz., semantic group denoting “聪明”. Word definition with the primitive word can not only solve the problem of “repetitive occurrence” in dictionary, but also demonstrate precisely the opposite aspects of semantic meanings between words and semantic groups respectively, which therefore help computers to understand natural language to some extent. Construction of semantic network gives a clear and vivid picture of the relationships between word meanings. So we may postulate that in the course of natural language understanding, computer may show us such features of word meanings shown in Feature 1 and Feature 2 by clicking on certain node in the semantic network. In addition, other features shown near the nodes in the network can

also be made use of, thereby activating all the necessary information to help computer to understand the word bank.

Obviously, there are limitations in this research: we have just touched the surface of the huge semantic group denoting psychological adjectives, i.e., the semantic group denoting "聪明". A complete semantic network for psychological adjectives needs much work in the future. What's more, we have to do some work about the numerical and logical language conversion before computers understand fully the semantic meanings that we have now probed into.

Acknowledgements. This article is supported by Wuhan University Doctoral Funding. Grant number: 201011102000026

References

1. Chao, Y.R.: Oral Chinese Grammar. Commercial Press, Beijing (1979)
2. Zhu, D.X.: Handouts on Grammar. Commercial Press, Beijing (2009)
3. Ma, J.Z.: Ma Jianzhong's Grammar. Commercial Press, Beijing (1898)
4. Li, J.X.: The New Chinese Grammar. Commercial Press, Beijing (1921)
5. Lü, S.X.: An Outline of Chinese Grammar. Commercial Press, Beijing (1942)
6. Wang, L.: Classification of Content Words in Chinese. Journal of Peking University 2, 55–69 (1959)
7. Zhang, B.J., Fang, M.: A Study on Functional Grammar in Chinese. Jiangxi Education Publishing Press, Nanchang (1996)
8. Shi, Y.Z.: Quantitative Features of Adjectives and Their Effects on Syntactic Operation. Chinese Language Teaching Around the World 2, 13–26 (2003)
9. http://baike.baidu.com/view/124566.htm (03/09/2010)
10. Mey, J.: Pragmatics: An Introduction. Blackwell Publishing, Oxford (2001)
11. Wierzbicka, A.: Semantic primitives and semantic fields. In: Adrienne, L., Eva, F.K. (eds.) Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization, pp. 209–227. Lawrence Erlbaum, Hillsdale (1992)
12. Li, B.J.: A Study on Metalanguage System of Chinese: An Analysis Project of Language Gene Atlas. Journal of Nanjing Normal University (Social Science Edition) 4, 140–147 (2002)
13. An, H.L.: A Study on Defining Metalanguage of Modern Chinese. Social Science Press, Beijing (2005)
14. Su, X.C.: A Study on Defining Metalanguage of Chinese. Shanghai Education Press, Shanghai (2005)
15. Collins CoBUILT English Dictionary Harper Collins Publishers, London, UK (2001)
16. Zhao, J.X.: A Study on Semantic Network of Psychological Adjectives in Mandarin Chinese, unpublished doctoral thesis. Nanjing Normal University, Nanjing, China (2006)
17. Zhang, J.Y.: Contemporary Linguistics in Russia. Commercial Press, Beijing (2003)
18. Modern Language Dictionary. Commercial Press, Beijing, China (2005)
19. Mei, J.J., et al.: Synonyms Thesaurus. Lexicographical Publishing House, Shanghai (1983)
20. Fu, Y.F.: Common Adjectives Classification Dictionary. Shanghai University Publishing Press, China (2005)
21. An, R.X.: Practical Chinese Adjectives Dictionary. China Standard Press, Beijing (1990)

Semantic Derivation of the Lexical Item *Yan/Mu* in Mandarin: A Cognitive Study

Xiangyun Qiu

Department of Taiwan Institute of Literature, National Changhua University of Education, Taiwan
chuss@cc.ncue.edu.tw

Abstract. The source of the semantic concept of body words is the body experience. With it we can construct new semantic concepts. *Yan/Mu* 'eye', as the primary sensory organ, is primarily relied upon to obtain the message to know the world. This article would like to combine the Lexical Semantics and Cognitive Metaphor to explore the lexical semantics of *Yan/Mu* in Chinese. The data are from dictionaries of Mandarin both online and printed in Taiwan. The theories applied include "Prototype" and "conceptual metaphor". I hope to construct a more clear system for the semantic derivation of the lexical item *Yan/Mu* in Mandarin.

Keywords: Lexical semantics, cognitive, semantic derivation, Yan, Mu.

1 Introduction

The meaning the body words represent is the source of the semantic concept. The concept the bodily words represent is the meta-concept in the cognitive world of human beings and plays a key role in the lexicon system [1]. Gestalt, with body experience as the original source, can help us to construct a new concept.

Among the body words, *Yan/Mu* is the primary sensory organ for human beings to grab message to get to know the world. Eye, the window of the soul, is called as "*mu*" in Ancient Chinese (commonly used in written language) and "*Yan*" in today's oral language. However, nowadays they exist side by side and are often used together, so here they will be discussed together.

2 Formation of *Yan / Mu* Words

Yan/Mu vocabularies are very common in the mandarin. According to the Revised Mandarin Chinese Dictionary of Taiwan Ministry of Education, there are 682 vocabularies composed of *Yan* and 563 vocabularies composed of *Mu* respectively [2], not to mention the derived complicated semantic expressions. We can classify the *Yan/Mu* vocabularies in the Chinese language and this is a categorization process

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 492–502, 2013.

for human beings to cognize the world [3]. Therefore, the compound words with *Yan/Mu* as morpheme are analyzed as follows:

2.1 Noun + *Yan/Mu*

Table 1. Compound words with *Yan/Mu* as morpheme

word-formation	-*Yan*	-*Mu*
(1) organ word + *Yan/Mu*	*Xin Yan* (mind's eye) *Du Qi Yan* (belly button) *Pi Yan* (asshole)	*Yan Mu* (Eyeliner, key point) *Mian Mu* (facial features) *Tou Mu* (head and eyes, boss) *Er Mu* (eyes and ears, spy) *Mei Mu* (brows and eyes, looks, prospect of a solution)
(2) animal word ＋*Yan/Mu*	*Ji Yan* (helosis) *Chong Yen* (wormhole) *Long Yan* (a fruit like dragon eye) *Zei Mei Shu Yan* (thievish-looking)	*Zhang Tou Shu Mu* (rat-eyed and buck-headed, Zeimeishuyan)
(3) plant word ＋*Yan/Mu*	*Ya Yan* (bud eye) *Liu Yan* (willow-leaf eye) *Xing Yan* (almond eye) *Tao Hua Yen* (peach-blossom eye, Woman winks) *Yen Hua* (have dim eyesight)	*Jie Mu1* (tree bulge wart)[1]
(4) thing word ＋*Yan/Mu*	*Quan Yan* (hole of spring) *Qiang Yan* (embrasure) *Zhen Yen* (needle eye, Sty) *Tai Feng Yan* (typhoon eye)	*Jie Mu2* (program) *Lan Mu* (column)
(5) Literary &language word ＋*Yan/Mu*	*Zi Yan* (wording) *Shi Yan* (poem eyes) *Wen Yan* (eye-catching point) *Hua Yen* (the meaning hidden in the words)	*Shu Mu* (book list) *Yun Mu* (rhyme catalog) *Pian Mu* (chapter heading) *Ti Mu* (topic) *Ju Mu* (a list of plays or operas)

[1] *Jie Mu* has two meanings: one is the bulge wart and the other is event program.

Table 1. (*continued*)

(6) quantifier word ＋*Yan/Mu*	*Kan Yi Yen* (a glance) *Yi Yan Jing* (a well) *Qian Li Yan* (thousand-mile eye) *Yi Ban Yi Yan* (beat in traditional Chinese music and operas)	(Wei Chi) *Yi Mu* (the cross point of lines) *Shu Mu* (amount) *Yi Mu Liao Ran* (be clear at a glance)
(7) character word＋*Yan/Mu*	*Zei Yan* (thievish eyes)	*Zi Mu* (Subheadings)
(8) abstract word ＋*Yan/Mu*	*Dian Yan* (electric eye) *Hui Yan* (insight) *Fa Yan* (a mind which perceives both past and future)	*Jia Mu* (price) *Ming Mu* (name) *Xiang Mu* (item) *Shui Mu* (tax items) *Zhang Mu* (account) *De Mu* (moral education course) *Ke Mu* (subject) *Ke Mu* (course) *Gang Mu* (outline) *Pin Mu* (name of things) *Tiao Mu* (article)

In the Revised Mandarin Chinese Dictionary [2], the meanings of *Yan* are listed as follows:

(1) visual organ of animals, such as *Yan Jing* (eyes)".
(2) hole, such as *Quan Yan* (hole of spring) .
(3) key point, such as *Jie Gu Yan* (critical point)".
(4) the place where there is no chess piece in Wei Chi.
(5) beat of dramas, such as *Yi Ban Yi Yan* (beat in traditional Chinese music and operas, the strong beat is called "*Ban*" and the weak beat is called "*Yan*").
(6) the unit counting how many times people look, such as *Duo Kan Liang Yan* (to look at two) .

While the meanings of *Mu* include:

(1) eyes, such as *Er Cong Mu Ming* (can hear and see well).
(2) terms and conditions, such as *Xiang Mu* (item) .
(3) articles in the front of the book, in order to facilitate the inquiry, such as *Shu Mu*
(book list).
(4) name and title, such as *Ti Mu* (topic). (5) leader, such as *Tou Mu* (boss).
(6) Level name in Biology, such as *Jie* (Categories) , *Meng* (phylum) , *Gang* (class) , *Mu* (order) .

From the above we can see although both *Yan* and *Mu* originally refer to eyes, with the diachronic extension of languages, their derivative meanings are not exactly alike. For example, there are more entity nouns composed of *Yan*, such as *Du Qi Yan* (belly button), *Quan Yan* (hole of spring) ; whereas more abstract nouns derive from *Mu*, such as *Jia Mu* (price) , *De Mu* (moral education course).

In addition, the derivative meanings of "*Yan*" center on key point, such as *Jie Gu Yan* (critical moment), *Shi Yan* (poem eyes), "*Mu*" emphasizes significant points in the above, such as *Tou Mu* (boss) , *Ti Mu* (topic) . Some of them even seem to be opposite. For example, *Yan* refers to something concave, such as *Quan Yan* (hole of spring) , *Yi Yan Jing* (a well) ; whereas *Mu* refers to something convex, such as *Jie Mu* which refers to the bulge wart. To take another example, in Wei Chi, the place where there is no chess piece is called "*Yan*" and the crossing point of each vertical and horizontal line is called "*Mu*".

2.2 *Yan/Mu* ＋Noun

(1) *Yan* + Noun: *Yan Li* (eyesight, discernment), *Yan Guang* (vision, insight), *Yan Se* (hint given with the eyes), *Yan Shen* (expression in one's eyes), *Yan Lian* (eyelid), *Yan Jie* (field of vision), *Yan Fu* (a feast of eyes), *Yan Zhong Ding* (a thorn in the flesh).

(2) *Mu* + Noun: *Mu Li* (eyesight), *Mu Guang* (vision, sight, view), *Mu Se* (eyesight), *Mu Ci* (table of contents), *Mu Lu* (catalogue), *Mu Di* (purpose), *Mu Biao* (target).

The two groups of words and expressions have something in common. For example, besides *Yan Li* (eyesight, discernment), *Yan Guang* (vision, insight) and *Yan Se* (hint given with the eyes), we can also say *Mu Li* (eyesight), *Mu Guang* (vision, sight, view) and *Mu Se* (eyesight). However, other [*Yan*+Noun]s seem to emphasize eyes as container, such as *Yan Lian* (eyelid) and *Yan Zhong Ding* (a thorn in the flesh); whereas [*Mu*+Noun]s stress the focus the eyes vision can reach, such as *Mu Di* (purpose) and *Mu Biao* (target).

2.3 Verb＋*Yan/Mu*

(1) Verb + *Yan*: *Chu Yan* (eye-catching, conspicuous, unpleasant), *Man Yan* (have one's eyes filled with), *Ru Yan* (pleasant to the eye), *Xing Yan* (catch the eye), *Yao Yan* (dazzling), *Qiang Yan* (eye-catching), *Ai Yan* (be an eyesore), *Zhuo Yen* (consider in a certain aspect)，*Fang Yan* (take a broad view), *Shun Yan* (pleasing to the eye), *Zhao Yan* (consider in a certain aspect), *Xian Yan* (conspicuous), *Yang Yan* (seductive), *Bi Yan* (close one's eyes, die), *Kai Yen* (widen one's view), *He Yan* (sleep, die), *Ci Yan* (offending to the eye), *Zha Yan* (blink), *Zhuan Yan* (in an instant), *Mei Zhang Yan* (blind), *Bu Qi Yan* (inconspicuous), *Gan Deng Yan* (look on in despair), *Kan Bu Shang Yan* (hold in contempt), *Mei Kai Yan Xiao* (smile from ear to ear), *Diu Ren Xian Yan* (make a fool of oneself).

(2) Verb + *Mu*: *Guo Mu* (look over so as to check or approve), *Chu Mu* (meet the eye), *Man Mu* (meet the eye on every side), *Ru Mu* (view), *Xing Mu* (striking), *Yao*

Mu (glaring), *Duo Mu* (dazzle the eyes), *Zhu Mu* (gaze at), *Yu Mu* (look over), *Ju Mu* (raise the eyes), *Zhang Mu* (open one's eyes wide), *Bi Mu* (close one's eyes, die), *Ming Mu* (close one's eyes in death), *Mang Mu* (aimless), *Xuan Mu* (dazzle), *Shi Mu* (sharpen one's eyes), *Fan Mu Cheng Chou* (fall out and become enemies), *Gua Mu Xiang Kan* (look at sb. with new eyes, admiration), *Li Li Zai Mu* (be still vivid in one's mind).

Some [Verb+*Yan*]s and [Verb+*Mu*]s can be replaced with each other, but the difference lies in that in many [Verb+*Yan*]s, eyesight is often a metaphor for attention, such as *Yao Yan* (dazzling), *Qiang Yan* (eye-catching), *Bu Qi Yan* (inconspicuous), *Kan Bu Shang Yan* (hold in contempt), and it is the same with some [Verb+*Mu*]s, such as *Xing Mu* (striking), *Yao Mu* (glaring), *Duo Mu* (dazzle the eyes), *Zhu Mu* (fix one's eyes upon), but few of them are words and most are four-part idioms. In addition, some eye motions further constitute time metaphor, such as *Yi Zha Yan* (a blink of the eye), *Yi Huang Yan* (in an instant), *Yi Zhuan Yan* (in an instant), *Zhan Yan Jian* (in an instant); whereas we can hardly see such semantic extension in [Verb+*Mu*]s.

2.4 *Yan/Mu* + Verb

(1) *Yan* + Verb: *Yan Cha* (out of sight), *Yen Chan* (be envious), , *Yan Kan Zhe* (do nothing but watch), *Yan Zheng Zheng* (helplessly).

(2) *Mu* + Verb: *Mu Song* (follow sb. with one's eyes), *Mu Du* (see with one's own eyes), *Mu Ji* (witness), *Er Ru Mu Ran* (bu influence by what one constantly sees and hears).

The differences lie in: many verbs in [*Yan* + verb]s are intransitive verbs, such as *Yan Cha* (out of sight), *Yen Chan* (be envious),whereas the verbs in [*Mu* + verb]s are mostly transitive verbs, such as *Mu Song* (follow sb with one's eyes), *Mu Du* (see with one's own eyes), *Mu Ji* (witness).

2.5 State Adjective＋*Yan/Mu*

(1) State Adjective + *Yan*: *Zui Yan* (eyes showing the effects of drink), *Leng Yan* (coldly), *Sha Yan* (be dumbfounded), *Zheng Yan* (squarely look); *Xie Yan* (squint at) *Qin Yan* (with one's own eyes), *Bai Yan* (contemptuous look), *Yan Re* (be jealous), *Da Xiao Yan* (unfair to treat people), *Xiao Xin Yan* (narrow-minded), *Lao Hua Yan* (presbyopia), *Shi Li Yan* (snobbish).

(2) State Adjective + *Mu*: *Nu Mu* (fierce stare), *Mang Mu* (aimless), *Ming Mu* (close one's eyes in death), *Xi Mu* (detailed catalogue), *Chen Mu* (stare angrily), *Yao Mu* (principal points), *Zong Mu* (superorder, comprehensive table of contents), *Ci Mei Shan Mu* (a benevolent and kind countenance), *Shang Xin Yue Mu* (feast one's eyes on sth.).

Besides literal meaning, the above [State Adjective ＋ *Yan*]s also have connotative meaning of emotion, mood or attitude. For example, *Zheng Yan* (squarely) means

respect; *Xie Yan* (squint at) means contempt; *Leng Yan* (coldly) means indifference and *Sha Yan* (be dumbfounded) means shock.

[Color word+*Yan*]s also have affective meaning. For example, *Bai Yan* (contemptuous look) means contempt; *Hong Yan* (jealousy) means envy; others like *Da Xiao Yan* (unfair to treat people), *Xiao Xin Yan* (narrow-minded), *Shi Li Yan* (snobbish) also show the attitude towards people.

Some [State Adjective＋*Mu*]s also have emotional meaning. For example, *Chen Mu* (stare angrily) means anger. Compared with [State Adjective+*Yan*]s, fewer [State Adjective+*Mu*]s have emotional meaning but many of them have the meaning of prominence or listing, such as *Xi Mu* (detailed catalogue), *Yao Mu* (principal points), *Zong Mu* (superorder, comprehensive table of contents).

2.6 *Yan/Mu*＋State Adjective

(1) *Yan* + State Adjective: *Yan Qian* (at present), *Yen Xia* (at the moment), *Yen Zhuo* (not recognize), *Yan Jian* (be sharp-eyed), *Yan Shu* (look familiar), *Yan Hong* (jealousy), *Yan Ba Ba* (look on with eager eyes), *Yan Zheng Zheng* (helplessly), *Yan Gao Shou Di* (have grandiose aims but puny abilities), *Jin Shou Yen Di* (have a panoramic view), *Yan Ming Shou Kuai* (nimble).

(2) *Mu* + State Adjective: *Mu Jin* (nowadays), *Mu Qian* (at present), *Mu Bu Jiao Jie* (have one's eyes open throughout the night), *Mu Bu Xie Shi* (look neither right nor left- but entirely absorbed), *Mu Bu Xia Ji* (too many things to see), *Mu Bu Zhuan Jing* (fix eyes on), *Mu Bu Shi Ding* (Illiterate), *Mu Zhong Wu Ren* (be supercilious), *Mu Kong Yi Qie* (Supercilious), *Mu Xuan Shen Mi* (dazzling).

In [*Yan/Mu* + State Adjective]s, the organ evolves into consciousness, namely eyesight or foresight, such as *Yan Jian* (be sharp-eyed), *Yan Ming Shou Kuai* (nimble), *Mu Bu Xia Ji* (too many things to see), but many of them are four-part idioms in the written language. Others like *Yan Hong* (jealousy) means infuriation or jealousy; *Yan Ba Ba* (look on with eager eyes) and *Yan Zheng Zheng* (helplessly) have such affective meaning as expectation or helpless. Besides, many [*Yan/Mu*+State Adjective]s mean space, and further "metaphor" abstract time, such as *Yan Qian* (at present), *Mu Qian* (at present) and so on.

3 Metaphor of *Yan / Mu* Vocabularies

Besides the literal meaning—visual organ, the above *Yan/Mu* vocabularies also have many other derivative meanings. One of the generated mechanism for semantic derivation is associative mapping of conceptual metaphor. Lakoff G. & Johnson M. pointed out that conceptual metaphor was the main way for human being to cognize the world and construct concepts, and people often rely on metaphor to refer to abstract things with concrete and familiar things so as to carry on brand-new concepts [4].

Conceptual metaphor, also called cognitive metaphor, can be divided into metaphor and metonymy. Metaphor is mapping between objects of different domains based on similarity. For example, *Ren Yan* (human eye) can be a metaphor source domain for *Zhen Yan* (needle eye) due to their similar shape; metonymy is replacement between objects of the same domain based on proximity, like functional replacement, *Ren Yan* (human eye) can be a metonymy source domain of *Shun Yan* (pleasing to the eye). Wu Shuqiong [5] said the mapping relationship of metaphor and metonymy can be shown as follows:

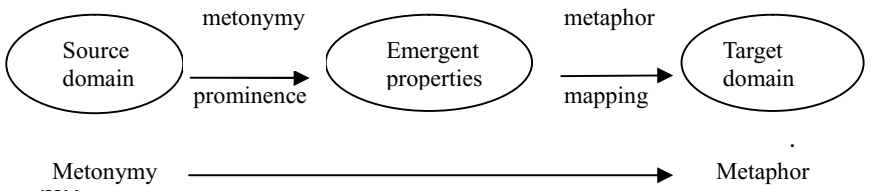


Fig. 1. The relational diagraph of metonymy and metaphor

In the following we will discuss metaphor and metonymy respectively.

3.1 Metaphor of Yan/Mu

Eyes metaphor vocabulary can then separate the following types: (1) Ontological Metaphor**,** (2) container metaphor, (3) Orientation (Space) metaphor [4], and (4) time metaphor. The metaphors performance as follows:

3.1.1 Ontological Metaphor

The above ontological metaphors formed through mapping from the source domain of *Yan/Mu* to other things are mainly mapping of different cognitive domains based on shape, location and function, such as:

(1) Shape metaphor: [Noun + *Yan*]s are mostly ontological metaphors based on shape, such as *Du Qi Yan* (belly button), *Pi Yan* (asshole), *Ji Yan* (helosis), *Quan Yan* (hole of spring), *Zhen Yan* (needle eye), *Qiang Yan* (embrasure), Through extending the shape of Yan (eye), The interwoven grid gap between cross can also be called *Yan*, such as *Wang Yan* (grid mesh), *Shai Yan* (sieve opening), and the void in Wei Chi the opponent cannot set to.

(2) Functional metaphor: there are also metaphors based on the function of *Yan* (eyes), such as *Dian Yan* (electric eye) meaning it has the visual function as the eyes.

(3) Property metaphor: there are metaphors based on the property of *Yan* (eyes), such as *Xin Yan* (mind's eye), *Jie Gu Yan* (critical moment), *Zi Yan* (wording), *Wen Yan* (eye-catching point) have the meaning of importance, and here *Yan* (eyes) has been extended from concrete domain to abstract domain.

The above are ontological metaphors, and in the following we will see other types of metaphors.

3.1.2 Container Metaphor

The basic elements of the image pattern—container include inside, outside and boundary. Eye is regarded as a container, and putting various information, knowledge and emotions in the container is metaphorized as the process of getting to know the world. As a container, eye has such attributes as deep/shallow opening/closing and entering/exiting [6], thus producing such expressions of container metaphor as *Ru Yan* (pleasing to the eye), *Man Yan* (have one's eyes filled with), *Yan Zhong Ding* (a thorn in the flesh), *Mu Zhong Wu Ren* (be supercilious), *Da Kai Yan Jie* (broaden one's horizon), *Ying Ru Yan Lian* (come into sight). Beyond that, *Qing Ren Yan Li Chu Xi Shi (*beauty is the eyes of the beholder), *Yan Jing Rong Bu Xia Yi Li Sha* (cannot bear a grain of sand in one's eyes) are also container metaphors.

3.1.3 Orientation (Space) Metaphor

Yan Qian (at present), *Ce Mu* (Sidelong glance,look askance at sb. with fear or indignation), *Kan Bu Shang Yan* (hold in contempt), *Yan Gao Shou Di* (have grandiose aims but puny abilities) are all metaphors of spatial orientation. People usually look up or aside when they look down upon others, so *Kan Bu Shang Yan* (hold in contempt) and *Xie Yan* (squint at) implies arrogance, contempt or disdain.

3.1.4 Time Metaphor

The words and expressions composed of *Yan* can also constitute more abstract time metaphors, such as *Yan Qian* (at present), *Yi Zha Yan* (a blink of the eye), *Yi Huang Yan* (in an instant), *Yi Zhuan Yan* (in an instant). Objects with a short spatial distance are also closer to each other in time, and therefore space domain can metaphorize time through further mapping. In Chinese language, we can use such metaphors in space as *Yan Di* (below one's eyes), *Yan Qian* (before one's eyes), *Mu Qian* (at present) to represent *Xian Zai* (now), *Ci Ke* (just now) in time [7]. Here *Yan* implies currentness [8].

3.2 Metonymy of *Yan/Mu*

Eyes are to watch things. Vision is the basis visual function and the vision is further extended to abstract meaning, namely judgment and observation ability. There is a similarity between the source domain and the target domain of metaphors, but proximity between those of metonymies. The metonymies of *Yan/Mu* vocabularies can be mainly divided into: metonymy of replacing whole with part, metonymy of replacing part with whole and metonymy of replacing part with part:

3.2.1 Metonymy of *Replacing Part with Part*

(1) Representing *vision* with eyes: *Fang Yan* (take a broad view), *Kan Yi Yan* (have a look at), *Lao Hua Yan* (presbyopia), *Jin Shi Yan* (near-sighted). Vision is one of the

elements to realize the visual function and the level of the vision directly determines the scope of visual involves. *Qian Li Yan* (thousand-mile eye), *Jing Shi Yan* (near-sighted), *Yan Jian* (sharp-eyed) are all metonymies of representing vision with *Yan* based on means-function relationship [9].

(2) Representing attention with eyes: the meaning of *Yan* is extended from vision to attention, such as *Re Yan* (eye-catching), *Xing Yan* (catch the eye), *Zhuo Yen* (consider in a certain aspect), *Yao Yan* (dazzling), *Bu Qi Yan* (inconspicuous).

(3) Representing judgment with eyes: people often perceive outward things, acquire outside information and make a judgment or choice with the visual organ [10]. Vision is the basis for visual function and its meaning is further extended as judgment and taste, representing judgment and taste with vision—mapping from concrete organ to abstract ability domain. The meaning of Kan (look at) is extended from visual perception to awareness and judgment in psychological level [11]. In the Chinese language, the words and expressions including Yan are all related to knowledge, intelligence and judgment, such as Yan Jie Gao (have one's sights set high), Yan Guang Du Dao (have a unique vision) [12], so Yan can be a metonymy for judgment, observation ability and taste. That is to say, when extending from concrete domain to abstract domain, the meaning of Yan—vision can also be extended as judgment, such as *Hui Yan* (a mind which perceives both past and future), *Yan Guang* (vision, insight), *Yan Zhuo* (not recognize), *Yan* Hua (have dim eyesight), *Xia Le Yan* (blind), *Kan Zou Yan* (make mistakes).

(4) Representing *the attitude towards things* with eyes: when people's attitude towards others or the outside is different, the expression in their eyes is also different. For example, when they look down upon others, they usually look up or aside and expose the white of the eyes, namely give them a supercilious look. *Yan* can be projected into attitude domain to express abstract concepts [7]. Thus formed *Yan Hong* (jealousy), *Bai Yan* (contempt), *Qing Yan* (favor, good graces). Besides, there are also *Leng Yan* (coldly), *Sha Yan* (be dumbfounded), *Da Xiao Yan* (unfair to treat people) to express people's mental attitude.

3.2.2 Metonymy of *Replacing Whole with Part*

(1) Representing *people* with eyes: *Qian Li Yan*—refers to one man with excellent eyesight;*Shu Yan*—traitors, *Yan Xian*—one who spies for sb. else, *Shi Li Yan*—refers to those who treat people according to their wealth and power, all of them are metonymies of replacing whole with part from another perspective. What metonymy involves is a kind of proximity and prominence relationship. People highlight one important part of the body in their thoughts and based on proximity, we can use eye to refer to the whole person. Based on the relationship of property-subject, eye can refer to people with certain characteristics.

(2) Representing *motion* with eyes: eye is a sense organ and has sensory ability which can be further extended as emotions, such as *Shun Yan* (pleasing to the eye),

Kan Dui Yan (love at the first sight), *Mei Kai Yan Xiao* (smile from ear to ear) means joyfulness, *Ci Yan* (offending to the eye) means dislike, *Nu Yan* (stare angrily) meas anger, *Yan Chan* (be envious) means dissatisfaction, *Yan Zheng Zheng* (with one's eyes wide open) means helplessness, and *Diu Ren Xian Yan* (make a fool of oneself) meaning shame. Eyes change with emotions and we can know people's mental attitude, inner life and emotion change through the expressions and movements of the eyes. For example, when we are happy, the corners of our eyes sag and the eyes flicker, thus producing *Mei Kai Yan Xiao* (smile from ear to ear); when we are surprised, our eyes are wide open and round, thus producing *Xing Yan Yuan Zheng* (almond-shaped eyes glaring round with rage); when we look down upon others, we do not look at them squarely but expose the white part of the eyes, thus producing *Zao Bai Yan* (been frowned upon). The change of the state of the eyes is caused by the change of people's mood and emotions [8], so people's mood can be metaphorized with the representation of the visual organ.

3.2.3 Metonymy of *Replacing Part with Whole*

Among *Yan/Mu* vocabularies, there are few metonymies of replacing part with whole and they can be divided into two kinds: one is to "metonymy" what you see with *Yan/Mu*, such as *Kan Yi Yan* (give a glance at)—referring to what you see with *Yi Yan* (a glance); the other is to metonymy the expression in one's eyes.

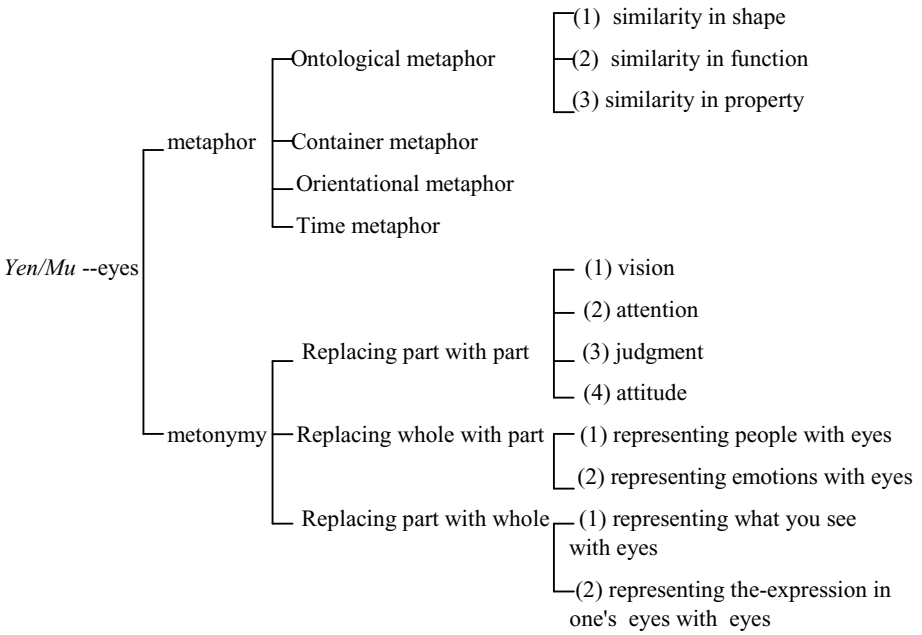


Fig. 2. The polysemy derivative diagram of metaphor and metonymy

4 Conclusion

From the above we can see the meaning of *Yan/Mu* vocabularies is abundant, the way how those meanings derive is diverse, and polysemy is closely related to metaphor/metonymy. Pan Yuhua [3] and Hui Jiaying [9] had listed their relationship derivative map. After reference, I draw the metaphor/metonymy relationship listed the semantic derivation relationship graph, and to summarized this article and use it as the conclusions :

This above shows the abundance meaning of eyes vocabularies and the semantic extension progess via metaphor/metonymy, and let us know the importance of body concept to congnitive concept construction.

References

1. Birong, H.: Semantic Research of Body Language. PhD Thesis. Shanghai International Studies University, Shanghai, Foreign Linguistics and Applied Linguistics (2010)
2. Ministry of Education. Revised Mandarin Chinese Dictionary, http://dict.revised.moe.edu.tw/
3. Pan, Y.: Semantic categories and metaphors of cognitive analysis of "Mu" word groups
4. George, L., Johnson, M., Shizen, Z. (1980); translation by Zhou, S.-Z. (2006): Metaphors We Live by. Linking Publishing Company, Taipei (2006)
5. Wu, S.: The Comparative Study of the Polysemous Networks of English Word "face" and Chinese "Lian" and "Mian" and Their Cognitve Motivation. Journal of China West Normal University(Philosophy & Social Sciences) (3), 88–95 (2009)
6. Jian, Z., Ping, C.: The metaphor of the "eye". Rhctoric Learning (2), 66–67 (2005)
7. Qin, X.: The Conceptual Metaphors of “eye” -A Comparative Study Based on the Corpus Between. Journal of Foreign Languages (5), 37–43 (2008)
8. Huang, S.: A study of body metaphor and metonymy in Taiwanese proverbs: taking “Taiwanese proverb dictionary” by Chen chuhsien as reference. National Cheng Kung University, Taiwan Literature, Taiwan. Master Thesis (2011)
9. Hui, J.: Comparative Study of Chinese and Japanese idioms-the body words: "eyes "and" heart ". Heilongjiang University, Japanese Language and Literature, Harbin. Master Thesis (2009)
10. Wu, W.: Study on the Visual Behavior Verbs of Modern Chinese. Ph D Thesis. Shandong University, Linguistics and Applied Linguistics, Jinan (2008)
11. Zhang, P.: The time structure, semantic extension and syntax of visual verbs in English and Chinese. Master Thesis. National Taiwan Normal University, Taipei, Teaching Chinese as a Second Language (2004)
12. Cao, F., Tsai, L., Liu, H.: Body and metaphor: the first interface of language and cognition. The Crane Bookstores, Taipei (2001)

A Study on Homophonic Puns from the Perspective of Semantic Field Theory

Chengfa Lu[1] and Yanli Li[2]

[1] College of Chinese Language and Literature, Wuhan University, Wuhan, China
lcfnit@126.com
[2] Nanchang Institute of Technology, Nanchang, China
liyanli1110@126.com

Abstract. Homophonic puns, a kind of frequently occurring trans-lingual rhetoric, have been discussed previously but unsatisfactorily in terms of their intrinsic operating mechanism, and of factors that influence their rhetorical effects. From the perspective of semantic field theory, this paper provides a new approach to describe the two terms mentioned above by taking into consideration the consistent core glosseme shared by homophonic word and mapping word, the quantity of the homophonic ingredients and mapping words, the abstract mapping paradigm between semantic fields constituted respectively by the homophonic ingredients and mapping ingredients, and the syntactic legality of homophonic words. All these factors are summed up coming into being the identification criteria for rhetorical effect of homophonic puns. According to the criteria, different types of homophonic puns are also discussed.

Keywords: homophonic puns, semantic field theory, rhetorical effects.

1 Introduction

Homophonic puns are among the various cross-linguistic rhetorical phenomena. Particularly, there are great amounts of homophones and semi-homophonic words in Chinese which make it very convenient to achieve homophonic puns; therefore, they are widely used in slogans, shop signs, literature, wisecracks and daily conversations. However, what are the intrinsic operating mechanisms of homophonic puns, and what factors affect the rhetorical effects? There have been a variety of linguists explaining these questions with a number of theories, but the explanations are not satisfactory. This article attempts to answer the questions by employing the semantic field theory, and to provide a theoretical guidance to the composition and application of homophonic puns.

2 Previous Studies on Homophonic Puns

The previous studies have employed a variety of theories, such as conceptual blending theory, relevance theory, metaphor theory, pragmatics, etc., to explain the interior

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 503–513, 2013.

functioning mechanisms of homophonic puns. However, these studies are mostly concentrated on the perceptive mechanism of homophonic puns, and they have not explicated what factors affect rhetorical effects of such puns. We will briefly introduce two of such explanations as follows.

2.1 Explanation by Conceptual Blending Theory

The conceptual blending theory is proposed by Fauconnier in 1997, which claims that there are at least four mental spaces: at least two input spaces (Input Space1 & Input Space2), generic space and blended space. Some common abstract structures shared by the input spaces constitute a generic space, and some interrelated elements between the input spaces have mapping relationships which are reflected in the blended space, and a new conceptualization of the structure—emergent structure—will come into being after many psychological activities such as deduction, analogy, the concept of clustering, knowledge framework, and etc [1-2]. Zhao has used this theory to explain homophonic puns [3]. Table 1 is his interpretation about "一桶天下" (*Yitong tianxia*; One barrel world, Unifying the World).

Table 1. Interpretation of "*Yitong tianxia*"

Generic Space	Input Space 1	Input Space 2	Blended Space
Event: Overlord conquering the world Objective: To unify the world	Visible context: 一统天下 (*Yitong tianxia*) Character: Overlord Instrument: Arms or weapons	Invisible context: 一桶天下 (*Yitong tianxia*) Character: Uni-President Enterprises Instrument: barrel of instant noodle	Emergent structure: The instant noodle of Uni-President Enterprises is so tasty that it can dominate the whole instant noodle market.

Conceptual blending theory provides a theoretical approach to comprehend the homophonic puns, which systematically demonstrates the whole interpretation process of the homophonic puns. However, this explanation has many limitations: Firstly, it gives no principles to determine the input spaces and the generic space. As a result, the practical operation of such explanation may seem too arbitrary. Secondly, the explanation causes the flattening of the rhetorical effects of homophonic puns because all the homophonic puns can be interpreted following the same pattern at the cost of neglecting different rhetorical effects of the homophonic puns. In addition, the theory only provides an approach to language understanding, but exerts little guidance value to the composition of other homophonic puns in language activities.

2.2 Explanation by Relevance Theory

Currently, there are many articles explaining homophonic puns with relevance theory. Sperber & Wilson, proposing relevance theory in 1986, assumes that relevance means people will seek the relationship between the newly-presented information and contextual assumptions when they comprehend the discourse, and under the same conditions, the greater the contextual effect there is, the stronger the relevance exists; and the smaller the effort for understanding is needed, the stronger the relevance is. Therefore, the degree of relevance usually depends on the relationship between the contextual effect and the effort to understand the discourse [4]. By using relevance theory, we can distinguish between different rhetorical effects of some homophonic puns. For example:

(1) 江山如此多娇，引无数英雄竞折腰。(The slogan of *Jianshan* Brand mosquito coil)
Jiangshan ruci duojiao, yin wushu yingxiong jing zheyao.
Land so beauty, makes countless heroes competitively bow.
This land is great in beauty, and has made countless heroes bow in homage.
(2) 闲妻良母。(The slogan of a washing machine)
Xian qi liangmu.
Idle wife good mother.
To relieve your wife and make her a good mother.

The line in case (1), cited from Chairman Mao's poem, is used as a slogan of *Jianshan* Brand mosquito coil. This homophonic pun is a little far-fetched because there is little relationship between the two terms of "*Jiangshan*", namely, in Chairman Mao's poem, "*Jiangshan*" means land (rivers and mountains, literally), while in the slogan it is just a brand name (*Jiangshan*, phonetically) of the mosquito coils. The rhetorical effect of this homophonic pun is limited. By contrast, in case (2), "闲妻良母" (*xianqi liangmu*, To relieve your wife and make her a good mother) maps with "贤妻良母" (*xianqi liangmu*, To be a good wife and a kind mother). It superimposes many meanings: the function of the washing machine is especially good, so it can enable the wife to get rid of the trivial, busy house chores, and make her as a "闲妻" (*xianqi*, idle wife). Moreover, the wife will have time to care for her husband and children, thus establishing herself as a "贤妻" (*xianqi*, good wife) and a "良母" (*liangmu*, good mother). It can be easily understood, and the context effect is fantastic. So the rhetorical effect of case (2) is much better than that of case (1).

The explanation by relevance theory also has its weaknesses: Firstly, it is subjective to judge the contextual effects and the effort to comprehend the discourse. The operability is poor when setting contextual effects and comprehensive efforts as the criteria for the degree of relevance. Secondly, it has not completely solved the problem of what factors are affecting the contextual effects and the effort made to understand the puns.

3 Rhetorical Effect Analysis of Homophonic Puns by Means of Semantic Field Theory

Homophonic puns occur when a phrase can be understood in two or more ways because another or other phrases share the same or similar sounds but bear different meanings. Homophonic puns always involve two kinds of words: one is the presented word called homophonic word, and the other is the implied called "mapping word". The parts that are shared by the homophonic word and the mapping word in the sound are called homophonic ingredients and mapping ingredients; the phrases respectively containing the homophonic word and the mapping word are called homophonic phrase and mapping phrase. Homophonic phrases and mapping phrases, homophonic words and mapping words, homophonic ingredients and mapping ingredients consist of semantic fields at different levels.

Semantic field is a collection of many words that share the same glosseme. Semantic field mainly refers to syntagmatic relation, such as relative semantic field, face semantic field; sometimes it also refers to paradigmatic relation, such as composite semantic field. Zhang distinguished ten structural relationships about the semantic field: synonymous structure, antonymous structure, hyponymy structure, class-defined structure, the whole-part structure, cross structure, sequence structure, structure of polysemy word formation, and composite structure [5]. The semantic field theory can desirously sort out the factors that influence the rhetorical effects of homophonic puns.

3.1 Consistency of Core Glosseme

The prerequisite of a word establishing itself as a homophonic pun is that the homophonic word and the mapping word share the consistent core glosseme, and they should not have any conflicting semantic features.

(3) 一桶天下。(The slogan for instant noodle of Uni-President Enterprises)
Yi tong tianxia.
One barrel all over the world.
One barrel dominates all the instant noodle market.

(4) 立肝见影。(The slogan for liver disease drug)
Li gan jian ying.
Erect liver see shadow.
To promptly see the liver drug effect.

In case (3), the mapping word "一统天下" (*yitong tianxia*) means that the overload conquers the whole world. The homophonic word "一桶天下" (*yitong tianx*ia) means Uni-President Enterprises can dominate the whole instant noodle market. The core glosseme of the two words is consistent, which is [+conquer/dominate]. However, case (4) has some problem. The mapping word "立竿见影" (*ligan jianying*) means to have effect instantly. The homophonic word "立肝见影" (*ligan jianying*) was used to mean that the drug could generate an instant effect. Although both of them have the

glosseme of [+instantly], the shared glosseme of [+shadow] is no harm for the mapping word while it is bad for the homophonic word—after taking the drug, the liver will show a shadow through chemical checkups. That means the drug will make the sick even worth. So this homophonic pun is not acceptable.

The consistency of the core glosseme discussed above is at the global aspect for the homophonic words and mapping words. If the corresponding ingredients between the homophonic word and mapping word share the same semantic features, rhetoric effects will be better. Take, for example, case (3), the corresponding ingredients of "桶"(*tong*, barrel) and "统" (*tong*, conquer) has no semantic relationship, so the rhetorical effect is limited. By contrast, in case (4), the corresponding ingredients of "闲"(*xian*, idle) and "贤" (*xian*, good) has a semantic relationship, namely, making the wife "idle" is a kind of "goodness". The rhetorical effect of case (4) is thus better than that of case (3).

3.2 Syntactic Legality of Homophonic Word

Whether syntax of the combination of the homophonic ingredient and other ingredients in the homophonic word is legal or not will affect the rhetorical effect of the homophonic pun. As in case (3), the combination of "桶" (*tong*, barrel) and "天下" (*tianxia*, world) is syntactically illegal, so this structure does not make any sense until we associate it with the mapping word. Some homophonic phrases are syntactically legal, such as case (2), the combination of "闲" (*xian*, to relieve) and "妻" (*qi*, wife) can constitute a verb-object structure, and it also generates the literal meaning. Because of the syntactic legality, this structure is more natural and smooth than the previous one.

3.3 Quantity of Homophonic Ingredients

If there are two or more homophonic ingredients in one homophonic phrase, or more than one homophonic words are used together, the rhetorical effect will be better. Examine the following examples:

(5) (上联)二三四五；(下联)六七八九；(横批)缺一少十 (an antithetical couplet written by a poor scholar)
(*Shanglian*) *Er san si wu*; (*Xialian*) *Liu qi ba jiu*; (*Hengpi*) *Que yi shao shi*
(Line 1 of a traditional couplet): Two three four five; (Line 2) Six seven eight nine; (Line 3, the horizontal wall inscription) Neither one nor ten (is present). Being shot of food and clothes.

(6) 喝酒必汾，汾酒必喝 (The slogan of *Fen* (brand name) wine)
Hejiu bi Fen, *Fenjiu bi he*.
Drink wine must *Fen*, *Fen* wine must drink.
If you drink, it must be *Fen* wine; on seeing *Fen* wine, you must drink it.

In case (5), the homophonic word "缺一少十" (*queyi shaoshi*) means lacking "one" and "ten", and the mapping word "缺衣少食" (*queyi shaoshi*) means lacking

"clothes" and "food". There are 2 groups of corresponding homophonic ingredients, namely, "一" (*yi*, one) and "衣" (*yi*, clothes), "十" (*shi*, ten) and "食" (*shi*, food). "一" (*yi*, one) and "十" (*shi*, ten) combines to constitute a figure semantic field; "衣" (*yi*, clothes) and "食" (*shi*, food) jointly achieve a semantic field of necessaries of life. The homophonic word and mapping word share a consistent glosseme of [+lack].The semantic structure of case (5) is shown in Fig. 1:

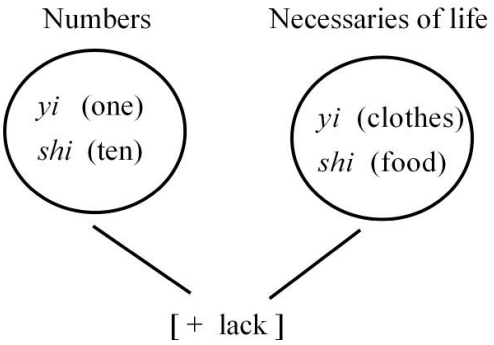


Fig. 1. Semantic structure of puns in case (5)

Case (6), however, demonstrates a weaker effect. The mapping phrase is "合久必分，分久必合" (*hejiu bifen, fenjiu bihe*) which means "After staying together for a long time, people will eventually be divided; long-time division will inevitably put people together." There are two sets of corresponding homophonic ingredients, namely, "喝酒" (*hejiu*, drink) and "合久" (*hejiu*, staying together for a long time), and "分" (*fen*, divide) and "汾" (*Fen*, a wine brand name). Each set of ingredients make up a semantic field of composite structure, but the homophonic phrase and mapping phrase, as well as homophonic ingredients and mapping ingredients, has no semantic relationship, so the rhetorical effect of case (6) is less than that of case (5).

3.4 Quantity of Mapping Words

A homophonic pun generally only has one mapping word, but there are also homophonic puns which may contain more than one mapping word, resulting in multiple interpretations:

(7) 元春、迎春、探春和惜春。(The names of the four Misses at Jia's Mansion in "A Dream of Red Mansions")
Yuanchun, *Yingchun*, *Tanchun he Xichun*.
The beginning of Spring, to welcome Spring, to explore Spring, to feel sorrowful on Spring.
The names of the four Misses at Jia's Mansion in "A Dream of Red Mansions" share a second syllable "春" (*chun*, Spring), while the first syllables come to achieve the sense as translated.

In this case, the homophonic words are the names of the four Misses at Jia's Mansion which constitute a character semantic field. Meanwhile, the homophonic words have two sets of mapping words and each of them compose a semantic field: 1) The chronological sequence semantic field contains four chronological stages of Spring. Namely, "元春" (*Yuanchun*) means the beginning of Spring; when the Spring starts, we should "迎春" (*Yingchun*, to welcome Spring) which implies that we are in the early Spring; Chinese people conventionally like to explore the Spring in the Mid-Spring when the Spring scenery is the best, therefore, "探春" (*Tanchun*) implies that we are in the Mid-Spring; "惜春" (*Xichun*) implies that we are in the late Spring because it is a commonplace that people feel sorrowful about late Spring in Chinese traditional literature. 2) "元迎探惜" (*Yuan-Ying-Tan-Xi*), the first syllable of each names, combines to establish a composite semantic field that maps "原应叹息" (*Yuan-Ying-Tan-Xi*) which means "It ought to be sighed", showing the author's attitude to the fate of the four Misses. The semantic structure of case (7) is shown in Fig. 2:

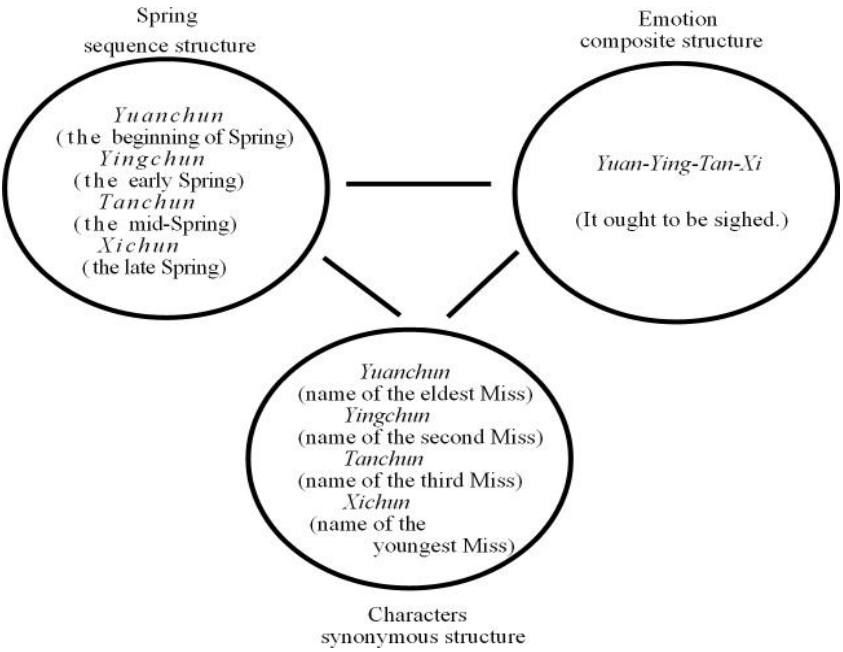


Fig. 2. Semantic structure of puns in case (7)

3.5 Relevance between Semantic Fields

If there is an abstract mapping paradigm between the different semantic fields respectively constituted by the homophonic ingredients and by mapping ingredients, it will

take less effort to associate and understand, thus improving the degree of relevance and rhetorical effects. See the following:

(8) 我失骄杨君失柳，杨柳轻飏直上重霄九。(Lines from the poem of "Butterflies love flowers—a reply to Shuyi Li")
Wo shi jiaoyang jun shi liu, *yang liu qing yang zhi shang chongxiao jiu.*
I lose beauty Yang (and) you lose Liu, poplar and willow blossoms slightly fly straight up to nine heavens.
I lost my lovely wife Kaihui Yang and you lost your husband Zhixun Liu. Yang and Liu may have gone up to the highest heavens just like the poplar and willow blossoms slightly flying up to the high sky.

In cases (8), the homophonic words "杨" (*Yang*, the family name of the poet's wife) and "柳" (*Liu*, the family name of the original poet recipient's husband) combine to develop a character semantic field. The mapping words "杨" (*yang*, poplar) and "柳" (*liu*, willow) achieve a plant semantic field. They share consistent glossemes of [+upstanding, +heroic]. Additionally, the metaphor of comparing trees to persons is common in Chinese classical poetry, which means that there is an abstract mapping paradigm between the character semantic field and plant semantic field leading to enhance rhetorical effects. The semantic structure of case (8) is shown in Fig. 3:

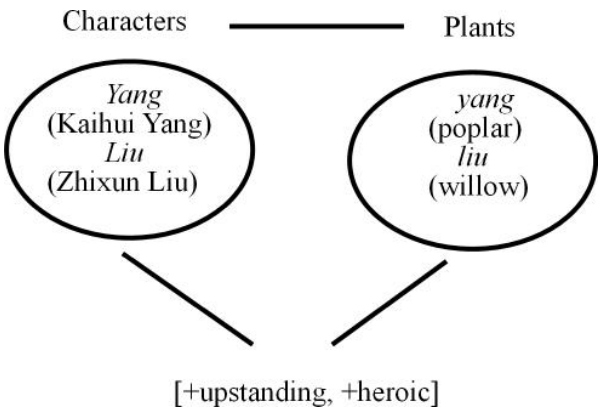


Fig. 3. Semantic structure of puns in case (8)

In fact, there are also mapping paradigms between the three semantic fields in case (7). Using "春" (*chun*, Spring) as a part of a female name is very popular in old China, and "Sorrow for Spring" is a traditional literary theme. "Sorrow" is thus the keynote of the four ladies' fates. Therefore, there are multiple semantic relationships between the three semantic fields.

4 Criteria for Identifying Rhetorical Effects of Homophonic Puns

The various factors discussed above can be summed up to establish the criteria for rhetorical effects of homophonic puns:

- The homophonic word and mapping word share the consistent core glosseme, and they should not have any conflicting semantic features.
- There should be more than one homophonic ingredient in a homophonic phrase, and the homophonic ingredients and mapping ingredients respectively constitute a semantic field.
- There should be an abstract mapping paradigm between semantic fields respectively constituted by the homophonic ingredients and mapping ingredients.
- The more the mapping words there might be, the better the rhetorical effect would be achieved.
- The corresponding ingredients between the homophonic word and mapping word share the same semantic features.
- The combination of the homophonic ingredient and other ingredients in the homophonic phrase is syntactically legal.

Among the criteria for rhetorical effects of homophonic puns, the first criterion is a necessary, and the others are to identify and grade the homophonic puns that may generate better rhetorical effects. Generally speaking, the more criteria a homophonic pun can satisfy, the better the rhetorical effect it will exert. Take the following, for example:

(9) 空对着，山中高士晶莹雪；终不忘，世外仙姝寂寞林。(Cited from Chapter 5 of "A Dream of Red Mansions")
Kong dui zhe, shanzhong gaoshi jingying xue; zhong buwang, shiwai xianshu jimo lin.
In vain face, in the mountain an eminent hermit with crystal-clear snow; Never forget, in the earthly paradise a fairy maiden amidst the lonesome forest.
To face in vain the eminent hermit in the high mountain with crystal-clear snow; not to forget the fairy maiden in the earthly paradise amidst the lonesome forest.

In this verse, "雪" (*xue*, snow) maps with "薛" (*Xue*, the family name of Baochai Xue), and "林" (*lin*, forest) maps with "林" (*Lin*, the family name of Daiyu Lin). "雪" (*xue*, snow) and "林" (*lin*, forest) combine to reveal a scenery semantic field; "薛" (*Xue*, the family name of Baochai Xue) and "林" (*Lin*, the family name of Daiyu Lin) combine to achieve a character semantic field. Meanwhile, "晶莹" (*jingying*, crystal-clear) and "寂寞" (*jimo*, lonesome) jointly constitute a trait semantic field. There are abstract mapping paradigms between these three semantic fields: the function of trait words is to modify the words of scenery, and the metaphor of comparing trees to people is commonplace in Chinese classical poetry. In the meantime, the corresponding ingredients of each semantic field have semantic links: snow is white and cold,

and Baochai Xue is noble and indifferent, so they share a consistent glosseme of "crystal-clear". "Forest" gives the impression of gloom, depression, and emptiness, while the fate of Daiyu Lin is sad and lonesome. The semantic features of the forest and Daiyu Lin are interlinked, thus the use of "lonesome" to modify the two entities is appropriate. Therefore, the homophonic pun in this poem has many semantic connections, resulting in multiple aesthetically everlasting experiences for the readers. The semantic structure is shown in Figure 4:

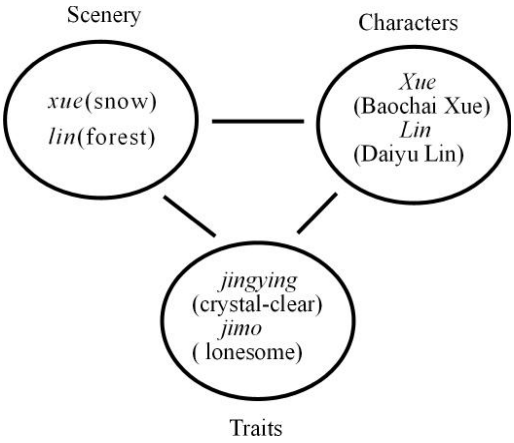


Fig. 4. Semantic structure of puns in case (9)

5 Concluding Remarks: Types of Homophonic Puns

According to the criteria for the rhetorical effects above, homophonic puns can be divided into different types.

- The stereotyped vs. implied types. Whether a homophonic phrase is a pun of the stereotype or implication relies in the legality of the phrase. The stereotype refers to a homophonic pun that contains homophonic ingredients incompatible with other elements in the phrase, like cases (3) and (4). The implication type refers to a homophonic pun that possesses a syntactically legal homophonic phrase collocation, such as cases (1), (2), (5), (6), (7) ,(8) and (9).
- The separate vs. coupling types. The separateness refers to a homophonic pun in which there is only one homophonic ingredient, as shown in cases (1), (2), (3) and (4). A homophonic pun that has more than one homophonic ingredient is called a coupling pun, such as cases (5), (6), (7), (8) and (9).
- The single vs. multiple types. Homophonic puns can be divided into singleness and multiplicity according to the quantity of the mapping phrases. Usually, a homophonic pun having only one mapping phrase is called a single pun, like cases (1), (2), (3), (4), (5), (6), (8) and (9). Some homophonic puns have more than one mapping phrase, which is called a multiple pun like case (7).

Generally speaking, the rhetorical effect of the implication type is better than that of the stereotyped type. The rhetorical effect of the coupling type is better than that of the separate type. The rhetorical effect of the multiple type is better than that of the single type. Therefore, if we compose a homophonic pun that is of the implied, coupling and multiple type, the rhetorical effects will be preferable.

Acknowledgements. This study is supported by the Fundamental Research Fund for the Central Universities (No. 2012111010203). We appreciate Mr. Guonian Wang and Miss Aiping Tu for their kind help with the proofreading. Two anonymous reviewers are acknowledged for their revision suggestions and comments.

References

1. Fauconnier, G.: Mappings in Thought and Language. Cambridge University Press, Cambridge (1997)
2. Fauconnier, G., Turner, M.: The Way We Think. Basic Books, New York (2002)
3. Zhao, Y.: Conceptual Blending Analysis of Advertisement Pun. Journal of Hunan University of Science and Engineering, 129–132 (2007)
4. Sperber, D., Wilson, D.: Relevance: Communication and Cognition. Blackwell, Oxford (1986)
5. Zhang, Z.Y., Zhang, Q.Y.: Lexical Semantics. The Commercial Press, Beijing (2001)
6. Luo, S.J.: The Cognitive Study on Parody in Advertisements. Foreign Language Research, 52–56 (2010)

The Semantic Relations of Internal Construction in NN Modifier-Head Compounds

Chong Qi

Université Paris Diderot-Paris 7 & CNRS CRLAO
cqparis7@yahoo.fr

Abstract. This present study is set to look at internal structure of NN modifier-head compounds found in Mandarin Chinese, under the framework of the Generative Lexicon (GL). The combinatorial possibility of NN compound words, which contain a single predicate, depends on the Qualia Structure of morphemes. We find three types of the semantic relation models between single morphemes (N): a. [appearance], b. [patient-agent] and c. [cause-effect]. The combinatorial possibility of NN compound words, which contain two predicates, also depends on the Qualia Structure of the morphemes. In this case, there is only one type of semantic relation model, i.e. [object-tool] semantic relation. In fact, these types of semantic relation models represent the generative model of NN compound.

Keywords: Chinese NN modifier-head compounds, Generative Lexicon (GL), Qualia Structure, Semantic relations model.

1 Introduction

Semantics and morphology always attach great importance to the subject study of the internal structure of NN modifier-head compound, from [6], [13], to recent [1]-[3] and [5] (249-253), [8]. All analyzed NN modifier-head compound in many kinds of languages, from the angles of syntactic and semantics, and their analysis methods and angles were innovated constantly. From their research it is possible to see that there are many ways to generate NN modifier-head compounds. However, there is a regularity to follow. The study of Chinese NN modifier-head compounds is very important; for example, [9], [12], [14]. However, their studies focus mainly on NN's syntax and semantic structure, instead of NN compound-generating factors, and we can see that some new methods in the fields of semantics and morphology were used in their research.

This paper tries to analyze the NN modifier-head compound[1] (for instance:[[菜cai vegetable$_{N1}$ - 刀dao knife$_{N2}$]$_N$ "kitchen knife", [车 che car$_{N1}$ - 祸 huo accident$_{N2}$]$_N$ "road accident", by investigating the semantic relationship[2] between the two Ns, in order to prove

[1] Beside modifier-head compound, of course, Chinese still has coordinative compounds, such as, 骨肉 gurou (flesh-blood, 'kindred'),岁月 suiyue (years-month, 'years') and so on. And coordinative compounds are not within the scope of this paper.

[2] We also use the term "relevance" to express the semantic relationship between the two Ns. This term means that, in the NN compound, the two Ns are relevant if certain semantic properties of one N are implied by the other.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 514–522, 2013.

the sematic factors in this kind of compound's generation. To be more specific, this paper analyzes whether the generation of this kind of compound's internal relation follows a certain semantic model.

2 Overview

The reason why NN modifier-head compounds are the focus of the study is that the category of word they forms is nominal, but their recognition for recessive verb V in the construction $N_1+(V_1)+N_2+(V_2)$[3], and its discussion on argument structure all take on its complexity.[4] Meanwhile, analyzing it from the same logical nature, the two Ns have a certain regularity and relevance, and this is the difference between the NV compound or the VN compound. However, the literature in this respect has not answered why the nouns collocated like as NN can be realized, and what is the process of this realization. Meanwhile, the literature definition for a compound word is not unified. Some literature involves words and phrases at the same time. We think lexicalized compounds have different characteristics from phrases; their semantic composition is very different from phrases. Therefore, in order to avoid a strong or weak problem of lexicalization, all the linguistic data in this study make reference to the 5th edition of *Modern Chinese Dictionary*'s NN compound double-syllable words;[5] for example, 菜刀caidao "knife", 车祸chehuo "car accident", 足球zuqiu "football", 石棉 shimian "asbestos", 电车 dianche "trolley bus", and so on. And these words can be formed by free or bound morphemes.

From the linguistic data analysis, we can see that there are two kinds of NN modifier-head compound word. The first one's two Ns belong to a genitive relationship, such as国法 guofa "national law", 象牙xiangya "ivory", 屋顶 wuding wuding "roof".[6] Although the second one is also the relation that N1 modified N2, the two Ns' relationship is not as complicated as the previous one: 花菜 huacai "cauliflower" (the predicate hidden should be [appearance], 车祸 chehuo "car accident" (the predicate hidden should be [cause], 马刀 madao "saber" (the predicate hidden should be two [ride] and [use]. The two Ns between the first kind of

[3] The V here represents the predicate contained in the compound, and the bracket represents the possibility that V may appear in different semantic types, and whether it could move in its apparent position. Eg.菜刀caidao [切qie "cut"$_V$菜cai "vegetable"$_N$的 de "of" 刀dao "knife"$_N$](knife to cut vegetables). Their semantic type is $V_1+N_1+N_2$.

[4] Qi Chong [12] recognizes and summarizes seven different kinds of Verbal predicates, according to the relationship between the verb and noun. He concludes twelve types of semantic collocation pairs.

[5] NN modifier-head compound originate from two sources: one comes into being from an abbreviated phrase, for example 衣架 yijia (clothing-shelf, "clothes hanger"); the other originates from concise idioms that have been passed down, such as "狗熊"(dog-bear, "bear").

[6] The other examples are as follows: 刀刃 daoren "cutting edge", 刀把 daoba "knife holder", 城门 chengmen "city gate", 门闩 menshuan "bolt", 头脑 tounao "brain", 身手 shenshou "skill"手掌 shouchang "palm", 手指 shouzhi "finger".

compound is very clear, and their analysis is very simple. Our study is aimed at the compound with a rather complicated semantic relationship namely, the second situation. This kind of compound also can be divided into the NN that contains a single predicate, and the NN that contains two predicates. As their characters are different, we will analyze them separately.

We found the second kind of NN modifier-head compound has two characteristics: A, this kind of compound describes an event; and B, the two Ns are the event participant. These two participants have an indispensable semantic relationship. On the other hand, they convey specific information, so it must be that the semantics of the two Ns attach some relevance to each other, and this kind of relevance makes the two Ns form an event. Based on these characteristics, we think a generative lexicon (GL)[7] analyzer is suitable for solving our problem. As [7] point out, first, it paid great attention to the compositionality and the collocation of the lexical semantics; that is, it emphasizes the semantic relationship of word's internal elements. Next, it can describe lexical semantics in different contexts (for instance, the Qualia structure), and this is what we need to analyze the internal structure of the NN modifier-head compound. In the meantime, the analyzing target of the GL is the opening discourse element, and the NN modifier-head compound belongs to the opening word set.

This study will use the GL to analyze the semantic relationship between N1 and N2, and establish a semantic model for generating a NN modifier-head compound. We will analyze a NN modifier-head compound which contains a predicate (Section 3), and then discuss the NN modifier-head compound that contains two predicates. (Section 4), and summarize of this kind of compound-generating model.

3 The NN Modifier-Head Compound That Contains a Single Predicate

Examples of this kind of compound are as follows:

(1) 花菜 huacai, flower-cabbage "cauliflower"
熊猫xiongmao/猫熊maoxiong, bear-cat/cat-bear "panda"
砂糖shatang, sand-sugar "granulated sugar"
石棉 shimian, stone-cotton "asbestos"
蒜泥 suanni, garlic-sludge "mashed garlic"
茶砖 chazhuan, tea-brick "brick tea"
马蜂 mafeng, horse-wasb "hornet"

(2) 菜农cainong, vegetables-farmer "vegetable grower"
茶农chanong, tea-farmer "tea grower"
车夫 chefu, car-driver "driver"
电工 diangong, electricity- worker "electrician"
鼓师 gushi, drum-master "drummer"

[7] See introductions about Generative Lexicon, in [10], [11] and [4].

马夫 mafu, horse-driver “groom”
乐师 yueshi, music-master “musician”
矿工 kuanggong, mine-worker “miner”

(3) 风扇 fengshan, wind-fan “fan”
车祸 chehuo, car-accident “car accident”
车辙 chezhe, car-rut “rut”
虫眼 chongyan, insect-eye “euglena”
虫灾chongzai, incect-damage “insect damage”
灯花 denghua, light-flower “candle light”
电光 dianguang, electricity-light “electric light”
风箱 fengxiang, wind-trunk “bellow”

The hidden predicate’s characteristic in example (1) should be [appearance], and its syntax structure should be [N_2(appearance)N_1] or [N1(appearance)N2].

The hidden predicate’s characteristic in example (2) should be [engage], and its syntax structure should be [N_2([engage])N_1], and for example (3) its syntax structure should be [N_2(cause)N_1]

From analyzing the GL’s frame for,example (1) compound word (i.e. huacai “cauliflower”), the two Ns’ relevance depend on the Formal role in its Qualia structure. Just as example (4) shows, because [花 hua]$_N$ “flower” and [菜 cai]$_N$ “vegetables’” demonstrated semantics that resemble their form. All the words in group (1) are this type.

(4)
$$\begin{bmatrix} \text{HUA} \quad \text{“flower”} \\ \text{QUALIA} = [\text{FORMAL}=x \;\;] \\ x' \approx x \end{bmatrix} \qquad \begin{bmatrix} \text{CAI} \;\; \text{“vegetable”} \\ \text{QUALIA} = [\text{FORMAL}=x'] \end{bmatrix}$$

In the type of compound in example (2) (Eg: 菜农vegetable grower), the two Ns’ relevance was demonstrated in the Agentive role of N1菜cai “vegetable” (its semantic feature is [planting]) and the Telic role of N2 农 nong “agriculture” (its semantic feature is [working in agriculture]). As the feature [working in agriculture]’s hyponym contains [planting], so its relevance is obvious. All the words in group (2) are type (5).

(5)
$$\begin{bmatrix} \text{CAI} \quad \text{“vegetable”} \\ \text{QUALIA} = \quad [\;\; \text{AGENTIVE} = x[\text{planting}] \;\;] \end{bmatrix}$$
$$\begin{bmatrix} \text{NONG} \;\; \text{“agriculture”} \\ \text{QUALIA} = \quad [\;\; \text{TELIC} = y[\text{working in agriculture}] \;\;] \end{bmatrix}$$
x = agent, y = patient

In (3), for example, 风扇 fengshan “fan”, the two Ns’ relevance in this word was demonstrated by the Agentive of N1 风 feng “wind” (its semantic feature is [air flow]

flow] and in the Telic role of N2 扇 shan “fan” (its semantic feature is [cause wind], as [air flowing] is the result of [cause wind]), so, their semantic feature is the cause and effect relationship. The words in group (3) are all model type (6).

(6)
$$\begin{bmatrix} \text{FENG} \quad \text{“wind”} \\ \text{QUALIA} = \begin{bmatrix} \text{AGENTIVE} = \text{x[air flow]} \end{bmatrix} \end{bmatrix}$$
$$\begin{bmatrix} \text{SHAN} \quad \text{“fan”} \\ \text{QUALIA} = \begin{bmatrix} \text{TELIC} = \text{y[cause wind]} \end{bmatrix} \end{bmatrix}$$
x = effect, y = cause

From the above examples of data, we can see that when the NN compound only has one underlying predicate, it shows mainly three different internal semantic models : i. [appearance], ii. [cause], iii. [cause and effect]. Meanwhile, we can see that a compound that contains a predicate will undergo its two internal Ns’ semantic operation in its Qualia structure, respectively.

4 The NN Modifier-Head Compound That Contains Two Predicates

As mentioned in Section 2, the Chinese NN modifier-head compound has a very complicated internal structure; namely, every N needs a predicate to realize relevance, so as to express some particular semantic information.

There are a lot NN compounds that contain two predicates: the words that use 马 ma “horse” as Modifier are as follows:

马刀 madao, horse-knife “saber”, 马枪 maqiang, horse-gun “carbine”, 马灯 madeng, horse-lantern “hurricane lamp”, 马褂 magua, horse-jacket “chinese jacket”[8], 马靴maxue, horse-boot “riding boots”, 马裤 maku, horse-drawers “riding breeches” and so on.

For example, the semantic structure of 马刀 madao “saber” (We simplify its definition of the dictionary) x+V_1(骑qi “ride”)+N_1马ma “horse”+V_2(用yong “use”)+(的 de “of”)+N_2刀 dao “knife”, x=[Human]

From the event logic angle seen two proposition are indicated: λxλyλz[RIDE’(x, y) & USE’(x, z)](x=human, y=horse, z=knife)。

This is a challenge for the GL system, because GL does not have the analyzing frame for these two propositions. Though we use type coercion of GL’s semantic analysis, we can’t obtain the semantic match for two N. Therefore, we will make some expansion for the GL’s framework in the following analysis, to obtain the comprehensive explanation for the NN compound’s internal semantic structure. Fradin (2008) made some adaptations to some parameters in GL, because he thought

[8] This is the buttoned Mandarin jacket of the Qing dynasty.

GL was only suitable for the semantic analysis in syntax environment, instead of suitable for the analysis of phrase semantics.

In the following text, we will take 马刀 madao "saber" as an example, and we will make a comparison with the NN compound word 牛刀 niudao, cattle-knife "butcher's knife", analyzing the expressive form of the NN compound that contains two predicates in GL. Our analysis means are just like (4), (5) and (6), (7) analyzes the head of the NN compound word 刀 dao "knife", (8) and (9) compare the analysis of the modifiers 牛 niu "cattle" and 马 ma "horse" respectively.

(7)

DAO "knife"
TYPESTR = [ARG = z artifact_tool]
ARGSTR = [D-ARG1 = y {object/animated}
D-ARG2 = x human]
QUALIA = [CONSTITUTIVE = {blade, knife holder…}
FORMAL = z
TELIC = cut_act (x,z,y)
AGENTIVE = make_act]

(8)

NIU "cattle"
QUALIA = [CONSTITUTIVE = mammal {…}
TELIC = { milk, plow, for the slaughter food(z,y) }]

(9)

MA "horse"
QUALIA = [CONSTITUTIVE = mammal{mane…}
TELIC = { drawn, plow, ride (x,w) }]

x = agent, y = patient, z = tool, w = patient

The above analyzing models indicate some major characteristics:

A. 刀 dao "knife(z)'s" Type Structure is [tool] (see (7)), and its Argument Structure includes the internal argument (y) and the external argument (x), so its Qualia Structure's Telic should be "cut, shear and some other activities", and it can have three arguments: x(agent), y(patient), z(tool), among which x(agent) is regarded as the Default Argument.[9]. The logic form here is goes like this: CUT(x,y)&USE(X,Z).

B. In the Qualia Structure of 牛 niu "cattle" and 刀 dao "knife", like (7) and (8), we can see the arguments of their Telic roles were shared by z(tool) and y(patient). 牛 liu "cattle" has three major Telic roles: the牛 niu "cattle" itself in [used as slaughtered slaughtered food] is the "patient", but the "tool"(z) can be same as the (z) in刀 dao "knife". In this situation, the Telic role of niu and dao is compatible, and their combination [niu-dao](cattle-knife "butcher") becomes a NN modifier-head

[9] That's the reason why we can say the sentence "这把刀切菜" Zhe ba dao qie cai. "This knife chops the vegetables".

compound, which contains a single predicate. This underlying predicate is the Telic role (cut, shear and some other activities) in刀 dao “knife”.

C. Compared to the semantic analysis of 牛 niu “cattle”, the Telic role of the Qualia Structure for马 ma “horse” has an important difference (see (9)). We can see that 马 ma ’s Telic role [riding] has two arguments: x(agent), w(patient), among which, x equal [human], and w [horse].[10] Compared with the semantics of 刀 dao “knife”, and the analysis, we can see that their Telic roles are not compatible, the only similarity is the [agent] argument(x). As马 ma “horse” and刀 dao “knife” both are internal arguments, and their combination can be realized by two predicates. The argument (x) [agent] is the Default Argument, which acts as a intermediate role to connect two propositions. It just as the analysis we made before, its expressive set should be (ride’(x, y)&use’(x, z)) (y=horse, z=knife). Therefore, the combination [ma-dao](horse-knife “saber”) is a NN compound which contains two predicates.

For the Chinese compound that contains two predicates, although GL can predict a compound’s combined possibility simply and concisely (such as (4),(5), and (6)), but this kind of special NN compound needs to add one more analysis process. Namely, it needs to select a suitable matching argument, and this is the reason why it is necessary to make an extension to the GL frame. However,, GL’s analysis of the compound’s combination and its explanation for the Qualia Structure is very effective.

This analysis also indicates that the morphemes’ combination between the Chinese NN compound is not arbitrary, and their internal semantic structure must satisfy certain conditions, so as to constitute a word well formed. From the analysis in this paper, we think the basic conditions to form the type of NN compound are that the two Ns’ Qulia Structure should have a corresponding correlation.

Not only does a Chinese NN compound have such special combination, but other languages also have similar situations. For example, in the English word “oil well”there is no corresponding point between the semantics of “oil” and “well”, and they also belong to the NN compound that contains two predicates.,They need to make use of the intermediate role [dig out oil] to make the semantic items in two Ns receive collocation. For another example, the French word *“sabre d’abordage”* saber-*prep.*- boarding “cutlass”. Here sabre *“sabre”* and *abordage* “boarding” can’t find a corresponding semantic that belongs to the NN compound and contains two predicates, so it must have an intermediate role [human]. And this word example is the same as the type we discussed above.

According to the above analysis, we can summarize that there are four semantic and syntactic types for Chinese NN modifier-head compounds.

i. appearance: $[\text{appearance}]_{N1(v)}$ $[\text{object}]_{N2(v)}$, see example（4）
ii. patient-agent: $[\text{patient}]_{N1}[\text{agent}]_{N2(v)}$, see example（5）

[10] We can analyze 马 ma‘horse’ as [tool], because in GL, ma “horse” belongs to the semantics of natural types. However, we know 水牛 shuiniu “buffalo”, 马 ma “horse”, 骡子 luozi “mule”, 驴 lü “donkey”, and some other domestic animals, were directly useful to human beings (characterized by creativity), and their existence is intentional, so they possess natural types and artificial types at the same time, and so they belong to complex types.

iii. cause-effect: $[cause]_{N1(v)}$ $[effect]_{N2}$, see example (6)
iv. object-tool: $[object]_{N1v1}[tool]_{N2v2}$, see examples (7) and (9)

5 Conclusion

This study looked at the internal structure of NN modifier-head compounds found in Chinese, under the framework of the Generative Lexicon (GL). The combinatorial possibility of NN compound words, which contain a single predicate, depends on the Qualia Structure of morphemes. We find three types of the semantic relation models among the single morphemes (N): a. [appearance], b. [patient-agent] and c. [cause-effect]. The combinatorial possibility of NN compound words, which contain two predicates, also depends on the Qualia Structure of the morphemes. In this case, there is only one type of semantic relation model, i.e. [object-tool] semantic relation. These types of semantic relation models represent the generative model of the NN compound.

This type of analysis also indicates that the morpheme of Chinese NN modifier-head compounds is not possessed arbitrarily, and their semantic relationship must reach a certain condition so can they form properly. From the analysis in this paper, we think the conditions for forming this type of compound are owing to the fact that the two Ns' Qualia Structure must have some corresponding collocation.

Although GL can predicate some well-formed compound easily, the Chinese NN modifier-head compound that contains two predicates needs to add one more process in the frame of GL analysis; namely, it needs to select suitable matching arguments or find an intermediate role. On the basis of the GL frame, this study established a set of intermediate frames which are suitable for analyzing complicated NN modifier-head compounda. The efficiency of this frame needs to be proved in a later study. This study proved that GL is very effective for analyzing compounds and making explanations in the Qualia Structure.

References

1. Arnaud, P.J.L.: Les Composés timbre-poste, Presses Universitaires de Lyon, Lyon (2003)
2. Bauer, L.: When is a sequence of two nouns a compound in English? English Language and Linguistics 2(1), 65–86 (1998)
3. Bisetto, A., Scalise, S.: The classification of compounds. Lingue e Linguaggio 2, 319–332 (2005)
4. Fradin, B. : Les adjectives relationnels et la morphologie. In : La raison Morphologique. Hommage à la Mémoire de D. Corbin, Fradin (dir.), pp. 69–92. John Benjamins, Amsterdam (2008)
5. Haspelmath, M., Sims, A.D.: Understanding Morphology, 2nd edn. Hodder Education, an Hachette UK Company, London (2010)
6. Jackendoff, R.: Morphological and Semantic Regularities in the Lexicon. Language 51, 639–671 (1975)
7. Johnson, M., Busa, F.: Qualia Structure and the Compositional Interpretation of Compounds. In: Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons, pp. 77–88 (1996)

8. Montermini, F.: La composition en italien, in La composition dans une perspective typologique. In: Amiot, D. (ed.), pp. 161–187. Artois Presses Université (2008)
9. Packard, J.L.: The morphology of Chinese: A linguistic and cognitive approach. Cambridge University Press (2000)
10. Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge (1995)
11. Pustejovsky, J., Bouillon, P.: Aspectual Coercion and Logical Polysemy. In: Pustejovsky, J., Boguraev, B. (eds.) Lexical Semantics – The Problem of Polysemy, pp. 133–162 (2005)
12. Qi, C.: The cover verb in modifier-head compopunds of [N+N]. In: Word Meaning and Computing-—Proceeding of the 9th Chinese Lexical Semantics Workshop, pp. 136–144 (2008)
13. Warren, B.: Semanic Patterns of Noun-Noun Compounds. Acta Universitis Gothoburgensis, Göteborg (1978)
14. Zhu, Y.: Semantic word formation of chinese compound words. Peking University Press (2004)

The Generation of Syntactic Structure Based on the Spherical Structure of Lexical Meaning

Qingshan Qiu

School of Chinese Language and Literature, Hubei University, Wuhan, P.R. China
qiuqs313@163.com

Abstract. Taking "Apple" and "Red" as examples, and basing on the new description of lexical meaning structure, we hold that the lexical meaning structure was composed of three elements: referent meaning, attribute meaning and feature-value meaning. In the lexical meaning structures of different word-classes, overt and covert conditions of these three elements are different. According to the conditions, we can divide the lexical meaning structure into two sections, that is, indication meaning and implication meaning. A lexical meaning structure is composed of many minimum lexical meaning structures, and in a lexical meaning structure, all minimum lexical meaning structures share one indication meaning, so this lexical meaning structure is treated as a spherical structure of lexical meaning, and this indication meaning is regarded as the centre of the sphere, and implication meaning of this word is treated as outside of the center of the sphere. The spherical structure of lexical meaning is the foundation of the generation of syntactic structure, and the syntactic structure is the extending and intersecting of the spherical structure of lexical meaning.

Keywords: syntactic structure, lexical meaning structure, spherical structure of lexical meaning, cognitive projection.

1 Introduction

Starting from the 1970s, the relationship between the lexical semantics of verb and the syntactic realization of the lexical semantics is the focus of the linguistic research. Linguists come to realize that there is a regular projected relationship between the lexical semantics and syntax of verb. The verbs of same semantic class have the same syntactic performance, and the lexical semantics of verb has a decisive role in syntactic form. Based on syntactic projection of the verb semantics, the syntax-semantics interface theory was gradually formed and the study of this theory achieved gratifying results. [1-3]

We believe that the syntax-semantics interface is not just concerned about the lexical semantics of verb, and also concerned about the lexical semantics of other parts of speech, and the structure description of the lexical meaning has a great role in promoting the study of the syntax-semantics interface.[4] In this paper, taking the lexical meaning structure of the adjective "Red" and the noun "Apple" as examples, and taking the cognitive linguistics and syntax-semantics interface theory as the

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 523–531, 2013.

theoretical backgrounds, we pay more attention to the basic role of lexical semantics of nouns and adjectives in generating the syntactic structures. The lexical meaning structure is a spherical structure, and this is a new description of the lexical meaning structure.[4] About the process of the generations of syntactic structures based on the lexical meaning structures, this paper has a detailed description. The basic concepts and conclusions of this paper will certainly help NLP (Natural Language Processing).

2 Surface and Deep Traits of the Syntactic Structure Generation

2.1 Surface Traits of the Syntactic Structure Generation

Looking from the surface trait of dominant language symbols, the syntactic structure is made up of at least two words, and only one word can not form a syntactic structure. So the essential features of syntactic structures generation is the linear composite construction of many words. But only when the semantic deep features is fully satisfied, can we have ensured the linear composite construction of many words is usable.

2.2 Deep Traits of the Syntactic Structure Generation

Looking from the deep trait of recessive semantic information, the syntactic structure generation is a process that the lexical meaning's non-determinacy of the head word of a sentence is gradually eliminated and the semantic information of the sentence is gradually engendered. Taking the process of the utterance generation as example, Lu Ch. explained that the function of information is to eliminate the non-determinacy. Lu Ch. said: "When people to speech, they always at first say out a topic, and then put some comments on the topic. This is to say, in order to keep the audience in suspense, the speaker put forward a non-determinacy by a topic, and then eliminate the non-determinacy by comments in order to remove the audience's suspense."[5] We think that this view of Chuan Lu is right.

2.3 Summary

As a matter of fact, the premise of many words as dominant language symbols can linearly be composited each other is the lexical meaning structures of those words are consistent in the deep aspect of semantic information.[6] We know also that the consistent of the lexical meaning structures is based on the consistent of the element of the lexical meaning structures. So the premise and base of the syntactic structures generation based on the linear composition of many words are the elements of the lexical meaning structures are consistent each other. So there have syntax-semantics interfaces in the process of the generation of the syntactic structures, and the key

decisive factor affecting and restricting syntax-semantics interfaces is the elements of the lexical meaning structures and the fusion of the elements of the lexical meaning structures.

3 Descriptions of the Lexical Meaning Structures Based on the Cognition Structures

3.1 Minimum Cognitive Structure and Minimum Lexical Meaning Structure

Taking the syntax construction "Red apple" as an example, this paper only describes the lexical meaning structures of thing noun "Apple" and adjective "Red". The lexical meaning structures of the same kind word can be described on the analogy of this.

In the process of the cognition, people at first must choose an object as the target of the cognition, for instance, "Apple"; secondly, they must choose the different attributes, for example, "color, taste, form, weight, place of production, price", etc., as the way to understand the target of the cognition; finally, people must obtain some appropriate results (feature-value) from the different attributes. Otherwise, the whole process of the cognition is not completed, and people can not really understand "Apple" as the target of the cognition. So we believe that minimum cognitive structure must contain three elements: one certain cognition target (such as "Apple"), one selected attribute (such as "Color") as a cognitive way, one appropriate feature-value (such as "Red") as a cognitive result.[7-9]

We know that the words are the symbolic language used by people to understand the objective world, and this paper believes that there is a projected relationship between the lexical meaning structure and the cognition structure, and the lexical meaning structure is the result that the cognition structure projects onto the word as the language symbol.[10] Therefore, the smallest lexical meaning structure also contains three elements: one referent meaning,[11] one appropriate attribute meaning and one appropriate feature-value meaning. For the word "Apple", in its elements of the lexical meaning structure, the referent meaning is dominant element, known as the "dominant referent indication meaning"; and its attribute meaning and feature-value meaning are recessive element, respectively known as the "recessive attribute implication meaning" and "recessive feature-value implication meaning". Attribute implication meaning and feature-value implication meaning known as the implication meaning. As a result, the three elements of the lexical meaning structure of "Apple" can be divided into two levels: the indication meaning level and the implication level.

3.2 Description of Lexical Meaning Structure of the Word "Apple" and "Red"

We express the above three elements and two meaning levels of the lexical meaning structure of "Apple" as follows by Fig.1, and known as the spherical structure of the lexical meaning.

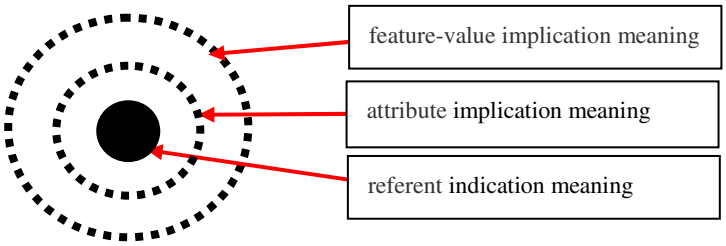


Fig. 1. The schematic diagram of spherical structure of lexical meaning of “Apple”

The solid circle in Fig.1 shows that the referent meaning of the word “Apple” is only one, and in the dominant indication state, and this referent meaning constitutes the indication meaning of spherical structure of lexical meaning; Two dotted circles show respectively that the attribute meaning and feature-value meaning of the word “Apple” there are many, and one attribute meaning can correspond one feature-value meaning, and also can correspond many feature-value meanings. The attribute meaning and feature-value meaning of the word “Apple” expressed by the two dotted circles are all in the recessive implication state,[12] and those attribute meanings and feature-value meanings of the word "Apple" constitute the implication meaning of spherical structure of lexical meaning.

We know that, “Apple” and “Red” as the object of people's cognition, but “Apple” is an independent cognitive object and “Red” is a dependent cognitive object, because the existence of “Red” must adhere to other entities, such as “Apple”, etc.. As a result, the indication meaning in the lexical meaning structure of the word “Red” does not indicates referent, but a feature-value of one attribute of the referent.

Therefore, based on the figure 1, the spherical structure of lexical meaning of the word “Red” can be expressed by the schematic diagram known as Fig.2.

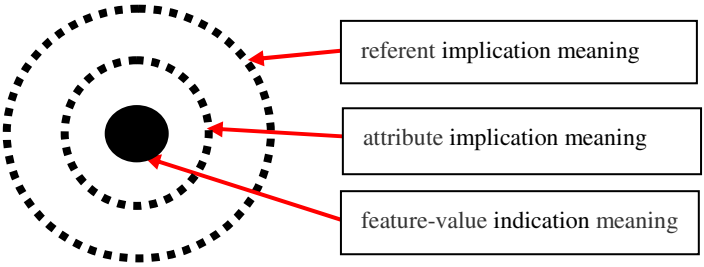


Fig. 2. The schematic diagram of spherical structure of lexical meaning of “Red”

Not same as the lexical meaning structure of the word “Apple”, in the lexical meaning structure of the word “Red”, the feature-value meaning is in dominant indication state, while the referent meaning and the attribute meaning are in the recessive implication state.

4 Generations of the Syntactic Structures Based on the Lexical Meaning Structures

4.1 Implication Meanings Play a Key Role in the Generation and Survival of the Syntactic Structures

According to the above description of the lexical meaning structure of the word "Apple" and "Red", we know that language symbols read by people in the syntactic structure actually first present their indication meaning. When people choose words to express a sense, they can select words on the basis of the indication meaning of these words, and then arrange these words into the right place by the minimum structure of the cognition and the minimum structure of the lexical meaning, and eventually form the syntactic structure that we are able to pass the information what we want to convey. When we try to understand the information conveyed by these syntactic structures, what we first understand is the indication meaning of these language symbols, and we also follow the composition principle of the minimum structure of the cognition and the minimum structure of the lexical meaning to get the appropriate implication meaning with all three elements.[13]

We know that though these implication meanings are in the recessive potential state, but it does not mean they do not exist, on the contrary, their presence play a critical limiting role in the generation and survival of the syntactic structure. It can be said that the implication meanings of lexical meaning structure play a key affect in the syntax-semantics interface. Because a single lexical meaning in the surface structure of dominant language symbols showing the indication meanings do not show more other semantic information, other semantic information exists in the inner of the lexical meaning by the state of the implication meaning. When the syntactic structures have been generated by the synthesis of words, these implication meanings naturally hide into the deep of the syntactic structure. Therefore, the implication meanings play a key role in the acceptability of the syntactic structure. In other words, the unity between the implication meanings and the indication meanings among lexical meanings is the key to the survival of syntactic structures.[4] If the combination of the syntactic structure between words is tenable, it requires that not only there has a linear arrangement relation between words of the surface of the syntactic structure, but also more critical is that an element of the lexical meaning can enter the spherical structure of the lexical meaning of another word, and become a constituent element of the spherical structure of the lexical meaning of another word. Otherwise, the two words do not constitute a linear combination relation.

4.2 The Process of the Generation of the Syntactic Structures

Look at the following syntactic combination structure:

- (1) 红苹果 (Hong pingguo; Red apple)
- (2) 坏苹果 (Huai pingguo; Bad apple)
- (3) 酸苹果 (Suan pingguo; Sour apple)

Any one of above three syntactic structures is composed of two words, and to analyze why these composite structures can survive, we can start to analyze from two aspects:

First of all, we can start to analyze these syntactic structures from the word "Red" as the modifier.[14] According to the model of the spherical structure of lexical meaning, we know that the word "Red" is a feature-value indication meaning of the attribute word "Color"(attribute implication meaning), and the word "Bad" is a feature-value indication meaning of the attribute word "quality"(attribute implication meaning), and the word "Sour" is a feature-value indication meaning of the attribute word "taste"(attribute implication meaning),[12] and that the word "Apple" is not only provided with the properties of all these attribute words "Red, Bad, Sour" as the implication meaning, but also the words "Red, Bad, Sour" as the feature-value indication meaning can become the feature-value implication meaning of the word "Apple". The above analysis shows two points: to begin with, the attribute implication meanings of all these words "Red, Bad, Sour" are included in the attribute implication meaning of the word "Apple". Secondly, the feature-value implication meaning of some specified attribute implication meanings of all these words "Red, Bad, Sour" are included in the feature-value implication meaning of some specified attribute implication meaning of the word "Apple". Therefore, based on the consistent of the attribute implication meanings and the consistent of the feature-value indication meaning and the feature-value implication meaning, the syntactic combination structures between words are tenable.

Secondly, we can start to analyze these syntactic structures from the word "Apple" as the modified word.[14] According to the model of the spherical structure of lexical meaning, we know that the referent meaning of the word "Apple" is a dominant indication meaning. Including the attribute meanings "color, quality, taste", etc., the "Apple" has many attribute implication meanings in its spherical structure of lexical meaning. And these attribute implication meanings all have at least one feature-value implication meaning, and these feature-value implication meanings may well are the feature-value meanings "Red, Bad, Sour", and so on. Therefore, the attribute of "Red" as the feature-value indication meaning is the same with the attribute of "Apple" as the referent indication meaning. That is to say, the attribute implication meaning of Red is included in the attribute implication meaning of "Apple". In addition, "Red, Bad, Sour", etc., as the feature-value indication meanings can play the part of the feature-value implication meaning of "Apple". Therefore, based on the consistent of the attribute implication meanings and the consistent of the feature-value indication meaning and the feature-value implication meaning, the syntactic combination structures between words are tenable.

In short, in the syntactic structure, the feature-value indication meaning of the word "Red" is highlighted, and its attribute meaning and referent meaning are implicated. While, in the syntactic structure, the referent indication meaning of the word "Apple" is highlighted, and its attribute meaning and feature-value meaning are implicated. As a result, as long as the spherical structure of lexical meaning of the word "Red" and the spherical structure of lexical meaning of the word "Apple" can intersect, which can guarantee that the syntactic structure produced by the composition of the words "Red" and "Apple" is able to survive. When the spherical structures of lexical

meaning of the words “Red” and “Apple” intersect, there have three intersections produced: the intersection of attribute implication meaning and attribute implication meaning; the intersection of feature-value indication meaning and feature-value implication meaning; the intersection of referent indication meaning and referent implication meaning. So, when we connect linearly these three intersections, it can generate two kinds of structural combination, one is the syntactic structure combination of (1)-(3) as above examples, while another is the syntactic structure combinations of (1a)-(3a) as below examples.

- (1a) 苹果红 （Pingguo hong ; Apple red）
- (2a) 苹果坏 （Pingguo huai ; Apple bad）
- (3a) 苹果酸 （Pingguo suan ; Apple sour）

Taking “Red apple” and “Apple red” as examples, we illustrate the above analysis and conclusions as follows:

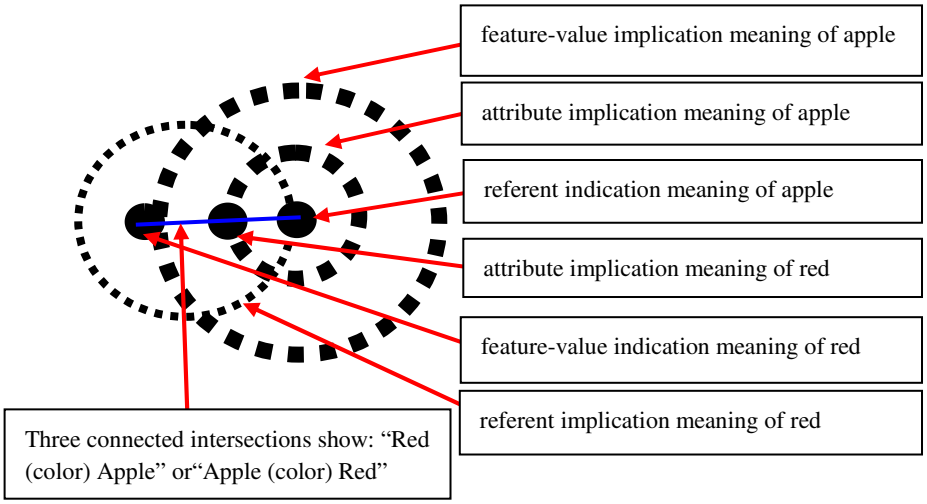


Fig. 3. The schematic diagram of syntactic structure generation of "Red apple" and "Apple red"

Can be seen from Fig.3, two thick black dotted circles denote the feature-value implication meaning and the attribute implication meaning of the word "Apple"; the solid centre of two thick black dotted circles denotes the referent indication meaning of the word “Apple”; and the spherical structure of lexical meaning of the word “Apple” consists of two thick black dotted circles and the solid centre of the circles. The thin black dotted circle represents referent implication meaning of the word “Red”, the referent indication meaning of the word “Apple” and the thin black dotted circle intersect to form a solid point, and this solid point can be seen as the referent indication meaning of the word “Apple” as well as can be seen as the referent implication

meaning of the word “Red”. A solid point on the attribute implication meaning of the word “Apple”, but this point denotes also the attribute implication meaning of the word “Red”. In fact, the attribute implication meaning of the word “Red” should also be a thin black dotted circle, but because the attribute implication meaning of the word “Red” only refers to the attribute “Color”, is unique, so the circle of the attribute implication meaning of the word “Red” can be reduced as a solid point. A solid point on the feature-value implication meaning of the word “Apple”, and this solid point not only is the feature-value indication meaning of the word “Red”, but also is the center of spherical structure of lexical meaning of the word “Red”. As a result, the spherical structure of lexical meaning of the word “Red” intersects the spherical structure of lexical meaning of the word “Apple” at the three solid intersections. The linear connection of the three intersections can generate two syntactic structures of “Red apple” (connecting from left to right) and “Apple red” (connecting from right to left).

5 Conclusions

The above analyses show that if the linear arrays of two words can be combined into a syntactic structure, the spherical structures of lexical meaning of these two words must be able to intersect. The method of mutual intersect is very simple, that is, three elements of the spherical structures of lexical meaning of one word must have at least one element can become the elements of the spherical structures of lexical meaning of another word. As long as there is one element of the spherical structure of lexical meaning of one word can enter the spherical structures of lexical meaning of another word, the spherical structures of lexical meaning of two words can intersect, and thus can generate a tenable syntactic structure.

In the process of language communication, we believe that a spherical structure of lexical meaning must continue to become specific objective, and the semantic information can be generated and be certain. The specific objective of the spherical structure of lexical meaning depends on the specific objective of the elements of the spherical structure of lexical meaning. The only method of the specific objective of the spherical structure of lexical meaning is the expansion of the spherical structure. That is to say, there have more spherical structures to intersect and enter mutually and continually, and until the composite syntactic structure is longer and longer. This means that more specific semantic information is passed on. Therefore, the essence of the syntactic structure is the expansion of the spherical structure of lexical meaning, and the spherical structure of lexical meaning is the basis of the generation of the syntactic structure. In the process of the expansion of the spherical structure of lexical meaning, based on the basic principles of abutment, each word all can generate a syntax-semantics interface together with others words.

Acknowledgements. This study is supported by the Youth Project of the National Social Science Foundation of China (Project No.: 12CYY057) and the Youth Project of the Humanities and Social Sciences Foundation of Hubei Provincial Department of Education (Project No.: 2011jytq013).

References

1. Shen, Y.: Syntax-Semantics Interface. Shanghai Education Press, Shanghai (2007) (in Chinese)
2. Pustejovsky, J.: The generative Lexicon. MIT Press, Cambridge (1995)
3. Levin, B., Rappaport, M.H.: Lexical Semantics and Syntactic Structure. In: Lappin, S. (ed.) The Handbook of Contemporary Semantic Theory, pp. 487–507. Blackwell, Oxford (1996)
4. Qiu, Q.S.: The Influence of Lexical Connotation on Sentential Ambiguity——Case Studies on "trousers" and "skirt". In: Proceedings of the 11th Chinese Lexical Semantic Workshop, pp. 97–101. COLIPS Publications, Singapore (2010)
5. Lu, C.: The Parataxis Network of Chinese Grammar, pp. 1–3. The Commercial Press, Beijing (2001) (in Chinese)
6. Zhao, S.J.: Decisive Function of Lexical Meaning on Grammar. Wuhan University Journal (Humanity Sciences) 2, 173–179 (2008) (in Chinese)
7. Liu, C.H.: The Study of Attribute Category of Modern Chinese. Sichuan Publishing Group, Chengdu (2008) (in Chinese)
8. Qiu, Q.S., Wang, X.Y.: The Implication Meaning of Grammatical Category of Chinese Lexical Meaning. Xiangfan University Journal (Social Science) 3, 52–55 (2009) (in Chinese)
9. Qiu, Q.S.: The Implication Meaning of Grammatical Attribute and Function of Chinese Lexical Meaning. Langfang Teachers College Journal (Social Science Edition) 3, 36–39 (2009) (in Chinese)
10. Jackendoff, R.: Semantics and Cognition. MIT Press, Cambridge (1983)
11. Dong, W.G.: Basic Types of Chinese Meaning Development. Huazhong University of Science and Technology Press, Wuhan (2004) (in Chinese)
12. Shen, K.M.: The Implication of Lexical Meaning. Chinese language Learning 5, 40–44 (1983) (in Chinese)
13. Fillmore Charles, J.: Frame Semantics and the Nature of Language. Annals of the NY Academy of Sciences 280, 20–32 (1976)
14. Fu, H.Q.: The Analysis and Description of the Lexical Meaning. Language & Culture Press, Beijing (1996) (in Chinese)

Cross-Linguistic Perspectives on Event Structure in Chinese下去xia qu(down-go) and Its English Equivalents

Ling Zhao[1] and Weidu Xiong[2]

[1] School of Foreign Languages and Literature, Wuhan University, 430072 Hu Bei, P.R. China
lingzhao2006@126.com
[2] School of Foreign Languages and Literature, South-central University for Nationalities, 430072 Hu Bei, P.R. China
kumane@163.com

Abstract. 下去xia qu(down-go) is a motion verb commonly discussed by many researchers. When it comes to cross-linguistic perspective, different types of motion event expressing received much attention previously, however, 下去xia qu was not detailed as an individual in the event structure. This essay investigates Chinese 下去xia qu and its English equivalents in event-driven structure from the angle of conceptual fundamental element on the basis of the past works. The head of a sentence is 下去xia qu or its English equivalents used to assemble an event together with entity. It holds that the syntactic change of 下去xia qu and its construction is projected from deep event conceptual structure, which mainly concerns agent and source. When the same event structure is described in its English equivalents' construction, it presents us with different syntactic forms.

Keywords: event, agent, starting point, equivalent, motion.

1 Introduction

In prototypical sense, motion is understood as a change of location of an object with respect to other object(s) successively from one point to another along a spatial extent over a period of time. Various researchers were and will be curious about motion event expression. They have put much energy into categorizing and subcategorizing those motion verbs of different languages.

1.1 Previous Work

Motion verb and entity (noun or nominal phrase) constitute motion event. The cognitive components of motion event structure are universal [1]. English and Chinese are not expressed in the same way.

下去xia qu(down-go) is a motion verb commonly discussed in many essays. When it comes to cross-linguistic perspective, motion event expressing has got much attention in each category, however, 下去xia qu is not studied in detail as an individual.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 532–539, 2013.

1.2 Our Work

This paper investigates Chinese 下去xia qu and its English equivalents in event-driven structure from the angle of conceptual fundamental element on the basis of the past works. The head of a sentence is 下去xia qu or its English equivalents used to assemble an event together with entity. It holds that the syntactic change of 下去xia qu and its construction is projected from deep event conceptual structure, which mainly concerns agent and source. When the same event structure is described in its English equivalents' construction, it presents us with different syntactic forms.

2 Starting Point

GuoZheng Xiao claimed that mastery of a language consists, inter alia, in mastery of a conceptual structured inventory of form-meaning pairs [3]. An event structure is a conceptual structure, just like a graph in which vertices represent different semantic uses or functions, and edges connect closely related uses. It is assumed that the graph structure is universal, and that language-specific categories always pick out connected sub-graphs of the universal graph. An event structure thus compactly represents what patterns of variation one may and may not expect to find in a given event domain, in terms of presumptively universal conceptual structure. So, we have to discuss fundamental elements in the event structure first in order to compare Chinese 下去xia qu and its English equivalent.

The object moves along the path. There is no question that Chinese can use Path Verbs alone to express motion events. Besides that, Chinese has Path Particles coexisting in the contemporary event expressions. Although Path Particles did grammaticalize from Path Verbs, this does not preclude them from belonging to a distinct category. 下去xia qu has Path Verb 去qu and Path Particle 下xia.

What we put forward next is the conceptual elements encoding of 下去xia qu in the motion event structure. Verb of motion 下去xia qu shows the movement is away from the speaker. It can be used as complex directional complement indicates a certain compound direction of the action, and gives a more specific description of the action.

(1) 下 去！
xia qu
down go
'go down'

Sentence 1 indicates a certain compound direction, which means: the agent both moves downwards and is away from the speaker. In the sentence, there is not any entity around 下去, and it still shows the direction. The starting point is identified by deictic demonstrative pronoun 这儿 zhe'er(here) and the finishing point is demonstrated by 那儿na'er(there). Though the agent and source are hidden, every one knows they can be added. If we add all the missing elements to Sentence 1, we can get Sentence 1':

(1') (你从 这儿) 下 (到 那儿)去！
(*ni cong zhe'er)* **xia** (*dao na'er*) **qu**
(*2s from here)* **down** *(to there)* **go**
('*you*) **go** (*away from here and*) **down (** *to there)*'

In the syntactic environment, the agent 你ni and two prepositional phrases like 从这儿 cong zhe'er and 到那儿dao na'er do not appear, however, they are needed in the deep event conceptual structure and understood from the context.

In the current event structure, motion is treated as a fairly abstract and general structure. The motion structure specifies that some agent starts out in one place (starting point), covers some space and ends up in some other place. The starting point here is the reference point of 下去xia qu, which is the location where the agent (the speaker, the listener, or the third party) occupies initially before it changes its location.

2.1 The First Starting Point

Huyang Qi noticed the Chinese 起点qi dian (starting-point), the starting point in 1998. The starting point usually has something to do with the speaker.

(1) 我 下 去！
wo xia qu
1s down go
'I go down'

In order to describe the situation cognitively, we view the first starting point as the position of the speaker.

2.2 The Second Starting Point

The starting point has a close relation to the listener.

(2) 你下 去！
ni xia qu
2s down go
'you go down'

We view the second starting point as the position of 你ni（you）, the listener.

2.3 The Third Starting Point

The starting point sometimes has nothing to do with the speaker and the listener:

(4) 李 明 下 去！
Li Ming xia qu

Li Ming down go
'Ming Li goes down'

In the sentence, 李明Li Ming is neither the speaker nor the listener. We view the third starting point as the position of the third party.

3 Motion

Talmy divides different languages into two categories, namely verb-framed languages and satellite-framed languages [7]. The former render PATH through the verb, whereas the latter express it by a particle. Chinese can not be simply labeled into these two categories, since it needs not only verb but also particle.

3.1 Voluntary Motion

When we add other verbs before 下去xia qu, we take 下去xia qu as the head while considering this case. The moving in this class is voluntary motion, which means an agent moves in a particular manner.

(5) 他跑 了 下 去。
ta pao le xia qu
3s run-PT down go (PT= past tense)
'he ran down'

跑下去pao xia qu lacks a spatial boundary, and is temporally non-bounded, since there is no co-occurrence syntactically. The whole clause is marked as bounded temporally by the past tense marker 了le. The starting point and the finishing point are both explicit. 跑pao (run) indicates the way of moving.

(6) 我 走 下 去。
wo zou xia qu
1s walk down go
'I walk down'

Similar to 跑pao, 走zou (walk) is also showing the way of moving. All the verbs before 下去xia qu in this category belong to mono-valence verbs with one entity, which needs only the agent.

3.2 Caused Motion

When we analyze the other group of verbs put before 下去xia qu, apart from the group of manner verbs, we categorize the other group as reason verb. The moving in this class is caused motion, which shows the reason that causes the moving of the agent:

(7) 他 放了 下 去。
ta fang le xia qu
3s put-PT down go (PT = past tense)
'he put down '

All the verbs before 下去xia qu in this category belong to bi-valence verbs with two entities. In Sentence (7), the agent here is animate being. The object is explicit. The agent causes the object to move down.

(8) 东西 吃了 下 去。
dong xi chi le xia qu
thing eat-PV-PT down go (PV= passive voice, PT = past tense)
'the thing was eaten '

In Sentence (8), the subject is not the agent and the agent here is explicit. The subject is moving because the agent causes it to move down. The subject indicates some non-intentional, typically non-human force.

4 Starting Point of 下去xia qu's English Equivalent

下去xia qu's English equivalent is "go down" in Collins Co-build English-Chinese Dictionary (2002) and Oxford Advanced Learners English-Chinese Dictionary (2004).

下去 xia qu means "moving downward and lower" [9]. Different from Chinese qu, "go" shows no direction.

(9) Once she did go down with him to the lands.
(10) Wear your safety helmet when you go down.

In Sentence (9), "down" and "to" are path particles. In Sentence (10), "down" is the path particle and shows the direction.

Chinese下去xia qu behaves like verbal phrase while下xia acts as a Path Verb and 去qu serves as a Path Particle. The下去xia qu's English equivalent, "go down", is a verbal phrase and only "down" is a Path Particle. It indicates a certain compound direction, which means: the agent both moves downwards and is away from the speaker.

(11) Go down and see who is at the door, please.

In the Sentence (11), there is not any entity around "go down", and it still shows the direction. The starting point is "here" and the finishing point is "there". Though the agent and the source are hidden, Sentence 11 is well-known that the sentence can be understood like Sentence 11':

(11') (you) **Go (** *away from here and*) **down (** *to the door)* and……

The italic words can be added according to the deep event structure while the bold words represent the syntactic form.

If only the finishing point appears, “to” is the marker of the finishing point. When the starting point and the finishing point both appear, “from” and “to” are the markers without “down”. The former indicates the starting point while the latter marks the finishing point:

(12) I want a rope that will go from the top window to the ground.

In Sentence (12), source and goal become obvious because of path particles. A different situation is reported in Chinese, which also uses path verb. In that case, if we do not analyze the motion event represented in the syntactic structure, it is not easy for us to access source and goal.

4.1 The First Starting Point

The same as Chinese, the starting point in English usually has something to do with the speaker:

(13) I’ll go down and answer the door!

In order to describe the situation cognitively, we also view the first starting point as the position of the speaker.

4.2 The Second Starting Point

The starting point has a close relation to the listener:

(14) You go down and see who it is!

Since the event structure is universal, we view the second starting point as the position of the listener.

4.3 The Third Starting Point

The starting point sometimes has nothing to do with the speaker and the listener:

(15) A cat goes from the top window to the ground.

In the sentence, “A cat” is neither the speaker nor the listener. We view the third starting point as the position of the third party. In this way, English and Chinese share the same motion event.

5 Motion

English shows a number of similarities to Chinese, however differs from Chinese in its specific ways of constructing many events syntactically and expressing their potential, and this is specially the case if we are dealing with event of motion. Chinese has been classifies by Talmy as a satellite-framed language, in which the satellites correspond to what are usually called "directional complements" [7]. However, as was pointed out by Tai[6] and Lamarre [5], Chinese does not fit very well in this category, which renders PATH through the particle. Chinese expresses Path by both verb and particle. Talmy uses the notion of "event-frame", which is constituted by "a set of conceptual elements and relationships that are evoked together or co-evoke each other" [8]. Motion is one of the central components.

5.1 Voluntary Motion

Another下去xia qu's English equivalent is "descend" [4], which needs no path particle in Sentence 16.

(16) It is easier to descend a mountain than to climb up it.

When we add other verbs before下去xia qu, the whole verbal phrase's English equivalents give us interesting evidence.

All the verbs before下去xia qu in this category belong to intransitive or mono-valence verbs with an entity. The agent here is lived being in Sentence 17.

(17) Let's rush down to the Exit.

跑下去pao xia qu's English equivalent is "rush down", in which下去xia qu is only represented by "down". We treat the event as a self-agentive or an autonomous motion event.

5.2 Caused Motion

We still take the下去xia qu as the head considering this case. All the verbs before下去xia qu in this category belong to transitive or bi-valence verbs with two entities.

When we analyze the other class of verbs put before下去xia qu, we categorize them as the class of reason. The moving in this is caused motion, which illustrates the reason that causes the moving.

(18) The fish was eaten.

吃下去了chi xia qu le's English equivalent is " was eaten". 下去xia qu here has no equivalent and there is no path word.

The syntactic location of an agent in the caused motion is not fixed:

(19) I ask him to eat the fish.
(20) He has the fish to eaten.
(21) He eats the fish.

In caused motion events, the agent may have several positions. Situation 1, the agent is put after the verb if it is in indefinite or generic structure as Sentence 19. Situation 2, the agent is put before the verb, and most of the time it is introduced by a marker 把ba if it is identifiable as Sentence 20. Situation 3, the agent is the subject in passive voice as Sentence 21.

6 Conclusion

Chinese and English express motion event structure differently [2]. We contend that Chinese 下 去 xia qu forms a semantic nucleus and share the same conceptual fundamental elements as their basic meaning with its English equivalents belonging to different categories. The syntactic change of下去xia qu and its English equivalents is projected from deep event conceptual structure.

In voluntary motion, Chinese下去xia qu has path verb and path particle. Some Chinese下去xia qu are translated into English path particles and others are interpreted as path verb. Even some Chinese下去xia qu is equal to "ϕ" in its English equivalent. In non-agentive motion, we note that in the cases when an inanimate agent appears before the verb and the verb phrase lacks any overt passive or causative marker in Chinese so we have to add the marker.

References

1. Berman, R., Dan, S.: Relating Events in Narrative: A Cross-linguistic Developmental Study. Lawrence Erlbaum Associates, Hillsdale New Jersey (1994)
2. Evans, V., Melanie, G.: Cognitive Linguistics: An Introduction, p. 157. Edinburgh University Press, Edinburgh (2006)
3. Xiao, G.Z., Xiao, S., Guo, T.T.: Mapping from Conceptual Gene to Semantic Fundamental Element. Journal of East China Normal University 43(1), 139–143 (2011)
4. WordNet, http://wordnet.princeton.edu/
5. Lamarre, C.: Directional Introducing Locative Phrases in Chinese. In: Proceedings of the 54th Annual Conference of the Chinese Linguistics Association of Japan, pp. 100–104. Kyoto University, Japan (2004)
6. Tai, H.Y.: Cognitive Relativism: Resultative Construction in Chinese. Language and Linguistics 4(2), 301–316 (2003)
7. Talmy, L.: Lexicalization Patterns: Semantic Structure in Lexical Forms Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon, pp. 36–149. Cambridge University Press, Cambridge United Kingdom (1985)
8. Ungerer, F., Schmid, H.J.: An Introduction to Cognitive Linguistics. Longman, London (1997)
9. Ji, Y.L., Henriette, H., Maya, H.: Children's Expression of Voluntary Motion Events in English and Chinese. Journal of Foreign Languages 34(4), 2–20 (2011)

YONG 用 as a Pro-verb in Taiwan Mandarin

Meichun Liu[1] and Ruiliang Xu[2]

[1] National Chiao Tung University,
Department of Foreign Languages and Literatures, Hsinchu, Taiwan
mliu@mail.nctu.edu.tw
[2] National Chiao Tung University,
Department of Foreign Languages and Literatures, Hsinchu, Taiwan
b93101048.flg98g@g2.nctu.edu.tw

Abstract. This study investigates the semantics and grammatical function, as well as the possible source and development of the emerging use of *YONG* as a pro-verb in Taiwan Mandarin. In the [*YONG* + NP] construction, the NP must be a referring Patient object which is directly affected by the physical action designated by the "replaced verb". This pro-V *YONG* may be a result of change induced by language contact with Taiwan Southern Min dialect. This change may involve relexification, grammaticalization or degrammaticalization, depending on the actual path of development of the pro-V *YONG*.

Keywords: pro-verb, co-verb, diachronic construction grammar, relexification, contact-induced grammaticalization, degrammaticalization, Taiwan Mandarin, Taiwan Southern Min.

1 Introduction

YONG (用) is a commonly used multi-function verb in Mandarin Chinese. There are different kinds of usage of *YONG*:

(1) 我可以用你的電腦嗎？
wo keyi yong nide diannao ma
I can YONG your computer QM.[1]
'Can I use your computer?'

(2) 我用湯匙喝湯。
wo yong tangchi he tang
I YONG spoon drink soup
'I use a spoon to drink the soup.'
= 'I drink the soup with a spoon.'

As shown in the examples above, *YONG* has the typical reading of 'use' in Mandarin Chinese. In (1), *YONG* is used as a main verb, while in (2), it is a co-verb introducing

[1] QM. is the abbreviation of Question Marker.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 540–550, 2013.

the Instrument *tangchi* 'spoon' (湯匙) for doing the main event *he tang* 'drinking the soup' (喝湯) of this sentence. However, there exists a quite different usage of *YONG* in Taiwan Mandarin (TM), as following:

(3) A:你可以幫我修車嗎？
A: *ni keyi bang wo xiu che ma*
you can help I fix car QM.
'Can you fix the car for me?'

B: 好，我來用。
B: *hao, wo lai yong*
OK I LAI YONG
'OK, I'll do it (fix the car).'

In (3), the meaning of *YONG* is not the typical reading of 'use', as that of (1) and (2), rather, it refers to the meaning of *xiuche* 'fix the car' (修車) in the previous question. That is, in (3), *YONG* behaves as a pro-form of the previously mentioned verb in the discourse. In other words, it works as a pro-verb which replaces the forementioned verb in the context. In some other cases, *YONG* can even be used independently without any "replaced verb" mentioned in the context, as in (4).

(4) 我昨天整天都在用報告，超累的。
wo zuotian zhentian dou zai yong baogao, chao lei de
I yesterday all-day all Asp.[2] YONG paper, super tired DE
'I wrote the paper all day long yesterday. I was exhausted.'

In (4), there is no any "replaced verb" in the previous or following context for *YONG* to refer to. The reading 'write' of *YONG* can only be obtained via pragmatic inference.

In sum, given the examples above, we can figure out that a special kind of *YONG* exists in TM. Thus, we may ask the following questions: what is unique about this special type of *YONG*? What are the semantics and grammatical function of this *YONG* in TM? What are the differences between this pro-verb *YONG* and other types of *YONG* (main V and co-V)? Why and how does this pro-verb *YONG* come into being in TM? Is this new usage a result of grammatical change? If it is, what kind of change it may be? Is it grammaticalization, degrammaticalization, or other types of change? These questions raise our interest in investigating this issue. Thus, in this current study, we try to deal with this issue and give possible explanation to these questions.

[2] Asp. is the abbreviation of Aspect Marker.

2 Approach

2.1 Theoretical Framework

Diachronic Construction Grammar [2], [5] is adopted as the main theoretical framework to realize the related patterns of *YONG*, since it holds that "change in grammatical organization can be adequately articulated only as a gradual conventionalization of patterns of understanding, in which morphosemantic structure, syntactic function, communicative function, and lexical meaning form an integrated whole." [2].

Besides, the Grammaticalization [1], [3] and Relexification [4] theories will be the theoretical base utilized to examine and account for the development and evolution of the pro-V *YONG* in Taiwan Mandarin.

2.2 Database

Due to the colloquial feature of pro-V *YONG* as being mostly used in casual speech, the main database utilized here is The NCCU Corpus of Spoken Chinese. Other internet resources such as personal blogs (e.g. Yahoo Blog, Wretch Blog), on-line search engines (e.g. Google Search), and the Bulletin Board Systems site (PTT), as well as about 4.5 hours of daily collected conversations are used to obtain more casual speech-like corpus.

3 Analysis

In the current study, [*YONG* + NP] construction is investigated.

(5) 他很喜歡用活動。
ta hen xihuan yong huodong
he very like YONG activity
'He likes holding activities.'

(6) 不要用垃圾在這裡。
buyao yong lese zai zheli
do-not YONG garbage at here
'Do not put/throw garbage here.'

YONG in the examples above has the function of pro-verb. The verb replaced by *YONG* indicates the Means or Manner used to fulfill the event in the clause. Thus, in (5), *YONG* may refers to an absent verb *banli* 'transact' (辦理) or *juxing* 'hold' (舉行); in (6), *YONG* may refers to an absent verb *baifang* 'put' (擺放) or *diu* 'throw' (丟). Note worthily, the object (*huodong* 'activity' (活動) in (5), and *lese* 'garbage' (垃圾) in (6)) of *YONG* must be the Patient of the replaced verb. That is, this replaced verb must have strong direct effect on the object. Whether the object is directly affected by the replaced verb depends on the discourse function as a judge. Consider the following examples:

(7) a. 他一直都在搞教育/搞破壞。
ta izhi dou zai gao jiaoyu/gao pohuai
he always all Asp. GAO education/ GAO destruction
'He has always been engaged in education/doing destruction.'

b. *他一直都在用教育/用破壞。
**ta izhi dou zai yong jiaoyu/yong pohuai*
he always all Asp. YONG education/ YONG destruction
'He has always been engaged in education/doing destruction.'

GAO (搞) is another well-known and commonly used pro-verb in Mandarin. In (7), the word *jiaoyu* 'education'(教育) and *pohuai* 'destruction' (破壞) are actually type referring in the discourse, not the directly affected individual objects by *GAO* or *YONG*. In other words, the two words *jiaoyu* and *pohuai* do not refer to a realized entity in the discourse. Therefore, they can not be construed with *YONG*.

As to the syntactic and semantic constraints on the "replaced verb" of pro-V *YONG,* since the replaced verb must have strong direct effect on the Patient object, it should basically be a transitive verb designating a physical action. The sentence pattern *ta yong de* 'He does/did it.' (他用的) can be utilized as a method to test this syntactic semantic property, as shown in the following examples:

(8) Prototypical Physical Transitive Verb

a. A: 誰摔破了盤子？
shei shuaipo le panzi
who break Asp. dish
'Who broke the dish?'

B: 他用的。
ta yong de
he YONG DE
'He did/broke it.'

b. A: 誰在敲桌子？
shei zai qiao zhuozi
who Asp. knock table
'Who is knocking on the table?'

B: 他用的。
ta yong de
he YONG DE
'He does/knocks it.'

(9) Stative Verb

a. A: 誰有這本書呢？
shei you zheben shu ne
who have this-CL[3] book QM.
'Who has this book?'

B: *他用的。
**ta yong de*
he YONG DE
'He does/has it.'

b. A: 誰需要這枝筆？
shei xuyao zhezhi bi
who need this-CL pen
'Who needs this pen?'

B: *他用的。
**ta yong de*
he YONG DE
'He does/needs it.'

[3] CL is the abbreviation of Classifier.

(10) Mental Activity Verb

a. A: 誰想要這份禮物？
shei xiangyao zhefen liwu
who want this-CL gift.
'Who wants this gift?'

B: *他用的。
**ta yong de*
he YONG DE
'He does/wants it.'

b. A: 誰思考過這問題？
shei sikao guo zhe wenti
who think Asp. this question
'Who had thought about this question?'

B: *他用的。
**ta yong de*
he YONG DE
'He had done/thought about it.'

c. A: 誰喜歡蘋果？
shei xihuan pingguo
who like apple
'Who likes apples?'

B: *他用的。
**ta yong de*
he YONG DE
'He does/likes it.'

(11) Perception-Cognition-Utterance (PCU) Verb

a. A: 誰聽過這首歌呢？
shei tingguo zheshou ge ne
who hear Asp. this-CL song QM.
'Who ever heard this song ?'

B: *他用的。
**ta yong de*
he YONG DE
'He did/heard it.'

b. A: 誰知道這個祕密呢？
shei zhidao zhege mimi ne
who know this secret QM.
'Who knows this secret?'

B: *他用的。
**ta yong de*
he YONG DE
'He does/knows it.'

c. A: 誰懷疑你的清白呢？
shei huaiyi nide qingbai ne
who suspect your innocence QM.
'Who suspects your innocence?'

B: *他用的。
**ta yong de*
he YONG DE
'He does/suspects it.'

d. A: 誰害怕蟑螂呢？
shei haipa zhanglang ne
who fear cockroach QM.
'Who fears cockroaches?'

B: *他用的。
**ta yong de*
he YONG DE
'He does/fears it.'

(12) Manipulative Verb

a. A: 誰要求你道歉呢？
shei yaoqiu ni daoqian ne
who ask you apologize QM.
'Who asked you to apologize?'

B: *他用的。
**ta yong de*
he YONG DE
'He did/asked it.'

b. A: 誰命令你離開呢？
shei mingling ni likai ne
who order you leave QM.
'Who ordered you to leave?'

B: *他用的。
**ta yong de*
he YONG DE
'He did/ordered it.'

(13) Copula Verb

a. A: 誰是老師呢？
shei shi laoshi ne
who is teacher QM.
'Who is the teacher?'

B: 他用的。
ta yong de
he YONG DE
'He is.'

b. A: 誰變成了新娘呢？
shei biancheng le xinniang ne
who become Asp. bride QM.
'Who became a bride?'

B: 他用的。
ta yong de
she YONG DE
'She did/became it.'

Based on the observation above, we can draw out an interim summary: *YONG* is a pro-verb which does not directly provide any event information. The verb replaced by *YONG* indicates the Means or Manner used to fulfill the event in the clause. This replaced verb must cause direct strong effects on the Patient object. In other words, the replaced verb is basically a transitive verb coding a physical event. Whether this verb has a direct strong effect on the object NP or not relies on the discourse context as a judge.

4 Discussion

Since in traditional Mandarin Chinese, there are only two verbal usages of *YONG*: main V and co-V, no pro-V usage. Then, we may ask a question: what is the source of this new use in Taiwan Mandarin? In Taiwan Southern Min dialect (TSM), there exist three kinds of verbal usage of *YONG*: main V, co-V and pro-V. Consider the following examples[4]:

[4] The transcription is presented in both Taiwanese Romanization, a commonly used spelling system of TSM, and in Taiwanese Han-character system, a widely used writing system of TSM in Chinese characters.

(14) main V
我這陣欲用車。
goa chit-chun beh iong chhia
I now would-like YONG car
'I would like to use the car now.'

(15) co-V
我用鎖匙開門。
goa iong sosi khui mng
I YONG key open door
'I use a key to open the door.'
(= 'I open the door with a key')

(16) pro-V
伊佮碗用置土腳。
i ka oa^{n} iong ti thokha
he Disp.[5] bowl YONG on floor
'He drops the bowl on the floor.'

The possible source of the pro-V *YONG* may be from TSM. The long period of language contact between Mandarin Chinese and TSM may result in grammatical change of *YONG* in TM. This contact-induced language change can be viewed and discussed from several perspectives. In the coming sections, the emergence of pro-V *YONG* in TM will be explored in light of these perspectives.

4.1 Relexification

Relexification is a kind of "relabeling" which is "a mental process that builds new lexical entries by copying the lexical entries of an already established lexicon and replacing their phonological representations with representations derived from another language" [4]. According to Lefebvre [4], relexification involves the following procedures: given a lexical entry in L_1 such as (17a), assign this lexical entry a second phonological representation drawn from another language L_2, yielding (17b). The original phonological representation is eventually abandoned. The resulting lexical entry in (17c) thus has the semantic and syntactic properties of the original lexical entry in L_1 and a phonological representation derived from the form in L_2.

(17)

a.	b.	c.
/phonology/$_i$ [semantic]$_i$ [syntactic]$_i$	/phonology/$_i$ /phonology/$_{j'}$ [semantic]$_i$ [syntactic]$_i$	/phonology/$_{j'}$ [semantic]$_i$ [syntactic]$_i$

[5] Disp. is the abbreviation of Disposal Marker.

In this respect, the current *YONG* in TM today may be an item which has the phonological representation of Mandarin Chinese *YONG* and the syntactic and semantic properties of the TSM *YONG*. That is, a new entry *YONG* is built in TM by copying the TSM *YONG* and replacing its phonological representations with that of Mandarin Chinese. Thus, the new pro-V use can be obtained in TM. This process can be schematized as the following:

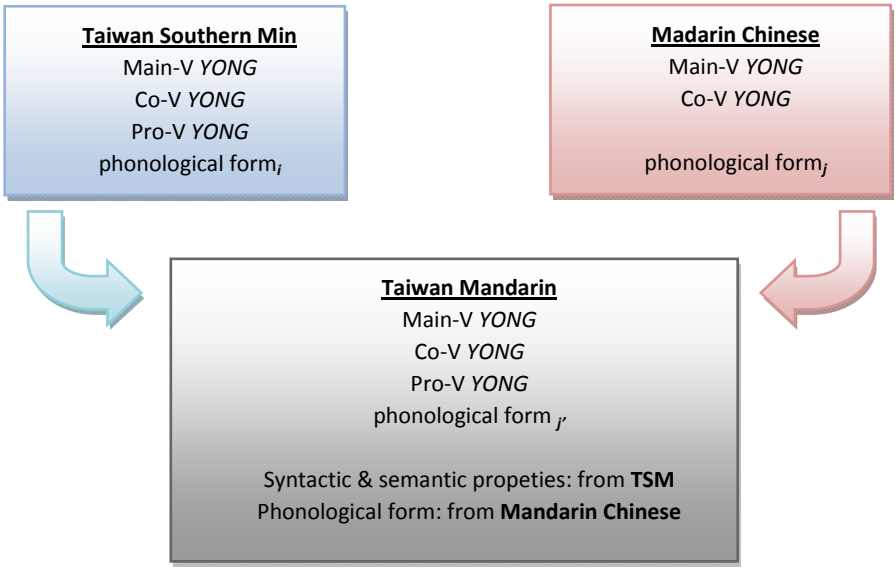


Fig. 1. Schema of relexification of pro-V *YONG* in Taiwan Mandarin

4.2 Contact-Induced (De-)grammaticalization

From a perspective of contact-induced language change, another question may be raised: what are the relations between pro-V *YONG* and other verbal usages of *YONG*? Is the emerging pro-V *YONG* in TM a result of grammatical change? To deal with this question, we can compare the three usages of *YONG* in terms of their degree of grammaticality. The main V *YONG* has its specific contentful meaning 'use, take advantage of', and behaves as a typical transitive verb. The co-V *YONG* works as a preposition introducing the Means or Instrument used to fulfill the main event in the sentence. The pro-V *YONG* refers to a certain "replaced verb" which has direct effect on the object. Based on the initial observation above, our focus will be on the comparison between the function-word-like co-V and pro-V use of *YONG*.

Brinton and Traugott defined grammaticalization as "the change whereby in certain linguistic contexts speakers use parts of a construction with a grammatical function. Over time the resulting grammatical item may become more grammatical by acquiring more grammatical functions and expanding its host-classes" [1]. Based on this

tenet, they claim that grammaticalization is a process which will lead to increasing productivity[6] and decategorization.

According to Brinton and Traugott [1], increasing productivity refers to increasing type frequency and host-class expansion[7]. Examining the [*YONG* + NP] pattern, almost all kinds of NP can be construed with co-V *YONG* and once it applies to this pattern, it will become an Instrument or Means for fulfilling the main event in the clause. However, only Patient NP can be construed with pro-V *YONG*. Moreover, this NP cannot be a non-referring entity. The following examples illustrate this:

(18) a. co-V + concrete NP
我用自行車上學。
wo yong zixingche shangxue
I YONG bike go-to-school
'I use a bike to go to school'
(= 'I go to school by bike.')

b. co-V + deverbal noun
我用走路上學。
wo yong zoulu shangxue
I YONG walk go-to-school
'I use walking to go to school'
(= 'I go to school by walking.')

c. co-V + non-referring NP
政府用教育提升人民知識。
zhengfu yong jiaoyu tisheng renmin zhishi
government YONG education improve people knowledge
'The government use education to improve people's knowledge.'
(= 'The government improves people's knowledge with education.')

(19) pro-V + non-referring NP
*他用教育很久了
**ta yong jiaoyu hen jiu le*
he YONG education very long Asp.
'He has been engaged in education for so long.'

In (18), the concrete NP *zixingche* 'bike' (自行車) and the abstract NP *jiaoyu* 'educaton' (教育) are interpreted as Instruments, while the deverbal noun *zoulu* 'walk'

[6] Brinton and Traugott [1]: "Items that grammaticalize become more productive in the sense that the grammaticalizing element occurs with increasingly large numbers of categories."

[7] Himmelmann [3]: "For example, when demonstratives are grammaticized to articles they may start to co-occur regularly with proper names or nouns designating unique entities (such as *sun*, *sky*, *queen*, etc.), i.e. nouns they typically did not co-occur-with before. This context-expansion could be called host-class expansion."

(走路) as the Means. However, in (19), the abstract non-referring NP *jiaoyu* can never be interpreted as a Patient, since it does not refer to an entity which undergoes direct effect by the action designated by the pro-V *YONG*. Thus, this sentence is ungrammatical. (18-19) indicates that co-V *YONG* has wider range of host-class and thus higher degree of productivity than pro-V *YONG* does.

Brintion and Traugott [1] cited Hopper's definition of decategorization as the process by which forms "lose or neutralize the morphological markers and syntactic privileges characteristic of the full categories Noun and Verb, and … assume attributes characteristic of secondary categories such as Adjective, Participle, Preposition, etc." To test the degree of categoriality of co-V and pro-V *YONG*, the [YONG + Result] pattern, which typically applies to main verbs, would be a useful tool.

(20) main V
為了實驗成功，他用壞了三台電腦。
weile shiyan chenggong ta yong huai le santai diannao
for experiment successful he YONG out-of use Asp. 3-CL computer
'For the successfulness of the experiment, he overused three computers and caused them out of use.'

(21) co-V
*他用壞了自行車上學。
**ta yong huai le zixingche shangxue*
he YONG broken Asp. bike go-to-school
'He goes to school by using and breaking the bike.'

(22) pro-V
他用壞了自行車。
ta yong huai le zixingche
he YONG broken Asp. bike
'He broke a bike.'

As (20-22) shows, main V and pro-V *YONG* can be construed with the resultative complement *huai* 'broken, out of use' (壞), but co-V *YONG* can not. This means that co-V *YONG* is less syntactically similar to main V *YONG* than pro-V *YONG* is. It shows that pro-V *YONG* is more like a prototypical verb than co-V *YONG* is. Thus, co-V *YONG* is more decategorized than pro-V *YONG*.

In sum, in terms of the two main criteria for grammaticatization - degree of productivity and decategorization, the degree of grammaticalization of the three kinds of *YONG* would be: co-V >> pro-V >> main V. Though the actual path of the development of pro-V *YONG* in TM is not clear yet, from these initial observations a tentative hypothesis can be made: If the pro-V *YONG* in TM is derived from main V *YONG*, it would be a result of grammaticization; if it is derived from co-V *YONG*, it would be a result of degrammaticalization.

5 Conclusion

The emerging use of *YONG* as a pro-verb in Taiwan Mandarin may result from relexification of the pro-V *YONG* in Taiwan Southern Min dialect pertaining to a contact-induced language change with TSM. This change may be viewed as either grammaticalization or degrammaticalization, depending on the actual path of the development of the pro-V *YONG* in TM. At this current stage, the pro-V *YONG* may only replace verbs of physical action and the object NP of pro-V *YONG* is limited to the Patient-Object of such an action. But further development is likely to happen given the increasing usage of the pro-V *YONG*. The case under study here presents an interesting scenario where an already existing verb is re-assigned with a new function, giving rise to a novel form-meaning pairing in the language system.

References

1. Brinton, L.J., Traugott, E.C.: Lexicalization and Language Change. Cambridge University Press, Cambridge (2005)
2. Fried, M.: Constructions and constructs: mapping a shift between predication and attribution. In: Bergs, A., Diewald, G. (eds.) Constructions and Language Change, pp. 47–80. Mouton de Gruyter, Berlin (2008)
3. Himmelmann, N.: Lexicalization and Grammaticization: Opposite or Orthogonal? In: Bisang, W., Himmelmann, N.P., Wiemer, B. (eds.) What makes Grammaticalization? A Look from its Fringes and its Components, pp. 21–44. Mouton de Gruyter, Berlin (2004)
4. Lefebvre, C.: The Contribution of Relexification, Grammaticalisation, and Reanalysis to Creole Genesis and Development. Studies in Language 33(2), 277–311 (2009)
5. Traugott, E.C.: The grammaticalization of NP of NP patterns. In: Bergs, A., Diewald, G. (eds.) Constructions and Language Change, pp. 23–46. Mouton de Gruyter, Berlin (2008)

On Computing Multi-predicate Sentences in Mandarin

Mengyue Yan[1,2]

[1] College of Chinese Language and Literature, Wuhan University, Wuhan, China
[2] College of Arts, Chongqing Normal University, Chongqing, China
yanmengyue1979@yahoo.cn

Abstract. This paper points out that the multi-predicate sentences in Mandarin, including serial verb construction, telescopicform and verb-copying construction and so on, are combined by simple clauses in discourse. Accordingly in the process of NPL, we can compute all these kinds of multi-predicate sentences by the method of processing the simple clauses which form them respectively beforehand. Therefore we can solve this problem more economically, the idea of this method is based on the theory of dependency grammar and the theory of grammaticalization across clauses.

Keywords: multi-predicate sentences, dependency grammar, grammaticalization.

1 Introduction

At present, multi-predicate sentences in mandarin are usually processed by engineering method and it is obviously effective in some respects. However, based on the analysis of the way by which multi-predicate sentences are formed, this paper suggests another intelligent solution on it. The new method seems more economical and is united in train of thought. So this paper tries to explore the grammar base of these sentences theoretically and the computational program on them, our aim is to assist sentence-computing of multi-predicate sentences in Mandarin.

2 Multi-predicate Sentences Generating Process

2.1 Basic Theory

According to grammaticalization theory [1], all kinds of multi-predicate sentences in Mandarin are formed with simple clauses. By the idea of dependency grammar[2], they can be represented by S (Pn), in which "P" stands for predicates, like verb or adjective, "n" stands for the number of arguments or so called “valence” which belong to P, "S" stands for sentence (here we refer it to multi-predicate sentences). And its value spans from 0 to 3. In some sentences, there are some circumstances constituents, such as time, location, quantity and modality (they are in the brackets of the sentences below) etc, and they are optionally co-exist with P’s arguments. Some simple examples of them as for:

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 551–558, 2013.

(1) P0 （昨天）下雨。
(zuotian) xiayu.
It rained yesterday.
(2) P1 花儿红/他走（一次）。
Huar hong/ ta zou (yici).
The flowers are red./ He went there once.
(3) P2 小明（在教室里）画画。
Xiaoming (zai jiaoshili) huahua.
Xiaoming is drawing in the classroom.
(4) P3 老师（乐意）给每位同学一本书。
Laoshi (leyi) gei meiwei tongxue yiben shu.
The teacher is happy to give a book to every student.

2.2 Sentences Properties

The sentences in Mandarin are very different in semantic meaning and grammatical function. Some sentences are independent, while some others are dependent. The former is not only self-sufficient in semantic and syntactic relations but also never exist by depending on the context. This kind of clauses usually contains a constituent which profiles the terminal or the result of action. [3] But the latter is quite the reverse. Here are two instances of the different kinds:

(5) 他看完那本书了。
Ta kanwan naben shu le.
He has written the book.
(6) 小明大声地笑着。
Xiaoming dasheng de xiaozhe.
Xiaoming is laughing loudly.

Instance (5) is a independent sentence, “看完 (kanwan)” is a causative construction, the verb “完 (wan)” shows the result of the verb “看 (kan)”. However, in (6), the functional word “着 (zhe)” is a progressive aspect marker in Mandarin, so “笑着 (xiaozhe)” is an action with no result or terminal, so (6) is a dependent sentence. It’s necessary to explain the distinctions between independent sentences and dependent sentences, although the distinctions between them are sometimes not absolute. Whether a sentence is independent is to certain extent decided by its context. [4] As for instance (7):

(7) 他有一个弟弟。
Ta you yige didi.
He has a young brother.

When we take (7)as a subsequent sentence, it is self-sufficient, just as in this sentence “我有一个姐姐，他有一个弟弟。(wo you yige jiejie , ta you yige didi.)” But when we take it as an initiative sentence in a certain discourse, sentence (7) is insufficient. Because the focus “弟弟 (didi)” which conveys new information has an attributive“一个” which emphasizes it, so it needs a subsequent sentence to make a

supplementary interpretation for it. Otherwise, it is isolated in semantic and makes people feel that there are still something to be said. Observe instance (8) below:

(8) 他有一个弟弟，（弟弟）在当兵。
Ta you yige didi , (didi) zai dang bing.
He has a young brother and he is a soldier.

Undoubtedly, the clause "他有一个弟弟 (ta you yige didi.)" in (8) is dependent and it highly depends on the subsequent clause "（弟弟）在当兵 ('didi'zai dangbing)" which makes the former clause's meaning independent. It is worth mentioning that subsequent clause is always dependent whatever it is. For example, in sentence (8), "（弟弟）在当兵 ('didi'zai dangbing)" is itself dependent. However in (8'), the subsequent sentence "出了国了 (chule guo le)" is independent.

(8') 他有个弟弟，出了国了。
Ta you ge didi, chule guo le.
He has a young brother and he went abroad.

3 The Forming Process of Multi-predicate Sentences

As we know it, there is not much morphology in the strict sense in Mandarin, so Chinese lacks syntactic constrains in grammatical units relatively. [5]In a certain discourse, when the relation between the initiative clause and the subsequent clause is distorted, they will be combined into one more short and economical form by some certain syntactic operations, just as overlapping, merging, deleting and insertion. [6] Some special sentences in Mandarin, such as serial verb construction, verb-copying construction and telescopicform and so on, are formed in this way.

3.1 Serial Verb Construction

Serial verb construction can be expressed as such a formula: SV1V2V3…Vn. Furthermore, there are no conjunctions between different predicates in it and all verbs in the sentence share the same subject. Serial verb construction are formed through this path: Firstly, the clause SV1, SV2, SV3,…SPn constitute a limited discourse, and all the verbs in it are non-causative. Secondly, the pauses between these clauses are canceled, then speaker logically or reasonably co-ordinate these V1, V2, V3 … Vn together and make them share the same subject [7]. As the result, a serial verb construction sentence comes into being. For example:

(9) 方子盯着王阿姨看。
Fangzi dingzhe wangayi kan.
Fangzi is staring at aunt Wang.
(9') 方子j盯着王阿姨 k，方子j看王阿姨k。
Fangzi j dingzhe Wangayi k, Fangzi j kan Wang ayi k.
Fangzi j is staring at aunt Wang k, Fangzi j looks aunt Wang k.

In other words，(9) is combined by the two clauses in (9'). the procedures of this operation are: step1 : canceling the pause in sentence(9'); step2: deleting "方子（fangzi）" and "王阿姨（wang ayi）" in the subsequent clause (because they are coreferent); Step3: clearing out the trace of operations.

step1: 方子盯着王阿姨，方子看王阿姨.
Fangzi dingzhe wangayi , fangzi kan wangayi.
Fangzi is staring at aunt Wang , Fangzi looks aunt Wang.
↓cancel pause
step2: 方子盯着王阿姨方子看王阿姨
Fangzi dingzhe wangayi fangzi kan wangayi.
Fangzi is staring at aunt Wang Fangzi looks aunt Wang.
↓delete coreferent constituent
Step3: 方子 j 盯着 王阿姨 k ej 看 ek
Fangzi j dingzhe wangayi k ej kan ek.
Fangzi is staring at aunt Wang
↓clear out trace
方子盯着王阿姨看
Fangzi dingzhe wangayi kan.
Fangzi is staring at aunt Wang

If we converse the procedures of these operations by which sentence (9) are formed, we shall compute (9) more intelligently. The procedures are: Firstly, divide serial verb sentence into simple clauses with one verb; And the second, compute them separately; And the third, synthesize them into one sentence. Like in the formula below:

盯着 (方子，王阿姨) ∧ 看 (方子，王阿姨)
Dingzhe (fangzi, wangayi) ∧ kan (fangzi, wangayi)
Stare (fangzi, wangayi) ∧ look (fangzi, wangayi)

In order to simplify procedures, we give an example with only two predicates, like (9). The computing procedures of multi-predicate sentences which contain more than one predicate verb, can be reasoned by analogy according to the procedures of operation in (9).

3.2 Telescopicform

Telescopicform is a special kind of multi-predicate sentences. When one verb in multi-predicate sentences has the property of [CAUSATIVE], in addition, the object of this causative verb (NPO) is at the same time the subject of the second verb (NPS), the two verbs share the same syntactic slot, a telescopicform sentence is formed. The formula is SV1 NPO/S V2, like instance (10).

(10) 这包松子使故事的尾声有了意味。
Zhe bao songzi shi gushi de weisheng youle yiwei.
This packet of pine nuts make the end of the story meaningful.
(10 ') 这包松子使[故事的尾声]j, [ej] 有了意 味。
Zhe bao songzi shi [gushi de weisheng] j, [ej] youle yiwei.
This packet of pine nuts make the end of the story meaningful.

In our model, sentence (10) can be computed on the basis of sentence (10'), according to the procedures of (9). But the first half of the clause "这包松子使故事的尾声 (zhe bao songzi shi gushi de weisheng)" is typically a dependent clause in semantic meaning, so it needs a subsequent clause "[ej] 有了意味([ej] you le yiwei)" to complete it. Because [j] "故事的尾声（gushi de weisheng）" and [ej] are co-referent, when the pause between two clauses is deleted and the trace is cleared out, we get sentence (10). (these parts of shadow mean co-referent)

Step1: 这包松子使故事的尾声，故事的尾声有了意味。
Zhebao songzi shi gushi de weisheng , gushi de weisheng youle yiwei.
This packet of pine nuts make the end of the story be meaningful.
↓cancel pause
Step2: 这包松子使故事的尾声 故事的尾声有了意味。
Zhebao songzi shi gushi de weisheng gushi de weisheng youle yiwei.
This packet of pine nuts make the end of the story be meaningful.
↓delete co-referent constituent
Step3: 这包松子使[故事的尾声] j ej 有了意味。
Zhebao songzi shi gushi de weisheng j ej you le yiwei.
This packet of pine nuts make the end of the story be meaningful.
↓clear out the trace
这包松子使故事的尾声有了意味。
zhebao songzi shi gushi de weisheng you le yiwei.
This packet of pine nuts make the end of the story be meaningful.

The fact needs to be pointed out is that the first clause "这包松子使故事的尾声 (zhebao songzi shi gushi de weisheng)"is not only dependent in semantic but also dependent in syntactic. In fact, from the respect of syntax, the independence of a causative sentence levels difference. That is to say, causative property of different verbs is a continuum. [8] With the decreasing of the causative property of a verb, the independence of SVNP is rising. Make a comparison between (10) and (11) (12), you will find it.

(10)这包松子使故事的尾声有了意味。
Zhebao songzi shi gushi de weisheng youle yiwei.
This packet of pine nuts make the end of the story be meaningful.
→a. ??这包松子使故事尾声。
zhebao songzi shi gushi de weisheng。
This packet of pine nuts makes the end of the story.

(11) 你派人去交涉一下 。
Ni pai ren qu jiaoshe yixia.
You ask someone to negotiate with them.
→ a. 你派人。
Ni pai ren.
You ask someone to.

(12) 他们逼他把牌亮开。
Tamen bi ta ba pai liangkai.
They force him to show his cards.
→ a. 他们逼他。
Tamen bi ta.
They force him to do something.

Because the causative property of the verbs “派 (pai)” and “逼 (bi)” is weaker than the verb “使(shi)” . Therefore, the sentences from (10a) to (11a), (12a) are gradually more acceptable and independent. All in all, telescopicform is one kind of multi-predicate sentences and it can be computed by the same procedures used in other multi-predicate sentences.

3.3 Verb-Copying Construction

The feature of verb-copying construction is: V1 and V2 have the same form and semantic meaning, and they share the same subject as well. Its form can be expressed as (S) VOVC. As a sentence, (S) VO is usually independent, VC ("C" is complementary) is the syntactic construction which shows the moving direction or the result of action. Observe the following sentences: (13) (14) (15).

(13) a. [你] j 读书, [ej] 读到哪里了。
[Ni] j dushu, [ej] dudao nali le.
You read book, where you have reach.
→b. 你读书读到哪里了。
Ni dushu dudao nali le.
Where you have reach in this book.
(14) a. [她] j 盼他， [ej] 都快盼疯了。
[Ta] j pan ta, [ej] dou kuai pan feng le.
She is long for him, she almost goes mad.
→ b. 她盼他都快盼疯了。
Ta pan ta dou kuai pan feng le.
The longing for him almost drives her mad.
(15) a. [他] j 吮毒液， [ej] 吮得嘴唇肿了。
[Ta] j shun duye, [ej] shunde zuichun zhong le.
He sucks the venom, his mouth goes swollen.
→b. 他吮毒液吮得嘴唇肿了。
Ta shun duye shunde zuichun zhong le.
He sucks the venom and his mouth goes swollen.

We can find that verb-copying construction is similar to telescopicform in the forming procedures. For example, they all contain a initiative clause "(S) VO" and a subsequent clause "VC", and they are independent and dependent respectively. Furthermore, their combining procedures are the same completely. For example, combining the clause “[他]j吮毒液” ([ta]j shun duye) and “[ej] 吮 得 嘴唇 肿了(shun de

zuichun zhong le)" through the same procedures mentioned above, so we get sentence (15)b. Conversely, the procedures by which we can compute it are as followed:

Step1:他吮毒液，他吮得嘴唇都肿了。
Ta shun duye , ta shunde zuichun dou zhong le.
He sucks the venom, his lips go swollen.
↓cancel pause
Step2:[他] j吮毒液 ［他］j吮得嘴唇都肿了。
[Ta] j shun duye [ta] j shunde zuichun dou zhong le.
He sucks the venom and his lips go swollen.
↓delete co-referent constituent
Step3: [他] j 吮毒液 [ej] 吮得嘴唇都肿了。
[Ta] j shun duye [ej] shunde zuichun dou zhong le.
He sucks the venom and his lips go swollen.
↓clear out trace
他吮毒液吮得嘴唇都肿了。
Ta shun duye shunde zuichun dou zhong le.
He sucks the venom and his lips go swollen.

4 Conclusions

It is a universal phenomenon that the status of the different syntactic structures is also different in a language. The most obvious fact is that there are distinctions between the nucleus syntactical structures and the marginal syntactical structures. The former ones are such structures as co-ordinate structures, adjunctive structures, subject-predicate structures and V-O structures. The latter ones include V-C structures, multi-predicate constructions and so on in Mandarin. [9]

Chinese has not morphology as indo-European language in strict sense, so as the syntactic and pragmatic factors both work on the sentences in a discourse, the nucleus sentences or clause are prone to combine with the marginal ones, so as the multi-predicate sentences in this paper. On the basis of the properties of multi-predicate sentences and the theory of information–parsing [10], we can compute them intelligently in the way that segregates a multi-predicate sentence into mono-predicate clauses and process them respectively beforehand.

References

1. Hopper, P.J., Traugott, E.C.: Grammaticalization. Cambridge University Press, Cambridge (2003)
2. Tesnière, L.: Elèment de syntaxe structurale. Editions Klincksieck (1959)
3. Givón, T.: Syntax: A Functional-typological introduction, vol. I. John Benjamins, Amsterdam (1984)
4. Chafe, W.: Meaning and the structure of language. University of Chicago Press, Chicago (1970)

5. Russell, K.: The ‘word’ in two polysynthetic languages. In: Alan Hall, T., Kleinhenz, U. (eds.). Studies on the Phonological Word. John Benjamins, Amsterdam (1999)
6. McCawley, J.: Lexical insertion in a transformational grammar without deep structure. In: Proceedings of the 4th Annual Meeting of the Chicago Linguistics Society (1968)
7. Lehmann, C.: Thoughts on Grammaticalization. LINCOM Europa, Munich (1995)
8. Lyons, J.: Semantics, vol. II. Cambridge University Press, Cambridge (1977)
9. Xue, H.W., Yan, M.Y.: The Grammaticalization, Structure of *Youqing* and the Property of *You*. Chinese Linguistics 34(2), 14–28 (2011)
10. Xiao, G.Z.: The study on Facts and theories of Chinese Grammar. Hubei People Press, Wuhan (2005)

The Description of *You* in Mandarin Based on the Concept-Semantic Approach of the WordGroup Model

Hongwu Xue

College of arts, Chongqing Normal University, Chongqing, China
xuehongwu386@sina.com.cn

Abstract. This paper points out the WordGroup model is a new approach in NPL, in which all the word's forms and semantic meaning can be derived from it, the chief way to realizing it is to describe the two-level approach, i.e. the conceptual-semantic method of unification. It can describe polysemy verb *you* in Mandarin economically and continuously. The performances of it are across words and phrases by its conceptual prototype, and its final representation is the net which is made up of the complex distinctions and shared feature sets. It can simplify the procedures of the discrete analysis which is based on different *you*, and they are still connected in semantic meaning and form.

Keywords: You, WordGroup model, concept-semantic approach.

1 Introduction

Generally speaking, the polysemous word *you* (means "there be" or "have") in Mandarin has such three kinds of usages according to its performances in morphosyntax and semantics:

1.1 Prefix

Being a prefix of nouns or verbs[1-2], it has two types, Ⅰa: "*you* N"(有N) and Ⅰb: "*you* V"(有V).

Ⅰa. "*you* N": *youzhou* (有周, *zhou* dynasty), *yousong* (有宋, *song* dynasty), and etc.

Ⅰb. "*you* V": *youqing* (有请, to invite), *youchong* (有忡, to be sad), *youming* (有鸣, to cry repeatedly and noisily), and etc.

The prefix *yǒu* is not independent and won't change the meaning and the function of "*you* N" and "*you* V". Its function is to add additional subjective meaning to the following word root.

(1) 有宋一代先后承受着北方少数民族政权的辽、金、元的威胁。

YouSong yidai xianhou chengshou zhe beifang shaoshumingzu zhengquan de Liao, Jin, Yuan de weixie.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 559–568, 2013.

Song dynasty was suffered from the intrusion from the Northern minority regimes of *Liao*, *Jin* and *Yuan* in sequence.

"yousong" is the stylistic variant of the noun "*song"*. The former carries more historical and cultural information comparing to the latter. The function of the prefix *you* is to express some subjective incremental meaning. Next example is of *you* V:

(2) a. 有请哥们! 看看我的时间明细。
Youqing gemen! Kankan wo de shijian mingxi.
Buddies, please see my time details.
b.有请哥们看看我的时间明细。
Youqing gemen Kankan wo de shijian mingxi.
Please Buddies to see my time details.

The verb *youqing* is the interpersonal variant of verb *qing*, prefix *you*'s function is to add polite meaning to verb *qing*. *youqing* is similar to *qing* in their syntactic function and basic meaning, because they both can be used as a independent sentence or construct a telescopicform pattern. In the same way, the adjective *youchong* expresses the meaning which is to a greater degree in sorrow. *youming* means that birds cry repeatedly and noisily. The abstract grammatical meanings of Ⅰ(a-b) are all subjective increment, these meanings arise from verb *you*'s semantic feature [+existence]. *Yousong* is originally a verb phrase, finally it is lexicalized into a noun and *you* is reanalyzed as a prefix. In this way, we can find out *youchong* and *youming* are parallel to the adjective *chong* and the verb *ming* in the semantic and grammatical function, the only difference between them is that the former is subjective.

1.2 Morpheme

Being a morpheme, *you* can be used to construct verbs, pronouns or adverbs[3-5]. In such cases, it has eight types in form:

Ⅱa. 有关(*you guan*, concerning to)
有利(*you li*, benefit);
Ⅱb. 设有(*she you*, to set up)
写有(*xie you*, to write)；
Ⅱc. 有些/点儿(*you xie/dianer*, little/some)
有所(*you suo*, certain/proper)；
Ⅱd. 有的(*you de*, some)；
Ⅱe. 有的是(*you de shi*, to have plenty of)；
Ⅱf. 有没有(*you mei you* , to have or not)[3]；
Ⅱg. 据有(*ju you*, to grasp)；握有(*wo you*, to hold)；
Ⅱh. 有着 (*you zhe*, to be being)

They are all the results of lexicalization of verb *you* and relevant components in the syntactic circumstances which have [＋existent] meaning. The different "*you"* in these forms have been grammaticalized in varying degrees , their grammatical meanings have also been subjectivized in varying degrees. In Ⅱa, *you* express the meaning of [＋subjective,＋increment]. For example, comparing to *li yu* (利于, benefit), we' ll

find *you li* has the grammatical meaning of subjective increment, while *li yu* has not.Ⅱb' s *you* express something which are in large quantity, in *xie you*, *you* can also be understood as a aspectual marker like *le* and *zhe* ("了/着", perfective /progressive aspect). Verb *xie* is the modifier of *you*, *you* is subjective. In Ⅱc, all forms are adverbials, which express the meaning of [+subjective], [+quantity bigger than normal]. *youde* is a indefinite pronoun and expresses the meaning of [+some]. *youdeshi* in Ⅱe is a verb and its meaning are [+existent]/[+possessive] and [+subjective]. *youmeiyou* is a compound word and its semantic meanings are [+existent]and [+confirmative]. As a positive trend phrase, it is equal to *you* in semantics and syntax. In Ⅱg, verb *ju* and *wo* are the modifiers of *you*, they express the meaning of [+possessive]. *youzhe* is a subjective bound verb, in which *zhe* is the subjective expressive affix, it comes from experiential progressive syntax. In a word, Ⅱa, Ⅱb and Ⅱg are productive in word-building , there are a large number of words like them, they amount to 430 or so. See examples below:

(3) 不吸烟有利健康。
Bu xiyan youli jiankang。
No smoking is beneficial to health.

(4) 安禄山握有重兵。
An Lushan woyou zhongbing。
An Lushan hold the heavily fortified.

(5) 墙上写有一条永久性标语。
Qiangshang xieyou yitiao yongjiuxing biaoyu.
There was written on the wall a permanent slogan.

(6) 天气有些/点儿冷。
Tianqi youxie/dianer leng.
It is a little cold.

(7) 教师工资有所增长。
Jiaoshi gongzi yousuo zengzhang.
Teacher's salary has somewhat risen.

(8) 他有的是时间/朋友。
Ta youdeshi shijian/pengyou.
He has plenty of time/friends.

(9) 你有没有见过这种车？
Ni youmeiyou jianguo zhe zhong che?
Have you ever seen this type of car?

It needs to be pointed out that the *youmeiyou* in (9) only distributes before "*jianguo zhe zhong che*(见过这种车)", and its grammatical function and semantic meaning confirm to the interrogative event of "*jianguo zhe zhong che*". It is different to verb phrase *youmeiyou*, for example:

(10) 你有没有这种车？
Ni youmeiyou zhe zhong che?
Do you have this kind of car?

In (10), *youmeiyou* is the paratactic construction which is made up of verb *you* and its negative form "*meiyou"* (do not have), and it only distributes before a NP, for example, "*zhe zhong che*", and its function form an alternative interrogative.

1.3 Syntactic Word

As has been investigated, the word of *you* have three to ten kinds of semantic and grammatical functions, we only list five basic usages here: Ⅲa.[+possessive]、Ⅲb.[+exist]、Ⅲc.[+happen/appear]、Ⅲd.[+assess/compare] and Ⅲe.[+nonspecific] so on. For example:

(11) 他有两本外文小说。
Ta you liangben waiwen xiaoshuo.
He has two foreign language novels.

(12) a. 唐朝有很多诗人。
Tangcao you xuduo shiren.
There are many poets in Tang dynasty.

b. 院里有棵树。
Yuanli you ke shu.
There is a tree in the yard.

(13) 他有病了。
Ta youbing le.
He is ill.

(14) a. 这条鱼有四斤。
Zhe tiao yu you sijin.
This fish weighs 4 kilograms.

b. 这条河有那条河深。
Zhetiao he de hui you natiao shen.
The river is as deep as that river.

(15) a. 有一天他没上课。
You yitian ta mei shangke.
One day he didn't go to school.

b. 有人爱评剧，有人爱越剧。
Youren ai Pingju, youren ai Yueju.
Someone like Ping opera, others like Yue opera.

The semantic meaning of *you* in (11), is of possession, and in (12) is of existence. The semantic meaning of *you* in (13), (14) and (15) seem to be different each other, in fact they is closely linked by the semantic feature [+existent], so their prototype semantic is still [+existence]. *You*'s semantic meanings of "happen"or"appear" are the [+existent] which is reinterpreted by the context (13). In the same way, the "assess" and "compare" semantic meaning of *you* in (14) and the "someone" and "nonspecific" of *you* in (15) are all the [+existence] which are reinterpreted in contexts too. In a word, they are all grammaticalized from *you*'s [+existent] semantic meaning and grammatical function.

In Mandarin, there is a new usage of you, follows as (16). It precedes "*shuoguo xiang ni shou xuefei*"(have ever said to collect tuition fees from you), because it is used frequently, so can be included into *you*'s usages. Its aspectual meaning is perfective. Besides this, it also has the grammatical emphatic meaning, the "emphasis" is increment of positive relation. So its feature of grammatical meaning is still [+exist] and [+increment]. For example：

(16) 我有说过向你收取学费吗？
Wo you shuoguo xiang ni shou xuefei ma?
Have I ever said to collect tuition fees from you?

The function of *you* emphasizes the event of "*shuoguo xiang ni shou xuefei* ". The *guo* is an experience Aspect marker, which is optional. The semantic feature of *you* is[+subjective], [+emphatic]and[+perfective]. When *you* distributes in interrogative sentence, its semantic feature is [+subjective][+infer]. The distinction between *you-meiyou* and *you* is the emphatic function of *you* is weaker.

As have been showed in (12) to (16) above, the semanteme [＋existent] or [＋possessive] of *you* is basic in its all kinds of usages, they are the *you'*s prototype semantic meanings and grammatical functions, the others semantic meanings and function are all the outcomes which interact between existence or possessive and the specific context.

The members of the three types of *you* above, are different only in the varying degrees on which the typicality of their semantic meaning shown. For example, Ⅰa is more typical than Ⅰb. but in all conditions the semantic meanings and grammatical representation of *you* in typicality is continuous. However, whether the semantic tagging or disambiguation in the sentence processing method of *you* at present is based on the analysis of section 1.3. All work on *you* seems to be discrete, Ⅰ(a-b), Ⅱ（a-g）and Ⅲ（a-f）, three types are not described and explained systematically, they do not show the relations between *you* and the relevant verbs *shi* (是,be) and *zai* (在, be on/at/in). Besides this, if we consider the problem of computer translation between Mandarin and ancient Chinese, they are not enough and economical and systematical.

2 Processing *You* in Systematically Computation

2.1 The Appropriateness of Semantic Description in Ontology

Semantic description in Ontology affects the efficiency of natural language processing (NPL). Extra complication and simplification would burden semantic recognition and sentence generation in computing. These burdens come from the complex program resulted by the over-described semantic component features on the object. The computer recognition is on the base of semantic mining, so how to use one or several appropriate operating model is the most important work in lexical-semantic data mining. These models should be able to describe the objects in a systematic and economical way and embody them in a linguistic information corpus. And it is important for the computing based on semantic resources especially.

2.2 WordNet Model and FrameNet Model

At present, one of the influential descriptive model on semantic resources is WordNet, using the way of describing the semanteme of word's sememe to construct a net relations between words or among a word's different sememe. For example, we can take the [+existence] as a node to set a transverse Synset {you zai shi}, but this is just a oversimplified description about the three words. Because it is a macroscopical way, It does not differentiate the three words in syntax and semantics and is simple on the sentence processing too. Another model is FrameNet, it describes you as 5 to 6 sememes like in section 1.3, and adds semantic role and syntactic feature bundle on the relevant NP. For example in Ⅲa, it can be described as following: NP1 is agent, its semantic features are [+location][+time], NP2 is object, its semantic features are [+entity][+quantity]. This is relatively accurate undoubtedly, but when we process you's form types in Ⅰ, Ⅱ and Ⅲ, we have to process the object part by part according to different semantic criteria. For example the xieyou in Ⅱb, we should go further to analyze its semantic meaning, analyze it as xiezhe/le (written/writing) and "write+ [+possessive] [+ existence]" in the case frame. Now we will not speak if this way is accurate, at least it is a too complicated in program. To increase the efficiency of sentence processing on the basis of WordNet and FrameNet, we must describe you appropriately and supply optimized semantic knowledge to computer. This includes such three features in Ontology at least:

- At least can combine WordNet and FrameNet in a certain degree.
- Objects can be included into a system in semantics and syntax.
- Make semantic description go across levels and syntactic relations, take a economical and clear path, keep consistency at work.

To process *you* which is with high frequency and has many varied usages, this paper proposed a concept of "WordGroup". It is the model which unified semantics and syntax, and it can show the complex relationships between the different variants and usages on the varying level. It is a high efficiency semantic Ontology Network. This processing way is appropriate and can avoid the over-simplification and over – complication.

3 *You*'s Description in WordGroup

3.1 The WordGroup Model

WordGroup is a lexical-syntactic category which includes phenomena of polysemy, synonymy and paronym. The members in it, have all kinds of relationships between near or far, which are derived from the root word in forms, semantics and syntax. The *you* in Mandarin can form a WordGroup, its members consist of prefix, morpheme and syntactical word. It can aggregate you's all variants and usages formed synchronically and diachronically.

3.2 The Concept–Semantic Two-Level Approach in Description

Concept is on the level of cognition, while semantics is on the level of linguistics including lexical senses and grammatical meaning [6]. On the level of concept, we can set a Synnet of {you shi zai} based on the concept of "existence" and find you's corresponding coordinate in the WordNet. Based on it, we can extract the basic semantic meaning and grammatical function which is "existence" and "possessive", they are outcome which interacts the concept of you and its distribution in syntax. The relation between the semantic of existence and possession, is the former is the prototype semantic, and the later is a experience gestalt of the former in physicals and cultures.

So we can link all forms of you in section 1 by the WordGroup, and describe its every member in a concept–semantic two-level approach appropriately, including tagging their syntactical distribution, case-frame and semantic features so on. At last, we get an accurate complex semantic component feature set about every member. The members can be linked by the shared semantic and grammatical features, which are nodes. The differences between the members can be distinguished by the distinctive semantic and grammatical components. In the way, the WordGroup of you forms a net which can be showed in a relation between nodes. In this net, the computer activates a hyponymy node, accordingly, the semantic and grammatical resources orderly are outputted. The net of *you* shows in the figures below：

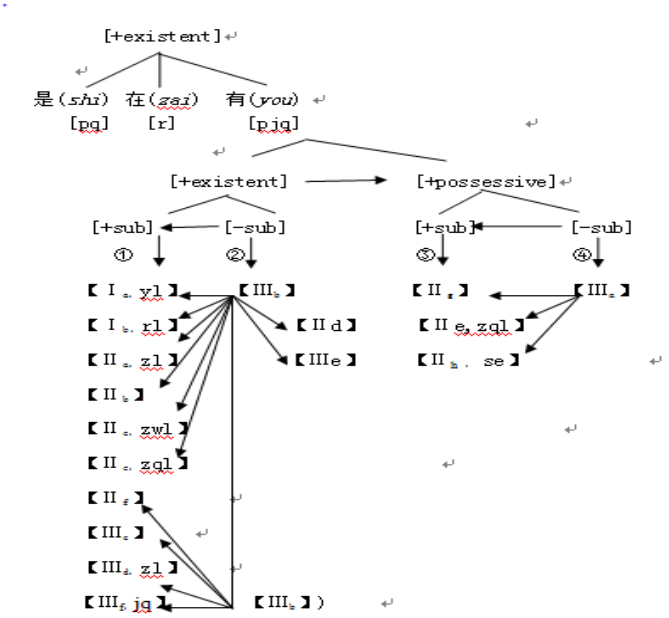


Fig. 1. You's semantic net

3.3 Notes of the Figures. 1

3.3.1

The symbol of [] embody hyponymy relationships, and【】embody the hypogyny semantic and grammatical features, lowercase show specific semantic features: sub-subjective, p-judge, r-orientation, y-colloquial, zl-increment, w-minim, q-emphasis, jq-more emphasis, se-subjective experience progressive, and arrow-head shows derived relation.

3.3.2

The semantic feature [+existence] is prototype conception of you, [+possessive] is not its primitive semantic but its experiential gestalt of [+existence]. Based on the [+possessive] is a important concept of experience in daily and grammar, it and the [+existence] are looked as the hypogyny concept of [+exist]. All the forms in ① and Ⅱd and Ⅲe in ② are all stemmed from Ⅲb, they have the [+existence] and [+subjective] semantic features. The forms of Ⅱg, Ⅱe and Ⅱh are stemmed from Ⅲa in④, their shared semantic features are [+possessive] and [+subjective].

3.3.3

Every usage of you have the same semantic basis and have themselves semantic distinctive features accordingly, for example, *yousong*'s semantic and grammatical features are [+existence], [+noun]and[+historical style], it can be distinguished from Song by the feature[+historical style], and it also can be distinguished *youqing* from the feature [+verb]; *xieyou* is equal to you, the distinctive between them are [+subjective] and the modifier *xie*. The rest of examples can be inferred by analogy.

3.3.4

you can be distinguished from the shi and the zai as following：the semantic and grammatical feature of shi is [+judgment], and zai is [+orientation]. The respect of subjective emphasis among the three is shi ＞you＞zai. For example：

(17) a. 桌上有本书。
Zuoshang you ben shu.
There is a book on the desk.
b. 桌上是本书。
Zuoshang shi ben shu.
It is a book that is on the desk.
c. 书在桌上。
Shu zai zuos hang.
The book is on the desk.

(18) a. 你在官在民，敢 报上来吗？
Ni zai guan zai ming, gan bao shanglai ma?
Dare you tell me that you are an official or a common people?

b. 你是官是民，敢报上来吗？
Ni shi guan shi ming, gan bao shanglai ma?
Dare you tell me that you are an official or a common people?

(19) a. 他是看了。
Ta shi kanle.
He has really seen it.

b. 他有看了。
Ta you kanle.
He has seen it.

The semantic meanings of the shi, zai and you in (17), (18) and (19), are equal basically. *you* shows that there is a existence relationship between the shu (book) and the zuozi(desk), while the zai shows the existence relationship between the two by the way of orientation. As for the degree of subjective emphasis, the shi is higher than the you, because the judgment of shi has the exclusive feature, while *you* has not. However, the common ground among the three is that they all have the semantic feature [+existent], because any judgment must be built on the foundation of the existent relationship.

4 Conclusions

The key problem of sentence processing in Mandarin is the semantic knowledge mining and tagging. This paper takes the both advantages of WordNet and FrameNet, and extends its objects further, set the WordGroup model. By it, we can describe any a word like you in an appropriate and continuous way. Its representation is the network which is consists of many semantic and grammatical features. Through activating in consequence the nodes of the word in WordGroup which have inner relationships and are hierarchical, we can completely output the objects which should be recognized by computer. So the paper's achievement can make up for the deficiency of discrete processing way before, and can enhance the efficiency of computer semantic recognition. This model is also valuable in dictionary compiling and language teaching.

Acknowledgements. This work is supported by the Humanity and Social Science Research Youth Fund of Ministry of Education of China under Grant NO.10YJC740116, and by the Doctor Fund of Chongqing Normal University under Grant NO. 11XWB012.

References

1. Xue, H.W., Yan, M.Y.: The property, semantic and function of *you* in Proper Noun of *You*M in ancient Chinese. Research in Ancient Chinese Language 95(2), 44–53 (2012)
2. Xue, H.W., Yan, M.Y.: The Grammaticalization, structure of *youqing* and the property of you. Chinese Linguistics 34(2), 14–28 (2011)

3. Xue, H.W.: The "V*you*"-Structrue in Modern Chinese. Journal of HUST (Social Science Edition) 22(4), 103–107 (2008)
4. Xue, H.W.: The grammaticalization of Chinese verb *Yousuo*. Research in Ancient Chinese Language 84(3), 27–30 (2009)
5. Xue, H.W.: The "V*you*"-Structrue in Modern Chinese. Journal of Ningxia University (Humanities & Social Science Edition) 32(2), 46–51 (2010)
6. Taylor, J.R.: Linguistic Categorization, 3rd edn. Oxford University Press, New York (2003)

The Study of the Structure "Verb+Numeral+Measure(le)" from the Perspective of Information Conveying Grammar

Dong Ouyang[1,2] and Xiaoming Hu[1,2]

[1] Center for Study of Language and Information, Wuhan University, Wuhan 430072, China
[2] College of Chinese Language and Literature, Wuhan University, Wuhan 430072, China
wdouyangdong@163.com, huxiaoming@whu.edu.cn

Abstract. The paper will study the structure"verb+numeral+measure(le)" based on the theory of Information Conveying Grammar[1]. Combined with the persuading and question-answering social intercourse contexts,this paper will discuss the following questions: (1)the dynamic types of the structure "verb+ numeral+measure(le)"; (2) Based on the theory of Information Conveying Grammar, this paper will also analyze three information structures of "verb +numeral+measure(le)". They are listed as following: one-level information structure which emphasizes quantity, two-level information structure in which action and quantity cannot be omitted, two-level information structure which emphasizes action and short timing. According to the above analysis, this paper aims at explaining the significant impact of dynamic information structure to the understanding of language forms and meanings. Moreover, Language forms and meanings restrict the information structure of language.

Keywords: verb+numeral+measure(le), autoregulated and unautoregulated verbs, semantic focus.

1 Introduction

"Verb+numeral +measure（le）"is a complex structure in Chinese, which draws much attention of grammar scholars. You[2] only gives a simple analysis. Ni and Xu[3]regard du shici ‘read ten times’ and qu yici ‘go once’ as a mixture of verb -object combination and subject-verb relation.Li[4]elaborates subjective quantity. Ma[5]analyses the grammar structure "verb+le+timing quantify+le" in terms of semantic features. He puts forward that the ambiguity in the structure "verb+le+timing quantify+le" is caused by the different semantic features of the verbs, which play a very important role in sentence structures. He thinks "verb" can be sorted into four types:

"Va" (like si ‘die’) Semantic features: [+finished, -continued, -condition];
"Vb" (like deng ‘wait for’) Semantic features: [-finished, -continued, -condition];
"Vc" (like kan ‘see’) Semantic features: [+finished, +continued, -condition];
"Vd" (like gua ‘hang’) Semantic features: [+finished, +continued, +condition].

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 569–577, 2013.

The condition will become more complex if the structure changes from "verb+le +timing quantifier+le "into "verb+numeral +measure (le)". For example, qu sanci ‘go three times’ is clear in meaning, while qu yici ‘go once’ is ambiguous, which may have different meanings with different stresses and semantic focuses. Therefore, it cannot be explained smoothly just from the aspect of verb’s semantic features.

We explain this phenomenon from the perspective of Information Conveying Grammar. The relevant theory holds that in a sentence, the part which loads the old information is called theme while the part which loads the new information is called rheme. The theme-rheme structure is a piece of information structure in a sentence. If a sentence is made up of three or more words, it contains usually multiple information structures. In specific study, the information structure can be acquired by using question-answering chains. Under the instruction of Information Convey -ing Grammar Theory, regarding the two types of social intercourse contexts of the structure "verb+numeral+measure(le) "as the studying background, that is the speaker’s persuasion to the listener and the answer to other’s questions, this paper attempts to analyze the dynamic types and information structures of "verb+numeral +measure(le) ".This structure can be used in the objective statement context. Whether the verb autoregulated or not and whether the quantity is virtual or not, the semantic focus is on the whole structure "verb+numeral+measure(le)". It is a two-level information structure in which action and quantity cannot be omitted . For example, Laowang si liangtian le. ‘Lao Wang has been dead for two days’,its information structure can be analyzed by the following answer- questioning chains: A. the question: Laowang zenme le? ‘What's wrong with Lao Wang?’ , the answer: Si le‘Been dead’. B. the question:Si jitian le? ‘How many days has he been dead?’, the answer: Santian‘three days’. The two answer-questioning chains stand for two pieces of information. "verb+numeral+ measure(le)" has only one type of dynamic type and information structure type if used in the object statement context. Therefore, it is not included in this paper when used in object statement context.

2 Reclassification of Verbs and Quantity

In the cognitive world of human beings,liang ‘quantity’ is involved in every object, event and property. For instance, an object includes geometric sense and number, an event includes action quantity and time quantum, and a property includes magnitude. These measuring factors gather up liang ‘quantity’, which is the logical category and reflects the real world. liang‘measure’ projects from real world to language and forms liang ‘quantity’ ’in linguistic category[6]. Language reflects the real world but not simply copies the real world, which can be found in measure words. For examples: (" ※"means false structures, the same below)

(1) Va:

①※si yitian‘die for one day’, ※ta yici‘collapse once’, ※chujia yici‘marry once’ （They cannot be used in the speaker’s persuasion to the listener.）

②si qitian le‘be dead for seven day’, shang qitian le ‘be injured for seven days’, chujia qinian le‘be married for seven years’

③si yitian le‘be dead for one day’, ta yici‘collapse once’, chujia yici ‘ be married once’ (When used as the answer to other’s question, they are valid)

(2) Vb:

①deng liangtian ‘wait for two days’, yang liangtian‘rest for two days’, ren liang tian ‘bear for two days’

②deng qitian‘wait for seven days’, yang qi tian ‘rest for seven days’, ren qitian ‘bear for seven days’

(3) Vc:

①kan liangtian ‘see for two days’, ting liangtian‘listen for two days’, jiao liang-tian ‘teach for two days’

②kan qitian‘see for seven days’, ting santian‘listen for three days’, jiao santian ‘ teach for three days’

(4) Vd:

①gua liantian‘hang for two days’, bai yitian‘display for one day’, diao liang-tian‘hang for two days’

②gua qitian‘hang for seven days’, bai santian ‘display for three days’, diao san tian‘hang for three days’

On one hand, we are customed to use "one/two"to express imaginary quantity. Under this condition, "verb+one/two+measure" indicates speaker’s persuasion to listeners and hopes listener can do the corresponding action. Taking (2)~(4)①for examples, the sematic focus and phonetic stress are on "verb". On the other hand, we can use "one/two + measure"to express real quantity, as (1)③, (2)①, (3)① and (4)①. Under this condition, the semantic focus will be put on"one/two +measure"or the whole structure. However, "three/four/five…+measure" can only express real quantity, like(1)~ (4)②.

Why (1)① are wrong structures when the speaker persuades the listener and wants the listener to do corresponding action? The reason is, if the listener is persuaded to act, the verb must be autoregulated, that is, this action can be done and controlled by the listener. But the verb in (1)① is unautoregulated and the speaker cannot require and persuade the listener to control the action that the verb means. Therefore, (1)① are false structures[7]. But why (1)③ are correct structures? That’s because (1)③ cannot indicate the speaker's persuasion to the listener, but can answer the speaker's question. Therefore, these are valid. "One/two+measure"expresses the timing quantity and momentum that the verb indicates. "One/two+measure" expresses real quantity.

Therefore, we divide roughly the structure "verb+numeral+measure(le)" into the following substructures (numeral A=three/four/five…, numeral B=one/two). The next part will discuss them in detail:

- verb(autoregulated)/(unautoregulated)+numeral A+measure.
- verb(autoregulated)/(unautoregulated)+numeral B+measure.

3 Verb(Autoregulated)/(Unautoregulated)+Numeral A+Measure

Firstly, if the sentence indicates the speaker's persuasion to the listener, the structure "verb(autoregulated)+numeral A+measure" is a two-level information structure in which the action that"verb(autoregulated)" expresses and quantity cannot be omitted. The speaker's semantic focus falls on the whole structure "verb(autoregulated) +numeral A+measure". For example: ("≌" represents the semantic meanings of the former and latter sentence are approximately the same. The same indication with the following "≌")

(5) Hai you qitian kaoshi, suoyi ni keyi zhunbei qitian.
'There are seven days left when the exam is coming, so you can only prepare it for seven days.'

Hai you qitian kaoshi, suoyi ni you shijian zhunbei, qixian shi qi tian.
'There are seven days left when the exam comes, you still have time to prepare but the deadline is seven days.'

Ni keyi zhunbei zhunbei kaoshi.
'You can have well prepared for the exam.'

(6) Yisheng jiao ni tang qitian.
'You should stay in bed for seven days as the doctor asked you to do this.'

Yisheng jiao ni tang chuangshang xiuxi, qixian shi qitian.
'You should stay in bed for seven days as the doctor asked you to do this. The deadline is seven days.'

Yisheng jiao ni tang chuangshang xiuxi .
' You should stay in bed and have a rest .'

In the example(5), This "qitian'seven days'" is a real quantity. It doesn't mean for "liutian'six days' ", or "batian'eight days'". "zhunbei qitian'to prepare for seven days' " tells us that this behavior can last for "seven days". We can acquire the information structure of "ni zhunbei qitian'you can prepare for seven days'"by using question -answering chains:

A. The question: Ni zai zuo shenme? 'What do you do?'
The answer: Wo zai zhunbei kaoshi. 'I am preparing for the exam.'
B. The question: Zhunbei duoshao shijian? 'How long time can you prepare for it?'
The answer: Qitian . 'Seven days.'

From the question-answering chains and the equal replacing sentences, we can see the action "zhunbei 'prepare'" and the quantity "qitian'seven days'" are all the infor -mation the speaker wants to convey. In A, "wo'I'" is the theme and "zhunbei kaoshi 'prepare for the exam'" is the rheme, so it is a piece of information. In B, "zhunbei kaoshi'prepare for the exam'" is the theme and "qitian 'seven days'" is the rheme, so it is a piece of information. Therefore, "zhunbei qitian 'prepare for seven days'" is a two-level information structure in which "zhunbei'prepare'"and "qitian'seven days'" cannot be omitted. This is the same to example (6).

Secondly, if the structure "verb(autoregulated)+numeral A+measure" is used in the context which answers other's questions, the semantic focus falls on "numeral A+ measure". For example:

(7) Jia:Ni qizi jiao ni deng ta duoshao shijian?
'How long time does your wife ask you to wait for?'
Yi:Deng santian. Santian.
'Wait for three days. Three days.'

(8) Jia:Yige xinqi shang jici ban?
'How many times do you go to work in a week?'
Yi:Shangqitian. Qitian
'Go to work for seven times. Seven times.'

In the conversation of example (7), the focus of Jia's question is"duochang shijian 'how long'". As for Yi's answer, "santian'three days'" is the new information and it replies Jia's question focus. "deng'wait for'" is the background knowledge which can be omitted in certain context. You can reply either"deng santian'wait for three days.'" or "santian'three days.'". "Deng san tian'wait for three days'" is a one-level information structure which emphasizes the quantity "santian'three days'". The example (8) is the same case.

Thirdly, the structure"verb(unautoregulated)+numeral A+ measure" can't be used as the listener's persuasion to the speaker. It can be used in the context which answers other's questions. The semantic focus is on "numeral A+measure." It is a one-level information structure which emphasizes the quantity. For example :

(9) Jia:Laowang si jitian le?
'How long has Lao Wang been dead?'
Yi:Si santian le. Santian.
'Have been dead for three days. Three days.'

(10) Jia:Ni qianbao diu jitian le?
'How many days has your wallet lost?'
Yi:Diu qitian le. qitian.
'Have lost for seven days. Seven days.'

(11) Jia:Lixiaojie chujia duoshaonian le?
'How many years have Miss Li been kept married?'
Yi:Chujia sannian le. Sannian.
'Have been kept married for three years. Three years.'

In the conversation of the example (9), the focus of Jia's question is "duoshaotian 'how many days'". In Yi's answer,"santian'three days'" is a new information which answers Jia's question. "Si'die'"is the background knowledge which can be omitted in certain context. Yi can reply either "si santian'have been dead for three days'" or "santian'three days'"."Si santian le'have been dead for three days'" is a one-level information structure which emphasizes the quantity"santian'three days'". It is the same case with the example (10) and (11).

The structure "verb(autoregulated)+numeral A+measure"can be used in two types of social intercourse contexts, the speaker's persuasion to the listener and the answer to other's questions. If used in the context where the speaker persuades the listener, the semantic focus of the structure "verb(autoregulated)+numeral A+measure" falls on the whole structure and it is a two-level information structure which emphasizes quantity and action. If used in the context where one answers other's question, the semantic focus of the structure is on "numeral A+measure"."Verb(autoregulated)" is the old information and it is a one-level information structure which emphasizes quantity. "Verb(unautoregulated)+numeral A+measure" can only be used in the context where one answers others' question. The semantic focus is on "numeral A+measure" and it is also a one-level information structure which emphasizes quantity.

4 Verb(Autoregulated)/(Unautoregulated)+Numeral B+Measure

"Verb(autoregulated)+numeral B+measure" is an ambiguous structure.

Firstly, as to(2)～(4)①, if "numeral B+measure" is imaginary quantity and the speaker requires or hopes the listener to carry out or implement the behavior repre -sented by the verb, then the phonetic stress is on "verb(autoregulated)". The "verb (autoregulated)" need to pronounce with stress while "numeral B+measure" is weakly stressed for it stands for a kind of virtual meaning. It is a two-level informa -tion structure which emphasizes the action and short timing. Please look at the example (12) to (16):

(12) Xiaoming,deng liangtian.
'Xiaoming, you had better wait for two days.'
≌Xiaoming,dengyideng.
'Xiaoming, you had better wait for some times.'

(13) Xiaoming,qu yici.
'Xiaoming, you had better go there .'
≌Xiaoming,ni qu yixia.
'Xiaoming, you had better go there.'

(14) Xiaoming,denglong gua liantian.
'Xiaoming, the lantern should be hung for two days.'
≌Xiaoming, ba denglong gu yi gua.
'Xiaoming, the lantern should be left hanging.'

(15) Xiaoming,yifu chuang liantian ba.'
'Xiaoming, you had better wear your coat for two days.'
≌Xiaoming,yifu chuang yixiaba.
'Xiaoming, ou had better wear your coat for some time.'

In the example (12), the speaker actually requires the listener "deng'waiting for'", not really "deng liantian'waiting for two days'", "liangtian'two days'" represents a virtual meaning and should be weakly stressed. The voice focus and semantic focus are the actions of "deng'waiting'" and "liangtian'two days'" can be omitted in certain context. "Deng liangtian'wait for two days'" is a two- level information structure which

emphasizes action and short timing. "deng liangtian‘wait for two days’" can be replaced by "dengyideng‘wait for a little while’". It is the same case with the example (13) to (15). The example (13) to (15) tell that the speaker persuades the listener to implement some kind of action. As for the quantity of action is not what the speaker concerns. Because repetitive verbs have the quality of short timing, thy can relieve the phonetic stress if used in imperative sentences[8]. Combined with what Zhu has expressed, the following can be concluded: Through transformational analysis, "numeral B+measure" stands for a kind of short timing or small quantity of action. It equals to virtual reference. This is related to Chinese people’s psychological thinking method.

Secondly, if "numeral B+measure" symbolizes the real quantity and is used in the persuading context, the speaker’s semantic focus is on the whole structure "verb(autoregulated)+numeral B+measure". The structure points out the time the action lasts and it is a two-level information structure which emphasizes action and quantity.It is the same case with the structure "verb(autoregulated)+numeral A+ measure "when the latter is used in the context where the speaker persuades the listener. The only difference is the quantity. The detailed analysis is omitted here.

Thirdly, as to the examples of (2)~(4)①, if used in the context where one answers others' questions, it emphasizes timing quality or quantity of action. The semantic focus is on "numeral B+measure"and it is a one-level information structure which emphasizes quantity. The"verb(autoregulated)" can be omitted in certain context. The examples are following.

(16) Jia: Xiaoming,ni deng jitian le?
‘Xiaoming, how many days have you waited for?’
Yi: Deng yi/liang tian. ≌Yi/liang tian.
‘Have been waited for one/two days.≌One/two days.’

(17) Jia: Xiaoming,ni tang chuangshang xiuyang jitian le?
‘Xiaoming, how many days have you stayed in bed recovering?’
Yi: Yang yi/liang tian. ≌Yi/liang tian.
‘Have stayed in bed recovering for one/two days.≌One/two days.’

(18) Jia:Denglong gua jitian le?
‘How many days have the lantern been hung for?’
Yi:Gua yi/liang tian le. ≌Yi/liang tian.
‘Have been hung for one/two days.≌One/two days.’

In the example (16), the focus of Jia's question is on the quantity of "deng‘waiting for’". The new information provided by Yi is the quantity like " yi/liang tian ‘one/two days’". This factual quantity is what Jia hasn't known. The "deng‘waiting for’in the answer is known information and can be omitted in certain context. Therefore, Yi can reply either" Deng yi/liang tian ‘wait for one/two days’"or " yi/liang tian ‘one or two days’". It is the same case for example (17)(18). According to the example (16) to (18), it is known that what Jia is concerned about is "numeral B+measure" in Yi's answer. It is also the information that Jia doesnot know.

"Verb(autoregulated)+numeral B+measure" can be divided into three dynamic types:"Verb(autoregulated)+numeral B+measure"and "verb(autoregulated)+numeral B+measure", verb(autoregulated)+numeral B+measure"("…"represents the semantic focus, the same below).

Finally, the structure of "verb(unautoregulated)+numeral B+measure" is used in the context that makes sense(the answer to others' questions). It is the same case with the structure "verb(unautoregulated)+numeral A+measure" which used in the context where the speaker persuades the listener. The only difference between them is quantity and it is also a one-level information structure which emphasizes quantity. The detailed analysis is omitted here.

"Verb(autoregulated)+numeral B+measure"can be used in two contexts. One is the speaker's persuasion to the listener and the other is the answer to other's questions. If used in the former context, when the "numeral B" stands for virtual sense, the semantic focus of "verb(autoregulated)+numeral B+measures" is "verb(autoregulated)", "numeral B+measure" stands for the short timing quantity of action, and it is a two -level information structure which emphasizes action and short timing. "Numeral B" is special and represents virtual sense. When "numeral B" represents real sense, the semantic focus of the structure "verb(autoregulated)+ numeral B+measure" is on the whole structure and it is a two-level information structure which emphasizes quantity and action. If used in the latter context, the semantic focus is on "numeral B+measure". "Verb(autoregulated)" is the old information, and "numeral B" means real sense which is a one-level information structure which emphasizes quantity. The structure "verb(unautoregulated)+numeral B+measure" can only be used in the context where one answers others' questions. The semantic focus is on "numeral B+measure" and "verb(unautoregulated)" is the old information which is a one-level information structure in which quantity cannot be omitted.

5 Conclusion

Language is used in communication. When facing a specific and solitary structure, we must have them returned to different specific contexts and study their dynamic types as much as possible. Therefore, we make the structure "verb+numeral+ measure(+le)" return to their specific contexts, persuasion and answering questions, and divide it into the following seven dynamic types:

I	Verb(autoregulated)+Numeral A+Measure
II	Verb(autoregulated)+Numeral A+Measure
III	Verb(unautoregulated)+Numeral A+Measure
IV	Verb(autoregulated)+Numeral B+Measure
V	Verb(autoregulated)+Numeral B+Measure
VI	Verb(autoregulated)+Numeral B+Measure
VII	Verb(unautoregulated)+ Numeral B +Measure

This paper analyzes three types of information structures: First, one-level information structure which emphasizes quantity (the dynamic types II, III, VI and VII); second, two-level information structure in which action and quantity cannot be omitted (the dynamic types I and V); third, two-level information structure which emphasizes action and short timing (the dynamic type IV).

Based on the typical contexts where the structure "verb+numeral+measure(le) " is used, this paper analyzes the dynamic types and information structures of this structure. This will shed light on the second language teaching, the processing of Chinese information and even the intelligent man-made interaction. Our study has verified a fact. That is the information structure of language has an effect on the understanding of language forms and meanings; the language meanings and forms (the paper refers to grammar forms and meanings like "autoregulated and unauto -regulated verbs") restrict the information structure. It embodies Xiao's Chinese Grammar Theory on " Three Worlds ".

Acknowledgments. This work has been supported by the Fundamental Research Fund for the Central Universities (201111101020002), the Major Projects of Chinese National Social Science Foundation (11&ZD189), the National Natural Science Foundation of China (61173095) as well as the National Natural Science Foundation of China (61202193).

References

1. Xiao, G.Z.: Theory of research on Chinese grammar. Central China Normal University Press, Wuhan (2001)
2. You, Q.X.: The Investigation and Analysis of the Degree of the Ambiguity. Chinese Language Learning 5, 16–19 (2000)
3. Ni, C.Y., Xu, N.W.: The Theory on Quantifier's and Momentum's Grammar Complementation. The Journal of Nanjing Normal University (Social Sciences Edition) 3, 124–125 (1995)
4. Li, Y.M.: "One v...numeral" Structure and Subjective Quantity Question. Yunmeng Journal 3, 70–73 (1999)
5. Ma, Q.Z.: Timing Quantifier Object and the Category of Verb. Journal of Chinese Language 2, 86–90 (1981)
6. Li, Y.M.: Subjective Quantity Cause. Chinese Language Learning 5, 3–7 (1997)
7. Ma, Q.Z.: Chinese Verb and Structure of Verb Character. Beijing University Press, Peking (2004)
8. Zhu, D.X.: Lectures on Grammar. Commercial Press, Peking (1985)
9. Xiao, G.Z.: The connotation constitution and developing choice of modern Chinese grammar in 21th century—expanding main part and enhancing two wings. The Journal of East China Normal University 3, 32–40 (2004)
10. Xiao, G.Z.: The information function and semantic value of adjective and dian–thinking on information conveying grammar (second). Chinese Teaching in the world 4, 39–46 (1999)
11. Xiao, G.Z.: The Excavation and Exploration of facts and theories in Chinese Grammar. Hu Bei People's Press, Hubei (2005)

Description of the Lexical Meaning Structure of Evaluated Speech Act Verb and Its Synset Construction

Shan Xiao

International Education College, China University of Geosciences, Hubei, China
wdxshan@yahoo.com.cn

Abstract. The domestic and abroad research of network of words which based on synset has several problems: less stringent structure, rough semantic particle size, Limited range of applications and so on. The paper considers the Parataxis Network based on 'Synset-Lexeme Anamorphosis' Method is one of the effective attempts to solve these problems. There have two fields of this network: One is Vertical Semantic Aggregation Network which discusses the construction of synset and semantic feature comparison of each Speech Act Verb. The other one is Horizontal Syntactic Combination Network which talks about grammatical feature of each speech act Verb of different synset and the choices of semantic roles on syntactic configuration. This paper introduced the 'Evaluation Class' of speech act Verb based on characteristic sense analysis, preliminary discussed the description of the lexical meaning structure and its synset construction rules.

Keywords: Synset, Evaluated Speech Act Verb, 'Synset-Lexeme Anamorphosis' Method.

1 Introduction

The arrival of the 21st century is the era of comprehensive development of Informationization. According to the extensive and in-depth application of Computer and Internet, network information has affected all the aspects of human life as one of the very important resources in today's society.

Semantic Web, the third generation of web we are in now, was coined by Tim Berners-Lee, the inventor of the World Wide Web and director of the World Wide Web Consortium (W3C) in 2001. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users. This web was hoped to have the ability of judgment and inference to solve the problems currently that the computer can not understand the semantics of web content and lower precision of useful information online. For our country, China is to use computer to realize the purpose of automatic processing on a variety of Chinese information online. 'It can not be calculated by the money that the benefit to our China's science and technology, culture and

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 578–586, 2013.

education, economic construction and national security by each step increased on Chinese information automatic processing' [9].

The construction and realization of Semantic Web need the support of natural language processing techniques. Our China's computer and language scientists also worked very hard on Chinese character processing for about 20 years. However, they found that Chinese information processing should change into paying more attention to how to tell enough language knowledge for machines to 'understand' and respond to complex human requests based on their meaning by most economical mode. So the meaning-based semantic resource construction emerges as the times required. The particularity of Chinese Vocabulary, such as the greater difficulty on automatic Chinese segmentation compared to the Alphabetic writing (e.g. English) or the similarity of Chinese vocabulary and sentence syntax rules, requires improving the processing technology of the vocabulary itself, especially the semantic. This improvement can also help improving the processing technology of the sentences for enhancing the computer automatic 'capacity of acquisition'.

2 Speech Act Verb and Idea of Parataxis Network Construction

2.1 About Speech Act Verb

Speech activity is one of the primary means of passing and exchanging information among people in society, which mapped to the vocabulary system is a series of Speech Act Verbs. The traditional linguistics studies of Speech Act Verb are often not isolated. They are basically attached to some other grammar content studies, such as Verbs Valence, Lexical Analysis. Meanwhile, the method and perspective of research at that time was limited. Some scholars just gave out the examples to illustrate what is the Speech Act Verb; some of them defined the Speech Act Verb by researching on its hyperonym 'Say'. There were very few scholars like Shouman Zhong,Dawei Liu, began to make a special study of that Verbs at the beginning of 21st century, but no one gave out an unified and comprehensive definition for it. We think the core semantic components of Speech Act Verb are all connected with the 'Speech' action. The figure is as follows.

Therefore, we may define the Speech Act Verb is a class of verbs that the speaker (speakers) use effective linguistic signs to express a certain intention or pass the information to the target listener (listeners).

As an important part of the human linguistics system, typical semantic network system research on Speech Act Verb is helpful for exploring the generation and understanding of the natural language. It is also useful for developing the formal description of semantic structure and providing plenty of stimulation on Computer Simulation of the Natural Language including the Verbs System for both linguistics and computer scholars.

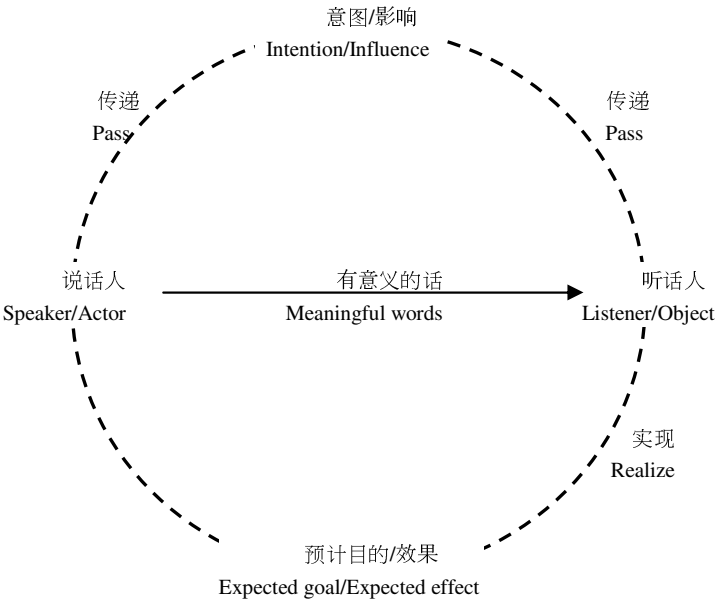


Fig. 1. Speech Action Form

2.2 Parataxis Network Construction

The foundation of Parataxis Network Construction is the 'Synset-Lexeme Anamorphosis' Method [10] which was advanced by Guozheng Xiao in 2007.His theory was built up on basic cognitive law of understanding the objects for people.

Cognitive prototype always exist when people begin to understand the objects and they are reflected in the vocabulary system constitutes the primary words. We called the position of these words 'Basic Lexemes' or 'Word-ontology', and the other words which called 'Non-basic Lexemes' are the variant relied on them. They are formed by different values of concepts in different attributes. Basic Lexemes and Non-basic Lexemes together constitute the Synonym Synset. So the key ideology of 'Synset-Lexeme Anamorphosis' Method is that the Semantic System of a language is a Synonym Synset which made up of Basic Lexemes and their Variants.

Speech Act Verb is one part of the Modern Chinese Vocabulary. According to the theory we mentioned just now, the Semantic System of Speech Act Verb is also a Synonym Synet which made up of Speech-act conceptual basic lexemes and their variants. There have two fields of this network: One is Vertical Semantic Aggregation Network which discusses the construction of synset and semantic feature comparison of each speech act verb. The other is Horizontal Syntactic Combination Network which talks about grammatical feature of each speech act verb of different synset and the choices of semantic roles on syntactic configuration.

Semantic feature comparison between Basic Lexemes and Non-basic Lexemes can show out their relations and oppositions thus can be used for describing the complicated semantic or grammatical relations among the words and synsets. To a great

extent, this method can satisfy the detailed requirements of semantic formal description of Natural Language Understanding, sum up the accurate language rules and build up large-scale semantic knowledge base to help the machine to deal with the Natural Language more scientific and faster like the human brain. The ultimate goal is for the realization of Human-Machine Cooperation and Human-Computer Dialogue.

2.3 Objects Classification and Case Study

Speech characteristics separate the Speech Act Verb from the other types of verbs obviously and basically. Therefore, we may classify the Speech Act Verb into 4 classes by Speech characteristics from the semantic structures.

- 'Interrogative Class': it emphasizes the behavioral reasons or conditions, such as 询问 (*xunwen*, enquire) , 征询(*zhengxun*, request)
- 'Interactive Class': it emphasizes the quantities and statue of actors, such as 交谈 (*jiaotan*, converse), 采访(*caifang*, interview)
- 'Object-point Class': it emphasizes the content of speech activities which point directly to the object (objects). According to the different speech behavior, we may divide the class into 3 kinds: one emphasizes the opinions or perspectives which called 'Evaluation' Class ,such as 点评(*dianping*, comment on); one emphasizes the actor's emotion which called 'Emotion' Class, such as 唾骂 (*tuoma*,spit on and curse);the other one emphasizes the information to be passed which called 'Inform' Class, such as 通知(*tongzhi*, notify).
- 'Imperative Class': it emphasizes the behavior to achieve some purpose, such as 请求(*qingqiu*, ask for)

Due to the length of limitation, we only choose the important and classical part 'Evaluation Class' as one case for studying. All the sentences related or verb examples come from 'Modern Chinese Dictionary' [4] unless particular marked.

3 'Evaluation Class' of Speech Act Verb

3.1 Semantic Features and Re-classification of EC

The Evaluation Class of Speech Act Verb is one kind of verbs which emphasize the opinions or perspectives of Actors. The most obvious semantic feature is 'evaluate', and '评价(*pingjia*, evaluate)'is its representative word. We know that different people have different views on same thing or person, so the contents of evaluation have personal emotion included. '高(*gao*,high)/好(*hao*,good)'belong to positive evaluation , while'低(*di*,low)/ 坏(*huai*,bad)'belong to negative evaluation. There are also some objectivity comments without any emotion. Therefore, we may classify the Evaluation Class of Speech Act Verb into 3 categories with some examples for each (Table 1).

Table 1. Categories of Evaluation Class of Speech Act Verb

Semantic categories	Examples
Neutral Evaluation	*评价*(*pingjia*, evaluate) *评论*(*pinglun,* discuss/comment on) *评说*(*pingshuo*, evaluate) *点评*$_1$(*dianping*, comment on)[1]
Position Evaluation	*称赞*(*chengzan*, praise),*赞叹*(*zantan*, highly praise) *赞美*(*zanmei* , eulogize), *夸奖*(*kuajiang*, praise)
Negative Evaluation	*恭维*(*gongwei*, flatter),*奉承*(*fengcheng*, flatter) *斥责*(*chize*, reprimand),*斥骂*(*chima*, scold) *责备*(*zebei*, blame)

We only approach the 'Basic lexeme – Variant' of Negative Evaluation Class of Speech Act Verb from different angles systematically in order to know by a handful the whole sack.

3.2 Semantic Relationship Type of 'Basic Lexeme – Variant' of NEC

As the section 2.2 points out, the Semantic the system of a language is a Synonym Synset which made up of Basic Lexemes and their variants. In accordance with the logical links, their semantic relationship can be concluded in 'Three Relationships and Four Matrixes'. Three relationships refer to 'Same-statue Relationship', 'Up and Down-statue Relationship' and 'Neighborhood-state Relationship'. Four Matrixes refer to 'Same Statue' 'Up Statue' 'Down Statue' and 'Neighborhood Statue'. As shown in Figure 2 below.

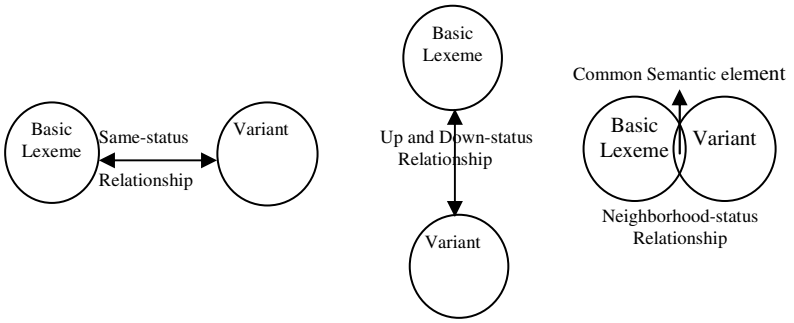


Fig. 2. 'Three Relationships and Four Matrixes' of 'Basic Lexeme – Variant' of EC

[1] '点评$_1$ (*dianping,* comment on)'means the second sense of this word in dictionary. Followings are the same.

For the Synset of Negative Evaluation Class of Speech Act Verb, it is in the 'Down Statue' of the whole Speech Act Verb Synset. Because of different semantic focus, it can be subdivided into different Synonym Synset. Every Synset also have its own basic lexemes and variants. They are equally suitable for 'Three Relationships and Four Matrixes'. The following describes in details.

Same-Statue Relationship

This means the basic lexemes and their variants only have the difference on word forms, but the connotation and the referent are the same. And they can be replaced each other in one same syntactic structure. So they have the same statue in one Synset.

'恭维(*gongwei*, flatter)' and '奉承(*fengcheng*, flatter)' belong to a synset which contain the meaning of ironical and negative evaluation. It sounds like kinds of praise by positive word form, but the deep meaning is that 'Going too far is as bad as not going far enough'. These two words can replaced each other in one same sentence as follows.

(1) 你别当面恭维我，我不相信你的话!

(ni bie dang mian gong wei wo , wo bu xiang xin ni de hua !)

You mustn't flatter me .I don't believe you!

(2) 你别当面奉承我，我不相信你的话!

(ni bie dang mian feng cheng wo , wo bu xiang xin ni de hua !)

You mustn't flatter me .I don't believe you!

Their semantic structures are the same expressed as follows.

'fengchen'/'gongwei':

[Behavior+Object+Praise/Caterto+Nice words+Intentionally+Actor]

They were mutual Same-statue variants and together constitute the 'Same-statue Synset'.

Up and Down-Statue Relationship

This means the basic lexeme is the concept of Superordinate, and its variant is the concept of Subordinate. Not only the word forms, but also the connotation and the referent have a lot of differences between the basic lexemes and their variants. The connotations of variants are implied into the basic lexemes and can be distinguished from each other by adding or changing the semantic features. Now we give out one example to explain more clearly.

Synset:{斥责(*chize*, reprimand)，责骂(*zema*,scold)，斥骂(*chima*,scold)，呵斥(*hechi*,scold in loud voice)，苛责(*keze* ,excoriate)，痛斥(*tongchi*,berate)}

This semantic of this synset is about negative evaluation for others' shortcomings, mistakes or faultss. The basic lexeme is '斥责(*chize* , reprimand)' which emphasis on '责(*ze*,blame)' . Each variant contains this basic semantic element '责(*ze*, blame)' and adds or changes distinctive features into further meaning according to 'Degree' or 'Intensity'. Such as '责骂(*zema*, scold)' and '斥骂(*chima*, scold)' are both have the basic lexeme but emphasis on '骂(*ma)*' which means criticize someone's

shortcomings, mistakes or faultss bluntly. ‘呵斥(*hechi*, scold in loud voice)’ focus on ‘呵（*he*）’ which means ‘say loudly’ in order to show the actor’s emotion by the voice. ‘苛责(*keze* ,excoriate)’ is not the common ‘blame’, its degree is much higher than the three mentioned which means ‘excessive’ or ‘severe’. ‘痛斥(*tongchi*, berate)’ emphasis the emotion intensity which is ‘痛(*tong*, sorrowful)’ and ‘a certain degree of angry ’. So their semantic structures are expressed as follows.

‘zema’/ ‘chima’ :[Behavior+ By Negative words + shortcomings/mistakes/faults+ bluntly + scold +Actor]

‘hechi’ :[Behavior+ By Negative words + shortcomings/mistakes/faults+loudly+ blame +Actor]

‘keze’:[Behavior+ By Negative words + shortcomings/mistakes/faults+ excessively + blame +Actor]

‘tongze’:[Behavior+ By Negative words + shortcomings/mistakes/faults+ sorrowful and with anger+ blame +Actor]

Certainly, some of the variants are not just have one different semantic feature. For example, compared with the synset we just talked about , ‘抨击(*pengji)*’ is also a negative evaluated speech act verb, but it is not included in because it means attacking someone or behavior in speech or writing, so the semantic features ‘by speech or writing’ and emphasis ‘ the object or its behavior’ are emphasized. Another example is ‘申斥(*shenchi)*’ which means blaming the subordinates rigorously, so the semantic features ‘to subordinates’ and ‘rigorously/harshly’ of this word are necessary.

Neighborhood-State Relationship

This means the semantics of basic lexemes and variants are overlapped. Meanwhile, the variants have their own specific and necessary semantic features. So we may say the relationship between basic lexemes and variants look like the neighbors. We also give out one example to explain more clearly.

Synset:{责备(*zebei*,blame)、责难(*zenan*,blame/reproach) 、贬责(*bianze*,censure) 、非议（*feiyi*, censure/disapproval） 、非难 (*fennan*,reprehend) 、数落$_1$(*shuluo*,reproach)、批驳(*pibo*,refute)}

The words in this synset also contain the same semantic element ‘责(*ze*,blame)’which refers to related features —— ‘shortcomings/mistakes /faults of objects’ and ‘complaints/negative comments’. But it is totally distinguished from the previous synset we mentioned. Such as ‘责难(*zenan*)’is the combination of ‘责(*ze*, blame)’and ‘难(*nan*, questioned faults and criticize)’,and the semantic feature ‘questioned and criticize’ can not be omitted. ‘非难 ’ not only includes the ‘责(*ze*, blame)’ but also has its own necessary semantic feature —— ‘责问(*zewen*, questioned)’.That means the actor questioned or inquired the objet while blaming , in order to make the objet feel shamed or bring into disrepute. Compared with other variants, ‘数落$_1$’ has another integrant semantic feature——‘数(*shu*, enumerate/list)’ which means denying other’s opinion or behavior by point out the shortcomings,

mistakes or faults one by one ,except the shared semantic element '责(*ze*,blame)'.So their semantic structures are expressed as follows.

'zenan' :[Behavior+ By Negative words + shortcomings/mistakes/faults+ blame and questioned and criticize +Actor]

'feinan' :[Behavior+ By Negative words + shortcomings/mistakes/faults+ blame and questioned +Actor]

'shuluo':[Behavior+By Negative words + shortcomings/mistakes/faults+ enumerate and blame +Actor]

Corresponding to the computer processing of lexical semantics, if we find out all the basic lexemes as the ontology of each synset, all the members (including the basic lexemes and variants) of synset can be activated and connected together by these three relationships. In this way, the speed and efficiency of dealing with the lexical semantics of computer will be enhanced greatly.

For example, if the computer want to identify the Speech Act Verb '斥责(*chize* , reprimand)'and some of the variants in this synset such as '责骂(*zema*,scold)' '痛斥(*tongchi*, berate)' '申斥(*shenchi*, scold the subordinates rigorously)', then it can active every node and connection relationship by semantic features of synset. The figure is as follows.

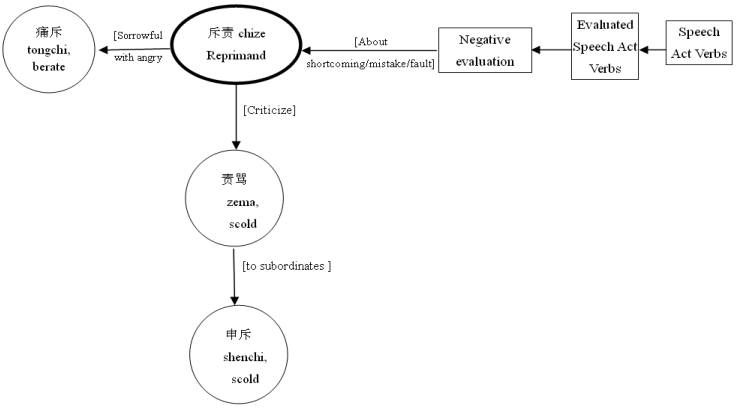


Fig. 3. Computer Identification of synset '斥责 (*chize*, reprimand)'

4 Conclusions

This paper introduced the Evaluated Speech Act verb based on characteristic sense analysis, discussed and explored the description of the lexical meaning structure and its synset construction rules according to the 'Synset-Lexeme Anamorphosis' Method. We draw the conclusion that the vocabulary system was made up of different synset. Each synset has its core basic lexemes and their variants by changing or adding one or more semantic features, and consists of the relation of 'Same-statue Relationship' 'Up and Down-statue Relationship' and 'Neighborhood-state Relationship'.

The variants can be classified into ‘Same-statue variant’ ‘Up and Down-statue’ and ‘Neighborhood-state’ accordingly. This Parataxis Network based on ‘Synset-Lexeme Anamorphosis’ Method can solve the problem such as less stringent structure, rough semantic particle size, limited range of applications of synset nowadays and meet the actual needs of the Computer Processing of Natural Language.

Acknowledgments. This work was supported by a grant from the Key Construction Program of the National (‘985’ Project) (No.985yk006), the Fundamental Research Funds for the Central Universities (CUG110836).

References

1. Ballmer, T.T.: Speech Act Classification: A Study of the Lexical Analysis of English Speech Activity Verbs. Springer, Berlin (1981)
2. Cui, X.L.: The Understanding and Recognize of Language. Peking Language and Culture Press, Peking (2001)
3. Chen, C.L.: Research on Syntactic and Semantic Attributes of Modern Chinese Verbs. Xuelin Press, Shanghai (2002)
4. Editorial office of Institute of Linguistics of Chinese Academy of Social Sciences: Modern Chinese Dictionary (Fifth edition). The Commercial Press, Beijing (2005)
5. Fu, H.Q.: Analysis and Description of Meanings. Language Publishing House, Beijing (1996)
6. Prabhu, N.S.: Second Language Pedagogy. Oxford University Press, Oxford (1987)
7. Zhan, R.F.: Semantics of Modern Chinese Language. The Commercial Press, Beijing (1997)
8. Zhong, S.M.: Semantic Cognitive Structure of English and Chinese speech act verbs. University of Science and Technology of China Press, Anhui (2008)
9. Xu, J.L.: Status and Prospects —— Issues on Chinese Information Processing and Modern Chinese. Chinese Language 6(2) (2001)
10. Xiao, G.Z.: Description of the Lexical Meaning Structure of the Verb ‘da’ and the Construction of its Synset —— A Study on “Synset-Lexeme Anamorphosis” Shared by Human and Compute: Chinese Computing Technologies and Related Linguistic Issues. In: Proceedings of the 7th International Conference on Chinese Computing, pp. 4–7. Electronic Industry Press, Beijing (2007)
11. Searle, J.R.: Expression & Meaning: Studies in the Theory of Speech Act. Cambridge University Press, Cambridge (1979)
12. Zhu, D.X.: Lectures on Grammar. The Commercial Press, Beijing (1982)

The Insights of Primitive-Primitive Structure into ELT through Task and Activity

Xiaohua Liang[1] and Guozheng Xiao[2]

[1] Foreign Languages School, Zhongnan University of Economics and Law
Wuhan, China 430073
[2] Center for the Study of Language and Information, Wuhan University,
Wuhan, 430072
Susan64@126.com, gzxiao@foxmail.com

Abstract. Task and activity serve as elicitors in ELT (English Language teaching) by both researchers and teachers. However, different researchers have different interpretation about these two concepts, which caused controversy in ELT for years. Xiao Guozheng puts forward the theory of Primitive-Primitive Structure to clarify the differences of these two terms and reveals the interrelationship between these two concepts through the analysis of their semantic meanings from the perspective of Primitive-Primitive Structure, which may give significant insights into ELT.

Keywords: task, activity, primitive, primitive structure, ELT.

1 Introduction

Task and activity are employed by both teachers and researchers to elicit students' language use [1]. For researchers, task and activity are used to document learners' language development while for teachers, task and activity are used as opportunities to develop learners' L2 language proficiency through communication [1]. However, controversy arises in teaching and research due to the diverse interpretation of the concepts of task and activity, and most researchers use task and activity interchangeably, although the argumentations are mostly over the communicativeness and non-communicativeness, openness and closedness, one-way or two-way communication. This obscurity and confusion impact teachers' task design and task implementation as well as their evaluation of classroom activities. The Primitive-Primitive Structure theory clarifies the confusion between these two concepts, which is hoped to offer great insights into ELT.

2 Task and Activity from a Cognitive Perspective

2.1 Task and Activity in the Western Teaching and Research

The "task-based" syllabus originated in the US military training in the 1950s [2] and developed to include, in the 1970s, communicative language teaching, and became

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 587–592, 2013.

the main attraction in applied linguistics. Tasks are used in ELT to develop the L2 language proficiency [1], and offer students more opportunities for communication. Most researchers take task as activity. For instance, Kumaravadivelu stated that these two terms are used interchangeably in his explanation. Ellis gave the similar interpretation in his talk and work. Many researchers defined these two terms in the same way, for instance, Littlewood [3] pointed out that tasks are activities in which students work purposefully towards an objective set by the students themselves or the teacher [2]. Willis defined task as activity where the target language is used by the learner for a communicative purpose, and he emphasized the communicative goal. Prahbu [4] pointed out that task is "an activity which requires learners to arrive at an outcome from given information through some process of thought", and he emphasized the process, information input and the task completion. Nunan [5] defined task as "activity in which learners communicate with the target language", and he emphasized the language use. Ellis [1] pointed out that task is a "work plan". At the same time he emphasized that like other teaching and learning activities, task has input and output in both oral and written forms. Although the foci of the definition given by the researchers are different, most of the researchers use task and activity interchangeably, i.e., task is an activity.

2.2 The Framework of Task

Ellis [1] formulated a framework of task in teaching in addition to defining the concept task. Through a great deal of research on tasks and a detailed description about these studies, Ellis [1] categorized the tasks, set the criterion and requirements. Based on these studies, he conceptualized the framework of tasks: tasks should have "goal", "input", "condition", "procedures" and "predicted outcome".

It can be seen from this framework that Ellis emphasizes on the task design by the teachers. This framework works as a guideline for the teachers in task design. However, the confusion about the two concepts of task and activity limited teachers' task design and task implementation as well as the utilization and evaluation of activities, as nowadays, teaching is more student-centered and learning-centered, where teacher-student and student-student activities, i.e., their performance of the task, are emphasized in the classroom, of which teachers' task design is only a part.

2.3 The Impact on ELT in China

Task-based language teaching was introduced in China, in the late 1990s, influenced by the western teaching model, and a great deal of research has been conducted in this area. Gong [2] noted that "task is a goal-directed activity people are engaged in their daily life, work and entertainment". As this is a western teaching model contextualized into the Chinese context, the definition of task and activity show western elements, i.e., "task is an activity…"

The obscurity and confusion of these two concepts showed great impact on teachers' task design and implementation, the utilization and evaluation of activities. Liang [6] conducted a 2-year classroom observation in an English immersion primary school

in Guangdong Province and collected 147 hours of audio as well as video recordings. Through the analysis, she found that the teacher designed and assigned the tasks very briefly. All the tasks were assigned as "talking about…", or "make a dialogue about …". One example from the data set is given below.

In Unit 2 of Grade 4 English textbook by People's Education Press, six pictures show that Zip sets the clock on hour earlier in order to wake up Zoom and get him out of bed earlier. The task the teacher assigned the students was to act out the dialogue and choose the best actor/actress and the best group. When interviewed after class, the teacher explained that the purpose of this task was to let the students recite the text and act it out when they could recite it.

The task seemed to be the same to all the students, i.e., to act out the dialogue. However, as the three groups from the dataset showed, the students' activities were completely divergent. The students in the first group were drawing a clock, which was used as a prop in their acting. Their performance was very successful, which led to the nominalization of the best group and the best actress. The students in the second group, who came from a lower level of language proficiency, practiced the dialogue three times very seriously, helping each other through suggestions, exemplifications, questions, illustrations, explanations about the goal, the order, the tone and the rhythm. However, when they came to the front, they stuttered and could only read the dialogue from the textbook, which led to the criticisms by the teacher. The third group joined the other group because one of them felt sick and went home. The boy in the third group fought enthusiastically to persuade the students in that group to give him the role, through exaggerations, repetition and misleading the other group members to believe that acting as the prop (the clock) was very interesting, and that he would like to be the clock if he had not been given the digital recorder for recording. In the end, he won the role as he had hoped, and was extremely excited in the performance and laughed all the way, which led to the criticisms by the teacher.

The description of these three groups shows the following three points: First, as the teacher used task and activity interchangeably, and he could not distinguish these two in his teaching or reflected on how to design the task effectively as long as his assignment led immediately to the students' activities. Secondly, due to the obscurity and confusion about the two concepts of task and activity, the teacher did not take into careful consideration of the interrelationship between task and activity and how tasks functioned well in the student activities. Lastly, the teacher's task contained an overt goal (acting well) and a covert goal (reciting the text), which the teacher did not state very precisely. At the same time, he neglected the multiple sources of reaching the purpose/goals of the activity. The teacher based his evaluation of student activities not on the process but on the outcome, which was shown through the teacher's discouragement on the students by his criticisms. Such frustration may hinder the students' English language learning and made them demotivated in their performance. Therefore, to clarify the concepts of task and activity and the semantic meanings of these two is of great significance in ELT.

3 Task and Activity from Primitive-Primitive Structure Perspective

3.1 Task and Activity in Chinese

Task and activity are defined in the Chinese dictionaries in nearly the same way by most Chinese scholars and researchers. Due to the limit of the layout, the definitions from two of the most influential definitions are selected as exemplification. Hu [7] defined task in The New Ancient and Modern Chinese Dictionary as "work assigned, responsibility undertaken", and activity as "goal-directed action and behavior". The definition of task in The Modern Chinese Dictionary compiled given by the Dictionary Editorial Office of the Language Institute of the Chinese Academy of Social Sciences is "work appointed to take, responsibility to undertake", and activity is "the action and behavior for certain purpose".

3.2 The Primitive Structure of Activity

Xiao [8-10] formulated the theory of Primitive-Primitive Structure when he found in his research on word-marking in computational linguistics the basic and core elements within each word when marked. He named these basic and core elements primitives and the interrelationship among these elements primitive structure

According to the theory of Primitive-Primitive Structure, the definitions of task and activity from the Chinese dictionaries indicate that both task and activity have two semes. For task, the first seme contains four primitives: superior, assigning, subordinate and work, and the second seme contains two primitives: subject and responsibility. For activity, the first seme contains three primitives: subject, purpose/goals and action/behavior, and the second contains two primitives: thing, functions or effects. According to the relationship between primitives, the primitive structure of task and activity can be annotated as follows:

Task:

1. [superior]^[assigning]^[subordinate]^[work]
2. [subject]^[responsibility]

Activity:

1. [subject]^[purpose]^[action/behavior]
2. [things]^[functions]v[effects]

Based on the primitives of task and activity, and the interrelationship between these primitives, the primitive structure of activity can be illustrated in the following figure.

As shown in the figure, activity is goal-directed, or task-driven/ problem-driven/ process-driven (students may find communicating with each other very interesting)/ outcome-driven (getting rewards), etc.. This means that one of the purposes of activity is completing the task, and there may be other purposes. The interrelationship between task and activity is revealed at the node "purpose/goals". The purpose/goals of the activities may be set by the students, or the teacher. The subject may be the teacher-student or student individuals or student-student such as pairs and groups.

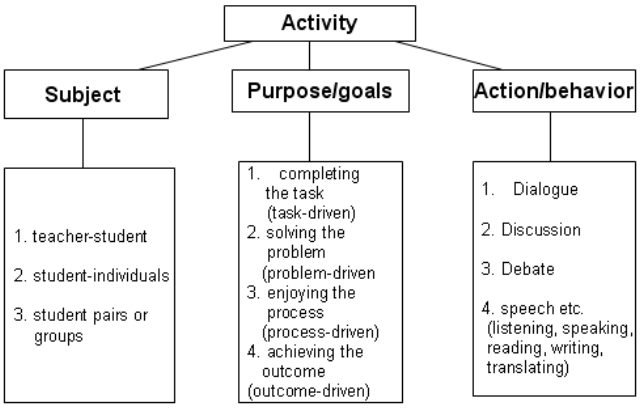


Fig. 1. Primitive structure of activity

In teaching and learning, task usually refers to pedagogical tasks. The superior often refers to the teacher, and the subordinate the students. Furthermore, work has clear requirements and goals, and assigning means that the teacher should inform the students the specific requirements in the process as well as the predicted outcome together with the mediations students can utilize in detail and with clarification. This meshes with Ellis's [18] framework, and further proves that Ellis's framework is not for the students' activities but for the teacher's task design. The annotation of the Primitive Structure of activity is very insightful to ELT.

3.3 The Insights into ELT

The primitive structure of activity, as shown in the figure above, gives insights into ELT, as it clears the confusion of task and activity, and clarifies clouds over the semantic meanings of these two concepts, i.e., task is directed to the purpose/goal of activity; teacher's task design is shown in the process of giving the information of the object and how to achieve the outcome by conducting the actions, through which students' agency emerges. Furthermore, the primitive structure of activity shows the multi-source objects, and task is only one of them. This reinforces the fact that teachers should design the task carefully and assign the task informatively and clearly, and employ student activities dynamically for the enhancement of teaching and learning.

4 Conclusion

The primitive structure of activity illustrated the interrelationship between task and activity, which clarifies the two confused concepts, and it is hoped to offer great insights into ELT. Task is different from activity, and it requires the teacher to design very carefully and assign clearly to the students to ensure that the students grasp the objects in their activities. The multi-source goal-directed activity may help the teachers to reflect on how to better employ the activities dynamically and how to strengthen the functions and effects of activity from multiple dimensions.

References

1. Ellis, R.: Task-based language learning and teaching. Oxford University Press, Oxford (2003)
2. Gong, Y.F., Luo, S.Q.: renwuxing jiaoxue (Task-based language teaching). People's Eduaction Press (2003)
3. Littlewood, W.: The task-based approach: Some questions and suggestions. ELT Journal 4, 319–326 (2004)
4. Prabhu, N.S.: Second language pedagogy. Oxford University Press, Oxford (1987)
5. Nunan, D.: Task-based language teaching: A comprehensively revised edition of designing tasks for the communicative classroom. Cambridge University Press, Cambridge (2004)
6. Liang, X.H.: Investigation on the mediation of activities into student peer talk in an English immersion context in China. Shanghai Foreign Language Education Press (2011)
7. Hu, Y.S.: Xinbian gujin hanyu da cidian (The new ancient and modern Chinese dictionary). Shanghai Lexicographical Publishing House (1991)
8. Xiao, G.Z.: Hanyu yufa de shishi fajue yu lilun tanshuo (The fact excavation and theory exploration of Chinese grammar) (2005)
9. Xiao, G.Z., Hu, D.: Xinxi chuli de hanyu yuyi zhiyuan jianshe xianzhuang yu qianjing zhanwang (Information processing analysis of the present situation of Chinese semantic resource construction and prospect). The Journal of Changjiang Academic 2, 86–91 (2007)
10. Xiao, G.Z., Ji, D.H., Xiao, S.: Ontology de leixing ji hanyu ciwang de ontology jiegou (The dichotomy of ontology and the ontology of Chinese wordnet structure). The Journal of Changjiang Academic 2, 115–117 (2011)

Verbal Empty Categories and Their Types in Mandarin

Aiping Tu[1] and Lei Zhang[2]

[1] Centre for Study of Language and Information, Wuhan University, Wuhan, China
tuaiping81@163.com
[2] Department of Chinese, Translation and Linguistics, City University of Hong Kong, Kowloon, Hong Kong
leizhang@student.cityu.edu.hk

Abstract. Empty categories are base-generated in the deep structure and do not have any overt expression in the surface structure. In this paper, we claim that Mandarin has verbal empty categories besides nominal empty categories, and further investigate their types. According to their semantic contributions, we argue that verbal empty categories in Mandarin mainly include two types: (a) the covert copular *shi* 'be' and other linking verbs such as *chengwei* 'become'; and (b) notional verbs. In the case that the empty verb is a notional verb, what the covert notional verb is, highly depends on the context.

Keywords: covert form, verbal empty category, deep structure, surface structure, semantic contributions.

1 Introduction

An expression may not present all constituents of deep structure according to the economy principle in some cases, which reflect the basic semantic relations, in the surface structure with overt phonetic forms. As a result, we can often see empty constituents which are base-generated in the deep structure but exist covertly in the surface structure, such as the *t*, *PRO* and *pro*, which occupy the positions of the nominal constituents in the syntactic structure.

It is generally acknowledged that empty categories have been studied with a main focus on nominal empty categories which occupy the position of the nominal constituents in the syntactic structure. Verbal empty categories, however, have been less discussed. Empty verbs are base-generated in deep structure and serve as the assigner of theta-roles and Cases, but take covert forms in the surface structure (symbolized as 'Ø'). Consider below.

(1) a. Jintian Ø xingqitian. (Type Ⅰ)
today Sunday
'Today is Sunday.'
b. Jintian Ø 2012 nian qi yue liu ri. (Type Ⅱ)
today 2012 year seven month six day
'Today is July 6th, 2012.'

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 593–602, 2013.

(2) a. Wo jia Ø liang ge haizi. (Type Ⅲ)
I family two CL child
'There are two children in my family.'
b. Tamen Ø laopengyou shide. (Type Ⅳ)
they old friend seem
'They look like old friends.'
c. Ø Da guniang le. (Type Ⅴ)
big girl SF
'She becomes a big girl.'
(3) Zhangsan Ø liang ge pingguo. (Type Ⅵ)
Zhangsan two CL apple
'Zhangsan has two apples.'

Sentences, like (1)-(3), which only have nouns but do not take verbs in the predicate are often called noun-predicate sentences. In the previous studies, most linguists explain it with terms like *omission*, *parataxis, predicator* and so on [1-5]. Under the framework of syntax, three types of sentence without a verb were discussed [6], in which type Ⅰ sentences are noun-predicate sentences, and type Ⅴ and type Ⅵ sentences are empty copula sentences and empty verb sentences, respectively. Moreover, it is suggested that the empty verb in a type Ⅴ sentence belongs to empty categories [7].

The phenomenon of verbal empty categories is mentioned by several scholars in recent years [8-13]. Among them, the term *verbal empty category* (abbreviated as VEC) was used to illustrate this phenomenon without proving the existence of VECs and investigating their types [13]. In this paper, we suggest that sentences in (1)-(3) contain typical VECs, and utilize evidences to show the existence of VECs. Furthermore, the VECs have been divided into two types: the covert copular *shi* 'be' and other linking verbs, as well as notional verbs.

2 The Existence of VECs in Mandarin

The phenomena that modifiers of verbs, tense auxiliaries, overt nominal expressions and obligatory arguments can stand alone without an overt verb provide the evidence of the existence of VECs. In this section, we will discuss the possibility and rationality of the existence of VECs.

2.1 The Attachment of Modifiers and Tense Auxiliaries to Verbs

Based on the presuppositional constraint on modification [14-15], the denotation of modified constituents cannot be empty.

Therefore, it can be deduced that the occurrence of verbal modifiers and tense auxiliaries require the existence of the modified verb overtly or covertly. A sentence composed of 'NP *le*' expresses the status considered changes. In which *le* selects a VP as its complement and indicates the change has already become a fact. In addition,

yijing is an adverb to modify VPs. Thus we can test whether the relevant sentences without overt verbs take covert verbs via inserting sentence final *le* and the adverb *yijing*. It turns out that the two particles can be inserted to the sentences considered. See below.

(4) a. Jintian yijing xingqitian le.
today already Sunday SF
'Today has already been Sunday.'
b. Jintian yijing 2012 nian qi yue liu ri.
today already 2012 year seven month six day
'Today has already been July 6th, 2012.'

(5) a. Wo jia yijing liang ge haizi.
I family already two CL child
'There have already been two children in my family.'
b. Tamen yijing laopengyou shide.
they already old friend seem
'They have already looked like old friends.'
c. Yijing da guniang le.
already big girl SF
'She has already been a big girl.'

(6) Zhangsan yijing liang ge pingguo.
Zhangsan already two CL apple
'Zhangsan has already had two apples.'

It can be seen that the insertion of modifiers and tense auxiliaries can prove the existence of VECs.

2.2 The Case-Feature Checking of Overt NPs

Nominal empty categories are covert obligatory arguments determined by the theta-role assigners that are denoted by VPs. In (2a), the overt constituents such as *jintian* 'today' and *xingqitian* 'Sunday' must accept the Case-feature checking, so there must be a Case assigner. The NPs in the subject position must get the nominative Case from the INFL which attached to the main verb, so it requires that there is a verb overtly or covertly in the sentence. See (7).

(7) Jintian $_{\text{VEC}}$ xingqitian.
today Sunday
'Today is Sunday.'

Now let's turn to the Case-feature checking of the NPs in the predicate position. When the NPs in the subject position equal to which in the predicate position, like (4), they have the same denotation. The covert *shi* 'be' can be interpreted as 'be equal to' here, it can assign the accusative Case to the NPs in the object position. Another type of NPs in the predicate position refers to the property when combining with a (covert) copula, like (7), acting as the predicate [16-17], so it certainly has not the requirement of Case assignment.

Therefore, it is not contradictory to admit the existence of nominal predicate sentences and admit the existence of covert copula. But only if the existence of VECs is accepted, the predicate composed of the second NP and copular can indicate the relevant property and assign Case to the first NP, and the first NP can get Case via Case-feature checking. In a similar fashion, the nominal constituents in other sentence also must meet the Case-feature checking and the fact there is a VEC in the sentence must be acknowledged.

2.3 The Theta-Role Assignment to Arguments

The theta-criterion requires a one-to-one correspondence between the argument and its theta-role. Nominal empty categories are covert obligatory arguments determined by the theta-role assigners that are denoted by VPs. Vice versa, it can be concluded that the NPs with theta-roles must have theta-role assigners.

Based on VP-internal subject hypothesis [18], all arguments get their theta-roles within the VP, and the specifier position of IP is only the landing-site of the subject after its movement.

The nominal constituents like the ones in (3) must be arguments with theta-roles; otherwise, they cannot be interpreted in Logical Form. Though a VEC has not an overt phonetic form, it can be inferred by the relevant arguments with theta-roles and can be filled out according to the context. Consider (8).

(8) Zhangsan (you_{VEC}) liang ge pingguo.
Zhangsan (has_{VEC}) two CL apple
'Zhangsan has two apples.'

However, the NPs in (1) and (2) are not arguments in the traditional sense. Moreover, the copula cannot assign theta-role. Thus, just as discussed in 2.2, the latter NPs form the predicate together with the copula (See section 3).

3 VECs Are Linking Verbs

Though a Linking verb cannot stand alone as a predicate because it has the function of linking, it can form a predicate combining with its complement. It has a meaning, however it cannot be a predicate by itself, and can be a predicate only when it forms a predicate together with its complement [19-20]. Many VECs in Mandarin are linking verbs.

3.1 The Covert *Shi* in Type Ⅰ Sentences

Sentence type I is composed of two NPs. The second NP expresses the property or type of the first one. Copula is the only type of verb either in its covert or overt form, which can occur in this type of sentence.

Copula is a special type of linking verbs. In some previous studies, the terms *linking verb* and *copula* have been interchangeably used and thus lead to confusions.

We consider that, comparing with other linking verbs, copula has no semantic contribution. The main copula is *be* in English [21-22], and *shi* in Mandarin with its variants such as *wei(为)*, *xi(系)*, *nai(乃)* and so on. The semantic representation of copular can be described in (9):

(9) ⟦be/shi⟧=λfλx[f(x)]

Thus, VECs in type I sentences are *empty copulas*, and the interpretation of this type of sentence is not dependent on the context. See below.

(10) a. Ta (shi_{VEC}) huang toufa.
she (is_{VEC}) yellow hair
'She has blond hair.'
b. Lu Xun, (shi_{VEC}) Zhejiang Shaoxingren.
Lu Xun, (is_{VEC}) Zhejiang province Shaoxing city person
'Lu Xun is from Shaoxing, Zhejiang province.'

In (10), the latter NPs express the type or property of the former NPs. Just as overt copula, empty copula cannot assign theta-role to NPs.

It was explained that Case assignment of English copula sentences with the raising analysis: Copula is a typical raising verb, whose subcategory selects a small clause/SC as its complement, and triggers the overt NPs concerned move to the subject position, and at the same time, satisfies the extended projection principle (abbreviated as EPP) [23].

Like English *be*, Mandarin *shi* here is a copula which serves as a linker. And the two NPs in the sentence are not of the same type: the former is the entity and the latter is the property which acts as the predicate.

3.2 The Covert *Shi* in Type Ⅱ Sentences

The two NPs in type Ⅱ sentences are equivalent, and the covert *shi* in the sentences can be replaced by the symbol '='. Consider (11).

(11) a. Jintian (shi_{VEC}) 2012nian qi yue liu ri. → Jintian = 2012 nian 7yue 6 ri.
today (is_{VEC}) 2012 year seven month six day→ Jintian = 2012, July 6th.
'Today is July 6th, 2012.' → Today = January 6st, 2012.
b. Dayan he, (shi_{VEC}) wo de baomu. → Dayan he=wo de baomu.
Dayan river (is_{VEC}) I DE baby-sitter → Dayan river = I DE baby-sitter
'Dayan river is my baby-sitter.' → Dayan river is my baby-sitter.
c. Beijing, (shi_{VEC}) weida zuguo de shoudu.
→ Beijing = weida zuguo de shoudu.
Peking, (is_{VEC}) great motherland DE Capital
→ Peking, (is_{VEC}) great motherland's Capital
'Peking is the Capital of the great motherland.'
→Peking = the Capital of the great motherland.

In the above sentences the function of *shi* is equivalent to the linking verb *equal* in English, and has its semantic contribution. Its semantic representation can be described as follows.

(12) ‖ *be / shi* ‖=λxλy [x=y]

It is suggested that the sentences like (11) should be regarded as the compound of two DPs which have the apposition relation [24]. One of them must move to the subject position to get the nominative Case, and the two DPs can share the same Case because of their apposition relation.

The analysis has differentiated predicative sentences from descriptive sentences in copula sentences, but has not solved the problem of theta-role assignment. Moreover, the appositives occupy only one syntactic position, and cannot form a sentence. We can observe it from the transformation in (11), because of the apposition relation between the two NPs, type II sentences do not require the context to help understand, and the VECs are empty linking verbs.

3.3 Covert Linking Verbs in Type Ⅲ-Ⅴ Sentences

Type Ⅲ-Ⅴ sentences are existential sentences indicating the presence, figurative sentences, as well as the 'NP *le*' sentences respectively. VECs in these three types of sentences do not require the context help to understand them, and can be interpreted as the linking verbs *exist*, *look like* and *become* respectively, Look at (13).

(13) a. Wo jia (cunzai$_{VEC}$) liang ge haizi.
I family (exist$_{VEC}$) two CL child
'There are two children in my family.'
b. Shafa shang (cunzai$_{VEC}$) yi dui yifu.
sofa on (exist$_{VEC}$) one CL clothes
'There are a pile of clothes on the sofa.'
c. Tian shang (cunzai$_{VEC}$) yi ge taiyang, shui zhong (cunzai$_{VE}$) yi ge yueliang.
sky on (exist$_{VEC}$) one CL sun, water in (exist$_{VEC}$) one CL moon.
'There is a sun in the sky and a moon in the sea.'

(14) a. Tamen lia (xiang$_{VEC}$) laopengyou shide.
they (look like$_{VEC}$) old friend seem
'They look like old friends.'
b. Haizi de lian (xiang$_{VEC}$) huar shide.
children DE faces (look like$_{VEC}$) flower seem
'Children's faces look like flowers.'
c. Gaosong de shanfeng (xiang$_{VEC}$) gudao yiban.
lofty DE peak (look like$_{VEC}$) island seem
'The lofty peak looks like an island.'

(15) a. (Ta$_{pro}$) (chengwei$_{VEC}$) da guniang le.
(she $_{pro}$) (become$_{VEC}$) big girl SF
'She becomes a young girl.'
b. Zhangsan (chengwei$_{VEC}$) lao bing le.
Zhangsan (become$_{VEC}$) old soldier SF
'Zhangsan becomes an old soldier.'
c. Ta (chengwei$_{VEC}$) daxuesheng le.
he (become$_{VEC}$) undergraduate SF
'He becomes a college student.'

Unlike the copula, but the same as *dengyu* 'to equal', the *cunzai* 'to exist', *xiang* 'to look like' and *chengwei* 'to become' have their semantic contributions and their semantic representation can be respectively described as follows.

(16) ||cunzai||=λy λx [exist (y)(x)]
(17) ||xiang ||=λy λx [look like(y)(x)]
(18) ||chengwei||=λy λx [become (y)(x)]

Therefore, it can be seen that covert verbs in type Ⅱ - Ⅴ sentences are linking verbs which can be called empty linking verbs [25-27].

4 VECs Are Notional Verbs

Mandarin VECs include not only covert linking verbs, but also notional verbs. Without the help of the discourse, the VEC in a sentence of type Ⅵ can be inferred by the theta-roles of its arguments. Usually it is the most possible one. Consider (19).

(19) a. Zhangsan (you_{VEC}) liang ge pingguo.
Zhangsan (ha_{VEC}) two CL apples
'Zhangsan has two apples.'
b. Nimen (he_{VEC}) pijiu, wo (he_{VEC}) baijiu
you (drink $_{\text{VEC}}$) beer, I (drink $_{\text{VEC}}$) white spirit
'You drink beer and I drink white spirit..'
c. Women mei ge xingqi ($\text{shang}_{\text{VEC}}$) si jie ke.
we each CL week (have_{VEC}) four CL class
'We have four classes each week.'

But if we put the type Ⅵ sentences in the specific context, the empty verb can be filled out according to the relevant context. For example, the empty verb in (19) can also be interpreted as other verbs. Look at (20).

(20) a. Zhangsan (mai_{VEC}) liang ge pingguo.
Zhangsan (buy_{VEC}) two CL apple
'Zhangsan has bought two apples.'
b. Nimen (mai_{VEC}) pijiu, wo (mai_{VEC}) baijiu
you (buy_{VEC}) beer, I (buy $_{\text{VEC}}$) white spirit
'You buy beer and I buy white spirit.'
c. Women mei ge xingqi (que_{VEC}) si jie ke.
we each CL week ($\text{absent}_{\text{VEC}}$) four CL class
'We missed four classes each week.'

Actually, it has diverse interpretations, for instance, the empty verb in *'ni liang ci, wo yi ci'*, can be one-place verbs, two-place verbs as well as three-place verbs, as shown below.

(21) a. Ni (ku_{VEC}) liang ci, wo (ku_{VEC}) san ci.
you (cry_{VEC}) two times, I (cry_{VEC}) three times
'You cry two times, and I cry three times.'

b. Ni (kan_{VEC}) (ta_{pro}) liang ci, wo (kan_{VEC}) (ta_{pro}) san ci
you (see_{VEC}) (him_{pro}) two times, I (see_{VEC}) (him_{pro}) three times.
'You have seen him two times, and I have seen him three times.'

c. Ni (wen_{VEC}) (ta_{pro}) liang ci ($wenti_{pro}$), wo (wen_{VEC}) (ta_{pro}) san ci ($wenti_{pro}$).
you (ask_{VEC}) (him_{pro}) two times question, I (ask_{VEC}) (him_{pro}) three times question
'You have asked him some questions two times, and I have asked him some questions three times.'

The verb *ku* 'to cry' in (22a) is a one-place verb, the verb *kan* 'to see' in (22b) is a two-place verb, and the verb *wen* 'to smell' in (21c) is a three-place verb. What's more, here the overt NPs can not only be interpreted as the subject but also as the object. Except for the overt NPs, there are empty pronouns. See below.

(22) a. Ni (lai_{VEC}) liang ci, wo (lai_{VEC}) san ci.
you ($come_{VEC}$) two times, I ($come_{VEC}$) three times
'You come here two times, and I come here three times.'

b. (Ta_{pro}) (kan_{VEC}) ni liang ci, (ta_{pro}) (kan_{VEC}) wo san ci.
he (see_{VEC}) (you_{pro}) two times, he (see_{VEC}) (me_{pro}) three times
'He has seen you two times, and he has seen me three times.'

c. Ta (wen_{VEC}) (ni_{pro}) liang ci ($wenti_{pro}$), ta (wen_{VEC}) (wo_{pro}) san ci ($wenti_{pro}$)
he (ask_{VEC}) (you_{pro}) two times question, he (ask_{VEC}) (me_{pro}) three times question.
'He has asked you some questions two times, and he has asked me some questions three times.'

The verb *lai* in (22a) is unaccusative, and the overt subject is the object in D-structure [28-30]. The semantic representations of an one-place predicate, two place predicate and three place predicate are illustrated in (23), (24) and (25), respectively.

(23) λPλx [P(x)]
(24) λPλyλx [P (y)(x)]
(25) λPλzλyλx [P (z) (y) (x)]

The empty verb in a sentence of type Ⅵ can not only be interpreted in isolation, i.e. be inferred from the theta-role of the overt NPs, but also be put in a specific context and be interpreted based on the context. To sum up, it has shown that there are covert verbs which assign theta-roles and Cases, and be called *empty notional verbs*.

5 Conclusions

In brief, revealing the nature of language requires comprehensive observations on the overt and covert forms of syntactic constituents. VECs are base-generated in the deep structure and have covert expressions in the surface structure, and can be inferred from the overt expressions.

Empty categories generally indicate nominal expressions like empty nouns (or pronouns). However, besides nominal empty categories, Mandarin also has VECs.

Some modifiers, tense auxiliaries (such as *le*), obligatory arguments with theta-roles and overt NPs with Cases require the existence of verbs. These verbs can be either overt or covert. When being covert, they are called VECs. The VECs in Mandarin can be the covert copulas, some other linking verbs and notional verbs, which differ in their semantic contributions.

Acknowledgments. This work has been supported by the Fundamental Research Fund for the Central Universities (201111101020002), the Major Projects of Chinese National Social Science Foundation (11&ZD189) as well as the National Natural Science Foundation of China (61173095, 61202193). We are indebted to many linguists including Guozheng Xiao, Paul Law, Liejiong Xu, Haihua Pan, Sze-Wing Tang, Dingxu Shi, Wenhe Feng and the PhD candidates and friends in City University of Hong Kong for the discussions in the writing of this paper. Needless to say, all errors are our own.

References

1. Lü, S.X.: The Examples from English and Chinese Grammar Comparison. Foreign Language Teaching and Research 2, 5–9 (1977)
2. Chao, Y.R.: A Grammar of Spoken Chinese. University California Press, Chicago (1968)
3. Qi, H.-Y.: Modern Chinese Phrases. East China Normal University Press, Shanghai (2000)
4. Xu, J.: Grammatical Principles and Grammatical Phenomena. Peking University Press, Peking (2001)
5. Zhang, H.Y., Tang, S.-W.: Licensing of Empty Categories and the Syntactic Variation of Copular Topic Sentences in Mandarin and Cantonese. Language Science 1, 58–69 (2011)
6. Tang, S.W.: Economy Principles and Chinese Verbless Sentence. Modern Foreign Language 1, 1–13 (2002)
7. Tang, S.W.: The Characteristics of the Empty Verbal Subordinate Clause. Journal of Chinese Language 1, 23–32 (2004)
8. Huang, C.T.J.: On Ta de Laoshi Dang-de Hao and Related Problems. Language Science 3, 225–241 (2008)
9. Tang, H.F.: The Empty Verb be in Chinese. Journal of Ocean University of China (Social Sciences Edition) 4, 77–80 (2009)
10. Feng, S.L.: Light Verb Movement in Modern and Classical Chinese. Language Sciences 1, 3–16 (2005)
11. Ng, S.: Processing Chinese Empty Categories. Ph.D. dissertation, The City University of New York (2009)
12. Walsh, M.D., Bunger, A.C.: Comprehension of Elided Structure: Evidence from Sluicing. Language and Cognitive Processes 1, 63–78 (2011)
13. Tu, A.P.: The Non-nominal Empty Categories in Mandarin Chinese. Modern Foreign Language (to be printed, 2012)
14. Lee, P.L.P., Pan, H.-H.: Focus and the Semantic Interpretation of Bu Sentence. Modern Foreign Language 2, 114–127 (1999)
15. Lee, P.L.P., Pan, H.-H.: Chinese Negation Marker Bu - not and Its Association with Focus. Linguistics 4, 703–731 (2001)
16. Heim, I., Kratzer, A.: Semantics in Generative Grammar. Blackwell, Massachusetts (1998)

17. Li, Y.H.A.: Plurality in a Classifier Language. Journal of East Asian Linguistics 8, 75–99 (1999)
18. Sportiche, D.: The Theory of Floating Quantifiers and Its Corollaries. Linguistic Inquiry 19, 425–449 (1988)
19. Horton, B.: What Are Copula Verbs? In: Casad, E.H. (ed.) Cognitive Linguistics in the Redwoods, pp. 319–346. Mouton de Gruyter, New York (1995)
20. Zhang, B.L.: The Identification Standards of Linking-Verbs. Language Teaching and Research 4, 48–54 (2002)
21. Chomsky, N.: Knowledge of Language: Its Nature, origin and Use. Praeger, New York (1986)
22. Crystal, D. (ed.): Modern Linguistics Dictionary, 4th edn., translated by Shen, J.-X. Commercial Press, Peking (2011)
23. Stowell, T.: What was there before there was there. Papers from the Fourteenth Meeting Chicago Linguistics Society. Chicago Linguistics Society, University of Chicago, pp. 458-471 (1978)
24. Liu, Y.A., Han, J.-Q.: On Copular Construction in English. Modern Foreign Language 2, 360–369 (2004)
25. Li, J.X.: New-edited Chinese Grammar. Commercial Press, Peking (1924)
26. Lü, S.X.: An Outline of Chinese Grammar. Commercial Press, Peking (1982)
27. Wang, L.: Modern Chinese Grammar. Commercial Press, Peking (1985)
28. Perlmutter, D.: Impersonal passives and the unaccusative hypothesis. Proceedings of the Berkeley Linguistic Society 4, 157–189 (1978)
29. Burzio, L.: Italian Syntax. Kluwer Academic Publishers, Dordrecht (1986)
30. Huang, C.T.J.: Thematic Structures of Verbs in Chinese and Their Syntactic Projections. Language Science 4, 3–21 (2007)

On the Core Elements in Sememic Description from the Perspective of Lexicographical Definition

Xinglong Wang[1,2,3]

[1] Center for Study of Language and Information, Wuhan University, Wuhan 430072
[2] College of Chinese Language and Literature, Ludong University, Yantai 264025
[3] Key Laboratory of Language Resource Development and Application of Shandong Province, Yantai 264025
wangxinglong100@163.com

Abstract. This paper, based on the structural trichotomy of the sememic description elements, mainly expounds the core elements in sememic description from the perspective of lexicographical interpretation. Firstly, it defines the core elements for words including nouns, verbs and adjectives. Secondly, it elaborates the syntax distribution of the core elements, including the gradation distribution of syntax and the location distribution of syntax. Finally, it discusses the divergence of the semantic relations among the core elements.

Keywords: sememic description, trichotomy, core elements, lexicography, definition.

1 Structural Trichotomy of Sememic Description Elements Based on Dictionary Definition

1.1 Introduction to Structural Trichotomy of Sememic Description Elements

If we are to segment the text of traditional dictionary definitions, we can break them down into some of the smallest and basic elements, which usually refer to single definition words and a few closely-knitted phrases. They constitute three modules: 1) the sememic description module in the core position — called Core Elements (Code F) — showing the essential core information; 2) the sememic description module in the main position — called Main Elements (Code S) — showing the necessary information of the trunk; and 3) the sememic description module in the modified position — called Marker Elements (Code T) — playing a connecting role and showing the necessary identification information. [1]

In addition, the elements in dictionary definitions which are usually some words (or phrases) could be called "Basic Elements". The above three modules of Lexicographical Sememic Description Elements generally can have their own Basic Elements, labeled respectively as F1, F2, … Fn (usually less in number than the other two); S1, S2, … Sn; and T1, T2, … Tn.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 603–611, 2013.

1.2 Selection and Establishment of Corpus

The corpus employed in this paper is described as follows. Firstly, the Modern Chinese Dictionary (5th edition) (MCD) is used as the corpus source. Secondly, with sememes as a unit, 500 nouns, 500 verbs and 500 adjectives are selected from the MCD randomly, and all these words to be interpreted are two-syllable Chinese words/phrases. Thirdly, the trial corpus is just formed based on these 1500 words, including of 8985 Basic Elements. Fourthly, these 1500 words are retrieved in a random order – every 11th page of MCD is employed as word source. The trial corpus is then marked and processed according to the structural trichotomy of the sememic description elements.

2 Definition of Core Elements

Core Element is the core module that contains the fundamental sememe. It is a synchronic concept and is different from Nuclear Sememe in Scholium, which is a diachronic concept [2].

Firstly, Core Element is the most essential for sememic description. Main Elements or Marker Elements, or even both, may not appear in the process of Sememic Description, but Core Elements have to be present.

Secondly, Core Elements can provide a reference to the improvement of the dictionary definitions. For example:

顶天：达到/F极限/S1或/T1极高/S2的/T2程度/S3

DingTian: dadao/F jixian/S1 huo/T1 jigao/S2 de/T2 chengdu/S3

DingTian: The degree of reaching its limit or the highest degree

As shown above, the Core Element is dadao (reach), a verb semem. But it also seems an adjective sememe with chengdu (degree). In fact, this term is a verb, thus it would be better to remove the Basic Elements de/T2 (of) and chengdu/S3 (degree)

Thirdly, definition of Core Elements can compensate for certain deficiencies of the Method of Sememic Feature Analysis. Check out the following [3]:

(BROTHER-IN-LAW: MALE(X) & (SPOUSE-OF-SIBLING-OF (X, Y) V

SIBLING-OF-SPOUSE-OF (X, Y)) MALE(X) & (SPOUSE-OF-SIBLING-OF (X, Y) V SIBLING-OF-SPOUSE-OF (X, Y))

How can the ambiguities of the “BROTHER-IN-LAW” be removed? Examine its Core Elements and its subtle meanings are clear:

Core Elements: SPOUSE → elder sister or younger sister’s husband

Core Elements: SIBLING → husband’s elder brother or younger brother

Core Elements of verbs and nouns are usually better to determine, but the adjectives are more complex.

2.1 Definition of Core Elements of Nouns

Core Elements of Nouns are usually the core nouns from the Expression of sememic description, and are mostly “specific” in terms of the “specific + differential” model.

If there exist multiple parallel structures of "specific plus differential", there will be more than one core noun, and then there are a number of Basic Elements. For example:

陡坡：和/T1水平面/S1所/T2成/S2角度/S3大/S4的/T 3 地面/F1 ；坡度/S5大/S6的/T 4 坡/F2

DouPo: he/T1 shuipingmian/S1 suo/T2 cheng/S2 jiaodu/S3 da/S4 de/T3 dimian /F1 ; podu/S5 da/S6 de/T po/F2

DouPo: a Ground with a large Horizontal Angle, or a large Slope

Core Elements of Nouns are not always in the position of "head" according to the structure of "attributive + head", and they could be in the object position. For example:

初度：原指/T1初生/S1的/T2时候/F1，后称/T3生日/F2为/T4初度/S2

ChuDu: yuanzhi/T1 chusheng/S1 de/T2 shihou/F1 , houcheng/T3 shengri/F2 wei/T4 chudu/S2

ChuDu: Initially, it means just the Time of birth, and then refers to one's Birthday

They could also appear in the subject position. For example:

斗筲：斗/S1和/T1筲/S2都/T2是/T3容量/S3不/T4大/S4的/T5容器/F1，比喻/T6气量/F2狭小/S5或/T7才识/F3短浅/S6

DouXie:dou/S1 he/T1 xie/S2 dou/T2 shi/T3 rongliang/S3 bu/T4 da/S4 de/T5 rongqi/F1, Biyu/T6 qiliang/F2 xiaxiao/S5 huo/T7 caishi/F3 duanqian/S6

DioXie: Dou (斗) and Xie (筲) are both small Containers, and they are also likened to Narrow-mindedness or Short-sightedness

The Part-of-Speech of Core Elements of a Noun and that of the word interpreted are usually the same, but there are also unequal examples:

编制3：组织/S1机构/S2的/T1 设置/F1 及/T2其/S3人员/S4数量/S5的/T3定额/F2和/T4职务/S6的/T5 分配/F3

BianZhi3: zuzhi/S1 jigou/S2 de/T1 shezhi/F1 ji/T2 qi/S3 renyuan/S4 shuliang/S5 de/T3 ding'e/F2 he/T4 zhiwu/S6 de/T5 fenpei/F3

BianZhi3: the organization Settings, and Fixing the number of the organization, also including Assignment of duties

The above noun elements – Settings, Fixing and Assignment – have their Chinese "equivalents" as 3 verbs (the underlined 设置shezhi, 定额ding'e and 分配fenpei).

2.2 Definition of Core Elements of Verbs

Core Elements of Verbs are usually the core verbs from the Expression of sememic description. There are a number of minimum-predicate structures (MPS), and generally a number of Basic Elements of Core Elements for Verbs. For example:

放债：借钱/F1 给/S1人/S2，收取/F2利息/S3

MPS 1 MPS 2

FangZhai: jieqian/F1 gei/S1 ren/S2, shouqu/F2 lixi/S3

FangZhai: Lend money to others, and charge interests

As to the Expressions of sememic description with the verbs containing predicated object, their predicated objects are generally seen as the Core Elements. For example:

被覆3：军事/S1上/T1指/T2用/T3竹/S2、木/S3、砖/S4、石/S5等/T4建筑/S6材料/S7对/T5建筑物/S8的/T6内壁/S9和/T7外表/S10进行/S11加固/F

BeiFu3: junshi/S1 shang/T1 zhi/T2 yong/T3 zhu/S2, mu/S3, zhuan/S4, shi/S5 deng/T4 jianzhu/S6 cailiao/S7 dui/T5 jianzhuwu/S8 de/T6 neibi/S9 he/T7 waibiao/S10 jinxing/S11 jiagu/F

BeiFu3: In the military field, it refers to reinforcing the wall and surface of facilities using some building materials such as bamboo, wood, brick and stone, etc

The Part-of-Speech of Core Elements of Verbs and that of the interpreted word are usually the same, but there are also unequal cases. For example:

分心1：分散/F1注意力/S；不/T 专心/F2

FenXin1: fensan/F1 zhuyili/S； bu/T zhuanxin/F2

FenXin1: Distracting one's attention, not Concentrating, inattentive

The above adjective elements – Distracting and Concentrating – have their Chinese "equivalents" as verbs (the underlined 分散fensan, 专心zhuanxin).

2.3 Definition of Core Elements of Adjectives

Core Elements for Verbs are usually the core adjectives from the Expression of sememic description. For example:

平等2：泛指/T地位/S 相等/F

PingDeng2：fanzhi/T diwei/S xiangdeng/F

PingDeng2：Refers to that the status is equal

The syntactic position of the core adjectives are more flexible. They can occur in the complement position. For example:

和谐：配合/S得/T 适当/F

HeXie: peihe/S de/T shidang/F

HeXie: cooperating in a proper way; harmoniously

The Part-of-Speech between Core Elements about adjectives and that of the interpreted word are usually the same, but there are also irregular cases. For example:

- Containing the words "形容(xingrong)", "样子(yangzi)" and "的(de)"etc.

  滴溜儿2：形容/T1很/T2快/S地/T3 旋转/F1 或/T4 流动/F2
  DiLiu'er2: xingrong/T1 hen/T2 kuai/S de/T3 xuanzhuan/F1 huo/T4 liudong/F2
  DiLiu'er2: to modify the speed of rotating or flowing quickly

- Satisfying the formula of "W is an F".

  潮红：两颊/S1泛起/S2的/T 红色/F
  ChaoHong: liangjia/S1 fanqi/S2 de/T hongse/F
  ChaoHong: Red flush appearing on the cheek

- Core Elements correspond directly to the non-adjective morpheme of the interpreted word.

应税：根据/T1税法/S1规定/S2 应当/F 交纳/S3税款/S4的/T2

YingShui: Genju/T1 shuifa/S1 guiding/S2 yingdang/F jiaona/S3 shuikuan/S4 de/T2

YingShui: of taxes that should be paid according to speculations of tax laws

- Core Elements have the properties of adjectives.

瘠薄：（土地） 缺少/F1 植物/S1生长/S2所/T1需/S3的/T2养分/S4、水分/S5；不/T3肥沃/F2

JiBo: (tudi) queshao/F1 zhiwu/S1 shengzhang/S2 suo/T1 xu/S3 de/T2 yangfen/S4, shuifen/S5; bu/T3 feiwo/F2

JiBo: (of soil) lacking of moisture and nutrients required for plant growth; not fertile

- The Expression of sememic description has the same grammatical function as adjectives.

促狭1：爱/S1捉弄/F人/S2

CuXia1：ai/S1 zhuonong/F ren/S2

CuXia1: Like to play tricks on other people

3 Grammatical Distributions of Core Elements

3.1 Syntax-Level Distributions of Core Elements

Some appear in the first level of grammatical structure. For example:

订立：双方/S1或/T1几方/S2把/T2商定/S3的/T3事项/S4用/T4书面/S5形式/S6（如条约、合同等）肯定/F下来/T5

DingLi: shuangfang/S1 huoT1 jifang/S2 baT2 shangding/S3 de/T3 shixiang/S4 yong/T4 shumian/S5 xingshi/S6 （ru tiaoyue、hetong deng）kending/F xialai/T5

DingLi: (parties concerned) affirm in writing the matters agreed

The first level: 双方或几方 把商定的事项用书面形式（如条约、合同等）肯定下来

The second level: 双方或几方 把商定的事项用书面形式（如条约、合同等） 肯定 下来

Others are present in the second level (or even lower level) of grammatical structure. For example:

平等1：指/T1人们/S1在/T2社会/S2、政治/S3、经济/S4、法律/S5等/T3方面/S6享有/S7相等/F待遇/S8

PingDeng1：zhi/T1 renmen/S1 zai/T2 shehui/S2、 zhengzhi/S3、 jingji/S4、falv/S5 deng/T3 fangmian/S6 xiangyou/S7 xiangdeng/F daiyu/S8

PingDeng1: Refers to the status or right that people can enjoy equal treatment in social, political, economic, legal and / or other fields

The first level: 人们 在社会、政治、经济、法律等方面享有相等待遇

The second level: 人们 在社会、政治、经济、法律等方面 享有 相等 待遇

3.2 Syntactic Position Distribution of Core Elements

Distribution of the Syntactic Position for Core Elements shows a wide variety of range scopes. There are six main syntactic positions as follows (the single underling shows a grammatical structure, and the double underlining shows the syntactic position of Core Elements):

Core Elements appear in the Head position of Attributive-Head structures. For example:

阿嚏：形容/T1打/S1喷嚏/S2的/T2声音/F

ATi: xingrong/T1 da/S1 penti/S2 de/T2 shengyin/F

ATi: Refers to the voice of sneeze

Core Elements appear in the Head position of Adverbial-Head structures. For example:

当红：（演员、文艺作品等）正/T走红/F

DangHong: (yanyuan, wenyizuopin, etc.) zheng/T zouhong/F

DangHong: (of performers, artistic works, etc.) Being popular

Core Elements appear in the Parallel positions of Coordinate structures. For example:

被动2：（事情）由于/T1遇到/S1阻力/S2或/T2干扰/S3，不能/T3按照/F1自己/S4的/T4意图/S5进行/F2

BeiDong2：（shiqing）youyu/T1 yudao/S1 zuli/S2 huo/T2 ganrao/S3, buneng/T3 anzhao/F1 ziji/S4 de/T4 yitu/S5 jinxing/F2

BeiDong2: an awkward situation of resistance or interference where things can not happen as expected

Core Elements appear in the Verb position of Verb-Object structures. For example:

报料1：向/T媒体/S1提供/F新闻/S2线索/S3

BaoLiao1: xiang/T meiti/S1 tigong/F xinwen/S2 xiansuo/S3

BaoLiao1: Providing news clues to the media

Core Elements appear in the Complement position of Verb-Complement structures. For example:

和谐：配合/S得/T适当/F

HeXie: peihe/S de/T shidang/F

HeXie: cooperating in a proper manner; harmoniously

Core Elements appear in the Adverbial position of Adverbial-Head structures. For example:

道地1：真正/F是/T1 有名/S1产地/S2出产/S3的/T 2

DaoDi1: zhenzheng/F shi/T1 youming/S1 chandi/S2 chuchan/S3 de/T2

DaoDi1: Really produced at the well-known place

4 Semantic Divergence of Core Elements

The following shows the distribution statistics for the 1500-word corpus.

Table 1. Distribution of Basic Elements of Core Elements

Items	Nouns	Verbs	Adjectives	Total
Single Basic Elements of Core Elements	384	294	209	887/59.13%
Multi-Elements of Core Elements	116	206	291	613/40. 87%

It can be seen from the above table that the proportion of Multi-Elements of Core Elements is 40. 87%, but the semantic convergence of Basic Elements of Core Elements is different. The different degrees of divergence can be grouped into 4 hierarchical groups (Basic Elements of Core Elements labeled with a smaller-size F, and Core Elements with a larger F):

4.1 Integrated and Interdependent

The Marker Elements capable of achieving this effect are usually "以（yi, with）" and "并(bing, and)", with the assistance of some other unmarked forms. For example:

摆拍：特意/S1布置/F1场景/S2，让/T1人物/S3摆出/F2一定/S4的/T2姿势/S5进行/T3拍摄/F3

BaiPai: teyi/S1 buzhi/F1 changjing/S2， rang/T1 renwu/S3 baichu/F2 yiding/S4 de/T2 zishi/S5 jinxing/T3 paishe/F3

BaiPai: arrange scenes and let someone pose for pictures

"布置(buzhi, arrange)", "摆出(baichu, pose)"、"拍摄(paishe, photograph) are interdependent and coherent.

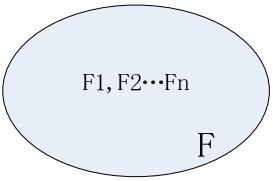


Fig. 1. The first gradient for the semantic divergence of Basic Elements of Core Elements

4.2 Separated and Correlated

The Marker Elements capable of achieving this effect are "或(huo, or)", "和(he, and)", "以及(yiji, plus)", etc., with the assistance of some other unmarked forms. For example:

部头：书/S的/T1厚薄/F1和/T2大小/F2（主要指篇幅多的书）

BuTou: shu/S de/T1 houbo/F1 he/T2 daxiao/F2（zhuyao zhi pianfu duo de shu）

BuTou: Thickness and size of books (usually considerable size)

“厚薄(houbo, thickness)” and “大小(daxiao, size)” are just part of the Core Elements for 部头 (BuTou).

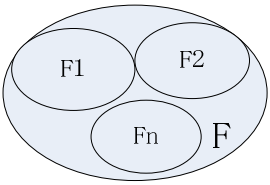


Fig. 2. The second gradient for the semantic divergence of Basic Elements of Core Elements

4.3 Cross-Referential and Complementary

The Marker Elements capable of achieving this effect mainly include “或 (huo, or)” etc. and some other unmarked forms. For example:

高看：看重/F1；重视/F2

GaoKan: kanzhong/F1; zhongshi/F2

GaoKan: Value or attach importance to

There is subtle semantic difference between “看重 (kanzhong)”, “重视 (zhongshi)” and “高看 (gaokan)”.

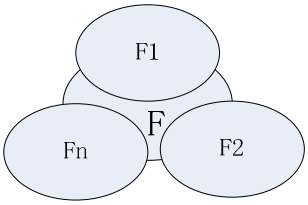


Fig. 3. The third gradient for the semantic divergence of Basic Elements of Core Elements

4.4 Independent and Correlated

The Marker Elements The Marker Elements capable of achieving this effect are “原指(yuanzhi, initially mean)”, “又指(youzhi, also refer to)”, “后多指(hou duozhi, later mainly mean)”, “后称(houcheng, later called)”, and “比喻(biyu, be likened to)” , etc. For example:

初度：原指/T1初生/S1的/T2时候/F1， 后称/T3生日/F2为/T4初度/S2

ChuDu: yuanzhi/T1 chusheng/S1 de/T2 shihou/F1， houcheng/T3 shengri/F2 weiT4 chudu/S2

ChuDu: Initially, it means just the time of birth, and then refers to one’s birthday

“时候 (shihou, time)” and “生日 (shengri, birthday)” are in closely related, but they also maintain their own independence.

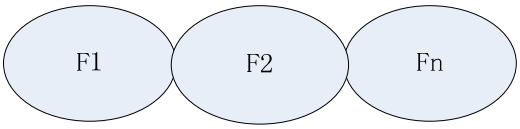


Fig. 4. The fourth gradient for the semantic divergence of Basic Elements of Core Elements

5 Conclusions

The 1500-word corpus designed in this study has facilitated the analysis of such word classes as nouns, verbs and adjectives from the perspective of lexicographic definition. Core elements and grammatical distributions of the three types of words are established and verified with examples. The semantic divergence between core elements are also grouped and exemplified.

This study is a preliminary step to further and better approaches of lexicographic definition, which in turn urges and pushes the development of methods and theories in Natural Language Processing, Semantic Web and other domains that could be so closely related.

References

1. Wang, X.L., Zhang, Z.Y.: The Method of Cutting the Glosseme's Elements into Three Components and Its Guidance of Value in the Dictionaries. Applied Linguistics 4, 92–99 (2007)
2. Wang, N.: The Exploring and Explaining of Chinese Etymological Meaning. Social Sciences in China 2, 167–178 (1995)
3. Fu, H.Q.: Analysis and Description of Meanings. Language Publishing House, Beijing (1996)

New Exploration into the Word Semantic Generation Mechanism Based on Word Representation

Shengjian Ni[1], Donghong Ji[2], Yibing Wang[3], and Fei Li[2]

[1] College of Chinese Language and Literature,
Wuhan University, Wuhan, China
hijackon@163.com
[2] Computer School, Wuhan University, Wuhan, China
Donghong_ji2000@yahoo.com.cn, kevin.lifei@gmail.com
[3] Third Faculty, Second Artillery Command College
cjt422@163.com

Abstract. As a hidden clue in the history of lexical semantics, word semantic generation mechanism (WSGM) based on word representation (WR) has not been studied independently in any paper. So far, the Generative Lexicon (GL) offered the most elaborate description of WSGM based on WR, yet it still has inadequacy. In this study, image content, especially image schemata (IS), is added to the WR of GL, which may be the most elaborate WR nowadays. It is shown that with the help of IS inference, there is possibility to perfect metonymic WSGMs and add metaphoric WSGMs to the WR of GL. It is pointed out that to consummate metonymic and metaphoric WSGMs, the senses of a target polysemous word should be induced beforehand with completeness and discreteness.

Keywords: lexical semantics (LS), word representation (WR), word semantic generation mechanism (WSGM), metonymy, metaphor.

1 Introduction

Customarily, WR refers to the representation of grammatical and semantic contents of a word with multiple levels, as the WR of GL (see Fig.1. next page). WR is usually dealt with when semantic structures or representations are studied[1]. Good WR can embody the WSGM (henceforth referred to as SGM) of language understanding and generation process. SGM can be understood as choosing the most suitable sense of a polysemous word in a specific context. Though there has been no paper specializing in SGM, a survey of the history of lexical semantics shows that SGM is an important hidden clue in it[2], used to express the creativity and productivity of a word (polysemy). After introducing the history of SGM, this paper points out the inadequacy of existing WR and introduces content of IS into the WR of GL to ameliorate its metonymic SGMs and embody metaphoric mechanism of SGM.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 612–620, 2013.

2 SGMs in Lexical Semantics

Cognitive viewpoint of historical-philological semantics and Paul's distinction between 'usual' and 'occasional' meanings has actually touched on SGM. The most direct and obvious resource of SGM is the projection rules of Katzian semantics [2]. To express the creativity of language use, Jackendoff's conceptual semantics considered the combination of linguistic and extra-linguistic knowledge[3] , using 'conceptual structure' as the interface between the two kinds of knowledge and trying to explain the flexibility of semantics through the interaction between the two [4]. However, in actual practice, Jackendoff has devoted more attention to the interface between syntax and semantics than to the flexible use of words or to the detailed description of the interplay between conceptual structure and extra-linguistic knowledge[2]. Furthermore, Jackendoff's distinction between linguistic and extra-linguistic knowledge is a kind of static division of long memory [2], revealing no dynamic interaction between the two in specific contexts. Two level semantics (TLS) [5] tried to manage the dynamic interaction, observing that senses of polysemous word can be differentiated by dividing knowledge into two levels: semantic form and conceptual structure. The former refers to formalized linguistic knowledge and the latter extra-linguistic. TLS attaches importance to the dynamic relation between context and words, paying attention to SGM and its main contribution lies in the stratification of factors concerning the understanding of word senses, which made the expression and explanation of meaning creativity possible.

Referring to the studies mentioned above, GL offered the most elaborate description of SGM based on WR[6-7]. Figure 1 is an example of WR of GL.

$$
(22)\quad \left[\begin{array}{l}
\textbf{drive} \\
\text{EVENTSTR} = \left[\begin{array}{l} E_1 = \mathbf{e_1{:}process} \\ E_2 = \mathbf{e_2{:}process} \\ \text{RESTR} = <_{\alpha} \end{array}\right] \\
\text{ARGSTR} = \left[\begin{array}{l} \text{ARG1} = \mathbf{x{:}human} \\ \text{ARG2} = \mathbf{y{:}vehicle} \end{array}\right] \\
\text{QUALIA} = \left[\begin{array}{l} \text{FORMAL} = \mathbf{move(e_2,y)} \\ \text{AGENTIVE} = \mathbf{drive\_act(e_1,x,y)} \end{array}\right]
\end{array}\right]
$$

Fig. 1. WR of the word 'drive' from GL [6][1]

Based on this kind of WR, GL offered many SGMs, with its creativity coming mainly from the combination of predicates and arguments and there are three kinds of SGMs[7]. The first kind is type matching: the type satisfying the function[2] is chosen

1 EVENSTR (event structure) concerns the classification of verbs and is familiar to Levin's classification of verbs (2007) but different from 'Events' from cognitive linguistic which are familiar to 'scripts' and 'frames', etc.

2 'Functions', including verbs and adjectives, are similar to but different from 'predicates' in logic.

from the coded argument directly. For example, the verb 'flow' requires that its first argument be liquid and if 'beer' owns the characteristics of a liquid, then 'the beer flows' is a meaningful structure. The second kind is type accommodation,[3] which is realized through lexical inheritance. For example, if 'car' can be the direct object of the verb 'drive', then 'no linguist drives a SUV' is acceptable, because 'SUV', as a hyponym of 'car', inherits the features of 'car'.

The last kind is type coercion. If the above-mentioned SGMs cannot explain the requirements of a function in an acceptable structure, there exists type coercion, which demands that the requirements of the function be imposed on an argument, forcing it to change its type; otherwise, there will be semantic abnormality [7]. Type coercion includes 'Exploitation' and 'Introduction'. 'Exploitation' extracts the prominent part of an argument to satisfy the function in certain context, which is useful for differentiating two independent or even contradictory senses combined in one word form. For example, 'book' has both the meaning of 'physical object' (physobj) and 'information'(info) and its meaning can be expressed as 'physobj•info'; thus, 'book' is a complex (or dot) object and has the following lexical conceptual paradigm (lcp): {physobj•info, physobj, info}, from which 'Exploitation' choose the suitable meaning for a predicate in a specific context. In the sentence 'John tear the book into pieces', 'tear' requires that the patient has the feature of 'physobj', and 'Exploitation' choose 'physobj' from the lcp to satisfy the requirement and there is no need to touch on 'physobj•info' or 'info', which embodies part-whole metonymy. 'Introduction' is a way of expanding an lcp to meet the requirements of a predicate and involves both analogy and metonymy. For example, 'read' requires its direct object to be a dot object and usually 'rumor' has only the meaning of 'info'; however, in 'We all read the rumor about the cook and the headmaster', because of the requirement of 'read', 'physobj' is introduced into 'rumor' which is expanded to be 'physobj·info', referring to 'book'.

SGMs of GL can explain more polysemous phenomena than before, but they mainly derive from the combination of predicates and arguments and lexical inheritance embodied in qualia structure; and they have many disadvantages, including: 1) extra-linguistic knowledge is deleted from conceptual semantics and TLS, thus GL can only explain a small part of metonymic SGMs and cannot express metaphoric SGMs; 2) in many occasions, GL's SGMs cannot really distinguish word senses. 'Exploitation' does not necessarily have the ability to differentiate senses, if we deem that 'physobj' and 'info' of the noun 'book' (physobj·info) are in fact two aspects of one sense but not two separate senses.

3 WR Expanded

According to cognitive science (CS), extra-linguistic knowledge is image system (image content) realized as schemata and word relations in language. Referring to CS, cognitive linguistics (CL), and discussion above, SGM achievements from LS, the author offers an expanded WR based on GL's WR in Fig.2. next page.

[3] Type accommodation embodies the metonymy of superordinate- hyponym interchange.

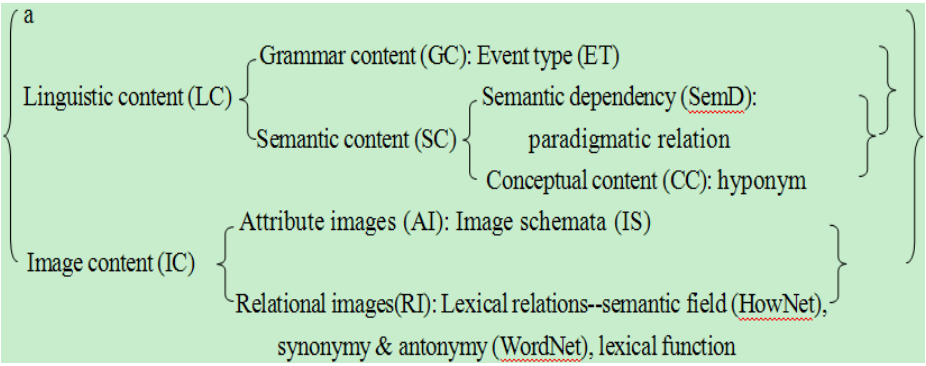


Fig. 2. A new WR

In Fig.2, ET adopts Levin's classification[8], which is more complete compared with the ET of GL and owns discreteness at the same time. SemD, coming from Mel'čuk & Polguère,etc.[9], substitutes argument structures of GL, with richer information. CC, combined with linguistic ontology, expresses lexical inheritance, with wider application scopes than qualia structure.[4] RI includes all kinds of lexical relations, which can be expressed by semantic field, etc. None of these contents are important in this paper and will be given no specification.

This new WR is different from that WR of GL mainly because IS is added to it, which makes it possible to supplement metonymic SGMs of GL and to provide metaphoric SGMs. Langacker[10] observed that CL tend to have a semantic understanding mechanism based on IS. One of the six principles of cognitive semantics is that cognitive model is based on schemata rather than on propositions[11] . Talmy (2000) summarized four principles of meaning construction of human language, the first of which is: meaning construction is based on IS....[12]. And the main meaning changes and their mechanisms-- metonymy and metaphor--are based IS, which is the key to the core task of LS--explaining polysemy [2][13]. Thus, IS is key to the understanding of polysemous words.

4 Illustration of Metonymic and Metaphoric SGMs Based on IS

The author once induced the senses of six Chinese characters, referring to corpus, linguistic ontology, and valency grammar, event types[8], qualia structure[6]. Here the author uses the senses of one of the six characters-- '包' (bao) to explain the possibility of supplementing metonymic SGMs and offering metaphoric SGMs based on the expanded WR. According to the author's study, '包' has the following

[4] The only means to realize inheritance in GL is type accommodation through the FORMAL of qualia structure, which is valid specially and mainly for artifacts. However, CC of this study can express all kinds of semantic inheritance.

21senses（for each sense, first comes the kind of taxonomy from HowNet[5], then the first primitive, and finally paraphrase as appears in a dictionary）：

ActSpecific 1 (ActS 1). [Include]. Wrap something up with paper, cloth or other thin and soft slices.

ActSpecific 2. [Include]. Surround.

Relation 1. [Concrete relation (CR)] (Different from attributive relation). [Situated]. Wrapped, contained.

Relation 2. [Abstract relation (AR)]. [Contain]. Contain.

Entity 1. [Artifact]. A flexible container with a single opening.

Entity 2. [Artifact]. A red paper bag with money in it which is used as a present or gift, etc.

Entity 3. [Physical]. Things wrapped.

Entity 4. [Abstract]. Contents of a contract.

Entity 5. [Physical]. Lump(s) on something, especially on a body.

Entity 6. [Food]. Food with the form of an oval.

Entity 7. [House]. Yurts made of felt.

Entity 8. [Information] (computer science). Virtual containers for storing information.

Number 1. [Motion number (MN)[6]]. (Refers to) an event of contracting.

Number 2. [Entity number (EN)]. Used as a modifier before words signifying things wrapped.

Actgeneral 1. [Bear]. Take on the whole task and see to finish it.

Actgeneral 2. [Bear]. Contract.

Actgeneral 3. [Do]. Wholesale.

Actgeneral 4. [Do]. Obtain or lend the full access of something or a place during certain period of time with/for money.

Modality. [Possibility]. Ensure.

Range. [Scope]. All.

Attribute value (AV). [Surname]. Bao.

Referring to the study of metaphor and metonymy in CL, the four criteria for differentiating metaphor and metonymy, and classification of metonymy[14], the taxonomy of HowNet[7], the 21 senses of 包(bao) are analyzed in detail and it is found that of the former 20 senses, ActSpecific1 (ActS1) is the core and most basic[8], the others deriving from it directly or indirectly through metaphor or/and metonymy. The relations of the former 20 senses of 包(bao) are shown in Fig.3., where A stands for

[5] HowNet. http://www.keenage.com/

[6] 13), 14), and 19), 20) cannot be found in HowNet.

[7] During the senses induction of 6 Chinese Characters, it is found that senses of '包' (bao)belongs to different conceptual branches of the taxonomy of HowNet, which is helpful in explaining the metaphoric and metonymic mechanism among different senses of '包'.

[8] Prototype theory can be used to explain both the structure (denotation) formed by a sense and the structure formed by senses of a polysemous word[15]and the basicness of a category or a senses can be determined by referring to four aspects[16].

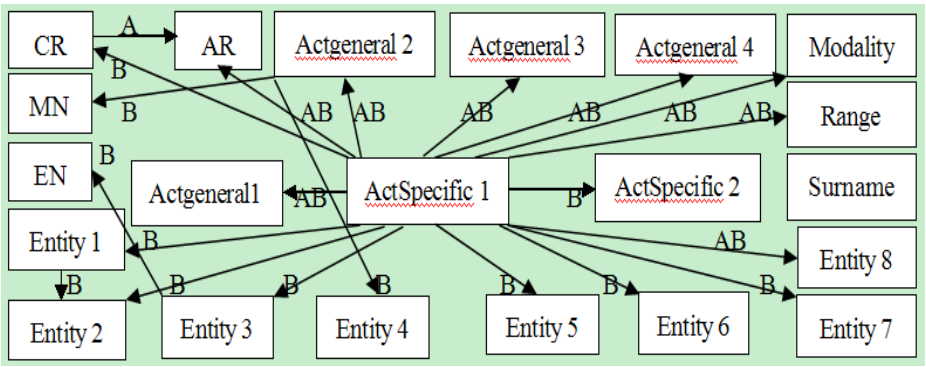


Fig. 3. Relations between senses of 包 (bao)

metaphor, B metonymy, and AB means that the meaning change involves both metaphor and metonymy; the arrowhead stands for the target sense while the other side for the source sense on which the target sense changed. ActS1 is the most basic sense and can activate the most complete schema[9], which includes 'wrap up something all around with hand(s) and paper, cloth or other thin and soft slices'. Based on this schema, many kinds of inference can be carried out[15] to realize all kinds of metaphoric and metonymic SGMs to obtain other related senses.

To obtain ActSpecific 2 from ActS1, the following steps may be needed: utilizing metonymy's trait of being based on contiguity, 'paper, cloth or other slices' is first replaced by their hypernym, which is then replaced by its hyponym with the similar characters as 'paper, cloth or other slices'. This process of inference may take place in the understanding of this sentence '骑兵朝敌人包抄过来' (The cavalrymen outflanked toward the enemy). Compared with ActS1 (a kind of CausetoMove according to the taxonomy of HowNet), the distance between the entities in ActSpecific2 (a kind of SelfMove in HowNet taxonomy) is bigger and there is clearer discreteness between them.[10]

CR is the resultative state of ActS1. To derive CR from ActS1, the metonymy of 'event' substituting 'result', which belongs to 'contiguity of behavior, event and process', is the main motivation[14]. AR is the result of CR projecting from a concrete field to a abstract one through metaphor. From ActS1 to Entity1, part-whole metonymy is the motivation which chooses the form of the resultative object of the event as the name of Entity1. Entity2 is the specification of Entity1 through hypernymy-hyponymy metonymy. Entity3 signifies the thing wrapped in the resultative object and there may be two metonymic transitions from ActS1 to Entity3: the event standing for the resultative object and the container (the wrapping) for the content (the wrapped). Entity4 can be considered as coming from ActGeneral2 through 'patient-for-event' metonymy. Entity5 highlights the form of the resultative object of ActS1 and the mechanism is also metonymy. Entity6 gets its name of '包

[9] According to the study of CS and CL, schemata are stored in the brain in structured, formalized way. Yet, they are expressed in linear way in this paper for convenience.

[10] For more information, see Peirsman & Geeraerts[14].

(bao)' because of the similarity between their shapes and the shape of the resultative object of ActS1 and the transitional mechanisms are metonymy and metaphor (based on similarity). Entity7 gets its name '包' for at least two reasons: on one hand, the top of them is similar to the resultative object of ActS1; on the other, they are closed all around-one characteristic of the resultative object of ActS1. Entity7 can be seen as deriving from Entity1 or ActS1 with narrower extension than Entity1, which also embodies metonymy. Entity8 signifies abstract or virtual things which can be seen as derived from Entity1 through metaphor or from ActS1 through both metonymy and metaphor. Treating dynamic events as static entities, we get MN, which is very common in Chinese. EN is obtained through 'container-for-content' metonymy. The mechanisms of obtaining MN and EN for '包(bao)' have not been discussed by scholars, for example, Peirsman and Geeraerts[14]and can be discussed in more detail if the reader is interested.

From ActS1 to ActGeneral 1, ActGeneral 2 and ActGeneral 3, metaphor based on part-whole metonymy plays the key role. For example, to get ActGeneral 1, part-whole metonymy is first activated to choose the meaning of 'all' from ActS1; then, recurring to context and activating the function of metaphor, '包 (bao)' can be used and understood as what ActGeneral 1 signifies. Range obtains its meaning through similar process as ActGeneral 1, which is used and expanded further to produce the modal meaning 'ensure'. When '包(bao)' expresses a surname (包2) , it has no relation with other 20 senses (包1). '包2' and '包 1' are homonyms.

So, there is possibility to unify the former 20 senses of '包' in one WR using ActS1 as the core , providing suitable contents of certain schema(ta) is/are offered. The WR of '包1' can be represented as in Fig. 4.

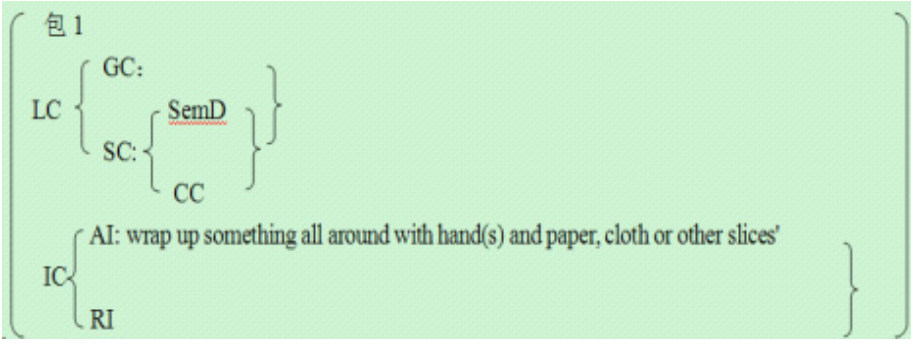


Fig. 4. WR of '包 1'

To highlight AI, other contents of WR of '包1' are not represented in Fig.4.. In principle, the contents of AI can not be completely expressed, as AI are usually gestalts. According to the explanation of the relations among the 20 senses of 包1 above, while representing contents of AI, one should consider at least: 1) characteristics of metonymy and metaphor; 2) whether the senses derived from the SGMs based on the schema have covered all the senses signified by the word form.

While explaining the sense relations for Fig. 3. above, we have in fact offered an illumination of SGMs for Fig. 4., from which we see that adding schematic content to WR provides the feasibility to materialize metonymic and metaphoric SGMs and thus we have improved the SGM of GL. Additionally, word relations expressed sufficiently, there is possibility to use statistic methods to decide on the SGM to be carried out and in turn the sense of a polysemous word form can be chosen in concrete contexts.

5 Conclusion

This paper unearths the SGM clue hidden in the history of LS, points out the weaknesses of exiting studies on SGM, and presents a new model of WR based on GL, making it possible to express metonymic and metaphoric SGMs based on WR if there is complete and discrete sense induction of the target polysemous word. During the study, some special metonymic SGMs are found in Chinese, which can be further studied in the future, in comparison with those in English. Existing studies usually represent the related senses of a word form separately in different WRs. However, this study shows that related senses of a word form can be represented in one WR if IC is introduced to WR. There are some disadvantages about this study, including: 1) lack of formalization of metonymic and metaphoric SGMs; 2) because of data deficiency, and the author's inability to grasp and assimilate linguistic knowledge, many existing studies are excluded from this studies.

Acknowledgment. The work is funded by the following projects: the Major Projects of Chinese National Social Science Foundation No.11&ZD189, National Natural Science Foundation of China No.61202193,61133012,61173062,61070082,and 61070243, and Youth Science and Technology Morning program of Wuhan No. 201150431105.

References

1. Vigliocco, G., Vinson, D.P.: Semantic Representation (2005), http://homepage.psy.utexas.edu/HomePage/Faculty/Griffin/independent/Vigliocco_Vinson2007_Semantics_chapter.pdf (retrieved in January 12, 2012)
2. Geeraerts, D.: Hundred years of lexical semantics. In: Vilela, M., Silva, F. (eds.) Actas do 1ºEncontro Internacional de Linguística Cognitiva, pp. 123–154. Faculdade de Letras, Porto (1999)
3. Jackendoff, R.: Conceptual semantics and Cognitive Linguistics. Cognitive Linguistics 7, 93–129 (1996)
4. Helbig, H., Informatik, F.: Knowledge Representation and the Semantics of Natural Language. Springer, Heidelberg (2006)
5. Bierwisch, M., Lang, E. (eds.): Dimensional Adjectives: Grammatical Structure and Conceptual Interpretation. Springer, Berlin (1989)

6. Pustejovsky, J.: The Generative Lexicon, pp. 60–142. Massachusetts Institute of Technology, Massachusetts (1996)
7. Pustejovsky, J.: Type Theory and Lexical Decomposition. Journal of Cognitive Science 6, 39–76 (2006)
8. Levin, B.: The Lexical Semantics of Verbs II: Aspectual Approaches to Lexical Semantic Representation, Course LSA.113P Stanford University (2007),
9. http://www.stanford.edu/~bclevin/lsa07asp.pdf (retrieved in April 14, 2012)
10. Mel'čuk, I., Polguère, A.: Dependency in Linguistic Description. John Benjiamins Company (2009)
11. Langacker, R.W.: Cognitive Grammar: A Basic Introduction, pp. 31–36 (2008)
12. Yin, W.: Probe into Cognitive Linguistics. Chongqing Publishing house, Chongqing (2005)
13. Zeng, X.: The Six Characteristics of Cognitive Semantic. Foreign Languages Research 5, 19–23 (2008)
14. Paradis, C.: Lexical Semantics. In: Chapelle, C.A. (ed.) The Encyclopedia of Applied Linguistics. Wiley-Blackwell, Oxford (to appear, 2013)
15. Hui, Z., Lu, W.: Cognitive Metonymy. Shanghai Shanghai Foreign Language Education Press, Shanghai (2010)
16. Geeraerts, D.: Theories of Lexical Semantics. Oxford University Press, New York (2010)
17. Ungerer, F., Schmid, H.J.: An Introduction to Cognitive Linguistics. Foreign Language Teaching and Research Press, Beijing (2001)

A Study on the Modal Particle "ne" and "ne" Interrogative Sentence from Information-Parsing Perspective

Tingting Guo[1] and Huili Zheng[2]

[1] College of Chinese Language and Literature, Language and Information Center, Wuhan University, Wuhan, China
guogaott@yahoo.com.cn
[2] Department of Modern & Classical Language, Saint Vincent College, Latrobe, PA, U.S.
huili.zheng@email.stvincent.edu

Abstract. This paper discusses the interrogative information of the modal particle "ne" and the function of "ne" questions. The property of "ne" is demonstrated by investigating the dynamic restrictive relationship between the form and information it expresses. The paper argues that the uses of "ne" are affected by the degree of prominence of the focal interrogative form. A coherent analysis is given to explain the different functions of "ne" questions according to the interrogative information carried by the modal particle "ne."

Keywords: particle, ne question, information-parsing interrogative information, focus.

1 Introduction

1.1 Existing Scholarship on the Modal Particle "ne"

Scholarly debates and discussions on the modal particle "ne" in interrogative sentence have a long history. The focus of the debates is whether or not "ne" expresses interrogative information. In general, there are two diametrically oppositional perspectives regarding this issue. Both perspectives emerged in the 1980s. One perspective, represented by scholars such as Lü Shuxiang [5], Zhu Dexi [13] and Lu Jianming [4], contends that the particle "ne" in interrogative sentence carries the interrogative information of the sentence and therefore should be regarded as interrogative modal particle. The other perspective, represented mainly by Hu Mingyang [2] and Shao Jingmin [8], argues that "ne" does not express interrogative information as evidenced by their respective examination of the grammatical meaning of "ne." The second perspective became the mainstream view in the 1990s. With the publication of Shao Jingmin's influential thesis A Study of Modern Chinese Interrogative Sentences, this perspective became more widely accepted. For instance, Jin Lixin [3], Xing Zhongru [12] and many others all entertained this view. However,

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 621–633, 2013.

despite its mainstream position, there were nonetheless dissenting voices. For example, Kimura Hideki [6] once challenged this perspective. From the twenty-first century onward, some scholars attempted to investigate this issue from a diachronic point of view. For instance, in his article "A Semantic Analysis of the Meaning of 'ne" and its Historical Evolution," Qi Huyang[7] reexamined the semantic meaning of the particle "ne" by surveying the grammar of "ne" as well as its major usages in modern Chinese, and concluded that in modern Chinese the basic function of "ne" is to bring forth interrogative mood.

Based on the existing scholarship on "ne" surveyed above it is clear that scholarly views are still split with regard to whether or not the particle "ne" expresses interrogative information. Therefore, we need to look into somewhere else to seek solution to this issue. The perspective of dynamic information transmission might allow us an alternative approach to reexamine the issue.

1.2 A New Approach to "ne"

According to Xiao Guozheng's theory of information-parsing information structure can restrict the self-sufficiency of semantic form to certain degree [10][11]. However, existing scholarship on "ne" has failed to pay due attention to the restrictions the information conveyed by the modal particle "ne" has on the semantic form of "ne" interrogative sentences. For instance, Shao Jingmin [9] once remarked that, "in non-yes-no questions since wh-questions, alternative questions and A-Not-A questions all include specific interrogative forms which already contain interrogative information, the use of 'ne' is not required. In other words, in these types of questions 'ne' is superfluous. It is merely a formal mark of non-yes-no questions whose presence or absence does not change the nature of interrogative sentence." As a result, the impacts and restrictions of the modal particle "ne" on the semantic form of interrogative sentence have been overlooked.

We therefore attempt to explore the information conveyed by the modal particle "ne" and the impacts of the information on "ne" interrogative sentence from a new approach informed by the restrictive relationship between form and information. It is hoped that this dynamic information-parsing approach can provide some solutions to the debates evolving around the "ne" issue.

2 Interrogative Information of "ne" and "Prominence of Focus"[1]

2.1 Modal Particle "ne" and Its Interrogative Information

Xiao Guozheng's theory of "language information restricting language form" [10][11] enables us to probe the requirement of "ne" in interrogative sentence from the

[1] "Prominence of focus" refers to the degree of the focal position of a specific interrogative form in a concrete context.

perspective of language information, and further to infer what kind of information “ne” contains.[2]

First let us take a look of the following examples:[3]

（1）a ……但是 作为 一个 总裁， 一个 领袖， 他 怎么
dànshì zuòwéi yígè zǒngcái, yígè lǐngxiù, tā zěnme
but as a CEO a leader he how
能够 把 自己 的 市场 预见性， 跟
nénggòu bǎ zìjǐ de shìchǎng yùjiànxìng gēn
be able to preposition oneself particle market prediction with
自己 的 团队 很好 地 融合 起来，
zìjǐ de tuánduì hěnhǎo de rónghé qǐlái
oneself particle team well particle integrate complement
怎么 做？
zěnme zuò?
how do

《对话 —— 走出 困境》
《duìhuà —— zǒuchū kùnjìng》
Dialogue----Out of Predicament

Sample sentence (1) is culled from authentic spoken language material. Although it consists of two wh-questions, the content of the second wh-question “zěnme zuò” (How to do it) is actually the same as that of the first wh-question “zěnme nénggòu……”(how will he be able to……) and the second wh-question merely serves to further complement the first wh-question. Now we might raise a question: What if

[2] Since scholars have had exhaustive discussions on “non-interrogation＋ne” sentence patterns and have agreed upon that “ne” carries interrogative information in this type of questions, this paper thus focuses on the more controversial “interrogation＋ne” sentence patterns and leave out the “non-interrogation＋ne” sentence patterns.

[3] The main language source material of this study is from the text versions of seven television interviews. These materials are spoken dialogue recorded in their original form. The seven texts are:

(1) *CCTV.1 Dialogue: Out of the Crisis*, CCTV Economy Department, “Dialogue” Column Group, Nanhai publishing company, July, 2004.
(2) *Acting: Interviews of Renown Chinese Writers*. Wang Meng, Zhang Jie et al., BaiHuaZhou Literature and Art Publishing House, August 2004.
(3) *Wonderful Truth*. Edited by Cui Yongyuan, China Photography Publishing House, January 2003.
(4) *Interviews of China’s Well-known University Chancellors*. Written by Li Qingchuan, China Federation of Literary and Art Circles press, January 2005.
(5) *Interviews of Masters of* Humanities. Mei Chen, China Federation of Literary and Art Circles press, January 2005.
(6) *Lu Yu: Tell Your Story.* “Lu Yu: Tell Your Story” Program. Liaoning People's Press, May 2004.
(7) *Super Interviews.* “Super Interviews” Program. Huayi press, July 2004.

we get rid of the second wh-question now that the two interrogative sentences express the same meaning and the contents of both sentences are repetitive. Let us take a look of the following example:

（1）b 但是 作为 一个 总裁， 一个 领袖， 他 怎么
dànshì zuòwéi yígè zǒngcái, yígè lǐngxiù, tā zěnme
but as a CEO a leader he how
能够 把 自己 的 市场 预见性， 跟
nénggòu bǎ zìjǐ de shìchǎng yùjiànxìng gēn
be able to preposition oneself particle market prediction with
自己 的 团队 很好 地 融合 起来？ #[4]
zìjǐ de tuánduì hěnhǎo de rónghé qǐlái? #
oneself particle team well particle integrate complement

Example (1) b is rid of the second wh-question “zěnme zuò” yet it is obvious that the acceptability of the sentence is not strong enough as the sentence seems to be incomplete. Why would we have such an impression? The answer might be sought from the perspective of information transformation: as far as this interrogative sentence is concerned, the interrogative pronoun “zěnme” (how) is both the information focus of the sentence and the carrier of interrogative information. Yet it is placed at the beginning rather than the end of the sentence where the focus is highlighted. Furthermore, as the sentence structure is quite long the focal position of “zěnme” is weakened in the flow of speech, and thus the interrogation implication of the sentence is implicit which makes the focus of the sentence unclear and the information structure unbalanced. Consequently the addressee expects the focus of the sentence to appear and this leaves one with the impression of incompleteness.

There are two ways to complete the information structure of the sentence and increase its acceptability. One way is to add a new VP to the end of the sentence as its focus to make the information structure of the sentence reach a new balance. For instance, the sentence can be rewritten as the follows:

（1）c但是 作为 一个 总裁， 一个 领袖， 他 怎么
dànshì zuòwéi yígè zǒngcái, yígè lǐngxiù, tā zěnme
but as a CEO a leader he how
能够 把 自己 的 市场 预见性， 跟
nénggòu bǎ zìjǐ de shìchǎng yùjiànxìng gēn
be able to preposition oneself particle market prediction with
自己 的 团队 很好 地 融合 起来 是
zìjǐ de tuánduì hěnhǎo de rónghé qǐlái shì
oneself particle team well particle integrate complement is
一个 很 关键 的 问题。
yígè hěn guānjiàn de wèntí
a very crucial particle question

[4] In this paper the symbol # indicates poor acceptability of the expressions in question.

What happens is that, by adding a new VP----“shì yígè hěn guānjiàn de wèntí,” (it is a crucial question) the sentence reads complete and its acceptability is increased. The new VP is placed right in the focal syntax position and as a result the interrogative pronoun “zěnme” loses its focal position in the sentence and accordingly is no longer the carrier of interrogative information. Now “zěnme” is devoid of its interrogative function and becomes merely a component of statement structure. However, this way of rewriting has changed the meaning of the original sentence.

The second way is to add carrier of interrogation information so as to reinforce the original focal position of the interrogative pronoun “zěnme.” This way the focus of the sentence is explicit and the sentence retains its original interrogative function. In order to achieve the desired results we can add a wh-question “zěnme zuò” as in the original example (1) a; or alternatively, we can add the modal particle “ne’ to the end of the sentence to reach the same goal, as in the example (1) d:

（1）a ……但是 作为 一个 总裁， 一个 领袖， 他 怎么
dànshì zuòwéi yígè zǒngcái, yígè lǐngxiù, tā zěnme
but as a CEO a leader he how
能够 把 自己 的 市场 预见性， 跟
nénggòu bǎ zìjǐ de shìchǎng yùjiànxìng gēn
be able to preposition oneself particle market prediction with
自己 的 团队 很好 地 融合 起来，
zìjǐ de tuánduì hěnhǎo de rónghé qǐlái
oneself particle team well particle integrate complement
怎么 做?
zěnme zuò?
how do

（1）d 但是 作为 一个 总裁， 一个 领袖， 他 怎么
dànshì zuòwéi yígè zǒngcái, yígè lǐngxiù, tā zěnme
but as a CEO a leader he how
能够 把 自己 的 市场 预见性， 跟
nénggòu bǎ zìjǐ de shìchǎng yùjiànxìng gēn
be able to preposition oneself particle market prediction with
自己 的 团队 很好 地 融合 起来 呢?
zìjǐ de tuánduì hěnhǎo de rónghé qǐlái ne?
oneself particle team well particle integrate complement particle

When comparing the two sentences it is clear that by adding wh-question “zěnme zuò” or the modal particle “ne” the sentence becomes complete. Furthermore, both forms convey the same information, i.e., seen from the viewpoint of information transmission the “ne” in this sentence is equivalent to “zěnme zuò.” In view of the above examples we may argue that the information carried by the modal particle “ne” in interrogative sentence equates with that carried by an interrogative form, and it follows that the information carried by “ne” in interrogative sentence is interrogative

information. Based on the theory of information transmission the inference process can be illustrated by the Fig below:

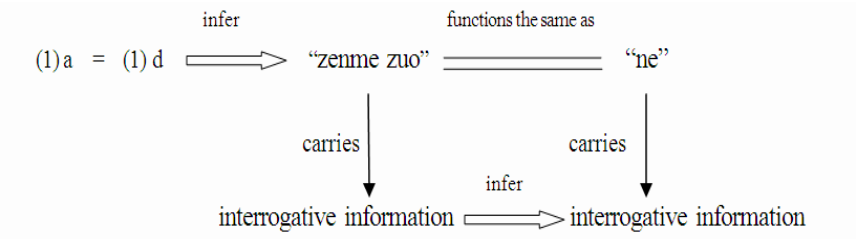


Fig. 1. The illustration about interrogative information of "ne"

There are quite a few similar examples, for instance:

（2）a 您 怎样 处理 这个 问题， 怎样 让 他
nín zěnyàng chǔlǐ zhègè wèntí , zěnyàng ràng tā
you how deal with this problem how let him
不 失于 个性 而又 能够 融合 在
bù shīyú gèxìng éryòu nénggòu rónghé zài
not lose personality while be able to integrate preposition
企业 文化 当中 呢?
qǐyè wénhuà dāngzhōng ne?
enterprise culture in the middle particle
《对话 —— 跨 世纪 的 柳传志》
《duìhuà —— kuà shìjì de Liǔ Chuánzhì》
Dialogue---- Cross-century Liu Chuanzhi

（2）b 您 怎样 处理 这个 问题， 怎样 让 他
nín zěnyàng chǔlǐ zhègè wèntí , zěnyàng ràng tā
you how deal with this problem how let him
不 失于 个性 而又 能够 融合 在
bù shīyú gèxìng éryòu nénggòu rónghé zài
not lose personality while be able to integrate preposition
企业 文化 当中 ？#
qǐyè wénhuà dāngzhōng ？#
enterprise culture in the middle

（3）a 高尔文 先生， 一个 企业 当时 处于
gāoěrwén xiānsheng , yígè qǐyè dāngshí chǔyú
Gao Erwen mister a enterprise at that time in
那样 一个 困境， 怎么 可以 在 这么 短
nàyàng yígè kùnjìng , zěnme kěyǐ zài zhème duǎn

such a predicament how could preposition such short
的 时间， 又 重新 回到 轨道， 重新 又
de shíjiān，yòu chóngxīn huídào guǐdào，chóngxīn yòu
particle time again once again back track once again again
让 大家 刮目相看 呢？
ràng dàjiā guāmùxiāngkàn ne ？
make people impress particle

《对话— 走出 困境》
《duì huà — zǒuchū kùn jìng》
Dialogue----Out of Predicament

（3）b 高尔文 先生， 一个 企业 当时 处于
gāoěrwén xiānsheng，yígè qǐyè dāngshí chǔyú
Gao Erwen mister a enterprise at that time in
那样 一个 困境， 怎么 可以 在 这么 短
nàyàng yígè kùnjìng，zěnme kěyǐ zài zhème duǎn
such a predicament how could preposition such short
的 时间， 又 重新 回到 轨道， 重新 又
de shíjiān，yòu chóngxīn huídào guǐdào，chóngxīn yòu
particle time again once again back track once again again
让 大家 刮目相看 ？#
ràng dàjiā guāmùxiāngkàn ？#
make people impress

By way of comparison it is clear that when rid of “ne” sentences of (b) group appear to be lacking in acceptability. All these examples are the same as (1) d in that we have to add the modal particle “ne” to increase interrogative information so as to reinforce the focal position of the original interrogative pronouns and make the information structure of the sentence appear complete. Therefore, for interrogative sentences with complex structure and less prominent interrogative forms, “ne” with interrogative information becomes an important mark to highlight the function of interrogative sentence.

2.2 The Degree of Prominence of Interrogative Form[5] Affects the Use of “ne”

From the above discussions we have reached the conclusion that “ne” carries interrogative information. However, it raises a new question that now that “ne” conveys interrogative information as well, would the use of “ne” repeat the interrogative information expressed by the original interrogative form of the sentence? Furthermore, in actual language material some interrogative sentences use “ne” whereas others do not. Why would it be this case and are there any regularities?

Our contention is that, seen from the requirement of information transmission, the degree of prominence of the interrogative form affects the use of “ne.” For instance,

[5] “interrogative form” refers to the language form which carries the interrogative information.

in the examples (2), (3) of 2.1 section, the interrogative pronouns “zěnyàng” and “zěnme” serve as interrogative form and none of them occupies the highlighted focal position. Moreover, their focal positions were further weakened as a result of the complex sentence structure following them. Consequently we might argue that these sentences have less prominent interrogative forms.

Let us take a look of some examples with more prominent interrogative forms:

(4) 李清川： 教育部 实施

lǐqīngchuān： jiàoyùbù shíshī

Li Qingchuan: The Ministry of Education implement

“长江学者 奖励 计划” 已经 有 四 五

“chángjiāngxuézhě jiǎnglì jìhuà ” yǐjīng yǒu sì wǔ

“Yangtze River Scholar Award Program” already have four five

年 了， 中山大学 也 因此 增设

nián le, zhōngshāndàxué yě yīncǐ zēngshè

year particle Zhongshan University also accordingly add

特聘教授 岗位。 情况 怎么样 ？

tèpìnjiàoshòu gǎng wèi 。 qíngkuàng zěnmeyàng ？

distinguished professor positions situation how

《中国 知名 大学 校长 访谈录 —— 岭南 一派》

zhōngguó zhīmíng dàxué xiàozhǎng fǎngtánlù —— lǐngnán yī pài

Interviews of Well-Known University Chancellors----The Lingnnan School

（5）穆涛： 你 认为 写作 一 部 长篇 小说

mùtāo：nǐ rènwéi xiězuò yī bù zhǎngpiān xiǎoshuō

Mu Tao you think writing a measure word long-length novel

过程 中 最 重要 的 是 什么？

guòchéng zhōng zuì zhòngyào de shì shénme ？

process in most important particle is what

《演技 —— 中国 著名 作家 访谈录： 贾平凹》

《yǎnjì —— zhōngguó zhùmíng zuòjiā fǎngtánlù： jiǎpíngwā 》

Acting: Interviews of Renown Writers in China: Ja Pingwa

（6）梅辰：…… 《教育短波》 一共 办 了

méi chén：……《jiàoyùduǎnbō 》 yīgòng bàn le

Mei Chen: Shortwave of Education altogether run particle

多少 年？

duōshào nián ？

how many year

《人文 大家 访谈录 —— 何兹全》

《rénwén dàjiā fǎngtánlù —— hézīquán 》

Masters of Humanties----He Ziquan

The interrogative forms in all the three examples are placed at the end of the sentence. According to the information-parsing theory the end of the sentence is the place where the focus is relatively highlighted. Therefore the interrogative forms in all the three examples are quite prominent.

Then what relationships does the degree of the prominence of interrogative form have with the use of "ne"? To answer this question we have compared the syntax position of the interrogative form "shénme" (what) in the two books: *Wonderful Truth* and *Interviews of Maters of Humanities*. The comparison statistics are as follows:

Table 1. The comparison statistics about the syntax position of the interrogative form "shénme"

Syntax position / Statistic sentence number / Use or no-use of "ne"	"shenme" as subject or component of subject (the prominence of interrogative focus is weak)	"shenme" as object or component of object (the prominence of interrogative focus is stong)
Sentences without "ne"	4 sentences	109 sentences
Sentences with "ne"	35 sentences	11sentences

Based on the result of the statistics we can draw a general tendency: when "shénme" is used as subject or component of subject sentences with "ne" outnumber those without "ne." when "shénme" is used as object or component of object and therefore is close to the end of the interrogative sentence where the focus is highlighted, "ne" is usually omitted. As with the above examples (4), (5), and (6), the interrogative forms are all placed at the end of the sentences where the focus is to be highlighted, and as a consequence it is more likely that "ne" is omitted. The sentences would read rather unnatural if "ne" was added.

Now we may come to the conclusion that the more prominent the interrogative form is, the less likely of the use of "ne."[6] The explanation is rather simple seen from the perspective of information transmission. Since interrogative form already occupies a strong interrogative focal position it has no need for modal particle "ne" carrying interrogative information to reinforce its focal function. And the interrogative information would appear superfluous if "ne" was added. This argument can be corroborated by the counter examples from (2) to (3) in 2.1section in which the sentences are more likely to use "ne." As the focal position of the interrogative forms is less prominent in these sentences, using "ne" can help reinforce its focal position and highlight the interrogative function and further keep the information structure of the whole sentence in balance. This tendency can be illustrated with the following Fig:

[6] This rule applies to the general situations of information transmission. However, it is a different matter if the speaker wants to add more subjective expressions on purpose.

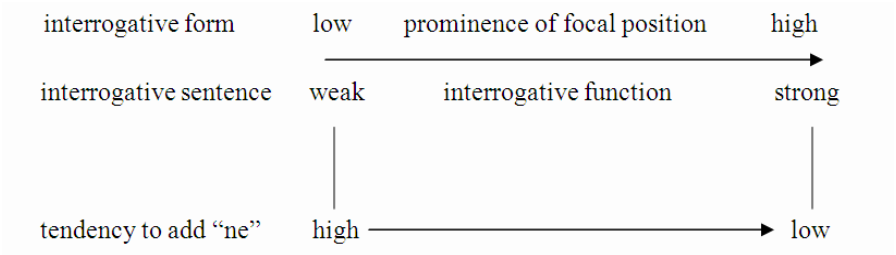


Fig. 2. The illustration of the tendency about the use of "ne"

3 The Impacts of the Interrogative Information of "ne" on the Function of Interrogative Sentence

As "ne" carries with itself interrogative information, when interacting with the interrogative forms originally included in the sentence it entails the sentence with some new function and affect the use of interrogative sentence. The impacts can be investigated mainly from the following two aspects:

3.1 "Ne" Adds Further Interrogative Mood to Interrogative Sentence

Shao Jingmin [8] once argued that adding further interrogative mood is precisely the derivative function the modal particle "ne" has in interrogative sentence. What we differ from Shao is that the implication of further interrogation is not carried by "ne" itself. It is a new function it entails to interrogative sentence when used in such a sentence. How is this new function of "further interrogation" generated? Is it the case that "ne" adds "further interrogation" to all non-yes-no questions?

In order to answer these questions we have to go back to the issue of the degree of the prominence of focus discussed above:

1. For interrogative forms with more prominent focus, sentences tend to have the function of "further interrogation" when "ne" is added. For instance:

（7）a 崔永元：　毕业　以后　会　去　干　什么？
cuīyǒngyuán: bìyè yǐhòu huì qù gàn shénme ?
Cui Yongyuan: graduation after will go do what
《精彩　实话 —— 同　在　蓝天　下》
《jīngcǎi shíhuà—— tóng zài lántiān xià》
Wonderful Truth——under the same blue sky
b 毕业　以后　会　去　干　什么　呢？
bìyè yǐhòu huì qù gàn shénme ne?
graduation after will go do what particle

（8）a 鲁豫： 那时候 每 个 月 给 你
lǔ yu: nàshíhòu měi gè yuè gěi nǐ
Lu Yu: at that time every measure word month give you
多少 钱?
duōshǎo qián ?
how much money
《鲁豫 有约： 说出 你 的 故事》
《lǔyù yǒuyuē: shuōchū nǐ de gù shì》
Lu Yu: Tell Your Story.

b 那时候 每 个 月 给 你
nàshíhòu měi gè yuè gěi nǐ
at that time every measure word month give you
多少 钱 呢?
duōshǎo qián ne?
how much money particle

By comparing the two groups of sentences it is evident that (b) sentences with “ne” has the implication of further interrogation than (a) sentences without “ne.” Why is the case? The interrogative forms of the three groups of sentences all are placed at the end of each sentence where the focus is highlighted, therefore the interrogative function is quite pronounced even in (a) sentences without “ne.” By adding the modal particle “ne” which carries with itself interrogative information the sentences acquire additional amount of interrogative information, and a direct result of it is the “cumulative value.” As a result of this “cumulative value” the interrogative information not only entails to the sentence an interrogative function, it also brings forth a new function----“further interrogation.” Therefore we can argue that the implication of “further interrogation” of “ne’ questions is not merely carried by “ne” but the result of the cumulative interaction between the interrogative information of “ne” and the strong interrogative information originally contained in the sentence.

2. Interrogative forms with less prominent focus will not generate “further interrogation” when “ne” is added. As the interrogative function of this type of interrogative forms is not sufficient and even risks to compromise the survival of the sentence, adding “ne” can only help supplement interrogative information and highlight the focus and consequently improve the acceptability of the sentence. As illustrated by the example sentences from (1) to (3) in 2.1 section, the acceptability will be compromised without adding “ne” or other interrogative forms. Thus as far as this type of sentences is concerned, adding “ne’ is required. “Ne” in this type of sentences only helps improve the basic interrogative function of the sentence and will not generate the new function of “further interrogation.”

3.2 Does "ne" Help Generate "Contrastive" Implication? Explanations from Information-Parsing Perspective

Some scholars contend that "ne' questions have the function of contrasting current situation with that mentioned earlier. For instance, Jin Lixin [3] has argued that the pragmatic function of "ne" is one of "putting in contrast of current situation according to the circumstances mentioned earlier." Let us take a look of the following examples:

(9) 明天 我 一整天 都 在家， 你 什么时候 来 呢?
míngtiān wǒ yīzhěngtiān dōu zàijiā，nǐ shénmeshíhòu lái ne ?
tomorrow I the whole day all at home you when come particle
(JinLixin 1996 example)

(10) 今天 没有 车子， 你 怎么 去 呢?
jīntiān méiyǒu chēzi，nǐ zěnme qù ne ?
today no car you how go particle
(JinLixin 1996 example)

We contend that the "contrastive" implication is not a function of "ne" by itself but generated by the context. "Ne" usually appears in the follow-up sentence of compound sentence, and more often than not it appears together with the conjunction "nà me."(then) Such contrastive context can mislead people to think that "contrasting" is a function of "ne." Then we may ask the question why "ne" question is usually used in follow-up sentence instead of beginning sentence. We believe that what is at work here is the rule of information transmission. For these sentences with interrogative form as follow-up sentence the previous sentence usually is statement sentence with strong declarative function, and the declarative information has an inertia within the flow of speech which distracts the addressee from receiving interrogative information contained in follow-up sentence. This is particularly the case with interrogative pronouns, alternative conjunction "hái shì"(or) and "A-not-A duplication." As these interrogative forms all have non-interrogative functions by themselves the interference of declarative information will weaken their interrogative function. Consequently the interrogative information of follow-up sentence appears to be not explicit enough. In situations like this the use of the interrogative modal particle "ne" is required to reinforce interrogative information and function. For example, when rid of "ne" from the above two examples the interrogative forms all can be treated as a part of a statement sentence.

(11) 明天 我 一整天 都 在家， 你 什么时候 来
míngtiān wǒ yīzhěngtiān dōu zàijiā，nǐ shénmeshíhòu lái
tomorrow I the whole day all at home you when come
都 可以。
dōu kěyǐ。
adv. okay

(12) 今天 没有 车子， 你 怎么 去 很 成 问题。
jīntiān méiyǒu chēzi，nǐ zěnme qù hěn chéng wèntí 。
today no car you how go very become problem

4 Conclusion

By investigating the dynamic restricting relationships between semantic form and semantic information of “ne” sentence and expounding on the property of “ne,” this paper argues that “ne” is an interrogative modal particle carrying interrogative information. It further points out that the degree of prominence of the interrogative forms affects the use of “ne,” and concludes that the more prominent the interrogative form is, the less likely the use of “ne;” conversely, the less prominent of the interrogative form the more likely the use of “ne.” this study further examines the interrogative information carried by the modal particle “ne” and provides an integrated explanation on the “further interrogation” and “contrasting” function of “ne. This study has demonstrated the inseparable relationship between lexical meaning and sentence function from the perspective of information transmission.

Acknowledgments. This study was supported by the following funds: Hubei Province Social Science Fund Project, “Constructing the Chinese Question-Answering Model” (project number: [2009]129); Youth Project, College of Humanities, Wuhan University, “Research of Modern Chinese Interrogative Sentences: Toward a Study of Chinese Information Processing.”

References

1. Guo, T.T.: Grammatical Restrictions of Information Structure on ‘X-neg-X’ Mark. Yangtze River Academic (4) (2009)
2. Hu, M.Y.: Modal Particles and Interjections in Beijing Dialect. Zhongguo Yuwen (5-6) (1981)
3. Jin, L.X.: On ‘ne’ in Interrogative Sentence. Language Teaching and Linguistic Studies (4) (1996)
4. Lu, J.M.: On Interrogative Modal Particles in Modern Chinese. Zhongguo Yuwen (5) (1984)
5. Lü, S.X.: Interrogation. Positive. Negative. Zhongguo Yuwen (4) (1985)
6. Kimura, H.: Pragmatics of Transmission Function of Chinese Modal Particles. In: Anthology of Studies of Chinese Language from Early Modern and Modern. Beijing Language College Press, Beijing (1993)
7. Qi, H.Y.: Analyses of ‘ne’ in Terms of Meaning and Historical Evolution. Journal of Shanghai Normal University (Philosophy and Social Science Edition) (1) (2002)
8. Shao, J.M.: The Function of the Modal Particle ‘ne’ in Interrogative Sentence. Zhongguo Yuwen (3) (1989)
9. Shao, J.M.: A Study of Modern Chinese Interrogative Sentences. East China Normal University Press, Shanghai (1996)
10. Xiao, G.Z.: A Thesis on the Chinese Grammar. Huazhong Normal University Press, Wuhan (2001)
11. Xiao, G.Z.: Chinese Grammar: Factual Discovery and Theoretical Exploration. Hubei People’s Press, Wuhan (2005)
12. Xiong, Z.R.: The Meaning of ‘ne’ in Interrogative Sentence. Journal of Anhwei Normal University (Social Science Edition) (1) (1999)
13. Zhu, D.X.: Notes on Grammar. The Commercial Press, Beijing (1982)

A Metonymic Approach to the Cognitive Mechanism of Chinese Lexical Meaning

Yanfang Liu

School of Foreign Languages, Zhongnan University of Economics & Law, Wuhan, China
Cherry34260@yahoo.com.cn

Abstract. The traditional study takes historical and social factors as the main reasons for the development of lexical meaning, while the cognitive study argues that the development of lexical meaning is motivated, instead of being arbitrary and independent of human beings' cognition and experience. Metonymy, as one of people's cognitive ways of thinking, is one of the approaches to and the internal factors for the development of lexical meaning. This paper attempts to explore the cognitive mechanism of Chinese lexical meaning from the perspective of metonymy.

Keywords: metonymy, cognitive mechanism, lexical meaning.

1 Introduction

The study of lexical meaning is, in essence, to explore how and why lexical meaning develops. The traditional study focuses more on the semantic description with examples, lack of a comprehensive summary of the specific rule. Although historical and social factors are the main reasons for the development of lexical meaning, they are simply external ones, which definitely make the study on the mechanism of lexical meaning rather inadequate. The experts on cognitive semantics have argued that the development of lexical meaning originates from language users' cognitive way of thinking. As a matter of fact, historical and social factors can only show the necessity of the development of lexical meaning, and it is cognitive factors that can account for its internal mechanism.

As one of the main modes of semantic transfer, metonymy is, to a great extent, neglected by modern linguistics, especially lexicology and lexical semantics. According to cognitive linguistics, metonymy and metaphor are essentially the same, which are both conceptual tools with a cognitive domain activating another one. In fact, metonymy and metaphor are both related and different. They are overlapped and interacted to certain degrees, which makes them difficult to be clearly distinguished. However, there remains the main difference that metaphor involves the mapping between different cognitive models, while metonymy in the same cognitive model, that is, a concept in the same cognitive model is used in place of another one so that the latter is highlighted. Although some scholars, like Bréal, Trier, Ullmann, and

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 634–640, 2013.

Jakobsom, have discussed the role of metonymy in the development of lexical meaning, metonymy is still considered to be subordinate to, even attached to metaphor in the overall development of lexical meaning, as it has been long claimed that metaphor includes synecdoche, metonymy, hyperbole, litotes and euphemism. [1] In the early cognitive studies on language of the 1980s, although Lakoff & Johnson emphasized the importance of some types of metonymy, metaphor has been occupying the dominant position with researchers.[2] Not until recent years has the importance of metonymy in the development of lexical meaning gradually drawn people's attention. Actually, as early as the nineteenth century, some western scholars, like Bréal, have discussed the role of metonymy in the development of lexical meaning, who once believed that almost all the narrowing, extension, upgrading and downgrading of lexical meaning are the result of metonymy. [1]

As a cognitive mechanism, metonymy plays a significant role in the development of lexical meaning, owing to which the meanings of a large number of words are extended by means of conceptual mapping. Nowadays, more and more scholars have agreed that most metaphors are based on metonymy, and that metonymy has become one of the basic mental processes for meaning extension, possibly more basic than metaphor. [3] Therefore, metonymy makes one of the fundamental justifications for lexical meaning extensions, and possibly turns out to be the most basic cognitive mechanism.

2 The Importance of Metonymy on Lexical Semantics

According to cognitive semantics, metonymy is not only a way of thinking, but also the important component of language, which has great influence on the study of lexical semantics. The traditional study of lexical semantics and world knowledge are inadequate to understand and clarify lexical meaning. In recent years, the study of lexical meaning based on cognitive semantics has derived lexical meaning from human perception and experience. It is believed that human beings, in the process of experiencing the real world, have gradually constructed a category, concept and way of thinking, established the cognitive structure, and then acquired meaning. It seems that metonymy is a crucial cognitive mechanism for this kind of experience.

The semantic justification for man-made words can be classified as "metaphoric" and "metonymic". Ullman has made a further claim that there is not a single language without concerning metaphor and metonymy, which are inherent in the basic structure of human language. [4] Like metaphor, metonymy is a principle for language, active in the process of a large number of semantic transfers. It was once taken as a figure of speech, a beautiful literary technique of expression.[5] Leach also put metonymy as a type of semantic transfer, equally important as metaphor. [6] Sweester explored the importance of metonymy in lexical meaning from the perspective of cognitive psychology, and argued that linguistic categorization depends not only on our different naming for the existing world, but also on our metonymic perception of the real world. [7] Taylor also believed that metonymy is one of the most basic ways of meaning extensions, and possibly even more basic than metaphor, and that both in the

past and at present metonymy (compared with metaphor) has been rarely discussed. Taylor further discussed the three ways of lexical meaning extensions based on metonymy. [8] In fact, words, too many to count in language, extend their lexical meaning with metonymy.

Nowadays, the semantics experts have generally believed that metonymy, as important as metaphor, is the common way of thinking for all human beings, and an important means for people to understand the existing world and name all the living things. As a result, a new probe into metonymy has important significance on the study of lexical semantics. Object naming indicates that there exists an "intermediary" between reality and language; that is, the objects or phenomena in the real world activate human's sense organ, thereafter people begin to perceive them, and choose certain fixed perceptual center from a variety of discrete perceptual materials. When selecting the appropriate expression, people tend to emphasize a certain aspect of objects, such as characteristic, shape, material, or function. And the cognitive mechanism of metonymy describes the lexical meaning by highlighting one aspect of the object. With objects' or concepts' various attributes, people's metonymic way of thinking often focuses more on the ones which are most prominent, easy to remember and understand. In this way, conventional metonymies can be solidified and become a part of lexical meaning which are included in the dictionary. Solidified metonymic meaning will turn out to be the tool for multilevel categorization, playing an important role in the forming of complex interconnected category network.

In recent years, metonymy has gradually become an important topic for the studies of lexical semantics and cognition, which has been considered as an indispensible cognitive tool for people to construct the concept system. Gardenfors has, in particular, pointed out the importance of metonymic operations in the conceptualization of meaning. He viewed the meaning of language as the mapping of language symbols onto certain mental entities. [9] In human's mind, meaning is attached to perception, and metonymic mapping is to use the word with the meaning of default construal (original meaning) to activate another totally different construal meaning (target meaning). In this process, the original meaning and the target meaning should belong to the same category, and we may infer their relationship according to the metonymic principle.

As a cognitive mechanism, metonymy helps words to change and develop in a particular way, which is the key to understanding the transfer, extension and narrowing of lexical meaning. That is to say, metonymy presents a challenge for the arbitrariness of language. Lexical meaning is largely justified, which is, to a certain extent, influenced and restricted by people's metonymic way of thinking as well as social and cultural background knowledge. There are varied justifications for metonymic lexical meaning, and metonymic relations can be derived from the closeness of things and symbols, or from cultural traditions and particular experience. Furthermore, particular experience may bring forth particular proximity; to be specific, people may think of the thing you see or hear, and associate with what one think of. In fact, the study of metonymy has constantly change people's understanding of lexical meaning that lexical meaning is constructed in the metonymic way of thinking rather than in an arbitrary manner.

3 The Metonymic Mechanism for Lexical Meaning Development

It has been argued that being uncertain, dynamic, fuzzy, and on-line, lexical meaning is people's dynamic construal for the actual situation. People tend to depend on a large amount of experience knowledge for the further metonymic reasoning, thus determining the appropriate lexical meaning. Metonymy, involving "contiguity" and "salience", is the cognitive process in the same "domain" or "idealized cognitive model" described as the overall knowledge for certain experience, characterized concepts, and semantic knowledge structure. [10] In this very cognitive process, a conceptual entity provides a mental access into another one. When getting to know the real world, people, based on their experience, tend to understand a more familiar object first, and then something similar so that when a new concept emerges, they will try to extend the lexical meaning of the existing concept instead of inventing a brand-new word.

Next, I attempt to clarify the process of metonymic cognition with the semantic system and the formation process of the Chinese word "臭 chòu (smelly)". The basic meaning of "chòu" is "a distinctive odor that is offensively unpleasant". With time going on, it has been given other lexical meanings due to metonymic cognition. For example:

- 不要摆臭架子。

bú yào bǎi chòu jià zi.
(Don't put on frills.)

- 市长背了一篇又臭又长的欢迎词。

shì zhǎng bèi le yì piān yòu chòu yòu cháng de huān yíng cí.
(The mayor recited a long and tedious speech of welcome.)

- 卖国贼遗臭万年。

mài guǒ zéi yí chòu wàn niǎn.
(The ignominy of the traitors will never be forgotten.)

Although "chòu" in the above sentences has no relation with the smell, they are closely related to the basic meaning of "chòu". In the examples, the reason why "chòu" can be understood is that it highlights "the quality extremely bad", which can either activate the metonymic target or provide the mental access into the target. The lexical meaning of the word "chòu" is based on the relevant knowledge experience, like "bad", "unpleasant", "offensive", and the principles of conceptual contiguity and salience are the internal mechanism for its lexical meaning.

Metonymy is the powerful cognitive tool for the conceptualized world, because metonymy makes it possible for us to conceptualize the object by relating it with another one; that is to say, people tend to substitute one part of an object well known or easily perceived for the entire object or other parts. Lakoff believed that metonymy is a cognitive phenomenon which defines and interprets the relationship between the parts and the whole in the same "idealized cognitive model", and that owing to the contiguity we can perceive other parts or the whole based on one part, or otherwise, which makes it possible to create and understand new lexical meanings. [2] Take the following phrases as an example --- "丢脸diū liǎn (lose one's face)", "铁青着脸 tiě qīng zhe liǎn (black in the face)", "拉长着脸 lā chāng zhe liǎn (draw a long face)", "两面三刀 liǎng miàn sān dāo (have two faces)", "板着脸 bǎn zhe liǎn (keep a

straight face)", "愁眉苦脸 chóu méi kú liǎn (a face as long as a fiddle)". Because we enjoy the same "reference domain" or "idealized cognitive model", we are able to perceive the entire object or other parts according to the characteristics and attributes of "脸 liǎn (face)", our most familiar body organ - face. Langacker took the metonymic relation between the parts and the whole as a phenomenon of reference point. [11] In this sense, metonymy is a phenomenon of reference point, and the entity a word refers to serves as a reference point, which provides the mental access into the target to be described or expressed and also leads the reader to the very target. In the above examples, "liǎn" plays the role of reference point, and all the phrases make the "reference domain" or "idealized cognitive model" as a whole, which can be interpreted with our mental accessibility of "liǎn". Furthermore, because of the same "reference domain" or "idealized cognitive model", the salience of "liǎn" may lead the reader to the target meaning of these phrases.

Words usually reflect the conventional imagery. People may construct a scene for language expression in a particular way, or may highlight one aspect and weaken the others. In the process of human cognition, people, with the help of conventional knowledge, tend to categorize the cognitive phenomenon, and form a variety of frame knowledge, the so-called "domain". The knowledge provides the basis for understanding or reasoning. For example, the conventional knowledge of "头 tóu (head)" involves components, shape, size, function, etc.. The expressions related to "tóu" almost have the same conventional knowledge in world culture so that people, based on conceptual contiguity, can turn to metonymy to reason out their meaning.

Word itself has no meaning, which is simply a prompt to construct meaning. When we understand expressions, we are not trying to interpret what the words are saying, which turn out to mean nothing without the effective cognitive process and knowledge. [12] Therefore, metonymy is a means to endow language symbols with meaning. When the relation incurred by metonymy has been widely accepted by the society, the metonymic meaning of the word will become fixed and conventional in the language system. The basic meaning of "窄 zhǎi (narrow)" is "not wide, especially in comparison with length or with what is usual". The reader, with the metonymic way of thinking, may understand other meanings of this word without too much cognitive endeavor --- "narrow-minded, lacking tolerance or flexibility; not having enough money for living". When the meanings of "zhǎi" are fixed in the dictionary, metonymy evolves into conventional metonymy.

With lexical meaning extending in a metonymic way, contiguity is characteristic of the meanings, which also reflects human's salient way of thinking. Salience may lead to extension or narrowing of lexical meaning. For example:

- 我们不聘用长头发。

wǒ mén bú pìn yòng cháng tóu fà.
(We don't hire longhairs.)

- 白宫正保持沉默。

bái gōng zhèng bǎo chí chén mò.
(The White House isn't saying anything.)

- 公共汽车在罢工。

gōng gòng qì chē zài bà gōng.
(The buses are on strike.)

In the above sentences, "cháng tóu fà" should be understood as "the person with longhairs", "bái gōng" as "the white house government", and "gōng gòng qì chē" as "the bus drivers". From the original referent to the contiguous one in the same cognitive domain, metonymic mapping change the meaning of "cháng tóu fà", "bái gōng" and "gōng gòng qì chē". Panther & Thornburg interpreted "contiguity" as the contingent relationship, as the result of readers' experiential association, which shows the experiential basis of metonymy, being a guide to the cognitive study on lexical meaning.

4 Implications

Metonymy is the product of human cognition development, as well as a necessity to understand the objects in the real world. The creativity of human brain lies in that it may perceive and name new objects with the help of the existing objects or language forms, which is the result of the development of human's metonymic cognition. Only with metonymic cognition is it possible for humans to store, memorize and express knowledge, in-formation, and words in an effective manner, and then further perceive and understand the real world.

By recognizing the nature of metonymy and its significance for the development of lexical meaning, it can help both foreign language teachers and learners to understand the word meaning more effectively. Furthermore, the importance of conceptual metonymy theory in the development of lexical meaning can also provide guidance for dictionary compilation. The dictionary complier can state specifically the process of metonymic cognition, making the lexical meanings integrated, thus helping the language learners to memorize vocabularies much easier.

References

1. Traugott, E.C., Dasher, R.B.: Regulating in Semantic Change. Cambridge University Press, Cambridge (2005)
2. Lakoff, G., Johnson, M.: Metaphors We Live By. The University of Chicago Press, Chicago (1980)
3. Taylor, J.R.: Category Extension by Metonymy and Metaphor. In: Ren, Dirben, Porings, R. (eds.) Metaphor and Metanymy in Comparison and Contrast. Mouton de Gruyter, Berlin (2002)
4. Ullmann, S.: Semantics: An Introduction to the Science of Meaning. Basil Blackwell, Oxford (1962)
5. Waldron, R.A.: Sense and Sense Development. Andre Deutsch Ltd., London (1979)
6. Leech, G.N.: A Linguistic Guide to English Poetry. Longman Group Ltd., Harlow (1969)
7. Sweester, E.: From Etymology to Pragmatics. Cambridge University Press, Cambridge (1990)
8. Taylor, J.R.: Linguistic Categories: Prototypes in Linguistic Theory. Clarendon Press, Oxford (1995)

9. Gardenfors, P.: Some tenets of cognitive semantics. In: Allwood, J., Gardenfors, P. (eds.) Cognitive Semantics: Meaning and Cognition, pp. 21–25. John Benjamins Publishing Company, Amsterdam (1999)
10. Radden, G., Kovecses, Z.: Toward a theory of metonymy. In: Panther, K.-U., Radden, G. (eds.) Metonymy in Language and Thought. John Benjamins Publishing Company, Amsterdam (1999)
11. Langacker, R.W.: Reference-point Construction. Cognitive Linguistics 4, 13–16 (1993)
12. Turner, M.: Reading Minds: The Study of English in the Age of Cognitive Science. Princeton University Press, Princeton (1991)

A Study on Measure Adjectives from the Perspective of Semantics

Wei Wang

Division of Chinese, Nanyang Technological University, Singapore 637332
wang0462@e.ntu.edu.sg

Abstract. There is a special category in adjectives, which is named as measure adjective. It is used to describe linear dimensions of physical objects in space, and to state the age of something physical. Measure adjectives share the general characteristics with other types of adjectives, but also have their unique characteristics. Most importantly, these words have peculiar distributions, and different usage frequencies, etc. In this study, from the perspective of semantics, the semantic features of measure adjectives will be analyzed, so the inherent reasons will be revealed why the distributions are peculiar and usage frequencies are different.

Keywords: Measure adjective, Semantics, Macro-scale measure adjective, Micro-scale measure adjective.

1 The Definition of Measure Adjectives

As special category in the adjectives in Chinese studies, measure adjectives have been investigated by many researchers. Lu [1] defined the measure adjectives as that these adjectives are used to describe the scale, and have the semantic feature of [+scale], these adjectives include '大da big,小xiao small,长chang long,短duan short,宽kuan wide,窄zhai narrow,远yuan far,近jin close', etc. The standard in his study was if these measure adjectives can be used in the frame 'A+ (了le particle) + quantitative numeral', and in this frame they should have the meaning of deviation. According to his study, '大da big, 长 chang long, 高gao high/tall, 宽kuan wide,厚hou thick, 深 shen deep, 粗cu thick, 重zhong heavy, 远 yuan far, 快kuai fast, 晚（迟）wan (chi) late, 贵 gui expensive, 多duo more' and '小xiao small, 短duan short, 低（矮）di (ai) short, 窄zhai narrow, 薄 bo thin, 浅 qian shallow, 细 xi slender, 轻qing light, 近jin near, 慢man slow, 早zao early, 贱（便宜）jian (pian yi) cheap, 少shao less 'could all be considered as measure adjectives.

Li [2] defined the 'spatial quality' that it is used to measure the dimensions of the objects including length, height, depth, distance and thickness, as well as the area, volume and the distance between the objects. He thought the measure adjectives should have the semantic feature of [+spatial quality]. These measure adjectives are used to indicate the dimensions including length, height, depth, distance and

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 641–652, 2013.

thickness, area and volume of objects, as well as the distance between objects. These measure adjectives are mainly 8 pairs:

大da large: Superior to the average or the counterpart for comparison in volume, area, quantity, power and strength. (opposite to小xiao small);

小xiao small: Inferior to the average or the counterpart for comparison in volume, area, quantity, power and strength. (opposite to大 da big);

长chang long: The distance between two objects is large. (opposite to短duan short, the meaning of its antonym is neglected hereafter)

宽kuan wide: The width is large; range is large. (opposite to窄 zhai narrow);

高gao high/tall: The distance from the bottom to the top is large. Far from the ground. (opposite to低di low or矮ai short);

远yuan far: The span in space and time is large (opposite to近jin close)

深shen deep: The distance from the top to bottom or that from exterior to the interior is large (opposite to浅qian shallow.)

粗cu thick: the cross-sectional area of rod is large (opposite to 细xi thin)

厚hou thick: the distance between two parallel surfaces is large (opposite to薄bo thin)

From the studies above, it can be seen that the definition on measure adjectives by Li [2] is stricter than that by Lu [1]. For example, Li didn't classify 'fast, heavy, late, expensive, more' and 'slow, light, early, cheap, less' as the measure adjectives, and he believed that only the adjectives having the semantic feature of [+spatial quality] can be classified as measure adjectives.

In this study it is believed that this pair of adjectives 'heavy, light' is related to the volume of object, hence it should be considered to have the semantic feature of [+spatial quality]. Meanwhile, 'fast, late, expensive, more', 'slow, early, cheap, less' can be used in the frame of 'A+ (了le particle) + quantitative numeral', and in this frame they have the meaning of deviation, so these 5 pairs should also be considered as measure adjectives. However, they have different distributions and semantic characteristics from other nine pairs of measure adjectives, which have the semantic features of [+spatial quality], these differences will be addressed in this study. Therefore, according to prototype theory, these nine pairs of measure adjectives with [+spatial quality] can be classified as the typical measure adjectives, while 'fast, late, expensive, more' and 'slow, early, cheap, less' are classified as the atypical measure adjectives.

According to the frame defined by Lu [1] , it is found that the measure adjectives have the meaning of deviation in the frame of 'A+ (了le particle) + quantitative numeral', such as 'big, long, high, wide, thick, deep, thick, heavy, far, fast, late, expensive, more'.

A	大了一平米	长了三公分
	da le yi ping mi	chang le san gong fen
	large particle one square meter	long particle three centimeters
	one square meter larger	three centimeters longer

	高了三公分	宽了两公分
	gao le san gong fen	kuan le liang gong fen
	high particle three centimeters	wide particle two centimeters
	three centimeters higher	two centimeters wider
	厚了一公分	深了二十公分
	hou le yi gong fen	shen le er shi gong fen
	thick particle one centimeter	deep particle twenty centimeters
	one centimeter thicker	twenty centimeters deeper
B	小了一平米	短了三公分
	xiao le yi ping mi	duan le san gong fen
	small particle one square meter	short particle three centimeters
	one square meter smaller	three centimeters shorter
	低了三公分	矮了三公分
	di le san gong fen	ai le san gong fen
	low particle three centimeters	short particle three centimeters
	three centimeters lower	three centimeters shorter
	窄了两公分	薄了一公分
	zhai le liang gong fen	bo le yi gong fen
	narrow particle two centimeters	thin particle one centimeter
	two centimeters narrower	one centimeter thinner

As discussed by Lu [1] , the measure adjectives in Group A mean ‘excess’, and those in Group B mean ‘insufficient’.

2 Classification of Measure Adjectives

Lu [1] believed that the measure adjectives are all monosyllabic. They can be classified into two groups, one group describing larger scale, and the other group for the smaller scale. Huang and Shi [3] believed that the measure adjectives are classified as the ‘active’ and ‘passive’ groups. Sun [4] classified the two groups as ‘macro-scale’ and ‘micro-scale’, and her classification is used in this study.

Macro-scale measure adjectives and micro-scale measure adjectives are quite different in their usage frequencies and distributions. We can use the frame ‘How + measure adjective’ to tell the difference between the macro-scale measure adjectives and micro-scale measure adjectives.

A	How high /tall	How big
	How long	How fast
	How wide	How heavy

B	How short	How small
	How short	How slow
	How narrow	How light

The measure adjectives in group A are macro-scale. The questions in group A have no presupposition. 'How high' is used to ask the height, and the height is not known, therefore, the answer can probably be very high or very short.

The measure adjectives in group B are the micro-scale. The questions in group B have presupposition, therefore, the answer for the question of 'how short' is only "very short", or is to describe how short it is, while it cannot be 'not short'. Of course in some special contexts, the presupposition is wrong, and the answer can be 'not short', in this study the common context is discussed.

3 The Analysis in Semantic Features of Measure Adjectives

Measure adjectives are used to describe linear dimensions of physical objects in space, and to state the age of something physical.

Lu [1] studied the possible sentence patterns of measure adjectives. In this study it is found that the macro-scale measure adjectives and micro-scale measure adjectives are very different in term of the distributions in these sentence patterns. Some sentence patterns can be used in both macro-scale and micro-scale measure adjectives, but other sentence patterns can be used only in the macro-scale measure adjectives. This means that both macro-scale and micro-scale measure adjectives not only have the common semantic features, but also have their peculiar semantic features. In this study, the semantic features of measure adjectives will be analyzed, and the inherent reasons will be revealed for the peculiar distributions.

Cruse [5] studied several antonyms of adjectives in English including the measure adjectives, as seen in the figures below.

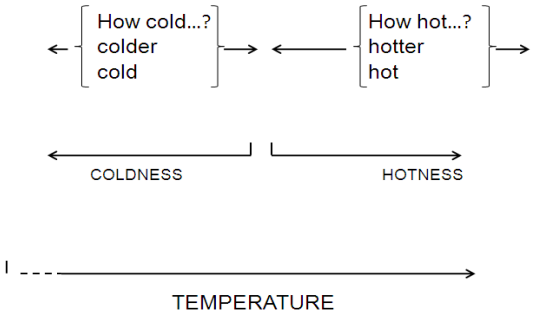


Fig. 1. Antonyms exemplified by 'hot' and 'cold'

Cruse analyzed the semantics of three different groups of antonym adjectives. As is seen in Fig.1, there are two scales underlying the pair 'cold, hot'. Since nothing that is colder can be hot, and nothing hotter can be cold, then we believe that there is no overlap between the scales of GOLDNESS and HOTNESS. The difference between hot and cold is in a sense absolute, rather than relative.

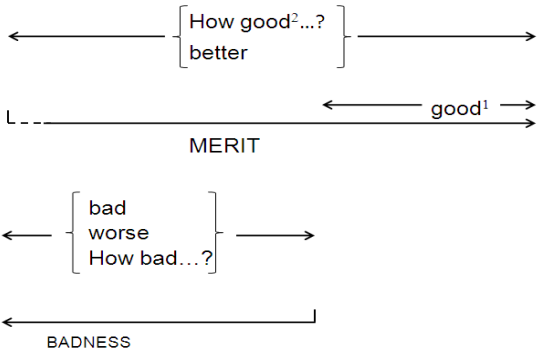


Fig. 2. Antonyms exemplified by 'good' and 'bad'

As is shown from Fig.2.that the adjectives 'good' and 'bad' stand for the subjective evaluations. There are two scales underlying the pair 'good, bad'. The scale of BANDESS must overlap the scale of MERIT (over which good^2 operates), but not extend into the region on the MERIT scale covered by good^1.

So is in Chinese. 'Good' can cover all the levels in the scale, and can also be a sub-section. But 'bad' can only be a sub-section. Therefore, in the sentence of 'is it good?' the answer can be 'good' or 'not good', while in the sentence of 'is it bad', the answer can only be 'bad' or 'very bad'.

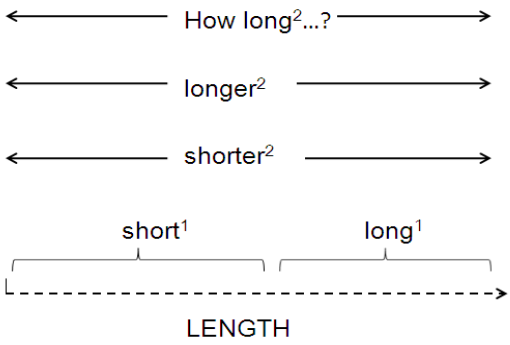


Fig. 3. Measure adjectives exemplified by 'long' and 'short'

Fig.3.shows the measure adjectives which we will focus on. Here 'long, short' are taken as example, there is a single scale underlies the pair 'long, short', and there is not cut-off point between 'long' and 'short'. The difference between long and short is in a sense relative. The macro-scale measure adjective 'long[1]' is also used to describe the full scale. In addition, the micro-scale measure adjective 'long[2]' can also be the sub-section of the full scale. As mentioned by Zhang [6], compared with the micro-scale measure adjective 'short[1]', the 'short[1]' in the scales is only a sub-section of 'long[2]'. Chinese may be the same with English, but in some cases Chinese may not be the same, and the main difference lies in that there are no comparatives in Chinese, while there are some sentence patterns which are used to describe comparison. In these sentences, the word senses of 'long' and 'short' may vary.

From the perspective of semantics, measure adjective 'long' has two word senses: 'long[1]'is used to describe the partial scale. Besides the semantic feature of [+ scale], it also has the semantic feature of [+ partial scale], which is opposite to 'short'. 'Long[2]' is used to describe the full scale, which has the semantic feature of [+ full scale] besides [+ scale]. The micro-scale measure adjective 'short' has only one word sense, and it also has the semantic feature of [+ partial scale], which is opposite to 'long[1]'. The two types of measure adjectives have both common semantic features and unique semantic features. According to their difference in the semantic features, the peculiar distribution of the two measure adjectives in the sentences will be further explained.

3.1 In the sentences which describe the full scale, only the macro-scale measure adjectives with [+ full scale] semantic feature can be used, while the macro-scale measure adjectives with [+ partial scale] and all the micro-scale measure adjectives cannot be used.

For example: the first sentence ---- 有+数量词 + ( )
you + shu liang ci + ( )
is + quantitative numeral+( )

有两平方米大
you liang ping fang mi da
is two square meters large
有三米长
you san mi chang
is three meters long

It is believed that '粗 cu thick' can also be used in this sentence, for example,

他的腰有三英尺粗。
ta de yao you san ying chi cu
he particle waist is three feet thick
His waist perimeter is three feet

However, the atypical macro-scale measure adjectives cannot be used in this sentence, such as 'fast, late, expensive, more '.

Here is the second sentence----- ( ) +数量词
( ) + shu liang ci
( ) + quantitative numeral

This sentence has ambiguities, if this is considered as the 'Subjective+Verb' structure, it means the full scale, hence only the macro-scale measure adjectives with [+full scale] can be used.

长江长6300公里。
chang jiang chang 6300 gong li
the Yangtze River long 6300 kilometers
The Yangtze River is 6300 kilometers long.
那水塔高40米。
na shui ta gao 40 mi
that water tower high 40 meters
That water tower is 40 meters high.

Lu [1] believed that only these six macro-scale measure adjectives 'long, high, wide, thick, deep and heavy' can be used in this sentence, while 'big, thick, far, fast, late, expensive, more' and micro-scale measure adjectives cannot be used in this sentence. Dong [7] gave a reasonable explanation that the object has three dimensions with three scales of length, width and height. The measure adjectives 'long, wide, high' are the expressions of three-dimensional space. 'Thick, deep' can be considered as the same dimension as 'high', i.e. 'long, wide, high', 'long, wide, thick', 'long, wide, deep' can reflect the three-dimensional space equivalently. 'Large, thick' are used to describe the 'appearance' of objects (area, volume, etc), not the simple linear distance.

It is more difficult to describe the appearance of objects than to describe the linear distance. Therefore, in sentences with accurate description, 'big, thick' and atypical measure adjectives cannot appear.

Here is the third special phrase ----- "名词+形容词" 估量短语。
"ming ci + xing rong ci "gu liang duan yu
'Noun +Adjective' (NA) evaluation phrase

In modern Chinese, when answering 'how +adjective', NA evaluation phrase can be used besides 'quantitative numeral'. Dong[7] mentioned that only the macro-scale measure adjectives can be used in (NA) evaluation phrase, such as 'big, thick, long, wide, high, thick, deep' etc. For example:

多大?	八平米大	核桃大
duo da	ba ping mi da	he tao da
How big?	eight square meters big	one walnut big

多长?	三米长	筷子长
duo chang	san mi chang	kuai zi chang
How long	three meters long	one chopstick long

多宽?	二尺宽	柜门儿
Duo kuan	er chi kuan	gui menr kuan
How wide?	two meters wide	one cabinet gate wide

多高?	一米高	桌子高
duo gao	yi mi gao	zhuo zi gao
How high?	one meter high	one table high

From examples above, it is found that these sentences need to describe the full scale. Two types of answers have different meanings. Quantitative numeral, for example, eight square meters, describes the accurate full scale, (NA) evaluation phrase describe the approximate scale.

Therefore, only the macro-scale measure adjectives with the semantic feature of [+full-scale] can be applied in this sentence, and those macro-scale measure adjectives with [+partial scale] and micro-scale measure adjectives cannot be used.

3.2 In the sentences which describe partial scale and deviation, only the macro-scale measure adjectives with [+partial scale] and micro-scale measure adjectives can be used.

For example this sentence ---- ( ) +数量词
( ) + shu liang ci
( ) + quantitative numeral

When this is considered as the 'Verb+Objective' structure, the macro-scale measure adjectives with [+ partial scale] and micro-scale measure adjectives can both be used. The measure adjectives with [+ partial scale] means excess, and the micro-scale measure adjectives mean insufficiency, but the macro-scale measure adjectives with [+ full scale] cannot be used.

长两厘米（述宾结构）
chang liang li mi (VO)
long two centimeters (VO)
two centimeters longer (VO)
短两厘米（述宾结构）
duan liang li mi (VO)
short two centimeters (VO)
two centimeters shorter (VO)

For another sentence --- 动词 + ( ) +了
dong ci + ( ) +le
Verb + ( ) + particle

Lu [1] believed that this sentence can have two meanings: one was that the results are realized, the other one is that results deviate from what is expected. In the second meaning, only the macro-scale measure adjectives with [+ partial scale] and micro-scale measure adjectives can be used. For example:

他这个坑儿挖深了，再填点土。
ta zhe ge kengr wa shen le, zai tian dian tu
he this classfier cave dig deep particle, again put a little soil
He digs this cave too deep, put some soil back
她的头发剪短了，再长点就好了。
ta de tou fa jian duan le, zai chang dian jiu hao le
she particle hair cut short particle, again long a little good particle.
Her hair is cut too short, the better if it is longer

3.3 From Fig. 3 it can be found that the comparatives 'longer' 'shorter' in English have the semantic feature of [+ full scale]. In Chinese there are no comparatives, but there are some sentence patterns which are used to describe comparison. For example:

这根绳子很短，就比那根长一点。
this gen sheng zi hen duan, jiu bi na gen chang yi dian.
this classifier rope very short ,only than classifier that long a little
This rope is very short, only a little longer than that one
这根绳子很长，就比那根短一点。
zhe gen sheng zi hen chang, jiu bi na gen duan yi dian.
this classifier rope very long, only than classifier that short a little
This rope is very long, only a little shorter than that one

Since something that is longer can be short, and something shorter can be long, here 'a little longer' we don't know if it is really long or short, while it means that it tends to 'long' in the full scale. Meanwhile 'a little shorter' does not necessarily mean 'short'. In the comparative sentences, both the macro-scale adjectives with semantic feature of [+ partial scale] and micro-scale adjectives can be used, while in this case the measure adjectives don't describe an absolute scale, but a relative scale.

Here is a sentence -----不+ （ ）
bu+ ( )
not + ( )

Lu [1] proposed that this sentence can have two meanings: one is that "not" is to deny what the measure adjectives describe, for example:

不大~大
bu da ~ da
not big ~ big

不小~小
bu xiao ~ xiao
not small ~ small

不长~长
bu chang ~ chang
not long ~ long

不短~短
bu duan ~ duan
not short ~ short

不高~高
bu gao~gao
not high ~ high

不低~低
bu di ~ di
not low vs low

不重~重
bu zhong ~zhong
not heavy ~ heavy

不轻~轻
bu qing~ qing
not light ~ light

不快~快
bu kuai, ~kuai
not fast ~ fast

不慢~慢
bu man~man
not slow ~ slow

In this case the macro-scale measure adjectives with the semantic feature of [+ partial scale] and micro-scale measure adjectives can all be used, and they describe the absolute scale. 'Not long' means 'short', 'not high' means 'low'.

This sentence also has the meaning of deviation. The second meaning not + ( ) is to deny the sentence ( ) +了.

( ) +le.

( ) + particle.

'Not long' does not necessarily mean 'short', so it describes a relative scale.

这根木棍儿长了。
zhe gen mu gunr chang le
this particle stick long particle
This stick is too longer.

这根木棍儿不长。
zhe gen mu gunr bu chang.
this particle stick not long
This stick is not long.

这跟木棍儿短了。
zhe gen mu gunr duan le
this particle stick short particle
This stick is too shor.t

这跟木棍儿不短。
zhe gen mu gunr bu duan
this particle stick not short
This stick is not short.

那灯吊得高了。
na deng diao de gao le
that lamp hang particle high particle
That lamp is hanged too high.

那灯吊得不高。
na deng diao de bu gao
that lamp hang particle not high
That lamp is hanged not high.

In the examples above, the sentences in left hand side actually have the comparative meanings. The macro-scale measure adjectives with semantic feature of [+partial scale] and micro-scale measure adjectives can be used in the sentence of '( ) + particle'. The sentences in the right hand side mean the opposition to the left corresponding sentences, which also have the comparative meanings. The macro-scale measure adjectives with semantic feature of [+ partial scale] and micro-scale measure adjectives can also be used. However, in these two sentence patterns, the measure adjectives describe the relative scale. For example,

That lamp is hanged too high. ----- That lamp is hanged not high.

We are not sure if the lamp is hanged 'high' or 'low'. 'High + particle' means approaching a little to 'high' in the length scale, 'not high' does not necessarily mean 'low' in this case.

4 Conclusion

Measure adjectives are a special category in adjectives in Chinese, and they have caused the attention and exploration from many researchers. From the viewpoint of semantics the measure adjectives in Chinese are investigated in this paper.

Firstly, the inherent characteristics of measure adjectives are studied. We believe that measure adjectives are mainly used to describe the scales of objects in space and time. In the meanwhile, we also adopt the criterion by Lu[1] in evaluating the measure adjectives, and try to find if these measure adjectives can be used in the format of 'A + (了 le particle) + quantitative numeral ' with semantics of deviation. On this basis, we select 13 pairs of measure adjectives, and divide them further into core measure adjectives and atypical adjectives, by judging if these adjectives can describe the scales in space and time.

Secondly, we carry out a detailed analysis on the characteristics of measure adjectives from the viewpoint of semantics. From the analysis, the macro-scale measure adjectives and micro-scale measure adjectives differ in terms of usage frequencies and distributions, and this difference comes from the different characteristics in their semantics. The macro-scale measure adjectives have two word senses: one has the semantic feature of [+ full scale], the other has the semantic feature of [+ partial scale], which is opposite to 'short'. However, the micro-scale measure adjectives have only one word sense, and have the semantic feature of [+ partial scale], which is opposite to 'long'. Therefore, in the sentences describing full scale, only the macro-scale measure adjectives with semantic feature of [+full scale] can be used, while in the sentences describing partial scale, deviation and comparison, the macro-scale measure adjectives with semantic feature of [+ partial scale] and micro-scale measure adjectives can all be used. From this analysis, it should be abundantly clear why the usage frequencies of macro-scale measure adjectives are higher than those of micro-scale measure adjectives, and we can also explain peculiar distribution of measure adjectives.

References

1. Lu, J.M.: Discussion in Measure Adjectives. Language Teaching and Linguistic Studies 3, 46–59 (1989) (in Chinese)
2. Li, Y.M.: The Study in the Quantity Category in Chines. Huazhong Normal University Press, Wuhan (2000) (in Chinese)
3. Huang, G.Y., Shi, Y.Z.: The Phenomena of Markness and Unmarkness in Adjectives in Chinese. Zhong Guo Yu Wen 6, 401–409 (1993) (in Chinese)
4. Sun, H.L.: The Distribution Imbalance of Quantifiable Adjectives. Journal of Qiqihar University (Philosophy and Social Science Edition) 9, 68–69 (2000) (in Chinese)
5. Cruse, A.: Lexical Semantics. Cambridge University Press, London (1986)
6. Zhang, G.X.: The Functional and Cognitive Study on Adjectives in Modern Chinese. The Commercial Press, Beijing (2006) (in Chinese)
7. Dong, X.M.: 'Noun + Adjectives' Evaluation Phrase. Chinese Teaching in the World 73, 76–82 (2005) (in Chinese)

The Systematic Characters of Synonymous Paradigm in Chinese

Dan Hu[1,2] and Hongping Hu[1]

[1] School of Foreign Language, Zhongnan University of Economics and Law, Wuhan, 430073
[2] Center for the Study of Language and Information, Wuhan University, Wuhan, 430072
kean@whu.edu.cn, hhp0728@hotmail.com

Abstract. If we want to improve the accuracy of semantic computing, the micro differences of meaning between synonyms cannot be neglected. At present, for the available lexical semantic resources represented by WordNet, much research on the macro semantic relations between concepts have been done, but the description of the micro differences of meaning between synonyms is not enough. This paper reveals four systematic characters of synonymous paradigm in Chinese, according to which the corresponding variables can be designed to make a systematic formalized description of micro differences of meaning between synonyms.

Keywords: synonym, synonymous paradigm, systematic characters of semantics, micro description of semantics.

1 Introduction

The semantic description of synonyms has been receiving much concern with the rapid development of natural language processing in recent years. However, due to the limitations of the theoretical system and the processing technologies of the lexical semantics of natural language in traditional linguistics, this problem has not been solved satisfactorily.

From the late 1980s, WordNet, a study which primarily aims to the human's psycholinguistic principle of memorizing vocabulary has been carried out widely as a knowledge engineering, and widely used in natural language processing. In the semantic knowledge representation system of WordNet, the synonymous paradigmatic relationship is quite remarkable. The basic structural units of WordNet are synsets, that is, sets of synonyms. Generally speaking, elements in the same synset constitute a synonymous paradigmatic relationship with each other, and a concept is represented by a synset of synonyms. [1] From a more theoretical point of view, each synset can express a lexicalized concept in human language. However, concepts are not represented by logical items in WordNet; instead, they are represented by a list of dictionary entries which can be used to express the concept. It reveals such a fact: it is synsets, but not dictionary entries nor any single ones of their meanings that is involved in the construction of most of the semantic relations in WordNet. Therefore,

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 653–662, 2013.

all the varied semantic relations discussed in WordNet are in fact the relations between synsets.

When defining the meaning of a concept node in WordNet, the common meaning of all the synonyms in a synset are described, while the differences between words in the synset are generally neglected. For example:

answer, reply, response *- a statement (either spoken or written) that is made in reply to a question or request or criticism or accusation.*

For the three words in the same synset (answer, reply, response), just a definition "*a statement (either spoken or written) that is made in reply to a question or request or criticism or accusation*" is used to describe their common meaning which is also the meaning of the concept node "reply", while the differences between the three ones are neglected. In fact, it is quite necessary for natural language processing to describe the differences between synonyms in detail so as to provide the machine with finer semantic knowledge for a better natural language understanding. Therefore, how to describe the commonness and deference between synonyms becomes a key linguistic issue during the construction of a lexical semantic knowledge base.

This paper hence aims to explore a theory and method of describing the micro semantic differences between synonyms by observing the systematic characters of synonymous paradigm in Chinese.

2 The Systematic Characters and Structures of Synonymous Paradigm

As a typical way of words clustering in human language, synonymy shows a strong systematic character. This systematicness exists not only among synsets, but also among the individual items within a synset.

The systematicness among synsets mainly exists in paradigmatic dimension, and generally manifests itself as various semantic and logic relations between the concepts expressed by these synsets, such as antonymous relation, hyponymy relation, *Is-a* relation, *Instance-of* relation and *Part-whole* relation. These are the typical semantic relations besides the synonymous relation described in different versions of WordNet in different languages which all originated from the English WordNet. They are also quite important theoretical contribution of WordNet. In WordNet, each synset forms an independent micro-semantic system, and all the synsets are linked into thousands of semantic chains with clear hierarchy by the hyponymy relation. All the chains are again interlaced with each other by more semantic relation (antonymous relation, *Part-whole* relation, etc.), thus to form a more complex system.

The semantic systematicness also exists inside a synset which is aggregated by the nuclear, a common or similar conceptual meaning of a group of synonyms. This semantic systematicness of distribution falls into four different structures, namely: the dichotomous contrast structure, the trisected symmetrical structure, the gradual quantitative-changing structure and the multiaxial discrete structure.

2.1 The Dichotomous Contrast Structure

When describing some certain concept, we often use a pair of contrast words to express the contrapositive features of the concept in a specific dimension, thus to form a dichotomous contrast structure.

The contrast of the individual property vs. collective property of a concept is a remarkable instance of this structure. This is a systematic difference shown by noun synsets, it is also named the contrast of specific property vs. general property or the one of specific property vs. abstract property. Some nouns can carry both collective and individual meaning, some others only carry collective meaning. For instance, the word "信 xin (letter)" can refer to both collective and individual meaning. We can say "一封信 yifeng xin (a letter)", "两封信 liangfeng xin (two letters)" and "很多信 henduo xin (many letters)". While "信件 xinjian (letters)" definitely refers to collective meaning. We can say "很多信件 henduo xinjian (many letters)" or "一堆信件 yidui xinjian (a pile of letters)" but not "一封信件 yifeng xinjian (one letters)" nor "两封信件 liangfeng xinjian (two letters)".

In modern Chinese, nouns in the "individual-collective" contrast structure often share a common feature: the individual noun does not merely refer to individual meaning specifically, but can also refer to collective meaning; while the collective one can only refer to collective meaning but never individual meaning. The synonyms in this structure just refer to the same concept, the only difference between them is contrast between individual meaning and collective meaning. More examples:{书 shu (book), 书籍 shuji (books)}, {河 he (river), 河流 heliu (rivers)}, {湖 hu (lake), 湖泊 hupo (lakes)},{纸 zhi (paper), 纸张 zhizhang (paper sheets)}, {布 bu (cloth), 布匹 bupi (piece goods)}, {船 chuan (ship), 船只 chuanzhi (vessels)}, {树 shu (tree), 树木 shumu (trees)}, {车 che (vehicle), 车辆 cheliang (vehicles)}, {马 ma (horse), 马匹 mapi (horses)}, {花 hua (flower), 花卉 huahui (flowers)}, etc.. In all these examples, the former ones in the synsets can refer to both collective and individual meaning, while the latter ones can only refer to collective meaning.

By observing the structure of these collective nouns, we find that their formation falls into two categories: one is composed of a nominal (morpheme) plus a quantifier (morpheme), such as "纸张 zhizhang (paper sheets)","布匹 bupi (piece goods)","船只 chuanzhi (vessels)",etc; the other is composed with two nominals (morphemes) which are in a synonymous relation, such as "书籍 shuji (books)", "河流 heliu (rivers)", "湖泊 hupo (lakes)", 树木 shumu (trees)",etc. There are an interesting phenomenon deserving much of our attention, that is, some collective nominal expression have both of these two structural forms, such as "书籍 shuji |书本 shuben (books)", "船只 chuanzhi |船舶 chuanbo (vessels)", etc.

Then for some concepts we have two pairs of individual - collective dichotomous contrast structures, and the two bi-tuples can be further riveted into a triple by the joint of the individual noun (which can also express collective meaning anyway) so as to form a larger synset. Such as: {书 shu (book), 书本 shuben (books)} and {书 shu (book), 书籍 shuji (books)} can be jointed into {书 shu (book), 书本 shuben (books), 书籍 shuji (books)}; {船 chuan (ship), 船只 chuanzhi (vessels)} and {船 chuan (ship), 船舶 chuanbo (vessels)} into {船 chuan (ship), 船只 chuanzhi (vessels), 船舶

chuanbo (vessels)},etc. These are in fact the extended forms of the individual - collective dichotomous contrast structure. Although there are three items in such a synset, it is still a dichotomous contrast structure semantically. The only difference is that the contrast does not exist between two single words, but between the individual word and the set of the other two collective ones, namely 书 shu (book) vs. {书本 shuben (books), 书籍 shuji (books)}, 船 chuan (ship) vs. {船只 chuanzhi (vessels), 船舶 chuanbo (vessels)},etc.

2.2 Trisected Symmetrical Structure

Words in a standard synset of trisected symmetrical structure form a triple, which respectively describes the positive, neutral and negative aspects of a same concept. The positive word and the negative one are built into a symmetrical geometric structure with the neutral one as the axis. This symmetry refers to the symmetry of semantics, but not of the structure or numbers of the words in the synset.

A synset in which the synonyms differ from each other in emotional color with a distribution of "commendatory-neutral-derogatory" mode is the typical representative of such a trisected symmetrical structure.

The emotional color is the semantic reflection of some certain human's emotion in language, which indicates our aesthetic, ethical or moral judgment of the world. Generally speaking, this judgment reflects the common sentiment of all mankind established according to some common human feelings for thousands of years, so emotional color is one part of lexical semantics which is social and stable. [2] The emotional color of synonyms is manifested as the difference between commendatory and derogatory orientation. The basic conceptual meaning of synonyms in this structure is the same, but the meaning of emotional color attached to them is different. It may be negative or positive or just neutral. The synonym with positive color is a commendatory word; the one with negative color is a derogatory word, while the one without any positive nor negative color is just a neutral word. This is a systematic difference existing in synsets of nouns, adjectives and verbs. Although whether these words should be regarded as synonyms still remains argument in theoretical linguistics, we advocate gathering them into synonymous set when building a computing-orientated network of words, because it agrees with the logical structure of the network of words, and can also reflect the semantic systematicness of the emotional color of synonyms.

In such a synset, the commendatory word and the derogatory word form a geometrically symmetrical structure with the neutral one as the axis. For examples, {鼓舞 guwu (inspire), 鼓动gudong (arouse), 煽动 shandong (agitate)}, {爱护 aihu (cherish), 保护 baohu (protection), 庇护 bihu (harbor)}, {团结 tuanjie (unite), 合作 hezuo (cooperate), 勾结 goujie (collude)}, {成果 chengguo(achievement), 结果 jieguo (result), 后果 houguo (consequence)}, {本领 benling (talent),才能 caineng (ability),伎俩 jiliang (trick)}, {谨慎 jinshen (prudent),小心 xiaoxin (careful),拘谨 jujin (overcautious)}, etc.

There are two variants of this structure, the one misses its axis of symmetry, and the other misses one of the symmetric items. For the former one, there are some synsets of a commendatory-neutral-derogatory structure in which the neutral words may

be missing, and for the later, either the commendatory words or the derogatory words missing.

The formation of this phenomenon is closely related to the concept itself expressed by the synonyms. People have various emotions and feelings: happiness, anger, sadness and joy; and human speech has various meaning and implication: respect, disdain, praise or depreciation. As a mirror image in language, words thus obtain emotional color for the reflection of people's rich and complex emotions. Some concepts themselves contain commendatory or derogatory connotation, which expresses some certain conventional value orientation of human. So it is quite difficult to express an opposite emotional color within the category of a concept which carries obvious emotional color. For example, some concepts carry a value orientation of "affirmation", "praise" or "love", so words for these concepts will inevitably bring such an emotional color, as "荣誉 rongyu (honor)", "威信 weixin (prestige)", "榜样 bangyang (exemplar)",etc. Within the category of these concepts, it is nearly impossible to find a word with derogatory meaning. While some other concepts carry a value orientation of "negation", "denouncing" or "hate" , so words for these concepts will correspondingly bring such a derogatory sense, as "懦弱 nuoruo (cowardice)", "夸耀 kuayao (boast)" , "企图 qitu (vainly attempt)",etc. It is also quite difficult to find a commendatory word within the category of these concepts. Neutral words may exist in the category of these concepts. They are only the weak expression for these concepts with strong value orientation, so as to apply these concepts to some context in which the emotional color needn't or can't be emphasized particularly. This is a supplementary semantic measure. Therefore, these words can only constitute synsets with their corresponding neutral words, with the missing of one of the symmetric items. Such as {荣誉 rongyu (honor), 名誉 mingyu (reputation)}, {威信 weixin (prestige), 威风 weifeng (power and prestige}, {榜样 bangyang (exemplar), 模范 mofan (model),样板 yangban (example)}, {懦弱 nuoruo (cowardly),软弱 ruanruo (weak)}, {夸耀 kuayao (boast), 夸张 kuazhang (exaggerate)}, {企图 qitu (vainly attempt),妄图 wangtu (vainly attempt),试图 shitu (try), 意图 yitu (intend)}, etc.

Since the emotional color of a word is the expression of people's subjective feeling, the word is inevitably subjective in a certain degree. Words expressing a same concept may be given different emotional color just because they are used for different objects. This emotional color is often given from the speaker's subjective anger. It means that in the speaker's point of view some objects should be complimented or depreciated, so different words, commendatory or derogatory, may be chosen for them. Some concepts are used to highlight the remarkable features of its referent in a particular aspect, these concepts themselves carry a strong semantic orientation, so there is no neutral words for them. Such as {果断 guoduan (decisive), 武断 wuduan (arbitrary)}, {顽强 wanqiang (tenacious), 顽固 wangu (stubborn)}, {赞扬 zanyang (praise), 阿谀 eyu (adulate)}, {聪明 congming (smart), 狡猾 jiaohua (sly)}, etc.

Such symmetrical structure lacks a formal axis of symmetry just because of the missing of a neutral expression, but in fact, this axis does exists semantically, it is the conceptual meaning of the synset. In other words, the axis of such symmetric structure is actually a virtual one, it is the missing neutral word which essentially should be there. This is a zero category in language. While the first variant with one of the symmetric

items missing is an unbalanced structure, which is in fact asymmetric. The missing symmetric item cannot be invented with a zero category. From a philosophical point of view, the imbalance of symmetry is a special form of symmetry. Both symmetric and asymmetric structure are existing in thousands in the world, so in this sense, the special unbalanced symmetric structure in language is consistent with the general rules of structure of the world. In fact, one of the important features of human language is its asymmetry, which may occur universally in any level of language, including pronunciation, word formation, syntax, semantics and pragmatic aspects. [3]

2.3 Gradual Quantitative Changing Structure

Within the category of a same concept, there may be a systematic quantitative change for some particular aspect of the object. This change is continuous which are mainly manifested as gradual semantic change in language, such as the semantic scope, the stress of semantic tone, the degree, etc. Synonyms expressing such a concept form a semantic continuum. This is the gradual quantitative changing structure of the distribution of synonyms. Here are some examples of the subclasses for this structure.

The Semantic Scope. In a synset, some synonyms may refer to a wider scope of something, and some others narrower. This is a slight difference shown by noun synonymy.[4] For example, in the synset {事情 shiqing (thing), 事件 shijian (incident), 事故 shigu (accident)},"事情 shiqing (thing)" refers to any object, feature, or event, its semantic scope is the widest; "事件 shijian (incident)" refers to something that happens, often something that is unpleasant, its semantic scope is narrower; "事故 shigu (accident)" refers event in which someone is hurt or killed, or something is damaged or destroyed, as a result of bad luck or carelessness specifically to the unfortunate things happen by chance, the range is smaller. In another example, in the synset {房屋 fangwu (building), 房子 fangzi (house),屋子 wuzi (room)},"房屋 fangwu (building)" refers to an architecture that has a roof and walls for human habitation, it has a widest semantic scope; "房子 fangzi (house)" generally refers to a separate building which includes several rooms, its semantic scope is narrower; "屋子 wuzi (room)" refers to a part of a building enclosed by walls or partitions, and with a floor and ceiling, its semantic is the narrowest. So the two synsets are of gradual quantitative changing structure. More examples are:{生活 shenghuo (life), 生计 shengji (livelihood), 生涯 shengya (career)}, {行为 xingwei (behavior), 行动 xingdong (action), 举动 judong (manner}, {灾难 zainan (disaster), 灾害 zaihai (calamity), 灾荒 zaihuang (famine)}, {时代 shidai (era), 时期 shiqi (period), 时候 shihou (moment),etc. A synset of this structure may contain more or less synonyms as its elements, and most of them contains only two, such as {新闻 xinwen (news), 消息 xiaoxi (message)}, {物资 wuzi (materials), 物品 wupin (goods)}, {故乡 guxiang (native land), 家乡 jiaxiang (hometown)}, {实验 shiyan (experiment), 试验 shiyan (test)},etc. Although these synsets contain two words, their semantics are not dichotomous contrast but in a continuous gradual systematic structure. Because though the semantic scope they referring to varies, the boundary between them is not absolutely clear, but fuzzy and gradual.

The Stress of the Semantic Tone. It refers to the semantic difference in degree. This feature is mainly manifested in synsets of adjectives and verbs, and more in verbs. Anyway we can also find very few noun synsets of this structure. For example, in {损坏 sunhuai (damaged), 毁坏 huihuai (destroy)} and {强调 qiangdiao (emphasize), 夸张 kuazhang (exaggerate)},the semantic tone of the later is much stronger than the former one. More examples are {违背 weibei (disobey), 背叛 beipan (betray) },{误解 wujie (misunderstand), 曲解 qujie (distort)}, {危害 weihai (harm), 迫害 pohai (persecute)}, {辩论 bianlun (debate), 争论 zhenglun (dispute)}, {轻视 qingshi (contempt), 鄙视 bishi (disdain)}, {请求 qingqiu (request), 恳求 kenqiu (pleading)}, {失望 shiwang (disappoint), 绝望 juewang (despair)}, {批评 piping (criticize), 批判 pipan (animadvert)}, {崇拜 chongbai (adore),崇敬 chongjing (revere)},{小气 xiaoqi (stingy), 吝啬 linse (miserly)}, etc. In modern Chinese, there are some monosyllabic verbs originally, and then with the development of the language, some disyllabic synonyms expressing the same concepts are produced. The difference between them is just the stress of the semantic tone .Such as {进 jin (enter), 进入 jinru (walk into)}, {加 jia (add), 增加 zengjia (add)}.The expression "进入教室 jinru jiaoshi (walk into the classroom)" and "增加一件衣服 zengjia yijian yifu (put up a cloth more)" have a stronger semantic tone than "进教室 jin jiaoshi (enter the classroom)" and "加件衣服 jia jian yifu (add a cloth)".Generally speaking, in these synsets, the disyllabic words have a stronger semantic tone than the monosyllabic ones. More examples are {买 mai (buy), 购买(purchase)}, {量 liang (measure), 测量 celiang (measure)}, {请 qing (ask), 邀请 yaoqing (invite)} (Qingci Gao,1985).

Most of these synsets contain only two synonyms, and few of them have more than two elements, as:{缺点 quedian (shortcoming),错误 cuowu (mistake),毛病 maobing(defect) }, {优良 youliang (good), 优秀 youxiu (excellent), 优异 youyi (outstanding)}, {秘密 mimi (secret), 机密 jimi (confidential), 绝密 juemi (top secret)}, {请求 qingqiu (request), 恳求 kenqiu (pleading), 哀求 aiqiu (implore)}, etc.

2.4 Multiaxial Discrete Structure

A concept has multiple attributes, and each concept is a multidimensional system composed with all the attributes. Among them the essence attribute of the concept locates at the core position, just like an atomic nucleus. It is the decisive attribute of the concept which differ it from the other similar or related concepts. In human language we can often use more than one word to express a same concept, each one of them definitely carries the essential attribute of the concept. At the same time, they are used to emphasize or highlight different aspects of the concept in some certain dimension. The common semantic attribute of all the words in a synset is the essential attribute of the concept they express, while the aspects they emphasize or highlight in different dimension is the variation between them. Words in the synsets of the dichotomous contrast structure, the trisected symmetrical structure and the

gradual quantitative changing structure discussed above are synonyms differ from each quantitatively in a same dimension of a concept. There remains another systematic synonymous structure in language, in a synset of which each word reflect a different attribute of a concept in a different dimension at a different axis, thus to form a multiaxial discrete structure with the essential attribute of the concept as the core.

Difference of semantic focus of synonyms is the typical representative of this structure. In a synset, each synonym may focus on a different semantic aspect of a concept. The difference of semantic focus is a major difference of synonyms in modern Chinese.

Generally speaking, the semantic focus of the synonymous nouns is often manifested in the characteristics of things they refer to. For example, in the synset {才能 caineng (capacity), 才华 caihua (literary or artistic talent), 才智 caizhi (intelligence), 才干 caigan (competence)}, " all the words have the common meaning "ability". "才能 caineng (capacity)" focuses on someone's ability of doing something and his mastering of knowledge and skills; "才华 caihua (literary or artistic talent)" focuses on someone's talent on literature or art; "才智 caizhi (intelligence) " focuses on someone's ability of judgment, discrimination and invention; "才干 caigan (competence)" focus on someone's ability of work.

The semantic focus of the synonymous verbs is often manifested in the manner and way of action. For example, in the synset {收罗 shouluo (gather), 网罗 wangluo (net), 搜罗 souluo (collect)}, words have the common meaning "try to put people or things together". "收罗 shouluo (gather)" means to get together in a common way, without paying much effort; "网罗 wangluo (net)" means to gather something exhaustively, just like fishing with a net; "搜罗 souluo (collect)" means to gather something by seeking and searching, generally harder than "收罗 shouluo" and with a smaller scale than "网罗 wangluo".

The semantic focus of the synonymous adjectives is often manifested in the features and state of the object they refer to. For example, in the synset {陡峭 douqiao (sheer), 峻峭 junqiao (precipitous)}, words have the common meaning "the mountains are high and steep", while "陡峭 douqiao (sheer)" means that the angle of the slope is very sharp which is nearly vertical; "峻峭 junqiao (precipitous)" emphasizes that the mountain is high and dangerous.[5]

Semantic focus is an important means to show the slight difference of lexical meaning from the characteristics, the attributes and the state of an object or from the method, the manner, the direction and the result of an action. It is these differences that constitute the attributes of a concept in multiple dimensions, so to form a rich and colorful multi-dimensional image of the concept. Despite all apparent changes, the core attribute of a concept just remains essentially the same. All these semantic focuses are just attached to the essential attribute of additional attributes of the concept. This is the typical feature of synsets of multiaxial discrete structure.

3 Design of the Variables for the Description of the Systematic Semantic Features of Synonyms

The differences between the elements of a synset are often manifested in more than one characteristic dimension, and each dimension reflects respectively one attribute of a particular aspect of the concept. The attribute of each dimension can be represented as a variable, the type and the value range of the variable varies according to the paradigmatic semantic structure of the element in this dimension.

The four types of synonymous paradigmatic semantic structure discussed above are corresponding to three types of data.

The attribute variables for the dichotomous contrast structure have only two values: "Yes" or "No", so they belong to logic variables. For example, for the attribute "individual | collective", we can name the attribute variable as "is_individual" (means whether it is individual or not). For an individual word, the value "T" (means true logically) is assigned to the variable, while for a collective one the value is "F" (means false logically). Similarly, the "bounded | unbounded" and "definite | indefinite" attributes are also belong to logic variable, whose names may be defined as "is_bounded" and "is_definite".

The attribute variable for the trisected symmetrical structure can be defined as an integer variable, whose value range is among the three decimal integers {-1, 0, 1},"0" for the axis of symmetry,"-1" and "1" for the two symmetric items respectively. For example, for the attribute variable expressing emotional color of a concept, whose name can be defined as "orien_emo", we can assign the value "0" to the neutral word, "1" to the commendatory word and "-1" to the derogatory word.

The attribute variable for the gradual quantitative changing structure is also an integer variable, whose value range is a positive integer greater than or equal to "1".The variable value of this attribute for the word at the starting point of the gradual axis, which is the attribute benchmark word, is assigned as "1". And the value of the other words in the same synset can be assigned among {2, 3, 4, 5...} according to the degree of their semantic scope or the stress of their semantic tone.

For the multiaxial discrete structure, the value range of the attribute variable is an irregular discrete system, so it should be defined as a string variable which is described with a text directly. The value of this variable is a small text fragment unit in natural language, which can be a word or a short phrase. For the convenient machine processing, we stipulate that the value can not be a sentences or an ambiguous expression.

4 Conclusion and Future Work

The systematic characters of synonyms exist both in syntagmatic and paradigmatic aspect. In paradigmatic axis this systematicness is manifested as the distribution of the conceptual meaning of the synset and the various additional meaning of each word. While in syntagmatic axis it is manifested as the distribution of the collocating words and the collocating ways. Here our discussion focuses only on the systematic character of synonymous paradigm and the design of variables for its formulated description. For the next step we will make a further study on the methods and the steps of the description, and the systematic character of the synonymous syntagmatic distribution is also in our near future work plan.

Acknowledgments. This work is Sponsored by the Research Foundation for Humanities and Social Science of MOE, P.R.C.(Grant No. 09YJC740060) and the Major Projects of National Social Science Foundation of China (Grant No. 11&ZD189).

References

1. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38, 39–41 (1995)
2. Yang, Z.L.: A Study on Dynamic Word Color (Dongtai Cicai Yanjiu). People's Publishing House of Shandong, Ji'nan (2003)
3. Shen, J.X.: A Book Review of Asymmetry and Markedness (Buduicheng He Biaojilun). Jiangxi Education Press, Nanchang (1999)
4. Gao, Q.C.: Synonymy and Antonymy (Tongyici He Fanyici). Shanghai Education Press, Shanghai (1985)
5. Xie, W.Q.: Synonyms (Tongyici). People's Publishing House of Hubei, Wuhan (1982)

The Fluid Food Feeding Verbs in *Jin Ping Mei*：喝 (he), 饮 (yin), 吃 (chi)

Wenhe Feng

College of Humanities, Henan Institute of Science and Technology, Xinxiang, 453003
wenhefeng@gmail.com

Abstract. This paper examines 喝 (he), 饮 (yin) and 吃 (chi) which perform as fluid food feeding verbs in *Jin Ping Mei*. The literatures show that 喝 (he) just begin to act as a fluid food feeding verb, but it can be followed by nouns which including abroad range of fluid foods. 饮 (yin) can only be followed by 酒 (jiu). 吃 (chi) can be followed by nouns which including abroad foods. Both 饮 (yin) and 吃 (chi)provide the substitute conditions for 喝 (he).

Keywords: Fluid food feeding verb, Object, Synonyms.

1 Introduction

In this paper, fluid foods refer to wine, tea, soup, porridge, water, etc. The feeding verbs are limited to action verbs, which cause food move from mouth to throat, and stomach. This paper examines the three main fluid food feeding verbs in *Jin Ping Mei*[1], 喝 (he, drink), 饮 (yin, drink) and 吃 (chi, eat/drink), which is always related to the fluid food, such as wine, tea, water, soup. The fluid foods usually act as object of the fluid food feeding verbs. The paper mainly investigates the application between these verbs and their objects. In the study, the objects of verbs include overt and covert form, i.e. objects on the object position (usually directly behind the verb) as the overt objects, contrarily objects not on the object position or omissions as covert objects. The context of the verb is distinguished from three language styles: dialogue, statement, title. In some cases, whether verbs and their objects are next to each other, i.e. whether other words being inserted between the combination of the verbs and their objects, is also investigated. What's more, the investigation of the three verbs in sequential are presented, and based on which a conclusion was given.

2 喝 (he, drink)

There are 53 occurrences of 喝 (he) in *Jin Ping Mei*, 50 of which are used in words such as 吆喝 (yao he, cry out), 喝令 (he ling, cry to adjure), 喝彩 (he cai, cheer). Only 3 occurrences of 喝 are used as fluid food feeding verbs. Consider below.

[1] All the language materials are from CCL. The edition of *Jin Ping Mei* is the Chongzhen edition.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 663–672, 2013.

2.1 酒 (jiu, wine)as the Object

There is only 1 token.

(1) 金莲 吩咐：“叫 你 姐夫 寻了 衣裳 来 这里 [喝] 瓯子 酒 去。”
Jinlian instruct let you brother-in-law look for clothe come here drink bottle wine go
Jinlian instructed: ‘Let your brother-in-law look for some clothes, and come here to drink a bottle of wine’

2.2 茶 (cha, tea) as the Object

There is only 1 token.

(2) 西门庆 道：“你 不 吃, [喝] 口 茶儿 罢。 我 使 迎春 前头 叫 个 小厮, 接 你 娘 去。”
Qing Ximen say you not eat, drink mouth tea Modal Particle I let Yingchun front call a servant, pick up your mother go
Qing Ximen said: ‘If you do not eat, drink a cup of tea, please. I let Yingchun go to the front, and order a servant to pick up your mother.’

2.3 汤 (tang, soup) as the Object

There is only 1 token.

(3) 西门庆 道：“我 心里 还 不 待 吃, 等 我 去 [喝] 些 汤 罢。
Qing Ximen say I heart in still not want eat, wait me go drink some soup Modal Particle
Qing Ximen said: ‘I don’t want to eat, wait me to drink some tea, please.’
In the three examples above, all objects of 喝 (he, drink) are in overt form and used in dialogue.

3 饮 (yin, drink)

There are 186 occurrences of 饮 (yin) in *Jin Ping Mei*,[2] 7 of which act as nouns with other character, such as 饮食 (yinshi, diet, and 5tokens), 饮馔 (yinzhuan, drink and food, 1token) and 馔饮 (zhuanyin, food and drink, 1token). 2 of which are in the causative usage (e.g. 饮马, yinma, cause horse to drink). Others (176) are all used as feeding verbs.[3]

[2] There is one token: ‘there are few family named Wu, only a short man who sells pastry, named Dalang Wu.’(This is retrieved by饮, it is the 65th) I think it’s wrong.

[3] The token below is not included. 檐滴露、竹风凉, 拚剧【饮】琳琅。
This is retrieved by 饮, it is the 148th, whose meaning is for research.

3.1 酒 (jiu, wine) as Object

There are 175 occurrences of 饮 (yin) take 酒 (jiu, wine) as the object. 酒 (jiu, wine) is always used as overt object, mostly (107 tokens) in the form of 饮酒 (yinjiu, drink wine), in which 饮 (yin)is directly followed by 酒 (jiu, wine). See (4)-(7).

(4) 西门庆 道：“他 又 不［饮酒］，不 消 邀 他 去。”

Qing Ximen say he yet not drink wine, not need call him go

Qing Ximen said: 'He does not drink wine, so it is not needed to ask him to go.'

(5) 轮到 月娘 跟前，月娘 道：“既 要 我 行令，照 依 牌谱 上［饮酒］：……”

turning to Yuening front, Yuening say if let me bid, also according to Chines Ci poem on drink wine

It turned to Yueniang, Yueniang said: 'If let me bid for drink, we drink wine according to the Chinese Ci poem.'

The examples above are used in dialogue, and the following are in statement.

(6) 吴月娘 留 他 同 众 堂客 在 后 厅［饮酒］，西门庆 往 人家 赴 席 不 在 家。

Yueniang Wu stay him with some guests at back living room drink wine, Qing Ximen go other home go banquet not at home

Yueniang Wu asked him to stay with some guests at the back living room (to drink some wine). Qing Ximen went to other's home for a banquet, so he was not at home.

(7) 席间 也 有 夏提刑、 张团练、 荆千户、 贺千户 一班 武官儿［饮酒］，鼓乐 迎接，搬演戏文。

the party during also there are Zhang sir, Zhang coach, Jing village head, He village head some military official drink wine beat drum welcome show operas.

During the party, there were some military officials to drink wine, they beat drum to welcome the guest, and showed some operas.

Sometimes 酒 (jiu, wine)is used as covert object of 饮 (yin,drink). See below.

(8) 王婆 道：“老身 得知 娘子 洪[饮]，且 请 开怀 吃 两 盏儿。”

granny Wang say I know lady drink heavily and please happily drink two cup

Granny Wang said: 'I know you can drink wine heavily, and please drink happily.

(9) 桌上 摆着 杯盘，妇人 拿 盏 酒 擎 在 手里，看着 武松 道：“叔叔 满［饮］此 杯。” 武松 接过 酒 去，一[饮]而尽。

on table set cup and plate, woman hold cup wine hold up at hand Look at Song Wu say brother completely drink this cup." Song Wu take wine go, drink all at once.

There are cups and plates on the table. Gasping a cup of wine, looking at Song Wu, the woman said: 'Brother, please drinks all cup of the wine.' Song Wu took the wine in hand, and drank it all at once.

The examples above are used in dialogue, and the below are in statement.

(10) 月娘 令 小玉 安放了 锺箸，合家欢[饮]。

Yueniang let Xiaoyu set cup and chopstick all the families drank happily.

Yueniang let Xiaoyu set some cups and chopsticks, then all the families drank happily.

(11) [饮] 够多时, 郑爱香儿 推 更衣 出去了, 独有 爱月儿 陪着 西门庆 吃酒。

(drink) a lot of time Aixianger Zheng apologize go to WC go out, only Aiyuer accompany Qing Ximen drink wine

Drinking for lots of time, Aixianger Zheng apologized for going to WC, and went out. Then only Aiyuer drank with Qing Ximen.

(12) 月娘 在 上房 摆酒, 郁大姐 供唱, 请 众 姐妹 欢[饮]了 一日 方散。

Yueniang at above room set wine, sister Yu sang, invite some sisters drank happily one day then dismissed.

Yueniang set wine at above room, sister Yu sang, the sisters drank happily for the whole day, and then went back.

It's worth noting that when 酒 (jiu,wine) acts as covert object, more lexicalized forms, such as 洪饮 (hongyin, drink heavily), 一饮而尽 (yiyinerjin, drink all at once), 欢饮 (huanyin, drink happily), tend to occur.

3.2 鸩药 (zhenyao, poison) as Object

There is only 1 token.

(13) 捉 奸情 郓哥 定计 [饮] 鸩药_武大 遭殃

catch amour Yunge devise stratagem drink poison Da Wu get burnt

To catch the amour, Yunge devised a stratagem, drinking poison, Da Wu got burnt.

In this case, 鸩药 (zhenyao, poison) is used as overt object. It is in the title of chapter 5, i.e. it is used in title. It's worth noting that in ancient Chinese, 鸩 (zhen) always refers to the poison within wine.

4 吃 (chi, eat/drink)

There is 845 occurrences of 吃 (chi, eat/drink) in *Jin Ping Mei*. In fact, 吃 (chi, eat/drink) is a feeding verb taking not only fluid foods, but also solid foods as the objects. The cases, 吃 (chi, eat/drink) taking object of fluid food can be listed as following.

4.1 酒 (jiu, wine)as Object

酒 (jiu,wine) is always used as overt object of 吃 (chi). There are 168 occurrences of 酒 (jiu,wine) follow to 吃 (chi)directly. See below.

(14) 金莲 便 问 来兴儿：“你 来 有 甚事？ 你 爹 今日 往 谁家 [吃酒] 去了？”

Jinlian then ask Laixinger you come have what thing your father today go whose home drink wine gone

Jinlian then asked Laixinger: 'What do you come for? Where did your farther go to have a drink?

(15) 金莲 道：“你 既 留人 [吃酒], 先 订下 菜儿 才 好。”

Jinlian say you if keep someone drink wine first set dish then well

Jinlian said: 'If you want to keep someone to drink, you had better order some dishes first.'

(16) 蕙莲 道：“爹 在房里 [吃酒], 小的 不敢 进去。等着 姐 屋里 取 茶叶, 剥 果仁儿 来。”

Huilian say father at home drink I not dare go wait sister room take tea strip nut come

Huilian said: 'My father is (drinking) at home, and I dare not to go into the room. So I'm waiting sister to take some tea from room, and come back to strip some nuts.'

The cases above are used in dialogue, and the following are used in statement.

(17) 两个 说笑了 一回, 不 [吃酒]了, 收拾了 家活, 归房 宿歇, 不在话下。

two people chatted and laughed for a while not drink wine cleaned up dinner set go to room sleep not to mention

They two chatted and laughed for a while, and did not drink any more, then cleaned up the dinner sets, and went to room to sleep. Need not to mention.

(18) 西门大姐 白日里 便 在 后边 和 月娘 众人 一处 [吃酒], 晚夕 归到 前边 厢房 中 歇。

Ximen sister daytime then at back with Yueniang other people together drink wine at night come front room in sleep

Then sister Ximen drank with Yueniang and other people together in the back room at daytime, and came back to the front room to sleep at night。

(19) 西门庆 一手 搂过 他 粉颈, 一递一口 和 他 [吃酒], 极 尽温存之态。

Qing Ximen one hand hugged her tender neck drink in turn with her drink wine great attentively

Qing Ximen hugged her tender neck with one hand at once, and drinks with her in turn very attentively.

酒 (jiu,wine) can follow 吃 (chi, drink) indirectly sometimes. Consider below.

(20) [吃] 得 酒浓 时…… 两个 丫鬟撤开 酒桌, 拽上 门 去了。

drink Auxiliary wine happily when two servant girl depart wine table close door go

When drinking happily, the two servant girls departed from the wine table, closed the door and went out.

(21) 次日, 西门庆 果然 治酒, 请过 花子虚 来, [吃]了 一日 酒。

the next day Qing Ximen as expected make wine invite Zixu Hua come drank one day wine

Just as expected, Qing Ximen made wine the next day, and invited Zixu Hua come to drink for a whole day.

In some cases, 酒 (jiu,wine)is covert object of 吃 (chi, drink). See (22)-(26).

(22) 以此 妇人 喜 他, 常 叫 他 入 房, 赏 酒 与 他 [吃]。

so woman like him often call him into room reward wine give him drink

So the woman liked him, and often asked him come into the room, and rewarded him some wine to drink.

(23) 那 王婆 陪着 吃了 几 杯 酒，[吃]的 脸 红红的，告辞 回家 去了。

granny Wang accompany drink several cup wine drink face red say goodbye return home go

Granny Wang drank several cups of wine with them. After her face turning red, she said goodbye and went home.

The cases above are used in statement, and the following are in dialogue.

(24) 妇人 笑 道：“干娘 来 得 正好，请 陪 俺 娘 且［吃］个 进门 盏 儿，到 明日 养 个 好 娃娃！”

woman smile say fosterer mother come DE well please accompany my mother for the moment drink one into the door cup till tomorrow give birth to one good baby

The woman smiled and said: ‘Fosterer mother, it is the time for you to come. Please drink a cup of wine with my mother, so that you can give birth to a good baby tomorrow.’

(25) 于是 拿 大 银锺 递与 李娇儿，说道：“二娘 好歹［吃］一杯儿。大娘，奴 不敢 奉 大 杯，只 奉 小杯儿 罢。”

then seize big silver cup give Jiaoer Li say aunt anyhow drink a cup. aunt I dare not give big cup only give small cup Modal Particle

The woman gave a big silver cup to Jiaoer Li, and then said: ‘Aunt, drink a cup of wine anyhow. Aunt, I dare not give you a big cup of wine, but only a small one.’

(26) 敬济 一壁 接 酒，一面 把 眼儿 斜溜 妇人，说：“五娘 请 尊便，等 儿子 慢慢［吃］！”

Jingji at one side receive wine at one side BA eye squint on woman say fifth mother please help yourself wait son/me slowly drink

While receiving the wine, Jingji squinted on the woman and said: ‘Aunt, please help yourself, let me drink slowly.’

4.2 茶 (cha, tea)as Object

茶 (cha, tea) is always overt tea, specifically it mostly directly combined behind 吃 (chi), with 31 tokens. See below.

(27) 那 婆子 笑 道：“官人，你 养的 外宅 东街上 住的，如何不 请 老身 去［吃茶］？”

the granny smile say officer you provide outside house on east street reside why not invite me go drink tea

The granny smiled and said: ‘Sir, the outside house you providing is on east street, why do not invite me to drink some tea?’

(28) 王婆 哈哈 笑 道：“我 又 不是 你 影射的，如何 陪 你［吃茶］？”

old woman Wang haw-haw smile say I also am not your image how accompany you drink tea

Granny Wang laughed and said: ‘I am not your image, how can I drink tea with you?’

(29) 须臾，摆下 茶，月娘 便 叫：“桂姐、 银姐，你 陪 他四个［吃茶］。”

For a moment set tea Yueniang then say Sister Gui Sister Yin accompany them four drink tea

For a moment, Yueniang set tea and then said: 'Sister Gui, Sister Yin, please accompany them four, and drink some tea.'

The cases above are used in the dialogue, and the following are in the statement.

(30) 吴舜臣 媳妇儿 郑三姐 轿子 也 先 来了，拜了 月娘 众人，都 坐着 [吃茶]。

Shunchen Wu wife third sister Zheng sedan chair also earlier come kowtow Yueniang people all sit drink tea

Shunchen Wu's wife and Sanjie Zheng also came earlier on sedan chair, then kowtow to Yueniang and others. Then all of them sat down and drank tea.)

(31) 他 娘子 让 进 众人 房中 去 宽衣服，就 放 桌儿 摆茶，请 众堂客 坐下 [吃茶]。

his wife let into people house go take off coat then set table lay out tea invite all guest sit down drink tea

His wife let all the people go into the house to take off their coats. Laying out the tea on the table, she invited all the guests sit down to drink some tea.

(32) 落后 潘金莲、李瓶儿 梳了头，抱着 孩子出来，都 到 上房，陪着[吃茶]。

later Jinlian Pan, PingerLi do up their hair, in arm child come out all come above room accompany drink tea

Later Jinlian Pan and PingerLi did up their hair, and went out with child in arms. All of them went to the above room, and drank some tea together.

Sometimes 吃 (chi, drink) can be followed by 茶 (cha,tea) indirectly. Consider below.

(33) 玉楼 道：“你 坐着 [吃]了 茶去。”

Yulou say you sit drink tea go

Yulou said: 'Sit down, and go to drink some tea.'

(34) 西门庆 [吃] 毕 茶，说道：“我 回去 罢，嫂子 仔细 门户。”

Qing Ximen drink over tea say I come back Modal Particle sister-in-law care door

After drinking tea, Qing Ximen said: 'I'll go back, please pay attention to the door.'

(35) 张四 见 说不动 妇人，到 吃 他 抢白了几句，好 无 色，[吃] 了 两清茶，起身 去 了。

Si Zhang see can't persuade woman instead incur she some censure very ashamed drink ASP two cup tea rise leave ASP

Si Zhang saw he can't persuade the woman. Instead, some censure incurred from her. He felt ashamed, drank some cup of tea, and rose to leave.

茶 (cha, tea) sometimes can also be used as covert object. See (36)-(37).

(36) 西门庆 叫道：“干娘，点 两杯 茶 来 我 [吃]。”

Qing Ximen call fosterer mother call two cup tea come I drink

Qing Ximen said: 'Foster mother, order two cups of tea for me to drink.'

(37) 只 见 大姐 走来，李瓶儿 让 他 坐，又 交 迎春：“拿茶与 你 大姑娘 [吃]。”

only see elder sister come Pinger Li let her sit also told Yingchun bring tea give your elder aunt drink

When seeing elder sister come, Pinger Li let her sit down, and told Yingchun: 'Bring some tea for your elder aunt to drink.'

The cases above are used in dialogue, and the following are in statement.

(38) 须臾，泡出 茶 来，桂卿、 桂姐 每人 递 了 一盏，陪着［吃］毕。

after a moment make tea come Guiqing Sister Gui everyone give ASP one cup accompany drink over

After a moment, some tea is made, both Guiqing and Sister Gui were gave a cup of tea. They accompanied with each other and drank the tea.

(39) 只 见 一个 小厮儿 拿出 一盏 福仁泡茶 来，西门庆［吃］了。

only see one CL servant bring one cup tea with furen come Qing Ximen drink SF

The only can be saw was a servant with a cup of furen tea, and Qing Ximen drank the tea.

4.3 '(X) 汤 (tang, soup) (Y)' as Object

In *Jin Ping Mei*, 汤 (tang, soup) was not used as the object of 吃 (chi, drink) independent. When 汤 (tang, soup) as object, it is always in the form of (*X*) 汤 (*Y*). Consider below.

(40) 不一时，吴大舅［吃］了 第二道 汤饭，走进 后边 来 见 月娘。

for a moment uncle Wu drink ASP second soup and meal go into back come see Yueniang

For a moment, uncle Wu drank the second soup and meal, and then went into the back room to see Yueniang.

(41) 来安儿 忙 走 向前，西门庆 分咐："到 后边 对 你春梅姐 说，有 梅汤~提 一壶 来 我［吃］。"

Laianer in hurry walk forward Qing Ximen instruct go backward to your Chunmen sister say have plum soup bring one teapot come I drink

Laianer walked forward in hurry. Qing Ximen instructed: 'Go backward, let your sister Chunmei bring me a teapot of plum soup for me to drink.'

(42) 王婆 道："大官人［吃］个 和合汤？" 西门庆 道："最好！ 干娘 放 甜 些。"

granny Wang say officer drink CL he-he-soup Qing Ximen say great fosterer mother make sweet some

Granny Wang said: 'Would you like to drink some he-he-soup?' Qing Ximen said: 'Great! Fosterer mother, please make it sweeter!'

It should be noted that in *Jin Ping Mei*, 吃 (chi), as feeding verb, whose object includes not only fluid foods, but also solid foods, such as porridge, rice, melon, fruit, vegetable, bird and meat. When different types of foods together using a single verb only, the verb must be 吃 (chi). See below.

(43) 只 落下 李铭 在 西 厢房，［吃］毕 酒饭。

only leave Ming Li at west room eat over wine and meal

In the west room, only Ming Li left, and finished the wine and meal.

(44) 春梅、 玉箫、兰香、 迎春 四 个，都 在堂客上边执 壶 斟 酒，就 立 在 大姐 桌头，同［吃］汤饭点心。

Chunmei, Yuxiao, Lanxiang, Yingchun four CL all on the living room hold pot pour wine right stay at elder sister side of table together eat soup, meal, and dim sum.

They four, Chunmei, Yuxiao, Lanxiang and Yingchun, all held pot and poured wine in the living room, right staying beside the table of elder sister. They ate soup, meal, and some cakes together.

(45) 西门庆 那日 没 往那去，月娘 分咐 玉箫：“房中 另 放 桌儿，打发 酒菜 你 爹［吃］。”

Qing Ximen that day do not go there, Yueniang instructed Yuxiao: in the room other set table make wine and dish your father eat

Qing Ximen did not go there that day, Yueniang instructed Yuxiao: ‘Set other table in the room, make some wine and dishes for your father to eat.’

The nature of 吃 (chi), taking-food-widely, in *Jin Ping Mei* is mostly inherited by modern Chinese.The distribution of the collection of fluid food feeding verbs and their objects in *Jin Ping Mei* are list in the following table.

Table 1. Fluid feeding verbs and their objects in *Jin Ping Mei*

proportion of the fluid food verb in the word form / object		喝 (he, drink) (3/50)	饮 (Yin, drink) (176/186)	吃 (chi, drink) (---/845)
overt object	in dialogue	*wine*、*tea*、*soup*	*wine*	*wine*、*tea*、*soup*
	in statement		*wine*	*wine*、*tea*、*soup*
	in title		*poison*	
covert object	in dialogue		*wine*	*wine*、*tea*、*soup*
	in statement		*wine*	*wine*、*tea*、*soup*
	in title			

5 Conclusions

Through the above analysis about the three main fluid food feeding verbs in *Jin Ping Mei*, the conclusions can be listed as following.

I. About 喝 (he, drink). 喝 (he, drink), which is widely used as a fluid food feeding verb in modern Chinese, just begin to use as a feeding verb in *Jin Ping Mei.* But it showed a wide range of applicability from the beginning. For example, the only 3 objects of 喝 (he, drink) in *Jin Ping Mei*, are *wine*, *tea*, and *soup* respectively. The three cases are all used in dialogue, which tell us that oral environment may be a hotbed of meaning differentiation or transfer.

II. About 饮 (yin, drink). 饮 (yin, drink), as fluid food feeding verb, whose object has the singularity, i.e. its object can only be wine. This is reflected not only in the high-frequency of direct linking between 饮 (yin, drink) and 酒 (wine) (107/176),

but also in that its covert object is unitary, i.e. its object can only be wine. This makes it possible to produce the word form, such as 洪饮 (Hong Yin), 欢饮 (Huan Yin), 一饮而尽 (Yi Yin Er Jin) whose objects undoubtedly being as wine.

III. About 吃 (chi, eat/drink). 吃 (chi, eat/drink), as fluid food verb, has broad applicability, its objects including *wine*, *tea*, *soup* and so on. So its covert objects are entirely dependent on context to determine. Compare with 饮 (yin, drink), it is more difficult to have the lexicalized form, such as 洪吃 (Hong Chi) and 欢吃 (Huan Chi). In addition, 吃 (chi, eat/drink) is actually a universal food intake verb, not just suitable to fluid foods. Generally 吃 (chi, eat/drink) has a pluripotent performance.

In summary, we can make the following speculation: The unitary object of 饮 (yin, drink), as a fluid food intake verb, make it difficult to bear too much function. On the contrary, the universal objects of 吃 (chi, eat/drink), make it bear too much function. In this case, it needs a new verb to bear the appropriate function of intake of fluid food. At the time, 喝 (he,drink) emerged as a new fluid food feeding verb. It is applicable to almost all the fluid food, but not any solid food, which makes it to be the dominant fluid food intake verb in future.

The synchronic study on synonyms in one specific book [1-2] will be advantageous for understanding the differences and similarities of synonyms.[4] Through the comparative study on one dimensional feature of the synonyms, in this paper the object of the verb, a clearer understanding of the personality performance of the word can be got, and the insufficient of sememe analysis [3-4] on word personality can be avoided. For example, although *wine, tea and water* share the common sememe, [+food] and [+fluid], they cannot be shared by the three fluid food verb, 喝 (he,drink), 饮 (yin,drink), and 吃 (chi,eat/drink). Without the study in this means, the specific differences and similarities of the synonyms can not be got. This kind of study is also beneficial to examine the historical transfer of word meaning.

Acknowledgments. This work has been supported by the National Natural Science Foundation of China (61273320, 61070243, 61173095 and61202193) ,the Humanities and social science projects of Henan Department of Education (2012-GH-080), as well as the Technology Research Starting Fund for High-level Talent of Henan Institute of Science and Technology.

References

1. Chi, C.H.: Synonyms in *Shiji*. Shanghai Ancient Books Publishing House, Shanghai (2002)
2. Zhao, X.Q.: Synonyms in *Han Feizi*. China Social Sciences Press, Beijing (2004)
3. Hu, D.: All concepts and Their Fomalized Description. China Social Sciences Press, Beijing (2011)
4. Shi, A.S.: Semantic Theory. The Commercial Press, Beijing (1993)

[4] Refers to Chi (2002:2), prefaced by Huang.

Three Directional Systems Involved in Verbs

Yuelong Wang[1] and Aiping Tu[2]

[1] Department of Chinese Studies, FASS, National University of Singapore, Singapore
yuelongwang@nus.edu.sg
[2] Centre for Study of Language and Information, Wuhan University, Wuhan, China
tuaiping81@163.com

Abstract. There are at least three directional systems involved in Chinese verbs. The first one is the objective direction system for the agent in the sentence. This system takes concrete substance as the reference point. The second is the subjective direction system for the speaker of the sentence, which takes the position the speaker stood as the reference point. The third one is the cognitive direction system taking the body container of agent as the reference point. Three systems are combined together to express complicated directional concepts. This paper analyzes the difference of three systems and their integrative expression, and proposes that the difference in cognitive direction can be used as entry-splitting standard for Chinese verbs.

Keywords: directional system, direction attribute, verbs, cognitive direction, entry split.

1 Introduction

It is widely recognized that verbs, especially action verbs, have directional features. Many scholars, such as Cui [1], Liu [2], Qiu [3] and Wang [4], have all studied verbs from the perspective of direction. However, most of their research is only descriptions about how the objective directions are embodied in Mandarin Chinese. They classified verbs only according to static directions, without considering word combination. Therefore, it is hard to illustrate selectional restrictions for words representing directions in the sentence. Moreover, their definitions about 'direction' are not very clear for confused different directional systems. Therefore, it is necessary to develop a more reasonable system for directions classification.

2 Clarification of Concept

Before embarking on a study of direction, it is necessary to clarify for some concepts concerned.

First, '*方向 Fang Xiang*' and '*方位 Fang Wei*' are two terms often used in related research, but there is no clear-cut distinction between them. According to Xing [5], '*方位 Fang Wei*' and '*处所 Chu Suo*' constitute the '*方所Fang Suo*' category.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 673–684, 2013.

Strictly speaking, the confusing use of '*方向 Fang Xiang*' and '*方位 Fang Wei*' is questionable. According to the definition of dictionary, '*方位 Fang Wei*' should consist of *direction* and *location*, and is by no means an equivalent concept for location. Therefore, we propose to use terms of '*方向 Fang Xiang*' and '*位置 Wei Zhi*' instead of '*方位 Fang Wei*' and '*处所 Chu Suo*'.

Direction and location are two closely-related but totally different concepts. Action always occurs in some location, but not always accompanied by objective direction. Therefore, it is unreasonable to classify any one of them according to the other one. For example, Liu [6] divided directional verbs into three subclasses according to their relation with start position and end position, pointing to start, pointing to end and no pointing. However, there are still many directional verbs, such as*开 kai* 'open' and*起 qi* 'up', which can express direction pointing without reference point.

Second, we should clearly distinguish 'direction' from 'movement'. Movement means location change of object in space. Therefore, it certainly has direction, but direction does not necessarily represent by movement. Some actions without location change can still have directions.

There are two methods to investigate direction inside of verbs. One way is to research how objective directions are encoded in language, such as the separation of normal verb and movement verb according to location change. The other way is to study what kind of direction is embodied in the field of verb. Some verbs can still have direction features even though they have no actual movement.

Third, we must clearly distinguish the term 'direction representation' from the term 'direction attribute'. Action always happened in some space. Therefore, verb is closely related to direction representation, but the representation of direction is not restricted only in verbs. Location word, preposition, directional verb and nouns are all possible options. The representation of direction often combines with the expression of location, which needs to be carefully distinguished. Nevertheless, Direction attribute is psychological direction, which belongs only to notional verbs.

3 Three Directional Systems Involved in Verbs

Direction representation in Mandarin Chinese is a complicated system. In our opinion, there are at least three distinct directional systems. The first one is the objective system for the agent in the sentence, which takes concrete substance as the reference point. The second is the subjective direction system for the speaker of the sentence. This system takes the position speaker stood as the reference point. The third one is the cognitive direction system taking the body container of the agent as the reference point. The performance of the three systems is different but often intermingled.

3.1 The First and Second Directional Systems

It is widely agreed that the most commonly used words representing directions are location words and directional verbs. They usually are regarded as one word, which

actually is the convergence of two directional systems. According to the difference of referring object, we can classify directional verbs into two categories.

Objective direction for the agent: 上 shang 'up', 下 xia 'down', 进 jin 'in', 出 chu 'out', 回 hui 'back', 过 guo 'across', 起 qi 'up', 开 kai 'open'

Subjective direction for the speaker: 来 lai 'come', 去 qu 'go'

Objective direction can be expressed by verb, noun, or directional verb. Not only actual displacement directions, but also abstract directions are involved. Some verbs without displacement, such as what Lu [7] mentioned *切 qie* 'cut', *炒 chao* 'fry', *煮 zhu* 'boil', *沏 qi* 'steep', *泡 pao* 'soak', *包 bao* 'wrap', *割 ge* 'mow', *剥 bao* 'peel', *剪 jian* 'trim', can have abstract direction features. Consider (1).

(1)a. 切下一块肉
Qie xia yi kuai rou.
Slice off one CL meat
'A piece of meat have been sliced off.'
b. 沏上一壶茶
Qi shang yi hu cha.
Brew on one CL tea.
'A pot of tea have been brewed.'
c. 剪下一段绳子
Jian xia yi duan shengzi.
Cut off one CL rope
'A rope have been cut.'

Here the directional verbs *上 shang* 'up', *下 xia* 'down' are the expression for abstract directions.

Subjective direction for the speaker has only two categories. One is getting closer to the speaker, the other one is getting away from the speaker. Subjective direction system takes the place the speaker stood as the referring point, which is represented by the directional verbs *来 lai* 'come' and *去 qu* 'go', and not necessarily has actual displacement.

Moreover, there is a third direction system for the agent, which is a psychological direction taking the body container as the referring point. This system only encoded in notional verbs and is often represented with the aid of directional verb that belongs to the first and second direction systems. Although it is similar to the first system for the agent, we should not confuse these two different systems. Action can have no objective displacement direction, but must have psychological direction.

In this section, we will mainly dedicate to the first and second systems. The following examples in (2)-(4) can be used to illustrate the difference.

(2)a. 孩子伸出手来。
Haizi shen chu shou lai.
Child stretched out hand come
'The child stretched out his hand.'
b. 孩子伸出手去。
Haizi shen chu shou qu.

Child stretched out hand go
'The child stretched out his hand.'
(3)a. 孩子缩回手来。
Haizi suo hui shou lai.
Child withdrew back hand come
'The child withdrew his hand.'
b. 孩子缩回手去。
Haizi suo hui shou qu.
Child withdrew back hand go
'The child withdrew his hand.'
(4)a. 孩子伸进来一只手。
Haizi shen jin lai yi zhi shou.
Child stretched income one CL hand
'The child stretched in a hand.'
b. 孩子伸出来一只手。
Haizi shen chu lai yi zhi shou.
Child stretched out come one CL hand
'The child stretched out a hand.'

出 Chu 'out', *回 hui* 'back' , *进 jin* 'in' and *来 lai* 'come' and *去 qu* 'go' used here are all the expressions for the directions of verbs but for different person. 出 *Chu* 'out', 回 *hui* 'back', 进 *jin* 'in' is for the agent *孩子hai zi* 'child'. What they expressed is the objective direction. Meanwhile, the direction *来lai* 'come' and *去qu* 'go' represented is subjective directions for the speaker, which is the subjective judgment of the speaker. Two direction systems encoded simultaneously in one sentence to express complicated direction concepts.

Sometimes, we can employ only one directional system. See (5) and (6).

(5)a. 孩子伸来一只手。
Haizi shen lai yi zhi shou.
Child stretched come one CL hand
'The child stretched out a hand.'
b. 孩子伸去一只手。
Haizi shen qu yi zhi shou.
Child stretched go one CL hand
'The child held out his hand.'
(6)a. 孩子伸进一只手。
Haizi shen jin yi zhi shou.
Child stretched in one CL hand
'The child stretched in a hand.'
b. 孩子伸出一只手。
Haizi shen chu yi zhi shou.
Child stretched out come one CL hand
'The child stretched out a hand.'

As mentioned above, *来 lai* 'come', *去 qu* 'go' and *进 jin* 'in', *出 chu* 'out' used here are different in the referring point. However, in the sentence using '来、去', the speaker chose the same position the agent stood. Therefore, Only the direction for the speaker was used and no actual direction expression. Meanwhile, in the sentence using *进 jin* 'in', *出 chu* 'out', only the objective direction expression was used, and the speaker did not express his own judgment.

The usage of *来 lai* 'come', *去 qu* 'go' in Mandarin Chinese is unique. Ma [8] named them *Subjective Categories* but still regards the disyllable directional verb as a unit. In our opinion, there are two kind of usage for *来 lai* 'come', *去 qu* 'go', notional verb and directional verb. *来 lai* 'come', *去 qu* 'go' acting as notional verb is distinct from *来 lai* 'come', *去 qu* 'go' acting as directional verb in the referring point and for different person. Moreover, they are different in pronunciation. *来 lai* 'come', *去 qu* 'go' for the speaker can be intoned, but not for the agent. As Shen [9] states, 'They are put at the end of the sentence to express the direction of the sentence. They are intoned and close to modal particles. It is reasonable to regard them as modal particles. They have direction meaning, but they are not directional complement. Lu [7] classified the directional complement behind the verb into three categories. The second group selected is just *来 lai* 'come', *去 qu* 'go' as we described here. Lu [7] described the difference of their representation in detail, but did not explain. In our opinion, the reason for their distinction of representation is that they are expressing different directions and for different referring person.

Liu [10] also regards *来 lai* 'come', *去 qu* 'go', *过来 guolai* 'come over', *过去 guoqu* 'go over', *住 zhu* 'hold / cease' as form words. It is necessary to clearly distinguish the word difference of two direction systems. *来 lai* 'come', *去 qu* 'go' for the speaker is more similar to what Liu labeled *Form Word*, which has only syntactic meaning, and cannot be the part of the sentence and be intoned.

Directional verb *来 lai* 'come', *去 qu* 'go' can also be used to express direction for the agent. That is because the speaker chose the same position the agent stood. Their pronunciation can be emphasized or not, which is related to rhythm, pragmatics and semantics.

The difference of two directional systems can affect the usage of the prepositions. For example, *往 wang* 'towards' and *朝 chao* 'towards' can be used to embody the difference of direction for the speaker or the agent.

(7)a. 汽车朝我开过来。 (√)
Qiche chao wo kai guolai.
Car towards I drive come over
'The car is coming towards me.'
b. 汽车朝我开过去。 (×)
Qiche chao wo kai guoqu.
Car towards I drive go over
'The car is going towards me.'
c. 汽车往我开过来。 (×)
Qiche wang wo kai guolai.
Car towards I drive come over
'The car is coming towards me.'

In these three sentences, only the first one can be used. Zhou [11] proposed that the reason for not using the third one is that the object of *朝 chao* 'towards' can be a noun representing person, and the object of *往 wang* 'towards' cannot be. However, we found an example in the CCL corpus as*她再也压抑不住自己，举步直往他奔去Ta zai ye ya yi bu zhu zi ji, ju bu zhi wang ta ben qu* 'She cannot hold herself any more, but ran towards him straightly'. Therefore, whether it is a noun representing person is not the reason for the distinction of two prepositions, but the direction system they expressed. The direction *往 wang* 'towards' used is for the speaker, and*朝 chao* 'towards' is for the agent. The reason for not saying*汽车往我开过来 Qi che wang wo kai guo lai* 'The car is coming towards me' is that the direction *往 wang* 'towards' represented is not compatible with the direction *来 lai* 'come' expressed. The reason for not saying*汽车朝我开过去 Qi che chao wo kai guo qu* 'The car is going towards me' is that the referring point *我 wo* 'I' is not compatible with *去 qu* 'go'.

Directional verb*上 shang* 'up', *下 xia* 'down' for the agent can also represent the difference of social relation. In Chinese mandarin, we prefer using*上 shang* 'up' to show respect. Therefore, in the action the subordinate did to the superior, we always use*上 shang* 'up' and *下 xia* 'down' for the opposite. See below.

(8)a. *交上来 Jiao shang lai* 'turn over'
b. *交上去 Jiao shang qu* 'turn over'
c. *发下来 Fa xia lai* 'deliver to'
d. *发下去 Fa xia qu* 'deliver to'

The distinction between the superior and the subordinate is accepted by all the people in the society, and is not changeable for individual opinion. Therefore, it is an objective concept and abstract usage. Ma[8] is wrong to regard *上 shang* 'up' and *下 xia* 'down' as subjective categories.

3.2 The Third Directional System

Even though the directional verb 出 *chu* 'out', 回 *hui* 'back' , 进 *jin* 'in' represented the difference of objective direction of verb *伸 shen* 'stretch', *缩 suo* 'withdraw', there are only two direction in psychological direction system, inward and outward, encoded for the verb 伸 *shen* 'stretch', 缩 *suo* 'withdraw'. Psychological directions take the body container as the reference point. We name this directional feature of verbs and their dimensions as direction attribute, which is restricted on notional verbs.

There is no correspondent between the representation for objective directions and the expression for directional attribute. The division of directional and unidirectional verb before is based on objective directions. However, all the notional verbs have directional attribute.

The evidence for some scholars classifying verbs into inward verb and outward verb is what we named directional attributes here, such as Li [12] mentioned 'built-in

directionalities'. However, if we investigate the co-occurrence of directional verb with verb according to this classification, we will confuse two different systems. Meanwhile, the psychological direction of verb can be different in different entries. Therefore, it is more reasonable to describe verb inside of each entry.

Although both the first and the third directional system are for the agent, the expression for the first system do not certainly take the position the agent stood as reference point. However, the psychological directions, inward and outward, only take the body container as the referring point.

3.3 Three Dimensions of Directional Attribute

Some verbs have obvious directional attribute expression in each entry. They can be use alone. However, some other verbs can have both inward and outward direction in different entries. Moreover, some verbs can have both inward and outward direction in the same entry. Therefore, their expressions for the direction need the aid of directional verbs or other verbs with overt psychological direction expression.

Inward and outward are psychological directions taking the body container as the reference point. It can be expressed by different words in the sentence. Moreover, there is mutual dimension for the verb representing actions with mutual outward referring for the participators. Verbs with mutual dimension are collective verbs.

Notional verbs with overt direction attribute can be classified into three categories.

First, verb with outward dimension.

卖 mài ❶ 动 拿东西换钱（跟"买"相对）：～房子｜把余粮～给国家。

卖Mài 'sell' ① v. change money with something [correspond with *买 mǎi* 'buy']: *~房子 fang zi* 'sell house' /*把余粮~给国家ba yu liang ~ gei guo jia* 'sell the surplus foods to country'

Other entries of verb*卖 mài* 'sell' have also outward dimension.

❷ 动 为了自己的利益出卖祖国或亲友：～国｜把朋友给～了。
❸ 动 尽量用出来，不吝惜：～劲儿｜～力气。
❹ 动 故意表现在外面，让人看见：～功｜～弄｜～俏。

② v. betray country or relatives and friends for one's own benefits: *~国 ~guo* 'betray country' /*把朋友给~了 ba peng you gei ~ le* 'betrayed one's friends'

③ v. try one's best, far from stingy: *~劲儿 ~jin er* 'gathering one's strength' / *~力气 ~li qi* 'gathering one's strength'

④ v. let something on the outside, and to be seen: *~功 ~gong* 'show-off one's contribution' / *~弄 ~nong* 'show-off' / *~俏 ~qiao* 'show-off one's beauty'

Second, verb with 'inward' dimension.

买 mǎi ❶ 动 拿钱换东西（跟“卖”相对）：～票｜～布｜卖出粮食，～进化肥。

买 Mǎi ‘buy’ ① v. change something with money [correspond with *卖 mài* ‘sell’]: *~票 ~piao* ‘buy ticket’ / *~布 ~bu* ‘buy cloth’/*卖出粮食，~进化肥 mài chu liang shi, ~ jin hua fei* ‘sell the food staffs, and buy some fertilizer’

Some verb can separately have inward and outward dimension in different entries. See below.

租 zū ❶ 动 租用：～房｜～了一辆汽车。

❷ 动 出租：这个书店开展～书业务。

租Zū ① v. *租用 zuyong* ‘hire’: *～房 ~fang* ‘hire a house’ / *～了一辆车 ~ le yi liang che* ‘hired a car’

② v. *租出 zuchu* ‘rent’: *这个书店开展～书业务 zhe ge shu dian kai zhan ~ shu ye wu* ‘This bookstore carries out the business of renting.’

These verbs need the aid of directional verb or other verb with overt direction attribute to designate the directional attribute, such as *借 jie* ‘borrow / lend’ and *租 zu* ‘hire / rent’. *借 jie* ‘borrow / lend’ can designate the outward dimension of psychological direction with the help of verb*给g ei* ‘give’ which has overt ‘outward’ direction attribute. *借 jie* ‘borrow / lend’ can also designate the inward dimension with the help of verb*到 dao* ‘gain’ which has overt inward direction attribute. *租 zu* ‘hire / rent’ can designate the directional dimension with the help of *出 chu* ‘out’ or *入 ru* ‘in’. The usage of*给 gei* ‘give’, *到 dao* ‘gain’, *出 chu* ‘out’ and*入 ru* ‘in’ acts as not only the expression for objective directions, but also the designation for the psychological directions. In a sentence, if the aiding direction indicators do not appear, ambiguities will occur.

(9) 我借了李明五十块钱。
Wo jie le Li Ming wushi kuai qian.
I borrow / lend SF Ming Li 50 yuan money
‘I borrowed / lent 50 yuan from / to Ming Li.’
S1. 我借给李明五十块钱。
Wo jie gei Li Ming wushi kuai qian.
I lend give Li Ming 50 yuan money
‘I lent 50 yuan to Ming Li.’
S2. 我从李明处借了五十块钱。
Wo cong Li Ming chu jie le wushi kuai qian.
I from Li Ming location borrow SF 50 yuan money
‘I borrowed 50 yuan from Li Ming.’

(10) 我今天租了一套房子。
Wo jintian zu le yi tao fang zi.
I today hire / rent SF one CL house
‘I hired / rent a house today.’
S1. 我今天租出了一套房子。

Wo jintian zu chu le yi tao fang zi.
I today rent out SF one CL house
'I rent a house today.'

S2. 我今天租下了一套房子。
Wo jintian zu xia le yi tao fang zi.
I today hire down SF one CL house
'I hired a house today.'

Thirdly, verb with mutual attribute.

Verb and it entries with mutual dimension in the directional attribute can be named collective verb. Yuan [13] labeled them as 'cooperate verb', while Tao [14] labeled them as 'mutual verb'.

Mutual direction means the action is completed by all the participators. For each participator, the direction dimension is clearly outward. In syntactic representation, two or more nouns linked by conjunction, plurality and noun introduced by preposition can be used. Consider below.

(11) a. 李明和王刚正在吵架。 (Linked by conjunction)
Li Ming he Wang Gang zhengzai chaojia.
Li Ming and Wang Gang in the process of quarrel
'Li Ming and Wang Gang are quarreling.'

b. 他们吵了很久才达成一致。 (Plurality)
Tamen chao le hen jiu cai dacheng yizhi.
They quarrel SF very long time then reach an agreement
'They quarreled a long time before the agreement.'

c. 李明正在和王刚吵架。 (Introduced by preposition)
Li Ming zhengzai he Wang Gang chaojia.
Li Ming in the process of with Wang Gang quarrel
'Li Ming is quarreling with Wang Gang.'

3.4 Dimension Change and Entry Split

The distinction in directional dimension is closely related to entry split of verb. In view of the blur of principles for word entry split, we propose to regard the dimension difference in psychological attribute as an important principle for entry split. That is easier to manipulate. Tao [14] described the form and features of mutual verbs in detail. These verbs can have different directional dimension in different entries. In addition, verb can evolve new direction dimension with the meaning change in usage. Therefore, it is necessary to add new entry for this verb. Generally, collective verb with mutual dimension need no direct object. If the verb that usually does not need object appears in the sentence with object, the direction dimension used might be different for splitting new entry. We can still use the verb '吵' for illustration. The definition of '吵' in the dictionary is as follows.

吵 chǎo ❶ 形 声音大而杂乱：～得慌 | 临街的房子太～
❷ 动 吵扰：～人 | 把孩子～醒了
❸ 动 争吵：两人说着说着～了起来 | 不要～，有话好好说。

吵 Chǎo ① adj. *声音大而杂乱 Sheng yin da er za luan* ‘Loud and disorderly’: *~得慌 ~ de huang* ‘very noisy’ */临街的房子太~ lin jie de fang zi tai ~* ‘the house next to street is too noisy’

② v. *吵扰 chaorao* ‘bother’: *~人 ~ren* ‘bothering’ */把孩子吵醒了 ba hai zi chao xing le* ‘to wake the baby’

③ v. *争吵 zhengchao* ‘quarrel’: *两人说着说着吵了起来 liang ren shuo zhe shuo zhe ~ le qi lai* ‘the talk between the two men ended with a falling out’*/不要吵，有话好好说 bu yao chao, you hua hao hao shuo* ‘Do not quarrel, keep cool’

As a collective verb, *吵 chao* ‘quarrel’ must need two or more participators. Therefore, the directional dimension of *吵 chao* ‘quarrel’ is mutual. However, in the sentence *我吵了他一顿 wo chao le ta yi dun* ‘I have criticized him a lot’, the verb *吵 chao* ‘criticize’ used distinct outward dimension. The search result for *我吵了他一顿 wo chao le ta yi dun* ‘I have criticized him a lot’ by Google is 46600 that mean a common phenomenon in language usage. In view of the new object added and the direction distinction, we can predict that new entry should be split for verb *吵 chao* ‘criticize’. Therefore, it is necessary to add a new entry for *吵 chao* ‘criticize’ in the dictionary as follows.

动 责骂，批评；
❹ v. *责骂，批评zema, piping* ‘criticize’

This entry is similar to the second entry of verb 吵 *chao* ‘criticize’ in the outward dimension but with different meaning. Therefore, we split a new entry. More similar examples can be found, such as *谈 tan* ‘talk’, *聊 liao* ‘chat’, *睡 shui* ‘sleep’. *谈 tan* ‘talk’, *聊 liao* ‘chat’, *睡 shui* ‘sleep’ are all collective verbs. However, they had clear outward pointing in directional attribute in the phrases of *谈对象 tan dui xiang* ‘let someone fall in love’, *聊网友 liao wang you* ‘chat with net friend’ and *睡女人 shui nv ren* ‘sleep with a woman’. Although currently there is no split in the dictionary, we can still feel the obvious difference.

Meanwhile, some verbs with apparent direction pointing can change their direction dimension to express collective actions as an intransitive verb. Still take the verb *打 da* ‘hit’ as example. *打 da* ‘hit’ means knocking or striking with hands or other things. Therefore, *打 da* ‘hit’ has clear outward direction dimension pointing to the object. In the sentence *他们俩打起来了 ta men lia da qi lai le* ‘they are fighting’, *打 da* ‘fight’ should have mutual dimension for the fighting meaning. The definition of *打 da* in the *Contemporary Mandarin Chinese Dictionary* is questionable for combing the ‘attack ‘and ‘fight’ meaning as one entry. Moreover, 打*da ‘reap’* can have inward dimension in some entries expressing acquisition meaning, such as *打水 da shui* ‘fetch water’ and *打鱼 da yu* ‘fishing’.

In the corresponding English expression, we may realize the difference of directional dimension more clearly. If we translate the sentence *他们俩打起来了 ta men lia da qi lai le* into English, the best translation should be ‘They are fighting’. The verb ‘fight’ has mutual dimension. Moreover, as the translation for *我打了他一顿*

wo da le ta yi dun, the verb should have outward dimension, the best translation should be 'I beat him'.

Some verb with single directional dimension will split out new direction with the actual usage. Therefore, it is necessary to add new entry. Consider below.

承包 接受工程、订货或其他生产经营活动并且负责完成。

承包 Chengbao 'undertake': to undertake the engineering , ordering , or other production and operation activity under contract

In this meaning, the verb*承包 cheng bao* 'undertake' has inward dimension. Along with the frequent usage of*承包出去 cheng bao chu qu* 'let something be undertaken', *承包 cheng bao* 'let something be undertaken' has apparent outward dimension. In view of the difference of direction dimension, it is necessary to establish a new entry to express this distinction. We can describe as following.

动 把工程、货物或其他生产经营活动交给别人并且负责完成。

v. 'let something be undertaken': 'let the engineering , ordering , or other production and operation activity be undertaken under contract

If a verb with single inward or outward dimension is used to express the same meaning but opposite direction, while there is no word replaceable, marks can be used to show the direction change. These marks can be*被 bei* 'a word used to bring in the actor or agent in a passive sentence' and*把 ba* 'a word used to bring in theme patient in a active sentence' in Chinese mandarin.

We can still use verb*打 da* 'hit' to illustrate the usage. *打 da* 'hit' has outward dimension in the entry meaning 'strike', and the expression frame often used for this meaning is 'agent + verb + recipient'. For example*妈妈打了小胖一下 ma ma da le xiao pang yi xia* 'Mama hit at Xiao pang'. If the positions of subject*妈妈ma ma* 'mama' and the object*小胖 xiao pang* 'Xiaopang' were exchanged, the same meaning cannot be expressed. If we still want to make the position of*小胖 xiao pang* 'Xiaopang' in front of*妈妈 ma ma* 'mama', marks for direction change were needed in the sentence. The correct expression should be*小胖被妈妈打了一下 xiao pang bei ma ma da le yi xia* 'Xiaopang was hit by his mother'. Another example is*吃 chi* 'perish'. Verb*吃 chi* 'eat' has outward dimension in the entry meaning 'eliminate'. If the opposite direction was needed, we can use the mark for direction change as*敌人的三个团被我军吃掉了 diren de san ge tuan bei wo jun chi diao le* 'Three regiments of the enemy were perished by our military'.

The expression for directional attribute is related to the symmetrical and asymmetrical phenomenon in language. If two words had the same meaning but were opposite in direction, these two words must be antonyms, such as*买 mǎi* 'buy' and*卖 mài* 'sell'. If there is no proper word for representing the same meaning but opposite direction, mark for direction change is needed. Symmetrical and asymmetrical phenomenon is worth further investigation.

4 Conclusion

Three directional systems involved in verbs are integrated as a complicated whole to express elaborate and accurate meaning. Distinguishing these three different systems will be meaningful for many applications.

First, it is helpful for dictionary compilation. The difference in directional attribute of verb can be used as an important principle for entry split. Second, it is useful for syntactic parsing. The distinction in directional attribute can be used to improve efficiency for syntactic parser based on feature selecting. In addition, it will also help the work of teaching Chinese as a foreign language. As we all know, directional system learning is a hard task for foreign students. Distinguishing three directional systems will greatly improve the efficiency of learning Mandarin Chinese for foreign students.

References

1. Cui, X.L.: The Understanding and Recognize of Language. Peking Language and Culture Press, Peking (2001)
2. Liu, N.S.: How to Expression Space Order in Chinese. Journal of Chinese Language 3, 169–179 (1994)
3. Qiu, G.J.: The Direction System of Verbs in Modern Chinese. Journal of Chinese Language 9, 59–70 (1999)
4. Wang, Y.: A Study on the Directivity of Verbs and the Directional Verbs Teaching. Peking Language and Culture University Press, Peking (2011)
5. Xing, F.X., Li, X.N., Chu, Z.X.: Times and Directions. In: Ma, Q.Z. (ed.) A Introduction to Grammar Research. The Commercial Press, Peking (1999)
6. Liu, Y.H.: The Grammatical Semantics of Directional Complement. In: The Research and Quest of Grammar 4, pp. 74–88. Peking University Press, Peking (1988)
7. Lu, J.M.: The Problem on the Position of Directional Complement, and Object after Verbs. Chinese Teaching in the World 1, 5–17 (2002)
8. Ma, Q.Z.: 'V lai/qu' and the Subjective Categories of Verbs in Modern Chinese. Language and Culture Researching 3, 16–22 (1997)
9. Shen, J.: To Observe the Relation between Phonetic and Grammatical Research from the Voiceless Phenomenon. In: Ma, Q.Z. (ed.) A Introduction to Grammar Research, The Commercial Press, Peking (1999)
10. Liu, S.X.: On the Research of Chinese Grammatical Categories. In: Ma, Q.Z. (ed.) A Introduction to Grammar Research, The Commercial Press, Peking (1999)
11. Zhou, X.B.: The Grammatical Attribute of Preposition and the Systematical Methods to Preposition Researching. Journal of Sun Yatsen University (Social Science Edition) 3, 109–115 (1997)
12. Li, Y.Z., Lu, J.M. (translated): The Order of Chinese Semantical Units. Foreign Linguistics 3, 33–39 (1983)
13. Yuan, Y.L.: On Quasi-bidirectional Verbs. Studies in Language and Linguistics 1, 12–25 (1989)
14. Tao, H.Y.: Mutual Verbs and the Sentences with Mutual Verbs. Sentence Patterns and Verbs, pp. 344–382. Language and Culture Press, Peking (1987)

A Semantic Study of Mandarin *Cai* as a Focus Adverb in Simple Sentences

Lei Zhang and Peppina Po-lun Lee

Department of Chinese, Translation and Linguistics, City University of Hong Kong, Kowloon, Hong Kong
leizhang@student.cityu.edu.hk,
ctpllee@cityu.edu.hk

Abstract. This paper examines the use of *cai* as a focus adverb in simple sentences. *Cai* as a focus adverb is supported by the three properties it demonstrated, of which König [1] argues to be the distinctive properties of focus particles. As a focus adverb, *cai* demonstrates the unique property of being flexible in its direction of association, with its associated item possibly being an element either to its right or to its left, or even the entire sentence. Focal mapping will then be triggered, giving the tripartite partition of Operator [Background] [Focus], in which focus is mapped to the nuclear scope and the rest of the sentence, with focus replaced by a variable, to the background. Besides, it is its direction of association that determines the scalar or non-scalar use of *cai*, though context still plays a role. In its scalar use, the quantificational domain of *cai* will be determined by the relevant scale.

Keywords: Focus adverb *cai*, Scale, Tripartite structure, Focal mapping, Focus association.

1 Introduction

Previous analyses like [2-7] examine different uses of Chinese adverb *cai*, which can be summarized as follows: (a) temporal use; (b) parametric use; (c) limiting use; and (d) emphatic use.[1] Based on its various uses, *cai* is analyzed either as an exclusive particle or as a scalar particle. Moreover, some studies have pointed out that the semantics of *cai* is closely related to focus, expectation and scalarity. However, the semantic relation of *cai* with focus and scales remains an issue for further investigation.

In this paper, we argue that the three uses of *cai* claimed in previous analyses, namely (a) the parametric use, (b) the limiting use, and (c) the emphatic use[2], can be

[1] The classification of different uses of *cai* varies according to different studies, and we basically adopt Biq's classification, cf. [2-3].

[2] Here the case that *cai* indicates 'the degree is high' is excluded from the emphatic use, because in this case *cai* is subsumed under the adverb of degree.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 685–695, 2013.

naturally derived if one considers *cai* to be a focus adverb.[3] Relevant examples are given below.[4]

(1) 他在 图书馆 才 能 打印 文章。[5]
Ta zai tushuguan *cai* neng dayin wenzhang.
he in library CAI can print paper
'He can print papers only in the library.'

(2) 张三 才 看了三 篇 文章。
Zhangsan *cai* kan le san pian wenzhang.
Zhangsan CAI see ASP three CL paper
'Zhangsan only read three papers.'

(3) A: 一起 去 吃饭 吧。
Yiqi qu chifan ba.
together go have-meal SFP
'Let's go for dinner together.'
B: 我 才 不 去 呢！[6]
Wo *cai* bu qu ne!
I CAI NEG go SFP
'I will not go!'

In this paper, we do not attempt to cover all uses of *cai*, and will focus on cases where *cai* occurs in simple sentences and serves as a focus particle. We aim to achieve the following goals: first, to demonstrate that *cai* is a focus particle; second, to probe into the semantic contributions of *cai* by providing the tripartite structures it triggered; third, to examine the direction of *cai*'s association; fourth, to investigate the relation between *cai*'s direction of association and scalarity; and finally, to account for the incompatibility of *cai* and the sentence-final particle (abbreviated as SFP) *le*[7].

[3] We consider that, in the case that *cai* takes the pure short-time-span meaning, namely it only signals 'just now', it plays a role of the adverb of time. In the case that *cai* indicates 'immediate past' and 'later than expected/stipulated' simultaneously, it has the dual-function: both a temporal adverb and a focus particle. For example, *Ta cai lai* 'He just came.' In the case that *cai* merely indicates that 'the degree is high', namely that it does not imply 'beyond the expectation of the speaker' at the same time, it serves as the adverb of degree. In the case that *cai* signals 'the degree is high' and 'beyond expectation' simultaneously, it has the dual-function: a degree adverb and a focus particle. For instance, *(Zuotian wo kan le yi chang bisai.) Na chang bisai cai jingcai ne!* 'Yesterday I watched a game. That game was very great!'Notice that, in the examples considered, *cai* has a single function or a dual-function is largely dependent on pragmatic factors such as the context.

[4] For easy illustration, here we suppose in (1) *cai* has the non-scalar use and interacts with *tushuguan* 'library'; in (2) it associates with the quantity *san* 'three'. In fact, for sentence (2), several elements are available to interact with *cai*. In addition, the position of focus will influence the interpretation of *cai*.

[5] Abbreviations used in this paper include – CL: classifiers; ASP: aspect markers; NEG: negative markers; and SFP: sentence final particles.

[6] In this case the particle *ne* generally co-occurs with *cai*. When *cai* is deleted, it is no need to use *ne*. Actually, usually the occurrence of the sentence final *ne* in sentences like *Wo bu qu ne!* 'I will not go!' is weird.

[7] To compare with the aspect marker *le* (called le_1) the SFP *le* is called le_2.

2 Literature Review

Previous studies such as [2-6, 8] have discussed various uses of the adverb *cai* and attempted to provide unified accounts for the semantics of *cai*.

Biq [2] points out that *cai* marks exclusive focusing and Biq [3] claims that the quantificational adverb *cai* is to mark denying-expectation focusing. Bai [4] argues that the discourse function of *cai* is to build up a relation between two units, which need not always have overt expressions. Lai [5-6] treats *cai* as a scalar particle and considers that its basic meaning is to reject expectation. Chen [8] holds that *cai* always relates to an element to its left and indicates 'later or more than expected'. In addition, Hole [7] is devoted to the parametric use type of *cai*[8] and demonstrates that in this case *cai* plays a role of negated existential quantification over alternatives entailed.

However, no consensus has been reached on the basic semantics of *cai*. Furthermore, the unified accounts proposed by previous studies, i.e. treating *cai* as a focus adverb or a scalar particle, cannot cover all the phenomena of *cai*. We hold that, in some cases where *cai* is claimed to be a focus adverb, it in fact demonstrates the use of a temporal adverb or a degree adverb as well. Moreover, *cai* does not always relate to a scale, either.

3 *Cai* as a Focus Adverb

In what follows, we will focus on the use of *cai* as a focus particle, and to be more specific, *cai* as an exclusive adverb.

In [1], König suggests that a focus particle has the following three properties: (a) a sentence with a focus particle entails the relevant sentence without the particle; (b) a focus particle has the quantificational force, which quantifies over the alternative set; and (c) a focus particle may include or exclude the alternative values that possibly satisfy the open sentence in which the focus is replaced by a variable. König proposes that based on property (c), focus particles are divided into two subclasses: exclusive/restrictive particles and inclusive/additive particles, which can be further tested by the entailment test and the quantification test, respectively.

To begin with, we will exemplify that being a focus adverb, *cai* can go through the entailment test and the quantification test given by König. A sentence with *cai* entails the corresponding sentence without *cai*, as revealed in (1), (2) and (3B) which entail (4), (5) and (6), respectively. It is thus obvious that *cai* passes the entailment test, in König's term.

(4) 他 在 图书馆 能 打印 文章。
Ta zai tushuguan neng dayin wenzhang.
he in library can print paper
'He can print papers in the library.'

[8] In the framework of [7], the parametric use type roughly refers to the case that *cai* interacts with a focus, which includes both the parametric *cai* and the limiting *cai* defined by [2-3].

(5) 张三 看 了 三 篇 文章。
Zhangsan kan le san pian wenzhang.
Zhangsan read ASP three CL paper
'Zhangsan read three papers.'

(6) 我 不 去。
Wo bu qu.
I NEG go
'I will not go.'

On the other hand, the asserted value as focus will introduce a set of alternatives. *Cai* operates on the alternative set and excludes the possibility that the alternatives satisfy the open sentence. *Cai* passes the quantification test, in König's term, as supported by (1), (2) and (3B). For (1), the asserted value *tushuguan* 'library' introduces an alternative set which includes possible places where he can print papers. With the semantics of *cai* these alternatives are eliminated. For (2), the asserted value *san* 'three' introduces a set of alternatives which are ordered on a quantity scale and located higher than the asserted value. *Cai* excludes the possibility that the alternative values like 'four' make 'Zhangsan read x papers' true. For (3B), the alternative proposition 'I will go to dinner with you all' is excluded.

Since *cai* can pass the two tests, it is not without grounds to consider it as an exclusive adverb.

4 The Basic Semantics of *Cai*

The exclusive adverb *cai* contributes a sentence in which it occurs mainly from the following two aspects: (a) A sentence with *cai* presupposes the corresponding sentence without *cai*; (b) A sentence with *cai* entails that none of the alternatives satisfy the open sentence under consideration. Consider (1) and (2) again.

For (1), this sentence presupposes that 'he can print papers in the library' and entails that 'he can print papers in the library rather than other places'. For (2), it presupposes 'Zhangsan read three papers' and entails 'Zhangsan did not read more than three papers'.

We assume that, the exclusive adverb *cai* will trigger a tripartite structure, which follows the rule of focal mapping, in the form of Operator [Background] [Focus]. Focus is mapped to the nuclear scope ([Focus]) and the rest of the sentence, with focus replaced by a variable, mapped to the restrictor ([Background]).[9] Consider (7), which gives the tripartite structures of *cai* in sentence (1).

(7) Focal mapping (*cai* associating with focus *tushuguan*)
Cai_s [ta zai s neng dayin wenzhang] [s=tushuguan]

[9] It should be noted that in the case where *cai* is related to a scale, the domain of quantification is restricted to the relevant scale.

5 The Direction of *Cai*'s Association

The focus adverb *cai* generally occurs before the predicate, as shown in (1), (2) and (3B). However, when *cai* associates with the subject, it occurs in a pre-subject position, and gives a limiting meaning to it, as shown in (8).

(8) 才 五 个 人 来 参观 艺术馆。
Cai wu ge ren lai canguan yishuguan.
CAI five CL person come visit art-museum
'Only five persons came to visit the art museum.'

It is observed that, the relative positions of *cai* with its associated element are various: *cai* can interact with either an element to its left as shown in (1), or an element to its right as shown in (2), or even the whole sentence in which *cai* is excluded as shown in (3).

When focus association does not occur, interpretation of *cai*-sentences depends on its direction of association. In the case of leftward association, the asserted value either determines the actuation of the relevant event or has the relevant property which is usually denoted by the predicate. In the case of rightward association, *cai* has the restrictive meaning, which implies that the asserted value is located in the lower position than stipulated or expected on the relevant scale. In the case of *cai* associating with the whole sentence, *cai* excludes the alternative proposition(s) introduced by the sentence in question.

Moreover, the direction of *cai*'s association will influence the interpretation of a *cai*-sentence. Look at (9).

(9) 小明 才 会 爬。
Xiao Ming *cai* hui pa.
Little Ming CAI can crawl

Image a scenario in which the speaker talks about the behavioral competence of a group of infants. Suppose *cai* associates with the verb *pa* 'crawl', *cai* is equivalent to another Chinese adverb *zhi* 'only' and has a scalar use. The verb *pa* will introduce a set of alternatives ordered according to the behavioral competence. On the scale the possible points are 'turn over', 'sit', 'crawl', 'toddle', 'run', etc.', and they represent different degree of competence of the infants. Here the sentence means 'Little Ming can only crawl and he can neither toddle nor run', with the assumption that both toddling and running are difficult to learn than crawling. On the other hand, there yet remains another possibility. Suppose the subject *Xiao Ming* is stressed, *cai* will interact with the focused subject and the sentence would have the interpretation of 'It is Little Ming but no other infants that has the capability of crawling'.

Here is one more example.

(10) 张三 六 点钟 才 到 中环。
Zhangsan liu dianzhong *cai* dao Zhonghuan.
Zhangsan six o'clock CAI come Central

Imagine a scenario in which Zhangsan and Lisi have planned to meet at some subway station along the Hong Kong Island Line. Assume they would meet at the Central station. Here *cai* associates with *liu dianzhong* 'six o'clock'. Since time points form a natural temporal scale, this implies that Zhangsan arrived at Central later than what is expected

or stipulated. There is still another possibility. Assume another slightly different scenario: Zhangsan and Lisi planned to meet at the station of Causeway Bay at six o'clock. Here *cai* is associated with *Zhonghuan* 'Central'. Under the assumption that stations on the subway line form a natural scale ordered according to spatial points, this indicates that the station that Zhangsan arrived at is still farther away from what is stipulated or expected on the subway line. According to the common knowledge of the Hong Kong Island Line, the excluded alternative stations include Causeway Bay and those stations which are closer to Causeway Bay than Central, namely Admiralty and Wan chai, and other stations which are farther away on the line and irrelevant under such a case.

6 The Relation of *Cai* with Scales

In the above section, we have already encountered sentences which demonstrate the scalar use of *cai*. In what follows, we will further examine the interaction of *cai* with scales and argue that *cai* can have both scalar and non-scalar use.

König illustrates that, depending on whether a scale is needed or not, focus particles can be divided into the following subgroups: (a) scalar particles like English *even*, which are always related to a ranking; (b) non-scalar particles like English *also*, which do not associate with a scale; and (c) particles which are only related to scales in certain cases, for example, English *only*.

It is controversial whether the adverb *cai* is always related to a scale or not. Some relevant studies are reviewed below. Biq [2-3] pointed out that in some cases *cai* distinctively relates to a scale. Bai [4] holds that in a construction without a subordinate clause marker, *cai* indicates a decreasing value, in which the starting point of the direction is the second unit which is posterior to the first unit on the scale considered. Lai [5-6] argues that *cai* is a scalar particle, which presupposes a change of state of the truth value of a proposition, and the asserted value of the change is located 'farther up' than expected on the scale under consideration. Hole [7] exemplifies that in the parametric use type *cai* does not always relate to a scale.

We claim that as an exclusive adverb, *cai* can have either a scalar use or a non-scalar use, which crucially depends on its direction of association and context.

As illustrated by previous analyses, in the case of leftward association, in many cases *cai* tends to interact with a scale. See below.

(11) 三 个 人 才 能 抬起那 张 写字台。
San ge ren *cai* neng tai qi na zhang xiezitai.
three CL person CAI can lift up that CL desk
'Only three persons can lift up that desk.'

(12) 很 聪明 的人 才 能 答对 这 道 题。
Hen congming de ren *cai* neng dadui zhe dao ti.
very smart DE person CAI can answer-right this CL question
'Only very smart persons can give the right answer to this question.'

(13) 他 十点 才 到 办公室。
Ta shi dian *cai* dao bangongshi.
he ten o'clock CAI come office
'He came to the office (as late as) at ten o'clock.'

For (11), *san ge ren* 'three persons' introduces a set of alternatives which are ordered according to the number of person(s) and located lower than the asserted value on the relevant scale. These alternatives are excluded due to the semantics of *cai*. Moreover, the values like 'four persons' in the higher position are irrelevant. For (12), assume that *cai* associates with *hen congming de ren* 'very smart persons', and an alternative set in which alternatives are ranked according to the degree of smartness of persons would be triggered. The values in the lower position such as 'a little smart' are excluded. For (13), the asserted value *shi dian* 'ten o'clock' introduces a set of alternatives which are ordered on a temporal scale and located lower than the asserted value. *Cai* excludes the possibility that alternative time points other than 'ten o'clock' would make 'he came to the office' true.

We suggest that, given proper contexts, *cai* in the case of leftward association is not related to a scale in some cases. Consider (14) and (15).

(14) 张三 才 是 组长。[10]
[Zhangsan]$_F$ *cai* shi zuzhang.
Zhangsan CAI be group-leader
'Zhangsan (but not someone else) is the group leader.'

(15) (抬 那 张 写字台，) 三 个 人 才 正 合适。[11]
(Tai na zhang xiezitai,) san ge ren *cai* zheng heshi.
carry that CL desk three CL person CAI exactly fit
'(To carry that desk,) only three persons are just enough.'

Assume (14) and (15) to be interpreted under the following scenarios. For (14), assume that the two speakers have the common knowledge that in this group there is only one group leader. The speaker uses (14) merely to reject another speaker's statement about who is the group leader, giving (14) an interpretation of 'Zhangsan rather than another person is the group leader'. Since only one value has the property of the predicate, it is difficult to say that there is an indirect quantity scale, which is ranked according to the number of persons that possibly satisfy the relevant property. Moreover, the context does not provide enough evidence to show that compared with another individual, e.g. Lisi, Zhangsan is more likely to be the group leader or is the more unexpected one. For (15), assume the scenario in which Zhangsan's friends are helping Zhangsan move house. They start with moving that desk. Lisi is very strong and he tries to move the desk by himself. But he finds that the desk is heavier than expected. Wangwu comes to help Lisi, but it is still a little heavy. Then Little Zhao joins in to carry the desk. They can carry the desk. Little Liu also offers to help carry the desk with the three persons. However, they have to pass through a narrow aisle, and four people plus that desk make the aisle too crowded. Therefore, it turns out that, that only three persons are just enough to lift up that desk and pass through the aisle.

The tripartite structures of *cai* in these two examples are given in (16) and (17).

(16) Focal mapping: Cai_x [x shi zuzhang] [x=Zhangsan]

[10] '[ ]$_F$' indicates elements which are in contrastive focus.

[11] Here *zheng* 'exactly' has the contribution 'neither more nor less than' due to its lexical meaning.

(17) Focal mapping: Cai_x [x zheng heshi] [x=san ge ren]

In the case of rightward association, the meaning given by *cai* corresponds to its limiting use defined by [2-3]. We agree with Biq that under such a case, *cai* requires a scale. However, unlike Biq, we do not consider that *cai* is always used to deny the expectation[12], as the basic meaning of *cai* is to mark that compared with the excluded alternatives, the asserted value is located lower on the relevant scale. This is exemplified in (18).

(18) 他 才 得 了 个“中”。
Ta *cai* de le ge ‘zhong’.
he CAI get ASP CL average
‘He only got a score of ‘average’.’

When *cai* operates on the whole sentence, it is more likely to give a non-scalar reading. For instance, *cai* in sentence (3B) is used to deny or exclude the affirmative proposition “I will go dinner with you all”. Since there are only two members in the set, namely the negative proposition “I will not go dinner with you all” and its affirmative counterpart “I will go dinner with you all”, it is difficult to say that the two propositions are ordered on a scale. Moreover, one may argue that what is involved is the scale of expectation, and sentence (3A) is taken as the expectation of Speaker A, but yet, sentence (3B), which is the assertion uttered by Speaker B, represents neither the expectation of Speaker A nor that of Speaker B.

From the above, we can conclude that in the scalar use of *cai*, the types of scale involved are diverse, which can be (indirect) quantity scales (cf. (2), (11)), degree scales (cf. (9), (12)), spatial scales (cf. (10), temporal scales (cf. (10), (13)), rating scales (cf. (18)) and so forth.

Unlike the non-scalar *cai*, the quantificational domain of the scalar *cai* is provided by and restricted to the scale which it adheres to. Note that since no focus association occurs here, the mapping involved is not focal mapping.

7 The Incompatibility of *Cai* with Le_2

Before concluding, one final issue to be discussed is the incompatibility of *cai* with le_2. In this section, we will account for such an incompatibility, which eventually further supports our analysis of *cai*.

As mentioned by [5-6] among others, *cai* cannot co-occur with le_2. This is illustrated in (19).

(19) a.*他 才 发表 了 三 篇 论文 了。
Ta *cai* fabiao le san pian lunwen *le*.
he CAI publish ASP three CL paper SFP
‘He has published three papers.’

[12] Let’s look at an example: *Ta cai yimiliu* ‘he is only one meter and sixty’. Image a situation in which *ta* ‘he’ refers to Zhangsan who is an adult and wants to be a security officer of Company A. The requirement on the height of a security officer proposed by Company A is that a security officer should be one meter and seventy or above. The speaker is familiar with both Zhangsan and the requirement. He utters this sentence to express that the height of Zhangsan does not attain the requirement. In this situation no expectation is needed.

b. *他 三 点钟 才 到 办公室 了。
Ta san dianzhong *cai* dao bangongshi *le*.
he three o'clock CAI come office SFP
'He came to the office at three o'clock.'

Lai [5-6] claims that, the incompatibility of *cai* with le_2 is due to the contradiction of the expectation between the adverb *cai* and le_2: The former indicates 'later than expected', whereas the later expresses 'earlier than expected'. Chen [8] posits a different analysis. He argues that *cai* signals the actual time that the event considered happened is later than what is expected, whereas le_2 indicates the actual time that the relevant event happened is earlier than the reference time. Moreover, Chen considers that in sentences like (19b) the reference time is the expected time. The grammatical meaning of *cai* conflicts with that of le_2, which renders the sentence under consideration ungrammatical.

However, these two analyses both have limitations. For Lai's analysis, in many cases, le_2 does not have an expectation-related interpretation, as revealed by examples given in (20) and (21), where neither of them relies on expectation for their interpretation. Therefore, taking 'contrary to expectation' to be part of the basic semantics of le_2 may render explanatory problems. On the other hand, unlike Chen's analysis, *cai* does not always imply 'later than expected', as shown in (1), whereas le_2 also allows the event in question happened at the reference time.

(20) A: 张三 呢?
Zhangsan ne?
Zhangsan SFP
'Where is Zhangsan?'
B: 他 去 买 东西 了。
Ta qu mai dongxi *le*.
he go buy something SFP
'He went out shopping.'

(21) 我 已经 知道 这 件 事 了。你 不 用 再 隐瞒 我 了。
Wo yijing zhidao zhe jian shi *le*. Ni bu yong zai yinman wo *le*.
I already know this CL thing SFP you NEG need again conceal me SFP
'I have already known this matter. You need not conceal it from me anymore.'

In order to account for the incompatibility of *cai* with le_2, first consider the semantics of le_2, which has long been an issue of great debate, cf. [9-14][13], etc. Basically we adopt the analysis proposed by Pan and Lee [12] that le_2 is an assertion operator and its basic meaning is to signal that the situation denoted by the relevant sentence is true

[13] Soh claims that le_2 indicates either 'change of state' or 'contrary to expectation' or both, and she describes the conditions under which each reading occurs. What is more, she provides a unified analysis of the two readings with the help of common ground and presupposition. Soh explains the restriction on the distribution of le_2 with downward-entailing quantifiers, and points out that the relevant analysis can be extended to cover the co-occurrence restriction of le_2 with *zhi* 'only'.

at or before the reference time. Moreover, we agree that interpretations like 'change of state' are derived from the core meaning of le_2.

We argue that sentences like (19) are ungrammatical for the following reason. The sentences in (19) both involve 'change of state', in which the old state changes to the new state, and usually the negative state changes to the positive one. However, due to the semantics of *cai*, the relevant change cannot happen in real world. Hence the sentences in question are unacceptable. For (19a)[14], le_2 operates on the relevant sentence and asserts that *Ta cai fabiao le san pian lunwen* 'He only published three papers' at or before the reference time. With the semantics of le_2, here involves a change from ¬*Ta cai fabiao le san pian lunwen* 'He did not only publish three papers' to *Ta cai fabiao le san pian lunwen* 'He only published three papers'. However, this kind of change cannot happen in real world. The reason is that generally speaking, as time goes on, the amount of papers that he published should increase instead of decreasing, and thus 'he did not only publish three papers' , whose meaning is equivalent to 'he published more than three papers', cannot change to 'he only published three papers'. In a similar fashion, sentence (19b) is ungrammatical, either. With the semantics of le_2, sentence (19b) asserts that he only came to the office at or before three o'clock. Assume that *cai* interacts with *san dianzhong* 'three o'clock', and there involves a change from ¬ *Ta san dianzhong cai dao bangongshi* 'He came to the office only at some time rather than three o'clock' to *Ta san dianzhong cai dao bangongshi* 'He came to the office only at three o'clock'. This change cannot occur in real world, which can be explained in the following way. When an event happened, the time that it happened was determined, and the aforementioned change simply cannot occur, which explains why the sentence is ungrammatical. On the other hand, assume that *cai* interacts with *bangongshi*, and there involves a change from ¬*Ta san dianzhong cai dao bangongshi* 'At three o'clock he did not only came to the office' to *Ta san dianzhong cai dao bangongshi* 'At three o'clock he only came to the office'. Imagine the scenario, in which he did things according to his plan of Monday morning: first, to withdraw some money in the bank; second, to go to get a book in the office; third, to go to return the book in the library; and last, to have lunch. In this situation, the negated state of the proposition *Ta san dianzhong cai dao bangongshi* means that he has done other thing(s) like returning the book at three o'clock. However, the positive state means that he did not do the things like returning the book at three o'clock, namely that his action falls behind what he planned. Once an event such as 'he returned the book' occurred, the change to the non-occurrence of the relevant result state cannot be true. Therefore, in this situation, (19b) is also unacceptable.

8 Conclusions

This paper has explored the use of *cai* in simple sentences, as a focus adverb. As an exclusive adverb, a sentence with *cai* presupposes the relevant sentence without *cai* and entails that none of the alternatives can fulfill the open sentence in question. The

[14] Here several elements are available to be associated with *cai*. For easy demonstration, suppose *cai* interacts with the quantity *san* 'three'.

direction of *cai*'s association is various: *cai* can associate with an element either to its left or to its right, or associate with the whole sentence. Moreover, being a focus adverb, when there is a focus in the sentence, *cai* would associate with the focus, with focal mapping triggered. Regarding the relation of *cai* with scale, *cai* has either scalar or non-scalar use, which is highly dependent on the direction of its association and context. The scalar *cai* is flexible in its scale selection, with its quantificational domain restricted to the scale to which *cai* adheres, while the quantificational domain of the non-scalar *cai* is crucially determined by context. Finally, we have explored the incompatibility of *cai* with le_2, and have argued that this is due to the semantic contradiction between the adverb *cai* and le_2.

Acknowledgments. Part of the results reported in this paper is supported by the RGC General Research Fund (GRF) CityU 146311 from the Hong Kong SAR government. The authors thus acknowledge the generous support of the relevant party. Sincere thanks also go to the anonymous reviewers for their invaluable comments. As usual, the authors alone are responsible for all potential errors that may exist in the paper.

References

1. König, E.: The Meaning of Focus Particles—a Comparative Perspective. Routledge, London (1991)
2. Biq, Y.O.: The Semantics and Pragmatics of *Cai* and *Jiu* in Mandarin Chinese. PhD dissertation. Cornell University (1984)
3. Biq, Y.O.: From Focus in Proposition to Focus in Speech Situation: *Cai* and *Jiu* in Mandarin Chinese. Journal of Chinese Linguistics 16, 72–108 (1988)
4. Bai, M.L.: Xiandai Hanyu *Cai* he *Jiu* de Yuyi Fenxi. Zhongguo Yuwen 5, 390–398 (1987)
5. Lai, H.L.: Rejected Expectations: The Scalar Particles *Cai* and *Jiu* in Mandarin Chinese. PhD Dissertation. The University of Texas at Austin (1995)
6. Lai, H.L.: Rejected Expectations: The Scalar Particles *Cai* and *Jiu* in Mandarin Chinese. Linguistics 37(4), 625–661 (1999)
7. Hole, D.P.: Focus and Background Marking in Mandarin Chinese—System and Theory behind *Cai*, *Jiu*, *Dou* and *Ye*. Routledge, Curzon (2004)
8. Chen, L.M.: On *Jiu* and *Cai*. Contemporary Linguistics 1, 16–34 (2005)
9. Chao, Y.R.: A Grammar of Spoken Chinese. University of California Press (1968)
10. Lü, S.X., et al. (eds.): Modern Chinese 800 Words. The Commercial Press (1980)
11. Li, C.N., Thompson, S.A.: Mandarin Chinese—A Functional Reference Grammar. University of California Press (1981)
12. Pan, H.H., Lee, Peppina P.L.: Mandarin Sentence Final-Le Is an Assertion Operator. Paper Presented at the 12th International Conference on Chinese Linguistics (2004)
13. Shi, D.X., Hu, J.H.: The Syntactic and Semantic Status of Sentence Final Particle *Le*. In: Yufa Yanjiu he Tansuo (13), pp. 94–112. The Commercial Press (2006)
14. Soh, H.L.: Speaker Presupposition and Mandarin Chinese Sentence-Final *–le*: A Unified Analysis of the 'Change of State' and the 'Contrary to Expectation'. Reading 27, 623–657 (2009)

Research of Contemporary Use of the Cultural Revolution Vocabulary

Tian-tian Zhang[1], Bin Li[1,2], and Liu Liu[1]

[1] Research Center of Language and Informatics,
Nanjing Normal University, 210097 Nanjing, China
dingningjin@hotmail.com,liuliu1989@gmail.com
[2] State Key Lab for Novel Software Technology,
Nanjing University, 210023 Nanjing, China
lib@nlp.nju.edu.cn

Abstract. Cultural Revolution Vocabulary is one of sociolinguistic research objects. This paper constructs the Cultural Revolution Corpus based on automatic word segmentation and part-of-speech tagging. Combining qualitative and quantitative analysis, we analyze the top 1,000 words having the max TF-IDF value and word highest frequencies during the Cultural Revolution. The top-100 words are thoroughly analyzed. In the framework of sociolinguistics, this paper explains the phenomena of words' disappearing, reuse and change of sense, and the color of words.

Keywords: the Cultural Revolution Vocabulary, the contemporary usage, dating terms, lexicology.

1 Introduction

The Cultural Revolution Vocabulary is one of the most important research objects of sociolinguistics. Many new words were created during the Cultural Revolution period (from year 1966-1976). Most of them disappeared before 1990s, but some of the disappeared words returned to use in 1990s.Therefore, a lot of researches discuss the dynamic development of the Cultural Revolution Vocabulary. But most of them explore the use of the Cultural Revolution Vocabulary during the Cultural Revolution and focus on the background and reference of some special words, which lead to the shortage of research discussed with the variation and development of the Cultural Revolution Vocabulary.

In recent years, scholars from mainland of China studying the Cultural Revolution Vocabulary tend to classify the words, such as the color words, extreme words, etc. However, this kind of research only focuses on a small scale of words. Corpus of the Cultural Revolution Vocabulary is not established and does not used during the research. So it is hard to find out the common features of the Cultural Revolution Vocabulary.

With the development of the Corpus linguistics, the constrained and chronological corpus can be used to study the use of the Cultural Revolution Vocabulary. We take

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 696–707, 2013.

the corpus of People's Daily (from year 1946-2000) for use which covers the Cultural Revolution period. Word segmentation and part-of-speech tagging are conducted under the ICTCLAS [1]. This paper mainly discusses the two questions: (1) the time distribution and the part-of-speech distribution of the Cultural Revolution Vocabulary. (2) Variation of contemporary use of the Cultural Revolution Vocabulary.

2 Related Work

Research of the Cultural Revolution Vocabulary is part of the researches of history of vocabulary by period. Influenced by studies of critical interpretation of ancient Chinese texts, the research methods of dating terms of our scholars always focus on explaining the vocabulary. Scholars described dating terms to study the situation of the whole words have appeared during the same decade, like [2]. Getting the help from explanation and description, the research of the Cultural Revolution Vocabulary is more and more plentiful.

Zhou and Li described the Cultural Revolution Vocabulary integrally [3]. They found the Cultural Revolution Vocabulary has four characteristics which are insulting, assaultive, mindless and non-standard. The description of the Cultural Revolution Vocabulary's characteristics is acute and exact.

Most of our scholars combined description and explanation to study different kinds of the Cultural Revolution Vocabulary. For example, from year 2006 to 2008, Diao studied traditional commendatory terms, loan words, popular words, extreme words and coined words appeared during the Cultural Revolution [4-8]. Using the combined method he could get characteristics of every kind of words but could not get generality of the whole Cultural Revolution Vocabulary.

From different angles, other scholars explained the generality of the Cultural Revolution Vocabulary but not popular. For instance, Jin studied the Cultural Revolution Vocabulary by theories of psycholinguistics [9].

There were little papers searching the Cultural Revolution Vocabulary by Corpus. But nowadays, with the development of corpus, we can study the Cultural Revolution Vocabulary objectively, entirely and deeply. Also we can find the characteristics of the contemporary use of the Cultural Revolution Vocabulary and try to explain the reason why the usage of them changing nowadays.

3 Data Sources

Data of this article comes from the database formed by automatic word segmentation and part-of-speech tagging of People's Daily (from year 1945-1999) which was done by ICTCLAS.

Based on about 800,000 words used by People's Daily during the Cultural Revolution (from year 1966-1976), ICTCLAS calculated the vocabulary usage frequency during the 54 years. According to the division different ages, a list of words used during 60 years of the Cultural Revolution was formed. Through this list, the nature of each word used in Cultural Revolution and their usage frequency in different 6 ages can be easily found out.

We used TF-IDF (Term Frequency, TF) algorithm to count word's frequency of the Cultural Revolution Vocabulary. Base on this algorithm we can do a chronological survey on the Cultural Revolution Vocabulary and analyze these words deeply.

4 Analyzing the Cultural Revolution Vocabulary

Before analyze the contemporary use of the Cultural Revolution Vocabulary, we arranged these words descending by their word's frequency in decades.

4.1 High-Frequency Cultural Revolution Vocabulary

After calculated the frequency of each Cultural Revolution Vocabulary, top-100 words which are with highest TF-IDF value and frequency during the Cultural Revolution were compiled. The contents of these top-100 Cultural Revolution Vocabulary are in the appendix in this paper.

Considering appearance times of these words, we found out words appeared before 1960s and noted them by brackets. Although these words with brackets were not new words, they were still used during the Cultural Revolution frequently.

4.2 Analyzing of the Top-100 Cultural Revolution Vocabulary

At first, we classified the first 100 Cultural Revolution Vocabulary by their parts of speech. After counting the number of each part-of-speech, we made a table including each word's part-of-speech and its corresponding number. Through this table, we can find the quantity of each part of speech.

Table 1. Words' Part of speech

Part of speech	Number	Part of speech	Number
Noun	30	Idiom	4
Name	16	Verb	4
Place name	14	None-verb	3
Time	9	Adjective	2
Conventional words	7	Aspectual expression	1
Adverbial verb	1		

Then we analyzed the contemporary use of these 100 words. We divided them according to the general variation of words.

Table 2. Diversification of word's changing

Words which are still used		Words which are no longer used
Meaning with changing	Meaning without changing	

Table 3. Meaning with changing of top-100 frequent Cultural Revolution Vocabulary

Rational meaning of the words		Extra meaning of the words	
Expand	2	Change	4
Transfer	1	Constant	1
Narrow	1		

The number of the top-100 Cultural Revolution Vocabulary which meaning without changing is 83. And the number of the top-100 Cultural Revolution Vocabulary which are no longer used is 11.

As may be noticed from these lists, we find out three obvious characteristics of the contemporary use of the Cultural Revolution Vocabulary:

1. **The number of the words used during Cultural Revolution which meanings do not change is majority:** The statistics showed the usings and meanings of most Cultural Revolution Vocabulary did not change.We realize the inheritance and development of the Cultural Revolution Vocabulary are stable. However, the words' frequencies of these words have changed.
   Frequencies of 83 words meaning without changing changed within time flying. We extracted 5 words from these 83 words randomly to prove this variation. These 5 words[1] are: 超级大国chao ji da guo(Noun),狠抓hen zhua(Verb),发展中国家fa zhan zhong guo jia(Conventional words),蹲点dun dian(Adverbial verb) and 意气风发yi qi feng fa(Idiom).
   Although firstly appearances of "超级大国chao ji da guo" and "发展中国家fa zhan zhong guo jia" were not during the Cultural Revolution, these two words were still used frequently during that time. In Fig.1., we can find out words' frequencies of "狠抓hen zhua" and "发展中国家fa zhan zhong guo jia" became higher after the Cultural Revolution. Meanwhile "超级大国chao ji da guo","蹲点dun dian" and "意气风发yi qi feng fa" were lower than before. The discrepancy of word's frequency shows diversity of the transformation of the Cultural Revolution Vocabulary.
2. **Extended meaning of words and rational meaning of words do not change at the same time:** There are only 5 words which meanings have changed in all 100 words. Among these 5 words, there are 4 words rational meanings have changed.

[1] Explanations in English of these words are in Appendix, the same hereinafter.

They are “牛鬼蛇神niu gui she shen”,“赤脚医生chi jiao yi sheng”,“支农zhi nong” and “妖风yao feng”.“造反派zao fan pai” is the only word which extended meaning has changed.

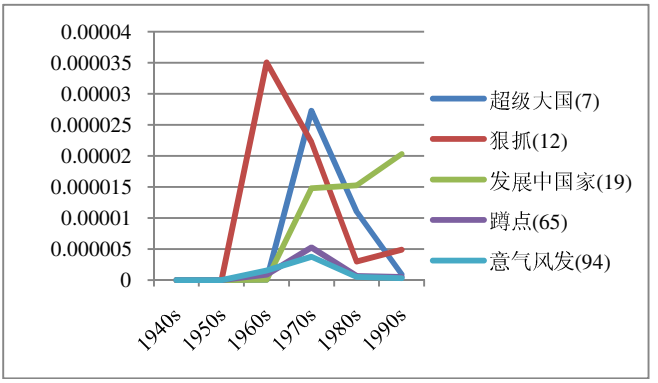


Fig. 1. Word’s frequency

3. **Some of the Cultural Revolution Vocabulary are no longer used:** In these 100 Cultural Revolution Vocabulary, 11 words are no longer used nowadays. They are “贫下中农pin xia zhong nong”,“革委会ge wei hui”,“走资派zou zi pai”,“红卫兵hong wei bing”,“四人帮si ren bang”,“造反派zao fan pai”,“红小兵hong xiao bing”,“三自一包san zi yi bao”,“多快好省duo kuai hao sheng”,“三支两军san zhi liang jun” and “讲用会jiang yong hui”. The main reason why these 11 words are no longer used is the word’s referring disappeared.

5 Reasons of the Transformation of Contemporary Use of the Cultural Revolution Vocabulary

Many factors influence the development and variation of language. Here we can analyze the Cultural Revolution Vocabulary from the angles of society which language relies on and of the system of language itself. And we can explore basic causes of the development of the Cultural Revolution Vocabulary and regular patterns of variation of these words.

5.1 System of Language

The developments and variations of languages depend on system of languages. Speech, lexicon and grammar are 3 main parts of a language. These 3 parts are interconnecting, interacting and balance among themselves in order to exert its communicative function in the best way. From this point of view, we can know that

the development and variation of words are influenced by other 2 parts of language. At the same time, vocabulary itself will change within years.

5.1.1 Development of Language Itself

Parts of the Cultural Revolution Vocabulary are still used these days. These words often appear in the books and movies which related to the Cultural Revolution. They can describe the Cultural Revolution vividly and convey these themes of the works. And people can choose several of the Cultural Revolution Vocabulary from the works and continue using them.

"Time will change everything; we have no reason to suppose that language is the exception" [10], like Ferdinand de Saussure said, language always changes and develops. Dynamic is one of the main characteristics of language. In the development and variation of a language, transformations of speech, lexicon and grammar are not poised. Among them the development and variation of lexis are the fastest.

Appearance of new words, disappearance of old words and reuse of old words can be thought as the most important performances of lexicon changes. These 3 phenomena can be considered as the adjustment of language itself. So the Cultural Revolution Vocabulary's appearance, disappearance and reuse can be seen as consequences of the development of language itself.

5.1.2 Development and Variation of Words

A word consists of an object that a word refers to or a concept needs to be explained, and the form which expresses the object or concept. The object that a word refers to or the concept that needs to be explained is the word's meaning and the word's form is its Chinese character. The word's gradual change has an intimate connection with the contents above.

A) The thing has disappeared

The reason why a lot of Cultural Revolution Vocabulary disappeared is things that words refer to disappeared. For example, "红卫兵hong wei bing" refers to "the counterrevolutionary group that Chief Qing Jiang and Biao Lin utilized a national and mass organization which mainly made up by university students and middle school students". Because this group has disappeared, its Chinese character "红卫兵hong wei bing" gradually disappeared and related word "红小兵hong xiao bing" also disappeared.

In the top-100 Cultural Revolution Vocabulary, the reason why "革委会ge wei hui","红卫兵hong wei bing","红小兵hong xiao bing","讲用会jiang yong hui","四人帮si ren bang" and "三支两军san zhi liang jun" have disappeared is the words refer to disappeared.

B) Conception is changing

Words formed during the Cultural Revolution usually have special meanings. For instance, "走资派zou zi pai" was an abbreviation of the powers supported the

capitalism and its color was derogative. Nowadays this word refers to the concept "the powers supported the capitalism" is considered as the capitalist leaders different from socialist leaders. Now this word's color is neuter. Since the transformation of the conception led conflict of this word's color, "走资派zou zi pai" is no longer used.

We can know that due to conceptual changes, some Cultural Revolution Vocabulary are no longer used. The reason why words "多快好省duo kuai hao sheng", "三自一包san zi yi bao", "贫下中农pin xia zhong nong" and "走资派zou zi pai" disappeared is this.

C) Replacement of Pattern

The objects or concepts that some Cultural Revolution Vocabulary refer to have been remaining after the Cultural Revolution, but the word's pattern was replaced. For example, "幼儿园you'er yuan kindergarten" appeared before the Cultural Revolution. This word influenced by the red political power and red thinking changed into "红幼班hong you ban kindergarten". Although the word's pattern changed, the meaning of the word did not change. When the Cultural Revolution ended, the pattern "红幼班hong you ban" was slowly replaced by "幼儿园you'er yuan kindergarten" or "幼稚园you zhi yuan kindergarten".

Replacement of pattern influences little on Cultural Revolution Vocabulary.

D) Partly remaining

We found the forth situation that the Cultural Revolution Vocabulary changed. The word's pattern did not change but the object or concept the word refers to has changed. The most typical word is "红宝书hong bao shu Ze-dong Mao's writings". Nowadays this word keeps word-formation "宝bao precious" with the emotion which is authoritative and precious. But the word now refers to monographs which are authoritative in some place replaces Ze-dong Mao's writings during the Cultural Revolution.

5.1.3 Disappearance of the Words Those are Anomalistic

According to Zhou, appeared during the Cultural Revolution, many Cultural Revolution words are anomalistic and nowadays are no longer used [11].

For example, a series of words in the form "革命ge ming revolution + X" are no longer used. "革命败类ge ming bai lei degenerates in revolution" is one of them. From the view of word's emotional coloring, 革命's color is neuter but 败类's color is derogatory. The word's emotional coloring of the two words is opposite. Now this situation is no longer appeared in Chinese.

5.2 Development of Society

The Cultural Revolution ended and people's life condition is changing. There are more and more new things arise. Language as a mirror reflects our society, lexicons as the materials of language fit and reflect the development of the society. Old things

were dying out and new things are appearing. These new things not only enrich our society but also influence person's thinking. Like the Cultural Revolution Vocabulary which represents revisionism and imprison were gradually dying out.

Besides, most of the Cultural Revolution Vocabulary that represent "battle" disappeared. And words that represent event during the Cultural Revolution like Three Disharmonies, Counter Flow in Feb and so on have died out too.

6 Summarization

Based on automatic word segmentation, part-of-speech tagging and ICTCLAS, from the view point of social linguistic we analyzed about 800,000 words which have used by People's Daily during Cultural Revolution (from year 1966-1976).We counted the part-of-speech and word's frequency of every Cultural Revolution Vocabulary. And we got the first 1,000 words having the max TF-IDF values and the highest frequencies, 96.6% of these 1,000 words are kept and only 3.4% of these 1,000 words disappeared. Then we analyzed the top-100 words of these 1,000 words in detail and found two main reasons why the Cultural Revolution Vocabulary have disappeared. These two main reasons are the system of language itself and development of society. In these two reasons the most important one is development and variation of words, including disappearance of what the word refers to, variation of the word's concept, replacement of a word's pattern and word-formation partly remaining. Word-formation partly remaining is the most special one.

Based on the achievements reached, we will do further studies. Firstly, we will expand material languages and enrich the corpus of the Cultural Revolution Vocabulary. Secondly, we want to find out the words which are lost during automatic word segmentation and part-of-speech tagging then add them into the corpus of the Cultural Revolution Vocabulary. Thirdly, seek the period when one of the Cultural Revolution Vocabulary first appeared accurately and find out word's etymology.

Acknowledgments. We are grateful for the comments of the anonymous reviewers. This work was supported in part by National Social Science Fund of China under contract 10CYY021, State Key Lab. for Novel Software Technology under contract KFKT2011B03, China PostDoc Fund under contract 2012M510178, Jiangsu PostDoc Fund under contract 1101065C, a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions(164320H107) and Jiangsu Research and Innovation Program for Postgraduates of Ordinary Colleges and Universities under contract CXLX12_0357.

References

1. Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS), http://ictclas.org
2. Wang, C.L.: The phenomenon of "Cultural Revolution" the Words Big Area Disappear. Journal of Yan'an Vocational & Technical Institute 23(2), 59–61 (2009)

3. Zhou, J., Li, G.X.: Strange languages were known in the time of misfortune. Journal of Xu Chang Teachers' College 17(1), 108–110 (1998)
4. Diao, Y.B.: Division of the Traditional Commendatory Words during "the Cultural Revolution". Journal of Esteem Liaoning University(Social Science) 8(4), 81–85 (2006)
5. Diao, Y.B.: Reviewing Loan words Used in the Period of the "Great Cultural Revolution". Journal of Anhui Institute of Education 25(1), 79–82 (2007)
6. Diao, Y.B.: The present Application and Reuse of the Popular Words used during the time of the Cultural Revolution. Journal of Hangzhou University (Social Sciences Edition) 6, 101–104 (2008)
7. Diao, Y.B.: Extreme Words Used in Cultural Revolution and Their Usage. Journal of Hengyang Normal University 29(4), 79–82 (2008)
8. Diao, Y.B.: Study on Coined words used in Cultural Revolution. Journal of Beijing Institute of Education 20(4), 42–45 (2006)
9. Jin, L.X.: Analyze of the Cultural Revolution Vocabulary from society and culture. New Argument 3, 21–26 (2001)
10. Saussure, F.: Cours de linguistiquegénérale. The Commercial Press, Beijing (1980)
11. Zhou, Y.: The Change of the Color Words' Meaning during "the Cultural Revolution". Journal of Tourism College of Zhejiang 3(1), 75–81 (2007)
12. Meng, J.: A corpus-based study of lexical periodization in historical Chinese. Literary and Linguistic Computing 25(2), 200–213 (2001)

Appendix: Top-100 Cultural Revolution Vocabulary

Word	Pinyin	Explanation	Word	Pinyin	Explanation
文化大革命	wen hua da ge ming	the Cultural Revolution	贫下中农	pin xia zhong nong	poor and lower-middle peasants
革委会	ge wei hui	revolutionary committee	走资派	zou zi pai	capitalist-roaders
一九七二年	yi jiu qi'er nian	the year of 1972	促生产	cu sheng chan	development economy
超级大国	chao ji da guo	the superpower	赤脚医生	chi jiao yi sheng	barefoot doctor
红卫兵	hong wei bing	the Red Guard	样板戏	yang ban xi	model opera
四人帮	si ren bang	the gang of four	狠抓	hen zhua	pay close attention
(赫鲁晓夫)	he lu xiao fu	Nikita Khrushchev	一九七三年	yi jiu qi san nian	the year of 1973
人民公社	ren min gong she	people's commune	一九七一年	yi jiu qi yi	the year of 1971

社队	she dui	production brigade	造反派	zao fan pai	the rebels
发展中国家	fa zhan zhong guo jia	developing country	领导班子	ling dao ban zi	leading group
文	wen	character	期	qi	stage
一九六九年	yi jiu liu jiu nian	the years of 1969	阮文绍	Ruan Wen-shao	Wen-shao Ruan
刘	Liu	Liu	和	he	and
朴正熙	Piao Zheng-xi	Zheng-xi Piao	斯里兰卡	si li lan ka	Sri Lanka
核武器	he wu qi	nuclear weapon	朗诺-施里玛达	Lang nuo-shi li ma da	Lon Nol - Shilimada
上山下乡	shang shan xia xiang	work in the countryside	坦桑尼亚	tan sang ni ya	Tanzanian
(少奇)	Shao-qi	Shao-qi	扎伊尔	zha yi'er	Zaire
(霸权主义)	ba quan zhu yi	hegemonism	多快好省	duo kuai hao sheng	better and more economical
不结盟	bu jie meng	no allies	克己复礼	ke ji fu li	controlling oneself
纳米比亚	na mi bi ya	Namibia	讲用会	jiang yong hui	preaching
施里玛达	shi li ma da	Shilimada	(亚非拉)	ya fei la	Asia, Africa and Latin America
阮文广	Ruan Wen-guang	Wen-guang Ruan	合作医疗	he zuo yi liao	cooperative medical service
红小兵	hong xiao bing	Little Red Guards	牛鬼蛇神	niu gui she shen	evil people
朝鲜	chao xian	South Korea	英·萨利	Ying·sa li	Ying·sali

一九七〇年	yi jiu qi ling nian	the year of 1970	(夺权)	duo quan	seize power
忆苦思甜	yi ku si tian	think over the good times	朗诺-施里玛达-山	Lang nuo-shi li ma da-shan	Lon Nol – Shilimada—Moution
津巴布韦	jin ba bu wei	Zimbabwe	北京市	Beijing	Beijing
(努克)	NuKe	Nuck	(科研)	ke yan	scientific research
邀请赛	yao qing sai	invitation game	诺罗敦·西哈努克	nuo luo dun·xi ha nu ke	Noro-dom·Sihanouk
苏联	su lian	USSR	时	shi	time
红彤彤	hong tong tong	all of a glow	(高棉)	gao mian	Cambodia
莫桑比克	mo sang bi ya	Mozambique	一九六六年	yi jiu liu liu nian	the year of 1966
蹲点	dun dian	stay at a selected grass-roots	妖风	yao feng	evil wind
豪情	hao qing	lofty sentiments	修	xiu	construct
麦贤得	mai xian de	Mai virtuous	三自一包	san zi yi bao	Three-Self-One-Packet
纳苏蒂安	na su di'an	Nasution	(国峰)	guo feng	guofeng
比绍	Bissau	Bissau	毛里塔尼亚	mao li ta ni ya	Mauritania
祝愿	zhu yuan	wish	一九六七年	yi jiu liu qi nian	the year of 1967
东帝汶	dong di wen	east Timor	战备	zhan bei	combat readiness
老大难	lao da nan	long-standing	一九六四年	yi jiu liu si nian	the year of 1964

一九六八年	yi jiu liu ba nian	the year of 1968	第三世界	di san shi jie	Third World
(巴勒斯坦)	ba le si tan	Palestine	恩维尔·霍查	En wei'er · huo cha	Enver·Hoxha
森潘	Sen pan	Senpan	天安门	Tian'An men	Tiananmen Square
苑化冰	Yuan Hua-bing	Hua-bing Yuan	差距	cha ju	gap
支农	zhi nong	support agricul-ture	讯	xun	news
私	si	private	(索马)	suo ma	Somalia
意气风发	yi qi feng fa	on one's mettle	三支两军	san zhi liang jun	three support's and two mili-tary's
交易会	jiao yi hui	trade fair	欧安会	ou'an hui	CSCE
(孔孟之道)	kong meng zhi dao	Confucius and Mencius	西索瓦·梅达维	xi suo wa · mei da wei	Sisowath·Mei Dawei
鼓足干劲	gu zu gan jin	strain oneself	抗震救灾	kang zhen jiu zai	earthquake relief

Measuring the Semantic Distance of the Near-Synonyms of Touch Verbs in Chinese

Siu Lun Au and Helena Hong Gao

School of Humanities and Social Science, Nanyang Technological University, Singapore
ausi0001@e.ntu.edu.sg, helenagao@ntu.edu.sg

Abstract. This study applies the semantic specification and decomposition methods to the analysis of the near synonyms of Chinese touch verbs, 触 *chù* "touch", 碰 *pèng* "touch", 拍 *pāi* "pat", 拭 *shì* "wipe", 弹 *tán* "flick", and 抚 *fú* "caress" and the measuring of the semantic distance among them. The semantic properties within the verb roots are classified as the basic semantic features for the analysis. The purpose of examining both semantic and syntactic features as the starting point for the analysis is to distinguish more effectively the differences among the near-synonyms when they are formed with other words into set phrases. The results of this research can contribute to the learning and dictionary compilation of near-synonyms in Chinese.

Keywords: lexical semantics, touch verbs, near-synonyms, semantic properties, semantic decomposition.

1 Introduction

Near-synonyms are expressions that are similar but not identical in meaning. Near-synonyms are an important part of lexicon in any languages. It gives rise to variations in expressions and also helps speakers to express their thoughts and describe things perceived from different perspectives. As such, to identify the semantic features of each near-synonym and distinguish the semantic relations among them becomes a meaningful and important task.

It has been hypothesized and accepted by semanticists that words are not the smallest semantic units but are built up of smaller components which are combined differently to form different words [1]. Research findings from experimental psychology generally support the assumption that in speech production the semantic representation of a lexical word activates not only the intended lexical item but also the semantically related words. The meaning of a word, as a whole and collective concept, can be decomposed into smaller components of meaning. The process of analysing the meaning of a word can be different based on the mode applied and consequently its result can be affected [2]. Therefore, to objectively create an unbiased model we can use one of the words in a group of near-synonyms as the focal point and decompose the semantic features from each member of the group to find out what types of semantic features are embedded in the lexicon and how we can sort them out categorically to reflect both individual and group features. After the decomposition and

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 708–717, 2013.

categorization of the features, we can use a statistical method to quantify and measure the features to show the semantic distances among the near-synonyms. This method is assumed to be able to systematically present the nature of lexical semantics in relation to the meanings of near-synonyms.

2 Aims / Objectives

Following the assumption that near-synonymous words have subtle differences in their meaning components that can be revealed by a semantic decomposition method as a first step comparison of the semantic features embedded in the words and the semantic distance between the words can be observed when the features can be analysed systematically, this study aims to distinguish the near-synonyms of touch verbs in Chinese by identifying the semantic features of the meaning components of the verbs and then using the features to measure the semantic distance among the verbs. The methods applied to decompose the semantic components and to compare the semantic and syntactic features among the verbs are to ensure that the differences between the near-synonyms are distinguished more effectively.

3 Previous Studies of Near-Synonyms of Physical Action Verbs

Studies of physical action verbs in Chinese have been found in recent years since the first study on physical action verbs conducted by Gao was published in 2001. However, few of them focus on the near-synonyms of physical action verbs as a domain-specific lexical category. Gao's studies [1, 3] made complete classifications of physical action verbs with detailed definitions and criteria for classifications. Gao's studies also included a specification system for measuring the semantic distance among the near-synonyms of physical action verbs. By specifying the action manners of physical action verbs as the key semantic components embedded in the verb roots, the author analysed the semantic properties of the near-synonyms of physical action verbs, such as, "verbs of cutting", "verbs of touching", "verbs of throwing", "verbs of putting", and "verbs of lying", etc. Differences between the members of each class are identified by their meaning components that are classified by the notions defined by Gao as Body Part Contact, Instrument, Force, Motion Direction, Speed, Effect, Intention, and Patient Object, etc. They are marked as categorical indicators embedded in the verb roots commonly shared by physical action verbs. These notions are applied as the principles in projecting the lexical semantic prominence in the classifications of a physical action verb's meaning components among its near-synonyms [3]. We will apply Gao's notions and methods to the analysis of the semantics of the class of touch verbs in this study.

4 Methodology

Our method for measuring the semantic distance requires random sampling of the target verbs from corpus data. A sample size of 1000 sentences was extracted. The detailed procedure for the analysis is given below:

1. 1000 sentences with 触 *chù* "touch", 碰 *pèng* "touch", 拍 *pāi* "pat", 拭 *shì* "wipe", 弹 *tán* "flick", and 抚 *fú* "caress" were randomly extracted from the CCL Corpus of Peking University[4].
2. Only the sentences that contained the six touch action verbs were used in the analysis.
3. The verb phrases of the touch verbs formed with other words and their syntactic patterns were collected from the corpus and classified according to the distribution ratio.
4. Based on the verb phrases and their syntactic patterns, the semantic features were deduced and classified as Body Part, Patient Object, Force, Speed, Motion Direction, Intention, Consequence, and Emotion. Meaning of verbs were also referenced from 《现代汉语词典》 *Xiàndài Hànyǔ Cídiǎn* "Modern Chinese Dictionary"[5,6].
5. To measure the semantic distance among the near-synonyms of touch verbs, each of the verbs was chosen as a basis or a head verb for comparison. The number of the semantic features shared between the head verb and its other class members were calculated and ordered hierarchically to reveal their relations in terms of the semantic distance.
6. In order to show the possibility of measuring the semantic distance in certain particular aspects among the near-synonyms, a few selected semantic features were applied to test and see if a similar relation among the near-synonyms could be revealed.

5 Extracting and Classifying Semantic Features

5.1 Semantic Decomposition

Based on the eight semantic features extracted as the basis for comparison, the meaning components of the six verbs are summarised below:

1. Body Part: This refers to the body part that involves in the physical action. 触 *chù* "touch" and 碰 *pèng* "touch" are unintentional actions with generally a non-specific patient object may not involve an actual physical contact, thus we can deduce that any body part can perform such an action. Hence, the semantic feature for both 触 *chù* and 碰 *pèng* is [any body parts]. As for 拍 *pāi* "pat", the Modern Chinese Dictionary defines it as "hit with the palm". We can conclude that the semantic feature for 拍 *pāi* is [palm]. As for 抚 *fú* "caress" and 拭 *shì* "wipe", both 用手指 *yòng shǒuzhǐ* "using finger" and 用手掌 *yòng shǒuzhǎng* "using palm" are found in the corpus data and thus we label 抚 *fú* "caress" and 拭 *shì* "wipe" as having the semantic feature of [finger and palm]. 弹 *tán* "flick" is a finger action and thus its semantic feature for this category is [finger].

2. Patient Object: The patient objects for the verb with 触 *chù* "touch" were found mostly non-specific. 接触 *jiēchù* "contact", 抵触 *dǐchù* "conflict", 触犯 *chùfàn* "break (laws, rules or regulations)", achieved the highest frequency but they do not involve any physical contact. As such, the semantic feature of 触 *chù* is

[-patient object]. As for 碰 *pèng* "touch", its patient object can be specific, (e.g., 皮球 *píqíu* "ball"), or non-specific, (e.g., 事情 *shìqíng,* "matter", 问题 *wèntí,* "problems"). Hence its semantic feature is [±specific subjects]. As for 拍 *pāi* "pat", 拭 *shì* "wipe", 弹 *tán* "flick", and 抚 *fú* "caress", the patient objects are mostly specific (e.g., 脸 *liǎn* "face"), hence their semantic feature is [+specific patient object].

3. Force: 触 *chù* "touch" is an action with a slight touch mostly by accident on something that is found unspecified. The force of 触 *chù* is thus uncontrollable and has the semantic feature labeled as [-force]. 碰撞 *pèngzhuàng* "clash", 碰硬 *pèngyìng* "go head on with",碰伤 *pèngshāng* "clash and injure", and 轻轻一碰 *qīngqīng yi pèng* "touch lightly" imply that 碰 *pèng* involves different degrees of force when combined with different words. Similarly, 拍 *pāi* "clap" is commonly followed by a complement such as 红 *hóng* "red", 肿 *zhǒng* "swollen", and 伤 *shāng* "injured" which implies a higher degree of force and 轻轻地拍 *qīngqīngde pāi* "pat lightly" implies that 拍 *pāi* can be a light action too. Thus the semantic feature for 碰 *pèng* and 拍 *pāi* is [+strong/weak force]. 拭 *shì* "wipe" and 抚 *fú* "caress" are often used together with the adverb 轻轻 *qīngqīng* "lightly". Hence 拭 *shì* and 抚 *fú* have the semantic feature [+weak force].《现代汉语词典》 *Xiàndài Hànyǔ Cídiǎn* "Modern Chinese Dictionary" defines that 弹 *tán* "flick" is to forcefully release the finger which is pressed and hit something by using the force generated, but the corpus data show that it can also be used to describe an action with a light force applied (e.g., 轻轻地 *qīngqīng* "lightly"). As such, the semantic features of *tán* are [+strong/weak force].

4. Speed: As 触 *chù* "touch" in the corpus data shows no specific patient objects, it would be hard to determine the action speed. In addition, there were not any adverbs of speed found together with 触 *chù*. Therefore, the semantic feature for 触 *chù* is [-speed]. 碰 *pèng* "touch" is a quick action. There are occurrences of 碰 used together with 着 but 着 in 碰着 *pèngzhe* actually does not indicate a continuous state of the action but rather a quick and usually unexpected action (e.g., 碰着大雨 *pèngzhe dàyǔ* "be caught in the rain"). The Modern Chinese Dictionary has a definition of 着 *zhe* as indicating the end of an action. As such, we can argue that 碰 *pèng* even when used with 着 *zhe* does not indicate a continuous state, but an action with a short duration, so the semantic feature of 碰 *pèng* is [+short duration]. As for 拍 *pāi* "pat", different degrees of force are involved in different contexts. For example, 拍苍蝇 *pāi cāngyíng* "swat a fly" is faster than the action of 轻轻拍他的肩 *qīngqīng pāi tā de jiān* "pat on his shoulder". Thus, we can say that 拍 *pāi* can be a slow action or a fast action and its semantic feature in this aspect should be [+short/long duration]. 抚 *fú* "caress" and 拭 *shì* "wipe", on the other hand, are generally done with a light degree of force to achieve a positive emotion and thus they are performed slowly. As such, they both have the semantic feature [+long duration]. 弹 *tán* "flick" does not show explicitly the speed of the action in the corpus data but figurative expressions like 弹指之间 *tánzhǐzhījiān,* "in the flick of a finger" and 弹指十年 *tánzhǐshínián* "ten years passed by like the flick of a finger" imply a quick completion of the action. Therefore, the semantic feature of 弹 *tán* is [+short duration].

5. Motion Direction: By motion direction we mean the motion direction of the agent together with or without the patient after the agent has a physical contact with the patient (e.g., 弹玻璃球 *tán píqíu* "flick a glass ball"). Only 抚 *fú* and 拭 *shì* have motion directions and the rest of the verbs have the semantic feature [-direction]. 抚 *fú* "caress" and 拭 *shì* "wipe" require the body part to be in contact with the patient object from the beginning point to the end of the action. 抚 *fú*, specifies a circular motion of the palm on an object and its semantic feature is thus labeled as [+to and fro direction]. As for 拭 *shì*, it is an instantaneous action; hence its semantic feature is [+single direction].

6. Intention: 触 *chù* "touch" is often used together with 不慎 *bùshèn* "unintentional/careless" to describe an unintentional action. 触电 *chùdiàn* "get an electric shock", 触雷 *chùléi* "step on a mine" and 触礁 *chùjiāo* "run on rocks in the sea" also imply that 触 *chù* is an unintentional action. Hence, the semantic feature of 触 *chù* is [-intention]. As for 碰 *pèng* "touch", 碰巧 *pèngqiǎo* "coincidentally", 碰壁 *pèngbì* " run into a wall (metaphor)" and 碰到问题 *pèngdào wèntí* "come across a problem" imply that the action of 碰 *pèng* is unintentional. 不小心碰 *bùxiǎoxīn pèng* "accidentally touch" is also a common combination. However, 轻轻一碰 *qīngqīng yi pèng* "touch lightly" implies that the force applied can be controlled. A controlled action is an intentional action. Hence, the semantic feature for 碰 *pèng* is [±intention]. As for, 拍 *pāi* "pat", 拭 *shì* "wipe", 弹 *tán* "flick", and 抚 *fú* "caress", the strength for each action is controllable. Hence, they all have the semantic feature [+intention].

7. Consequence: 触发 *chùfā* "trigger", 触怒 *chùnù* "anger", 触犯 *chùfàn* "break (the law or regulation)", 触电 *chùdiàn* "get an electric shock", 触礁遇难 *chùjiǎo yùnàn* "hit rocks in the sea and sink", and 触电伤亡 *chùdiàn shāngwáng* "get electrocuted and die" indicate that 触 *chù* has the semantic feature [+activate] (see its definition in Gao 2001, Chapter 4). The complements 伤 *shāng* "injured" and 破 *pò* "destroyed" of the verb 碰 *pèng* "touch" and the complements 红 *hóng* "red" and 肿 *zhǒng* "swollen" of the verb 拍 *pāi* "touch" indicate that they both have the semantic feature [+destruction]. However, as mentioned above, 碰 *pèng* and 拍 *pāi* can vary in the force degree and thus an action with a low degree of force may not result in destruction. Hence, its semantic feature is [±destruction]. As for 拭 *shì*, its patient object is 泪 *lèi* "tear", 汗 *hàn* "sweat", and 脏物 *zāngwù* "dirt", and often, it is used with the complement 去 *qù* "away". So, we can deduce that the semantic feature of 拭 *shì* is [+removal]. 抚 *fǔ* is always used with such adverbs as 怜爱地 *liánaìde* "affectionately", 温存地 *wēncúnde* "attentively", and 亲切地 *qīngqiède* "kindly". In the verb phrase formation, we find 抚慰 *fǔwèi* "console", 爱抚 *aìfǔ* "caress" and 抚爱 *fǔài* "caress" and thus the semantic feature of 抚 *fǔ* is [+consoling].

8. Emotion: 触怒 *chùnù* "anger", 触动 *chùdòng,* "emotionally touched"; 拍肩膀 *pāi jiānbǎng* "pat on the shoulder" and 拍桌子 *pāi zhuōzi* "slam on a table"

implies that both the positive and negative emotions can be involved and thus the semantic features of 触 *chù and* 拍 *pāi* are [+positive/negative emotions]. As for 碰 *pèng* "touch" and 弹 *tán* "flick", there is no evidence for any emotion to be involved, so their semantic feature is [-emotions]. 抚 *fǔ* "caress" and 拭 *shì* "wipe" have the semantic feature labeled as [+positive emotions] as the actions of 拭泪 *shìlèi* "wipe tears" and 抚慰 *fǔwèi* "console" are done with emotions involved.

5.2 Distribution of the Semantic Features

After extracting the semantic features from the corpus data, we are able to demonstrate the distributions of the semantic features among the near-synonyms of touch verbs as shown in Tables 1, 2 and 3 below.

Table 1. Patient object and body part feature distributions

	Patient Objects		**Body Parts**		
	Patient Objects	Specific	Palm	Finger	Other Body Parts
触 **chù "touch"**	-	-	+	+	+
碰 **pèng "touch"**	±	±	+	+	+
拍 **pāi "pat"**	+	+	+	-	-
拭 **shì "wipe"**	+	+	+	+	-
抚 fǔ **"caress"**	+	+	+	+	-
弹 **tán "flick"**	+	+	-	+	-

Table 2. Force degree and speed feature distributions

	Force Degree		Speed	
	Strong	Weak	Short Duration	Long Duration
触 chù "touch"	n/a	n/a	n/a	n/a
碰 pèng "touch"	+	+	+	-
拍 pāi "pat"	+	+	+	+
拭 shì "wipe"	-	+	-	+
抚 fǔ "caress"	-	+	-	+
弹 tán "flick"	+	+	+	-

Table 3. Intention, consequence and emotion feature distributions

	Intention	Consequences			Emotions	
	Intention	Activate	Destruction	Removal	Positive	Negative
触 chù "touch"	-	+	-	-	+	+
碰 pèng "touch"	±	-	±	-	n/a	n/a
拍 pāi "pat"	+	+	±	-	+	+
拭 shì "wipe"	+	-	-	+	+	-
抚 fǔ "caress"	+	-	-	-	+	-
弹 tán "flick"	+	-	+	±	n/a	n/a

5.3 Measuring the Semantic Distances

Using the data from Tables 1, 2 and 3, we can apply Gao's method [1] to calculate the semantic features shared by each of the touch verbs and then use one verb at a time as the head verb to compare its shared features with those shared by other members of the class. By ordering the verbs according to the number of the semantic features shared, we demonstrate the semantic distance among them in a hierarchical order. We begin with the verb 触 *chù* to show the ordering (Due to restriction of the page limit, Gao's method is simplified here):

触 *chù* "touch"→碰 *pèng* "touch"(7)→拍 *pāi* "pat"(5)→拭 *shì* "wipe"(3), 抚 *fǔ* "caress"(3)→弹 *tán* "flick"(2).

As we can see, by using 触 *chù* as the basis for comparison, 碰 *pèng* has 7 semantic features shared with that of 触 *chù*. As such, 碰 *pèng* is semantically closest to 触 *chù*. With 拍 *pāi,* 触 *chù* shares 5 semantic features, which makes it the next closest to 触 *chù*. Of 拭 *shì* and 抚 *fú*, 3 semantic features are shared with 触 *chù* and thus both are equally close to 触 *chù*. 弹 *tán* has the furthest semantic distance from 触 *chù* as they share only 2 semantic features. The higher a number a verb possesses, the closer its semantic distance is to the head verb in the hierarchical order. Taking each of the rest of the verbs as the basis for comparison, we demonstrate below the changes of the semantic distance among the class members.

碰 *pèng* "touch" → 拍 *pāi* "pat" (9)→ 弹 *tán* "flick" (7)→ 触 *chù* "touch" (5), 拭 *shì* "wipe" (5), 抚 *fǔ* "caress" (5)

拍 *pāi* "pat" →碰 *pèng* "touch" (9)→ 弹 *tán* "flick" (8) →拭 *shì* "wipe" (6), 抚 *fǔ* "caress" (6) →触 *chù* "touch" (5)

拭 *shì* "wipe"→ 抚 *fǔ* "caress" (8) →拍 *Pāi* "pat" (6)→ 弹 *tán* "flick" (5), 碰 *pèng* "touch" (5) →触 *chù* "touch" (3)

抚 *fǔ* "caress"→拭 *shì* "wipe" (8) →碰 *pèng* "touch" (5), 拍 *pāi* "pat" (5) →弹 *tán* "flick"(4) →触 *chù* "touch" (3)

弹 *tán* "flick"→碰 *pèng* "touch" (10) →拍 *pāi* "pat" (8) →拭 *shì* "wipe" (5) →抚 *fǔ* "caress" (4) →触 *chù* "touch" (2)

From the above ordering we can see that after changing the basis for comparison, the semantic relations among the verbs differ. This supports our hypothesis that if A and B are a pair of near-synonyms, when A = B, B might not necessary be equal to A. After changing the basis for comparison, a different semantic route is triggered in the semantic map. However, when constructing a sentence, we seldom use every semantic feature of one word as a basis to compare what our best choice of a word is in a certain context. As such, another way of comparing semantic distance would be just choosing a single or a few semantic features that are suitable for the context and see which word would best fit the context. For example, using [+positive emotion] as a focal point:

抚 *fǔ* "caress", 拭 *shì* "wipe" →拍 *pāi* "pat", 触 *chù* "touch" →碰 *pèng* "touch", 弹 *tán* "flick".

To create positive emotions, 抚 *fǔ* "to caress" and 拭泪 *shìlèi* "wipe tears" are the most possible type while 碰 *pèng* "touch" and 弹 *tán* "flick" are the least.

Taking [force] and [destruction] as the focal points for another example:

碰 *pèng* "touch"→ 拍 *pāi* "pat"→ 弹 *tán* "flick" →抚 *fǔ* "caress", 拭 *shì* "wipe"→ 触 *chù* "touch"

The effect of *pèng* can result in serious injury to a person (碰断门牙 *pèngduàn le ményá,* "break the front tooth") while 拍 *pāi* would at most cause a swollen spot on the body. As such, the force degree of 碰 *pèng* can be greater than that of 拍 *pāi*. By using Newton's first law of motion, *f=ma*, the action of 拍 *pāi* "slam" with a palm would definitely requires a greater mass than the action of 弹 *tán* "flick" with a finger. Therefore, the force degree of 拍 *pāi* is greater than that of 弹 *tán*. Relatively, 抚 *fǔ* and 拭 *shì* require lower degree of force. As for 触 *chù*, force is not involved. As such, the order of semantic distance can also predict which word is a better choice in a certain context if we focus on just a few semantic features to consider. However, the degree of force involved in different actions also differs when performed by different people. In the next section, we will discuss further the force feature.

5.4 Degree of Force

The degree of force applied to an action differs from person to person. Apparently, it is impossible to measure the degree of force required for the action of 碰 *pèng* or 抚 *fǔ* without any comparison. As such, we can only describe the force degree of an action relative to that of another action. From what we show above, only 碰 *pèng*, 拍 *pāi* and 弹 *tán* have the semantic feature [+strong/weak force]. The differences of these actions in force degree can be deduced when an adverb of degree is present in a sentence, as shown below:

A. Low Degree of Force
1. 他将那滴泪用食指轻轻弹去。
Tā jiāng nā dì lèi yòng shízhǐ qīngqīng tán qù
He used his index finger to (lightly) flick his tears away.

2. 小坡用头轻轻一碰 ，门就开了。
Xiǎobō yòng tóu qīngqīng yi pèng, mén jiù kāi le.
Xiaobo used his head to touch the door lightly and it was opened.

3. 我轻轻拍他的肩，低声问他为何不睡。
Wǒ qīngqīng pāi tā de jiān, dīshēng wèn tā wèi hé bú shuì.
I patted his shoulder and asked him why he didn't go to bed.

B. High Degree of Force
1. 手掌拍麻了。
Tā de shǒuzhǎng pāi má le.
His palms are numb from clapping.

2. 是你弹得太用力了。
Shì nǐ tán de tài yònglì le.
You flicked too hard.

3. 我将心爱的瓷茶壶盖头碰断了。
Wǒ jiāng xīnài de cí cháhú gàitóu pèngduàn le.
I (accidently) broke the lid of my favorite tea cup.

As we can see, any actions from group A would imply a higher degree of force than that from group B. Therefore, the use of adverbs with these verbs would help define the force degree involved. However, within group A, no matter how light one flicks (sentence A1), it would still be stronger than the action of touch on someone's head because to flick is to remove something light while to touch is to aim at making a physical contact. As for A2 and A3, the actions of 碰 *pèng* "touch" and 拍 *pāi* "pat" would require the same degree of force as both actions are generally done with the intention of making a physical contact.

As for group B, the force exerted by a palm is greater than that exerted by a finger. To *pèng*, "clash" until the tooth is broken would definitely require greater force than that of 拍 *pāi* "pat" or 弹 *tán* "flick". As such, the degree of force will differ in different contexts, which can be reflected with the presence of an adverb of force, an intention or some consequence of the action. For this kind of verbs with varied degrees of force involved, the semantic features of [intention] and [consequence] seem to be more helpful in distinguishing the degree of force.

6 Conclusion

From the sample data we extracted from the CCL Corpus, we realised that the syntactic patterns of the verb phrases formed by the touch verbs are closely related to their

semantic features. After extracting and classifying their semantic features, we can not only distinguish the members of this class of near-synonyms, but also reveal the semantic distance among them. However, it is noted that the data collected might not be reflective of the complete semantic extensions of the verbs but the method for analysis was found still reliable as it can allow us to analyse the semantic features more subjectively and more systematic.

The discussions of the method and the analysis are brief. More elaborations could be provided. A few problems that cannot be solved in this paper are listed below and we will work on them in the future.

1. The filtering process in step 2 mentioned in the methodology section needs to be improved More sentences should be extracted to replace the discarded sentences so as to keep the actual usable data about the same for the words analysed.
2. The criteria for the filtering varied in order to include as many phrases as possible but this approach ignored the distinctions between verb phrases and noun phrases formed by the near-synonyms. A better system of selection should be developed in the future.
3. The time of an action is not a bi-value function like [long/short duration]. Some verbs such as 接触 *jiēchù* "contact" do not have a clear indicator of the start or the end point of an action. It would be difficult to compare semantic features if we were too specific in defining the [time/duration] of a word. As such, we could only deduce a bi-value function when discussing semantic features in this paper.

To sum up, this study takes a systematic approach to analyse the lexical semantics of near-synonyms in Chinese. Touch verbs are the focus for the analysis which gives the end result of revealing the semantic distance among the class members. It is believed that the methods used for this study can be applied to the analysis of other classes of near-synonyms of different verbs, nouns and adjectives.

Acknowledgement. We wish to acknowledge the funding support for this project from Nanyang Technological University under the Undergraduate Research Experience on CAmpus (URECA) programme.

References

1. Gao, H.: The Physical Foundation of the Patterning of Physical Action verbs: A Study of Chinese Verb. Lund University Press (2001)
2. Yu, P.F.: Towards a Framed Analysis of Activities Sememes' Definition. The Center for Linguistics and Applied Linguistics of Guangdong University of China (2005)
3. Gao, H.H., Ouyang, S.: A. Feature-Based Algorithmic E-learning Tool for Learning Near-synonyms in Chinese. In: Proceedings of 2009 International Conference on Applied Linguistics & Language Teaching, pp. 585–596. Crane Publishing Co. Ltd. (2009)
4. CCL Chinese Corpus Database. Centre for Chinese Linguistics PKU, `http://ccl.pku.edu.cn:8080/ccl_corpus/`
5. Modern Chinese Dictionary, 5th edn. China Commerce and Trade Press (2005)
6. Chinese Dictionary. Cishu Chubanshe, Shanghai (1993)

Features, Improvements and Applications of Ontology in the Field of Sports Events during the Era of the Semantic Web

Juan Xiao[1] and Jing Chen[2]

[1] Department of Sports, Wuhan University, Wuhan 430072, China
973266289@qq.com
[2] Department of Foreign Language, Hubei University of Technology, Wuhan 430064, China
370958331@qq.com

Abstract. Domain Ontology is a cutting-edge hot topic during the era of the Semantic Web. This paper studies Ontology in the field of sports events. Firstly, it explains the overview of Ontology in the field of sports and Fundamental Ontology, and then makes a comparison between them. It proceeds to provide recommendations for the improvement. The characteristics of Ontology in the field of sports events are summarized and its logical triples interpreted. The intelligent application of Ontology in the field of sports events is consequently discussed.

Keywords: Ontology, Domain Ontology, Semantic Web, intelligent application.

1 Introduction

Ontology is a conceptual model describing the concept and the relationship between concepts, and it explains the concept by the relationship between concepts. It is proper to formalize human knowledge and information, by which it constructs a common information understanding mechanism between the user and the machine, and realizes the sharing of domain knowledge.

2 An Overview of the Ontology in the Field of Sports Events and Domain Ontology

2.1 Fundamental Ontology and Domain Ontology

Knowledge existing outside the human knowledge system belongs to the domain of fundamental knowledge. The Ontology is based on it is Fundamental Ontology. Fundamental Ontologies such as WordNet, FrameNet and CCD are widely used. The Ontology based on a special professional or Domain knowledge is Domain Ontology.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 718–727, 2013.

2.2 The Ontology in the Field of Sports Events

The Ontology in the field of sports events is a typical Domain Ontology. It collects some concepts and terminology in the field of sports events, and constructs their relationships.

3 Partial Comparison and Integration of the Ontology in the Field of Sports and Fundamental Ontology

3.1 The Base of Comparison and Integration

In this paper, Fundamental Ontology is based on *Synonymy Thesaurus* by Mei Jiaju and the Ontology in the field of sports events is based on *The Development of Sports Corpus and its Sports Lexical Study* by Chen Wei and *Chinese-English Sports Classification Dictionary* by Chen Naixin. On this basis, it initially develops an Ontology in the field of sports events.

3.2 Partial Comparison and Integration

For lack of space, we only choose some key words to compare and integrate partially the Ontology in the field of sports events and Fundamental Ontology.

3.2.1 From the Perspective of the First-Layer Node

The first-layer node of fundamental Ontology which corresponds to the Ontology in the field of sports events mainly focus on the categories of [man], [thing], [time and space], [abstract thing] and [sport], and nearly have no correspondence to the categories of [characteristic], [psychological], [activity], [phenomenon and condition],[relation][auxiliary vocabulary], and [complimentary vocabulary]. This shows that the semantical categories of the Ontology in the field of sports events have their focuses. The phenomenon that the categories of [organization] and [referee] correspond to the categories in Fundamental Ontology of [abstract thing] [activity] and [thing][action] shows their complexity. By comparing them, we know that there are some differences between each first-layer node of the two Ontology.

3.2.2 From the Perspective of the Second-Layer Node

We here only compare [personnel] of the Ontology in the field of sports events and [human] of Fundamental Ontology. Fundamental Ontology has no categories of [referee], and [assistant clerk] shows its defect, because these two categories are fundamental in fact, having a higher degree of generalization.

3.2.3 From the Perspective of the Third-Layer Node

We only contrast the words in [athletes] of the Ontology in the field of sports events with the words in [athletes] of fundamental Ontology. As is shown in the following, the first layer derives from the Ontology in the sports events [1], and the below layers are from fundamental Ontology [2]:

- 运动员 yundongyuan athlete
- 球员 qiuyuan player 发球员 faqiuyuan server 击球员 jiqiuyuan hitter 接球员 jieqiuyuan receiver 接发球员 jiefaqiuyuan receiver 守门员 shoumenyuan goalkeeper 跑垒员 paoleiyuan base runner 击球手 jiqiushou batter 投手 toushou pitcher 垒手 leishou corner man
- 拳击手 quanjishou typewriter 棋手 qishou chess player 选手 xuanshou contestant 种子选手 zhongzixuanshou seeded player 非种子选手 feizhongzixuanshou unseeded player
- 前锋 qianfeng forward 小前锋 xiao qianfeng small forward 后卫 houwei guard 中锋 zhongfeng center 替补1 tibu alternate 自由人 ziyouren libero 二传手 erchuanshou second pass
- 主队 zhudui host team 客场 kechang away games 击球方 jiqiufang Batting side 发球方 faqiufang serving side 接发球方 jiefaqiufang receiving side
- 运动员 yundongyuan athlete 选手 xuanshou contestant 健儿 jian'er strong man
- 种子 zhongzi seed 健将 jianjiang master
- 前锋 qianfeng forward 中锋 zhongfeng center 左锋 zuofeng left forward 右锋 youfeng right forward 中卫 zhongwei halfback 前卫 qianwei vanguard 左卫 zuowei left back 右卫 youwei right back 后卫 houwei guard 守门员 shoumenyuan goalkeeper
- 一传手 yichuanshou first pass 二传手 erchuanshou second pass 主攻手 zhugongshou main attacker
- 投手 toushou pitcher 捕手 bushou backstop 一垒手 yileishou first baseman 二垒手 erleishou second baseman 三垒手 sanleishou third baseman 左翼手 zuoyishou left wing man 右翼手 youyishou right wing man 中坚手 zhongjianshou backbone man 游击手 youji shou shortstop

We will not translate them again when referring to some of the above words and expressions therein.

There are 28 words in the node of [athlete] about the Ontology in the field of sports, and 27 words in the node of [athlete] about fundamental Ontology. The total number of the words in each Ontology is equivalent. They both have the 8 words as shown above.

Through Fundamental Ontology, we find that the Ontology in the field of sports events has its disadvantages, and the improvement are furnished as follows:

Firstly, he nodes are few and their capacity is too large to highlight the advantages of the concrete of Domain Ontology. For instance, there seems a lack of the terms of "左锋 zuofeng left forward", "右锋 youfeng right forward", "中卫 zhongwei halfback" and "前卫 qianwei vanguard", etc. As is usually known, sport is a national event and an issue of public concern, so the sports terms have a high degree of popularization. We should enlarge the size of the nodes properly during the construction of Fundamental Ontology.

Secondly, the structure level of few nodes is too general. For example, the terms of "前锋 qianfeng forward", "小前锋 xiao qianfeng small forward" and "后卫 houwei guard" can be disintegrated into different subsidiary nodes. [3]

By contrast, we find the disadvantages of the Ontology in the field of sports events. Likewise, through the Ontology in the field of sports events, we can also find the disadvantages of the Ontology as follows:

On the one hand, the nodes are few and their capacity is too large to include the terms with high a popularizing degree, such as “球员 qiuyuan player”, “棋手 qishou chess player” and “替补tibu alternate”, etc.

On the other, the subsidiary nodes are so many that some terms are suitable to appear only in the constructing process of the field Ontology, such as “一垒手 yileishou first baseman ”, “二垒手 erleishou second baseman ”, “三垒手 sanleishou third baseman ”, “左翼手 zuoyishou left wing man ”, and “右翼手 youyishou right wing man”, etc.

4 Characteristic of the Ontology in the Field of Sports Events

4.1 Dissimilarity of the Structure Level

Within the Ontology in the field of sports events, [man] is in the first floor and [sportsman] is in the third floor. It shows that the same nodes are in the different structure levels when they appear separately in Fundamental Ontology and the Ontology in the field of sports events.

4.2 Scientific of the Structure Level

The Framework for the design and distribution of nodes of the Ontology in the field of sports events is relatively professional and scientific. A very important reason is that the Ontology in the field of sports events can adopt the triple logic form of <concept, attribute, instance>, introducing the attribute, and yet the Fundamental Ontology is short of the attribute.

4.3 Diversification of the Nodes

The category “sportsman” has a high commonality. A small number of lower nodes are separated from the node of [sportsman] for Fundamental Ontology. On the contrary, a large number of lower nodes are separated from the node of [sportsman] for the Ontology in the field of sports events. This shows that the lower nodes of [sportsman] in the field of sports events are getting more, so as to provide the professional and disciplinary terms.

5 Logic Triples of the Node of Ontology in the Field of Sports Events

In theory, the nodes of Ontology in the field of sports events include concepts, attributes, instance and so on. So it could be expressed by the logic triples of <concept, attribute, instance>. Take the node of [athletes] from the Ontology in the field of sports events above, for example. The relationship between it and its lower nodes as shown below in Fig. 1.

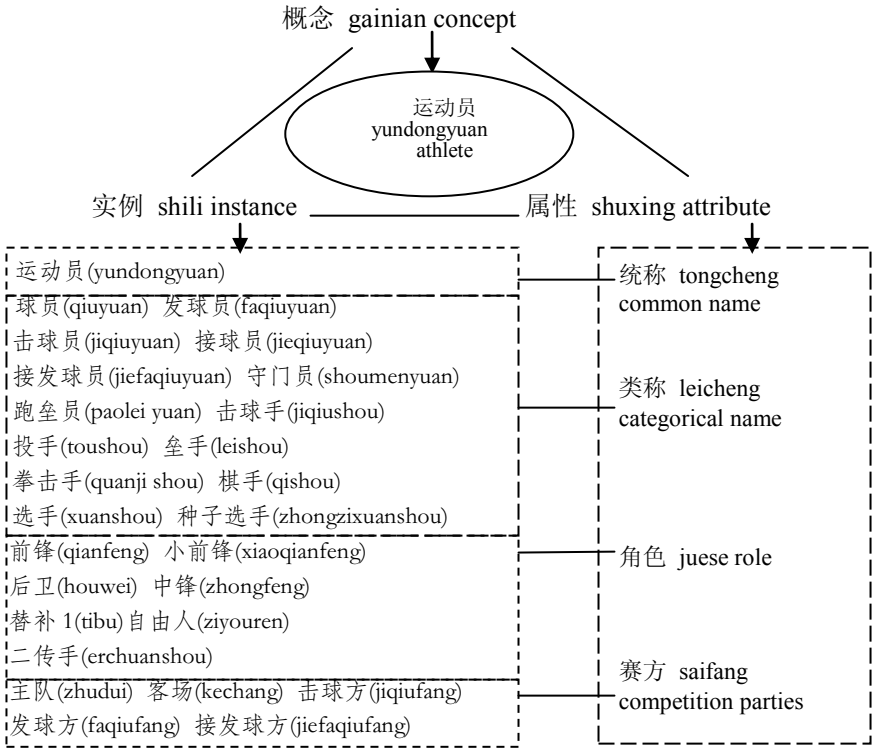


Fig. 1. The relationship between the node of [athlete] and its lower nodes from the perspective of the logic triples

Attribute is reusable. That is to say, attributes that appear in the upper node can also appear in the lower nodes. For example:

<athlete, role, baseman >< baseman, role, first baseman >< baseman, role, second baseman >< baseman, role, third baseman >

The construction of the Ontology in the field of sports events can be achieved by using it. For example:

[athlete] —— {role} —— [baseman] —— {role} —— [first baseman]
[second baseman][third baseman]

6 Intelligent Applications on Ontology in the Field of Sports Events

6.1 Improving the Level of Intelligent Retrieval in the Field of Sports Events

At present, there are three types of information retrieval: text retrieval, data retrieval and knowledge retrieval. In dealing with the sport category, the Ontology in the field of sports events has a much better concept hierarchy and supports performance for

logical reasoning, so it can be widely used in the intelligent retrieval for sport information and knowledge.

The following is a true text selected randomly from the webpage[1]:

6日晚，C罗头顶脚踢完成了本赛季西甲联赛的第4次“戴帽”，皇马7比1横扫奥萨苏纳。而算上曼联时代，C罗职业生涯的帽子戏法数已提高到了13次。C罗本赛季的联赛进球数提升到了13个，他用进球为自己刚刚拿到的欧洲金靴奖庆祝。

6 ri wan, C luo toudingjiaoti wancheng le ben saiji xijia liansai de di 4 ci daimao, huangma 7 bi 1 hengsao aosasu'na. er suanshang manlian shidai,C luo zhiye shengya de maozi xifa shu yi tigao dao le 13 ci. C luo ben saiji de liansai jinqiushu tisheng dao le 13 ge, ta yong jinqiu wei ziji ganggang nadao de ouzhou jinxue jiang qingzhu.

In this paper, the translation of this text is as follows, and we will not translate it again when referring to some words and expressions therein.

On the evening of (November) 6th, C Luo (Cristiano Ronaldo) won his forth hat-trick of the La Liga this season with his head and foot, and Real Madrid beat Osasuna by 7 to 1. Counting the Manchester United era, the number of C Luo's hat-tricks in his career has been increased to 13. C Luo's has gained 13 marks this season, and he just celebrated his European Golden Shoe with the last shoot.

The Internet user could usually enter the keywords or sentences for search:

6日晚本赛季西甲联赛谁进球？

6 ri wan ben saiji xijia liansai shui jinqiu?

Who scored the La Liga this season on the evening of (November) 6th?

6日晚西甲联赛哪个球队获胜？

6 ri wan ben saiji xijia liansai na ge qiudui huosheng?

Which team won the La Liga this season on the evening of (November) 6th?

Presently, the major search engines are still equipped with the keywords searching method. So the user often only enters the keywords of “进球 jinqiu goal”或“西甲联赛 xijialiansai La Liga”. Obviously, from the text above, the user cannot retrieve the correct information needed. But if the user makes full use of the Ontology in the field of sports events, he will obtain the information needed.

Through the structure tree above, we find that the node of [戴帽(daimao)] is the lower node of [进球(jinqiu)], and the machine/program will judge that“戴帽(daimao)” means“进球(jinqiu)”. We also find that [横扫(hengsao)] is the synonymy node of [获胜(huosheng)], and the machine will judge that“横扫(hengsao)” means“获胜(huosheng)”. Then it will give the answer as follows to the user through a series of computations.

6日晚本赛季西甲联赛谁进球？→ 6日晚本赛季西甲联赛C罗进球。

→ 6 ri wan ben saiji xijia liansai C luo jinqiu.

→ On the evening of (November) 6, C Luo goaled in the La Liga season.

6日晚西甲联赛哪个球队获胜？→6日晚西甲联赛皇马获胜。

→ 6 ri wan ben saiji xijia liansai huangma huosheng.

→ On the evening of (November) 6, Real Madrid beat the La Liga.

[1] See http://news.163.com/11/1107/14/7I919TO200014AED.html

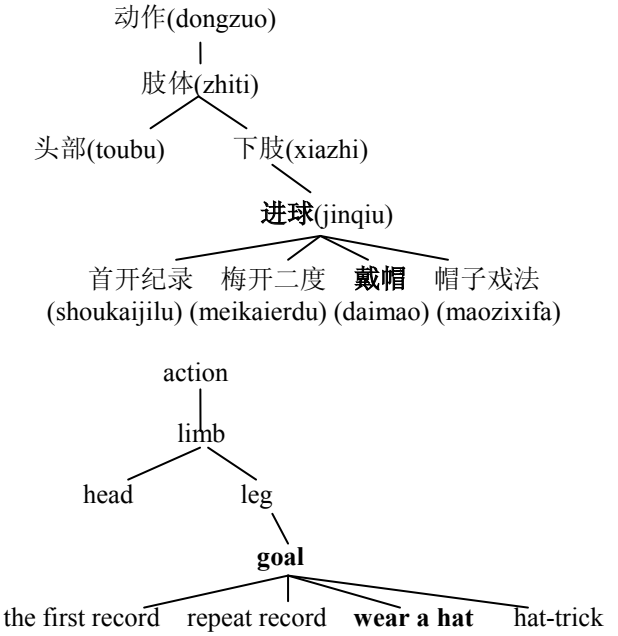


Fig. 2. The structural tree of the nodes of [动作 dongzuo action]

6.2 Promoting the Depth of Text Understanding in the Field of Sports Events

The event predicate can be regarded as a concept and has its attribute which can be granted a certain value. So each sentence can be expressed by the logic triples of <concept, attribute, instance> . For the event predicate, they all have the characteristic of carrying a particular argument, and the value of the characteristic is the predicate's semantic role of "agent", "patient, and so on. When we further analyze the text above, and find its structure can be expressed as follows:

<头顶脚踢(toudingjiaoti)，AGENT，C 罗(C luo)>
<进球(jinqiu)，AGENT，C 罗(C luo)>
<横扫(hengsao)，AGENT，皇马(huangma)>
<横扫(hengsao)，PATENT，奥萨苏纳(aosasu'na)>
<提高(tigao)，EXPERIENCE，帽子戏法数(maozixifashu)>
<获胜(huosheng)，AGENT，皇马(huangma)>
<提升(tisheng)，EXPERIENCE，进球数(jinqiushu)>
<庆祝(qingzhu)，AGENT，他(ta)>

Suppose the user continues to enter the following search request:
C罗为什么庆祝？
C luo wei shenme qingzhu?

Why did C Luo celebrate?

Because "他(ta)"和"C罗(C luo)" are of the co-referential relationship, "C罗(C luo)"和"皇马(huangma)" are of the controlled and controller relationship, a chain of the theme events has formed in Fig. 3 on the basis of the structure above.

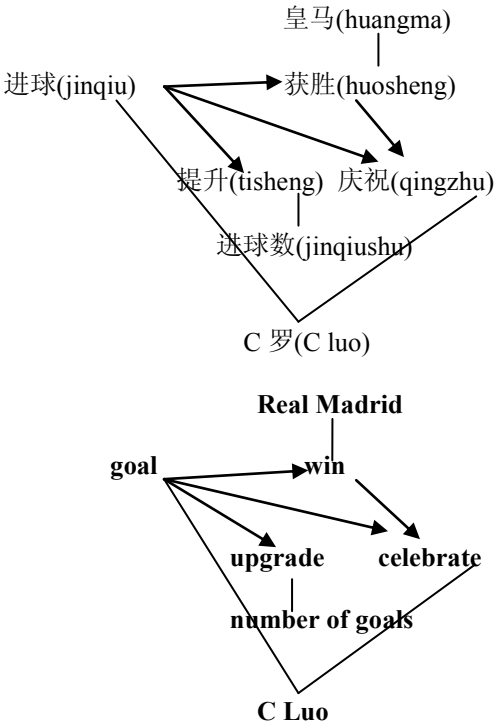


Fig. 3. The chain of the theme events

The chain of the theme events is effective for the computer to understand and to induce the web text information. The computer can judge that the central event of the text is "进球(jinqiu)", and that the other three events of "获胜(huosheng)", "提升(tisheng)" and "庆祝(qingzhu)" are all the chain reaction of the central events. Computer will return the result below:

C罗为什么庆祝？→因为C罗进球，皇马获胜，C罗进球数提升。

→ Yinwei C luo jinqiu, huangma huosheng, C luo jinqiu shu tisheng

→ Because C Luo goaled, Madrid won the match, and C Luo's goals increased.

6.3 Improving the Quality of Machine Translation in the Field of Sports Events

When the machine translates, we can mark each node of the Ontology in the field of sports events with a semantic code. Because Chinese and English use the same

semantic code, the corresponding relations between them will be clear. See another example below:

米利安每场比赛均是以替补出场。

Mili'an mei chang bisai jun shi yi tibu chuchang.

Miljan is a bench every game.

If we make the correct semantic code label in advance, then the machine would not translated "替补(tibu)"into "替补2(tibu)", which is used specially in "race" in sports events. So the only correct result is to translate "替补(tibu)"into "替补1(tibu)", which is used as "role" of sports events, and we get the result as shown in Fig. 4.:

体育赛事领域Ontology(tiyu saishi lingyu Ontology)

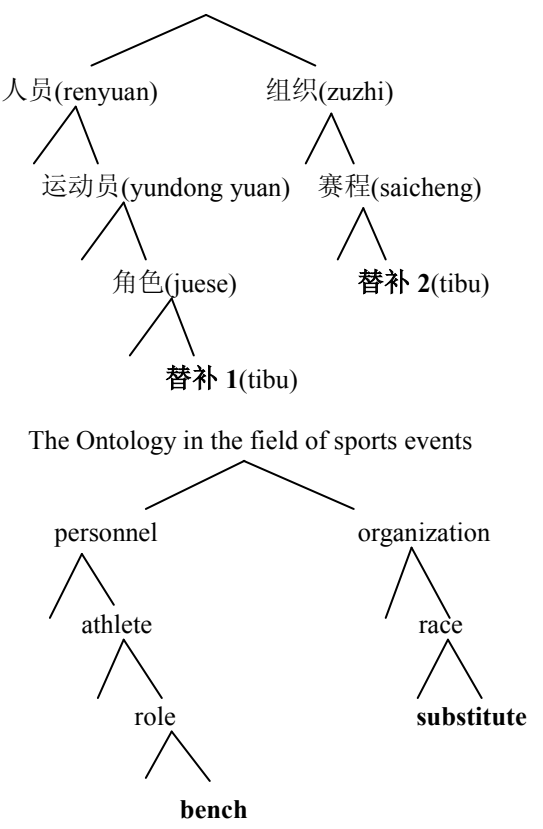


Fig. 4. Part of the Ontology in the field of sports events

6.4 Providing the Base of Web Information Sharing and Exchange in the Field of Sports Events

Ontology in the field of sports events is the base of sharing, and it exchanges the web information on the level of semantics. As we all know, XML and RDF not only provide grammar frameworks but also some semantic description. For example, the

XML fragment of <*Author* > *Jack* < / *Author* > shows that Jack is author, the RDF fragment of < *rdf : Description about* = *"http : //sports.sohu.com/ Home / Steven"* ><*s : Creator* > *Steven* < / *s : Creator* > < / *rdf : Description* > shows that Steven is the founder of the webpage of "http://sports.sohu.com/Home/Steven". From the point of <concept, attribute, instance>, the attribute scope of XML and RDF is lacking in specifications and restrictions, and is subject to change. For example, "Author" and "Creator" could be replaced by "Writer" in the text, and the result becomes <*Writer* > *Jack* < / *Writer* >< *rdf : Description about* ="*http* : *//sports.sohu.com/ Home / Steven*"><*s* : *Writer* > *Steven* < / *s* : *Writer* > < / *rdf* : *Description* >. In fact, for Ontology, "Author", "Creator" and "Writer" are of the same concept, and can share a same upper node.

For another example, the node of [instructor] is both appearing on the webpage of a basketball club website and a fitness equipment website, XML and RDF cannot well define it as "coach" or "equipment manual". At this time, the Ontology in the field of sports events can achieve disambiguation by judging that they should belong to different nodes, and the distance of nodes is rather long.

7 Outlook

The construction of Ontology in the field of sports is not a task that can be completed once for all. We should view it from the point of development because it has a bearing on the evolution of the Ontology. Therefore, it is a trend that the evolution of one Ontology in the field of sports could lead to another Ontology of the same domain.

References

1. Chen, W.: A study on the building of the sports register corpus's building and its event vocabulary. Journal of Nanjing Normal University, 38–41 (2007)
2. Mei, J.J.: Synonymy Thesaurus. Shanghai Dictionary Publishing House, Shanghai (1983)
3. Chen, A.H.: Sports Dictionary. Shanghai Lexicographical Publishing House, Shanghai (2000)

The Ordering of Mandarin Chinese Light Verbs

Chu-Ren Huang and Jingxia Lin

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong
churen.huang@inet.polyu.edu.hk, ctjlin@polyu.edu.hk

Abstract. Two or more light verbs are sometimes found to co-occur in both Taiwan and Mainland Mandarin Chinese, e.g., *jiāyǐ* and *jìnxíng* in *duì xuéshēng jiāyǐ jìnxíng yǐndǎo* 'to guide students', and a particular ordering is often preferred, e.g., *jiāyǐ jìnxíng* over *jìnxíng jiāyǐ*. This study argues that the order of the light verbs is closely associated to two kinds of information that the verbs specify, i.e. aspectual eventive information and argument information. On one hand, a light verb that denotes aspectual eventive information tends to occur before light verbs without such information. On the other hand, a light verb with less argument information is more likely to occur after other light verbs and closer to the event complement. The two semantic constraints form a general principle for the ordering of light verbs. The findings of this study can also contribute to a finer-grained classification of Chinese light verbs.

Keywords: ordering of serial light verbs, aspectual eventive information, argument information.

1 Introduction

Previous literature has not yielded a clear-cut and unified definition for the notion of "light verbs" (cf. [1-3]; among others). In Chinese as well, studies differ from each other in terms of the number of light verbs that Chinese has, cf. [4-7] and others. Taking a "loose" but comprehensive perspective, this paper defines light verbs as semantically impoverished verbs that may contribute information about event shape (e.g., beginning or ending of an event), but specify little about the kind of event under description (cf. [5]; [7]; among others). The event, i.e. the predicative content of a light verb construction, mainly comes from the event-denoting element that is taken as complement by the light verb. For instance, in the construction *jìnxíng tǎolùn* proceed-discuss 'discuss', the event of discussion is denoted by the complement *tǎolùn* 'discuss', whereas the light verb *jìnxíng* 'proceed' indicates a process aspect of the event. With such a definition, the light verbs discussed in this paper not only include the typical ones such as *jiāyǐ* 'inflict', *jìnxíng* 'proceed', and *zuò* 'do', but also the less discussed verbs such as *kāishǐ* 'start', *jiéshù* 'finish', and *jìxù* 'continue'.

In both Mainland and Taiwan Mandarin, two or even more light verbs may occur together in a complex predicate; such co-occurrences are not abundant, but they are sometimes found in governmental or institutional documents where the language is expected to be formal and standard, as in (1).

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 728–735, 2013.

(1) a. *Rénmínbì huìlǜ tiáozhěng bìxū jǐnshèn, yǒu bùzhòu,*
RMB exchange.rate adjustment must cautiously have step
jiànjìnde ***jiāyǐ jìnxíng***
gradually inflict proceed
'Adjustment of RMB exchange must be carried out in a cautious, step-by-step, and progressive way.' (http://www.gov.cn/zlft2011/zbzy.htm)

b. *Yòu'éryuán jiāng lùxù* ***kāishǐ jìnxíng*** *bàomíng gōngzuò*
kindergarten will successively start proceed registration work
'The kindergartens will begin registration successively.'
(http://jtjy.e21.edu.cn/content.php?id=9362)

Table 1. Google tokens of light verb sequences (2012-4-10)[1]

Light verb sequence	**Mainland**	**Taiwan**	**Example**
jiāyǐ jìnxíng inflict proceed	10,300	8,900	*jiāyǐ jìnxíng kòngzhì* inflict proceed control
jìnxíng jiāyǐ proceed inflict	8,060	1,500	*jìnxíng jiāyǐ chóngxiū* proceed inflict rebuild
jìnxíng zuò proceed do	523,000	35,400	*jìnxíng zuò wèijìng jiǎnchá* proceed do gastroscopy
zuò jìnxíng do proceed	285,000	16,200	*zuò jìnxíng jiǎyá xiūfù zhìliáo* do proceed denture repair treatment
jiāyǐ zuò inflict do	19,200	8,700	*jiāyǐ zuò gèzhǒng wéitiáo* inflict do all.kinds.of fine-tuning
zuò jiāyǐ inflict do	4,670	1,070	*zuò jiāyǐ qūbié* do inflict distinguish
kāishǐ jiāyǐ start inflict	55,700	7,330	*kāishǐ jiāyǐ zhěngdùn* start inflict rectify
jiāyǐ kāishǐ inflict start	1,310	157	*jiāyǐ kāishǐ jiǎnchá* inflict start inspect
kāishǐ jìnxíng start proceed	10,400,000	2,380,000	*kāishǐ jìnxíng bǐsài* start proceed game
jìnxíng kāishǐ proceed start	83,000	38,800	*gǎizào xiūshàn gōngchéng jìnxíng kāishǐ* reform repair project proceed start
kāishǐ zuò start do	18,200,000	2,320,000	*kāishǐ zuò shíyàn* start do experiment
zuò kāishǐ do start	18,300	18,800	NA

[1] Although the token numbers are based on raw data where invalid examples are also counted, the Mainland and Taiwan data consistently shows that a preferred ordering exists for each pair of the light verbs.

More importantly, Internet data shows that the light verbs tend to follow some preferred orders when they co-occur. Table 1 presents raw data from the Google searches for the co-occurrences of *jiāyǐ* 'inflict', *jìnxíng* 'proceed', and *zuò* 'do', and *kāishǐ* 'start' in Mainland and Taiwan websites; it shows that in both varieties of Chinese, although examples of both orders are found when two light verbs co-occur, a particular order is often preferred over the other: *jiāyǐ jìnxíng* inflict proceed, *jìnxíng zuò* proceed do, *jiāyǐ zuò* inflict do, *kāishǐ jiāyǐ* start inflict, *kāishǐ jìnxíng* start proceed, and *kāishǐ zuò* start do are all more frequently used than the reversed orders.

Furthermore, the frequencies in Table 1 give rise to an ordering hierarchy for the four light verbs, as given in (2).

(2) *kāishǐ* 'start' > *jiāyǐ* 'inflict' > *jìnxíng* 'proceed' > *zuò* 'do'

The hierarchy illustrates that when two of the four light verbs co-occur, their left-to-right order tends to follow the hierarchy from left to right, i.e. "*kāishǐ* 'start' + *jiāyǐ* 'inflict'/ *jìnxíng* 'proceed'/ *zuò* 'do'", "*jiāyǐ* 'inflict' + *jìnxíng* 'proceed'/ *zuò* 'do'", and "*jìnxíng* 'proceed' + *zuò* 'do'".

The orderings presented in Table 1 as well as the hierarchy in (2) indicate that there are internal differences among Chinese light verbs, and such differences may not only provide an explanation to the order of the light verbs, but also shed light on a better characterization and classification of the verbs. Therefore, a study is necessary to discover the factors that bring about the differences and the ordering.

2 Semantic Properties Determining the Order of Chinese Light Verbs

This study proposes that the relative order of Chinese light verbs is closely associated to two basic semantic properties of the verbs: aspectual eventive information and argument information. Aspectual eventive information refers to the information about the overall (viewpoint) aspectual shape of event structure, including inception, process, completion, and so on. Argument information refers to the information about the arguments or participant roles involved in the event, e.g., agent, goal, location, theme, and so on. Eventive information and argument information are provided with different representation in some theories of verbal semantics. For instance, according to the MARVS (Module-Attribute Representation of Verbal Semantics, [8]), eventive information and argument information are represented as Event Module and Role Module, respectively. Furthermore, more refined specifications on a particular event or argument are represented as Event-internal Attributes and Role-internal Attributes, which thus enables the differentiation of synonyms even when they share the same Event/Role Modules. The MARVS is able to provide a succinct but critical description of verbs, especially near synonyms. For instance, *tóu, zhì, rēng,* and *diū* all denote an event of throwing, but [9] argue, within the framework of MARVS, that these verbs can be differentiated based on their aspectual and argument information: in terms of aspectual information, *diū* 'throw' specifies a bounded process, whereas all other three verbs only denote activities with a starting point (Event Module); in terms

of argument information, *diū* 'throw' and *rēng* 'throw' do not specify goal information, whereas *tóu* 'throw' and *zhì* 'throw' specify direction and goal, and thus can take spatial NPs as direct object (Role Module); in addition, *tóu* 'throw' and *zhì* 'throw' can be further differentiated that the former specifies the kind of goal associated to the event, as it often collocates with container-like locative NPs (Role-Internal Attributes). The MARVS demonstrates the significance of the eventive and argument information that a verb specifies in determining the verb's syntactic distribution and behavior.

The rest of this section discusses the eventive and argument information specified by Chinese light verbs, and proposes the principle that determines the order of co-occurring light verbs.

2.1 Eventive Information of Chinese Light Verbs

It has been found in many languages that light verbs may contribute viewpoint aspectual information to the event described by the light verb construction ([10-11]; among others). Different viewpoint aspectual systems have been proposed by previous studies (cf. [8]; [11-13]), and this paper follows [8] for Chinese light verbs. [8] identify five atomic event structures ---- boundary, punctuality, process, state, and stage ---- and argue that these atomic event structures can combine and generate twelve event types with different degree of complexity. In Chinese, *kāishǐ* 'start', *jìxù* 'continue', *jiéshù* 'finish', and *jìnxíng* 'proceed' are among the light verbs that denote aspectual meanings. For instance, *kāishǐ* 'start' specifies inception. Evidence can be found with the types of complements that *kāishǐ* 'start' takes. For example, both *shuìzháo* 'fall into sleep' and *dǒngde* 'get to understand' are understood as instantaneous events, but *kāishǐ* 'start' is only compatible with *dǒngde* 'get to understand'. The incompatibility of *kāishǐ* 'start' and *shuìzháo* 'fall into sleep' is probably because *kāishǐ* 'start' denotes an inceptive meaning, but an event of *shuìzháo* 'fall into sleep', in contrast to an event of *dǒngde* 'get to understand', usually does not have a specific starting point (cf. [10]). Unlike *kāishǐ* 'start', the light verbs *jìnxíng* 'proceed' and *jìxù* 'continue' specify a process. For example, both *jìnxíng* 'proceed' and *jìxù* 'continue' are compatible with the continuative marker *xiàqù*, as in *bǐsài jìxù xiàqù* 'the game continues' and *diàochá jìnxíng xiàqù* 'the investigation continues', cf. *kāishǐ* 'start' that is rarely found with *xiàqù*.

We propose that when two light verbs occur together, the one with an aspectual meaning occurs before the verb without such a meaning. For instance, *jiāyǐ* 'inflict' does not clearly specify any aspectual information, as evidenced by the fact that it usually cannot be suffixed with aspectual markers (??*jiāyǐ-le* 'perfective'/*zhe* 'durative'/*guò* 'experiential'), so *jiāyǐ* 'inflict' is preceded by kāishǐ 'start' when they co-occur. Furthermore, if both light verbs specify some aspectual meanings, their order is determined by the logical relation of the aspectual meaning. For instance, although both *kāishǐ* 'start' and *jìnxíng* 'proceed' denote aspectual information, *kāishǐ jìnxíng* start proceed is preferred over *jìnxíng kāishǐ* proceed start because the former is consistent with the logical order that the inception comes before the process.

Nonetheless, the constraint of aspectual information is unable to determine all orderings that are found in Chinese light verbs. For instance, the constraint predicts that *jìnxíng* 'proceed' will occur before *jiāyǐ* 'inflict' as the former contains aspectual information, but as shown in Table 1, *jiāyǐ jìnxíng* inflict proceed is more frequently used than *jìnxíng jiāyǐ.* Therefore, some additional constraint is needed. Section 2.2 introduces the new constraint.

2.2 Argument Information of Chinese Light Verbs

As introduced in Section 1, light verbs are usually devoid of concrete meanings, and thus the information about the event under description usually comes from the complements taken by the light verbs. In Mandarin Chinese, the event-denoting complements can be deverbal nominals or event NPs, as in *jìnxíng dì yī cì huìyì* 'to have the first meeting' and *jìnxíng diàochá* 'to investigate', respectively. However, as observed in many previous studies, light verbs are semantically bleached to different degrees, and thus differ in terms of the types of complements they can take. Table 2 is a comparison of the semantic and syntactic features of the complements to *jìnxíng* 'proceed' and *jiāyǐ* 'inflict' based on previous studies ([7]; [14-15]; among many others) and corpus data.[2]

Table 2. Semantic and syntactic features of the event complements to *jìnxíng* 'proceed' and *jiāyǐ* 'inflict' (2012-4-14) ("?" marks that the use is unnatural)

Event complements		***jìnxíng* 'proceed'**	***jiāyǐ* 'inflict'**
Semantic features	Formal and spontaneous event (*fēnxī* 'analyze')	√	√
	Durative event (*fēnxī* 'analyze')	√	√
	Event involving interaction of the agent and patient (*gōutōng* 'communicate', *hézuò* 'cooprate')	√	??
Syntactic features	Transitive verb (*kòngzhì* 'control')	√	√
	Intransitive verb (*xiāoshòu* 'sell')[3]	√	???
	NP (*dìyī xiàng yìchéng* 'The first item on the agenda')	√	????
	VP (*wánshàn fúwù* 'improve service')	????	????
	Event complement at subject position (*huìyì zhèngzài jìnxíng* 'the meeting is in progress')	√	????

[2] Regional variations can be found in Taiwan and Mainland *jìnxíng* 'proceed' and *jiāyǐ* 'inflict', but they are not discussed in this paper as the differences are relatively minor.

[3] *Xiāoshòu* 'sell' is usually used as an intransitive verb in Mandarin Chinese, cf. *mài* 'sell' which is a transitive verb.

Table 2 shows that the event complements in *jiāyǐ* 'inflict' constructions are usually transitive and the events they describe do not involve any interaction between the agent and the patient, e.g., *gōutōng* 'communicate' and *hézuò* 'cooprate'. This indicates that *jiāyǐ* 'inflict' still maintains some of its literal meanings, that is, the event associated with *jiāyǐ* 'inflict' is usually "imposed" by an agent onto the patient. In contrast, *jìnxíng* 'proceed' is more semantically bleached as it selects a larger variety of complements.

In a double-light-verb construction, however, only the selectional restriction of the second light verb, i.e. the verb closer to the complement, must be satisfied. For instance, Table 3 is a summary of the semantic and syntactic features of the complements to *jiāyǐ jìnxíng* inflict proceed from Mainland and Taiwan Internet data (200 Baidu examples for each variety); the table shows that the types of complements to *jiāyǐ jìnxíng* inflict proceed are very similarly to those to *jinxing* 'proceed', cf. Table 2. For instance, intransitive verbal complements and complements denoting events involving interaction of the agent and patient are still found in the constructions despite the presence of *jiāyǐ* 'inflict', cf. Table 2.

Table 3. Semantic and syntactic features of the event complements to *jiāyǐ jìnxíng* inflict proceed (2012-4-14)

Event complement		***jiāyǐ jìnxíng* inflict proceed**	
		Mainland	**Taiwan**
Semantic features	Formal and spontaneous event (*fēnxī* 'analyze')	√	√
	Durative event (*fēnxī* 'analyze')	√	√
	Event involving interaction of the agent and patient (*miànduìmiàn de jiāoliú* 'face-to-face communication', *tǎolùn* 'discuss')	√	√
Syntactic features	Transitive verb (*kòngzhì* 'control')	√	√
	Intransitive verb (*tuīlǐ* 'deduce', *pèiduì* 'make a pair')	√	√
	NP (*yǒuxiào de tiáojié* 'effective regulation')	√	√
	VP (*Pínglián xuéqí chéngjī* 'assess semester academic grades')	NA	√
	Event complement at subject position (*Zīyuán de pèizhì jīběn shàng tōngguò shìchǎng jiāyǐ jinxing* 'Allocation of resources is usually carried out through the market')	√	√

Therefore, we propose that in order to optimally satisfy the selectional restriction between the light verb and the complements, the light verb that allows a larger variety of complements, i.e. the verb that is more bleached, tends to occur closer to the complement in a double-light-verb construction. In this sense, *jìnxíng* 'proceed' is expected to occur after *jiāyǐ* 'inflict' although it is more specific than *jiāyǐ* 'inflict' in terms of aspectual information, cf. Section 2.1.

2.3 The Ordering Principle

Sections 2.1 and 2.2 introduce the two important semantic properties of Chinese light verbs and their effect on the ordering of light verbs: (a) a light verb with eventive information tends to occur at the beginning of a light verb construction, and light verbs with different aspectual information follows the logical order of the aspect meaning; (b) a light verb that is more semantically bleached tends to occur closer to the complement. The two semantic constraints complement each other and form a general principle for the ordering of light verbs.[4]

The ordering principle is able to account for the hierarchy in (2): on one hand, *kāishǐ* 'start' precedes the other three light verbs because it specifies an inceptive meaning; on the other hand, although *jìnxíng* 'proceed' is also associated with an aspectual meaning, it is relatively more bleached than *jiāyǐ* 'inflict' and thus tends to follow *jiāyǐ* 'inflict', and it is relatively less bleached than *zuò* 'do' and thus tends to precede *zuò* 'do'. Furthermore, although the co-occurrence of three light verbs is rarely found in the corpus data, the principle is able to provide a unified account for their relative order. As illustrated in (3), the order of *kāishǐ* 'start', *jìnxíng* 'proceed', and *zuò* 'do' is consistent with the hierarchy in (2).

(3) a. *Yǐ yǒu zhìshǎo liǎngjiā* ***kāishǐ jìnxíng zuò***
already have at.least two.Classifier start proceed do
shì shāng jiāoyì de yánjiū hé chóubèi
city business trade Modifier research and preparation
'There are at least two brokerage firms who have begun the research and preparation of city business transaction.'
(http://archive.news.sina.com.tw/破冰-推進-做市商-441910)

b. *Gè zǔ tóngxué tǎolùn yīxià, tǒngyī-le yìjiàn,*
each group fellow.student discuss a.little.bit agree-Perfective opinions
zài ***kāishǐ jìnxíng zuò*** *shíyàn*
then start proceed do experiment
'Students of each group should have a discussion to reach an agreement, and then begin the experiments.'
(http://www.edudown.net/teacher/jiaoan/xqita/200607/7923.html)

3 Conclusion

This preliminary study investigated the relative order of co-occurring light verbs in both Mainland and Taiwan Mandarin Chinese. It proposed that the ordering is co-determined by the eventive information and argument information that the light verbs specify. The proposal of this study will also shed light on a better classification of

[4] Note that in Table 1, the frequency difference between *jiāyǐ jìnxíng* inflict proceed and *jìnxíng jiāyǐ* proceed inflict is much larger in Taiwan Chinese than in Mainland Chinese. This indicates that the constraint of argument information may play a more important role in Taiwan Chinese, but due to space reason, regional variations are not discussed in this paper.

Chinese light verbs. In future studies, we will provide a more detailed representation of the verbal semantics for each Chinese light verb in order to achieve a finer-grained classification. In addition, a more comprehensive comparison of Mainland and Taiwan light verbs will be carried out in order to describe the possible regional differences for better cross-strait communication.

Acknowledgments. We are very grateful to Ge XU for his help with the extraction of the Baidu data. This work was supported by PolyU project 1-ZV8E.

References

1. Jespersen, O.: A Modern English Grammar on Historical Principles. Allen & Unwin, London (1942)
2. Grimshaw, J., Armin, M.: Light verbs and θ-marking. In: Linguistics Inquiry, pp. 205–232 (1988)
3. Butt, M., Geuder, W.: On the (semi)lexical status of light verbs. In: Corver, N., van Riemsdijk, H. (eds.) Semi-lexical Categories, Mouton De Gruyter, Berlin (2001)
4. Yin, S.C.: Shilun Nianzhuo Dongci (A tentative study on bound verbs). In: Zhongguo Yuwen (1991)
5. Zhu, D.X.: Xiandai Shumian Hanyu de Xuhua Dongci he Mingdongci (Light verbs and verbal nouns in Modern literary Chinese). In: Selected Papers of Zhu Dexi. Commercial Press, Beijing (1999)
6. Li, L.D.: Xiandai Hanyu Dongci (Modern Chinese verbs). China Social Science Press, Beijing (1990)
7. Diao, Y.B.: Xiandai Hanyu Xuyi Dongci Yanjiu (A study on Modern Chinese abstract verbs). Liaoning Normal University Press, Shenyang (2004)
8. Huang, C.-R., Ahrens, K., Chang, L.-L., Chen, K.-J., Liu, M.-C., Tsai, M.-C.: The module-attribute representation of verbal semantics: From semantics to argument structure. In: Computational Linguistics and Chinese Language Processing, pp. 19–46 (2000)
9. Liu, M.-C., Huang, C.-R., Lee, C., Lee, C.-Y.: When endpoint meets endpoint: A corpus-based lexical semantic study of Mandarin verbs of throwing. In: Computational Linguistics and Chinese Language Processing, pp. 81–96 (2000)
10. Butt, M.: The Structure of Complex Predicates in Urdu. CLSI, Stanford (1995)
11. Allerton, D.J.: Stretched Verb Constructions in English. Routledge, London (2002)
12. Comrie, B.: Aspect. Cambridge University Press, Cambridge (1976)
13. Smith, C.S.: The Parameter of Aspect. Kluwer Academic Press, Dordrecht (1991)
14. Lü, S.X.: Xiandai Hanyu Babai Ci (800 words in Modern Chinese). Commercial Press, Beijing (1980)
15. Lu, F.B.: Duiwai Hanyu Changyong Ciyu Duibi Lishi (Comparative illustrations of common Chinese words and expressions). Beijing Language and Culture University, Beijing (2000)

Negation and Double-Negation of Chinese Oppositeness

Jing Ding and Chu-Ren Huang

Chinese & Bilingual Studies, The Hong Kong Polytechnic University,
Hong Kom, Kowloon, Hong Kong
amanda.ding@connect.polyu.hk
churen.huang@inet.polyu.edu.hk

Abstract. Oppositeness refers to the paradigmatic relationship of two words holding the contrast meanings. The fact that linguistic opposite differs from logical contrast has been discussed in theories, but has not been tested in practice. In this paper, we investigate three main subtypes of Chinese oppositeness, via using logical tests of negation and double-negation. Result shows that purely logical test does not always work on oppositeness, and indicates that contrast relations within different pairs also vary.

Keywords: oppositeness, negation, double-negation.

1 Introduction

As one of the fundamental paradigmatic sense relations, oppositeness is very common in everyday language use, and has been discussed in numerous semantic literatures. However it is easy to notice that the opposite relationships in language is different from the ones of logical contrast pairs: entailments such as negation work well for the later but might not be applicable to the former in cases like *happy: angry*, or *buy: sell*. But for opposite pairs like *dead: alive*, the logical entailments seem to be applicable. Hence, it is possible to guess that the different subtype opposites actually hold different kinds of contrastive relationships, which may at least be partly revealed by negation and double-negation from the aspect of logical tests.

The rest part of this paper is organized as following: Part Two is a literature review on how oppositeness is determined in general and how it is categorized into several subtypes in both English and Chinese; Part Three tests the negation and double-negation of some typical subtypes of Chinese oppositeness; Part Four describes the results and analyses the reason of their different performance in the tests; Part Five summaries the work of this paper and suggests future work.

2 Oppositeness in General and Its Main Varieties

2.1 Oppositensss in General

Oppositeness, sometimes also known as antonymy, is defined as the two members of a lexical pair holding the contrast meanings. After traditional categorization [7], the

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 736–744, 2013.

term antonymy is restricted to only the meaning contrasting pairs which does not exclusively dichotomy the domain, or, more precisely, the pair of "gradable, directionally opposed" ones [4]. And at the same time, oppositeness (roughly equal to opposition) is selected as the most general term in his and others' later work (for example, [2]). A natural way of finding the opposite is by asking the question: "What is the opposite of ...?" On the other hand, linguists with a more radical point might assert that any word pairs having the meaning difference(s) would be theoretically possible to be opposites. However, we adopt the view that the words "appear paradoxical" [2] would be called opposites in general.

Lyons [7] leaves an open answer to the wondering whether it is a universal human tendency to have the experience dichotomize or polarize in a two-word-pair. But it is very often to have the opposite relation holding between only two members, and any more-than-two clusters would be felt less canonity, such as *black: white: grey*. Hence, in this paper, we inherently focus on the two-word-pairs.

2.2 Main Varieties of Oppositeness

Any native speaker of English will naturally feel that the way of how *dead* contrasts to *alive* is not the same as the one of *buy* and *sell*, *up* and *down*, *wife* and *husband*, and so on. Naturally, under the general definition there are many different subtypes of oppositeness.

Lyons and Cruse agree that for the basic distinction within oppositeness is whether they are gradable and ungradable. The ungradable opposites are termed as "complementaries" [2, 6], or "binary antonyms"[1], while the gradable ones are called antonyms [2, 7]. For the rest opposites, Lyons [7] defines converse as these pairs like *buy: sell*, *husband: wife* and directional opposite for *come: go*, *up: down*, respectively. Cruse [2, 3] defines converse as a relational opposite to "express a relationship between two entities by specifying the direction of another along some axis", among other relational opposites.

Lyons' categorization [7] for opposition stops at the relatively early stage, distinguishing only the types of contrast, complementary, antonymy and converse. Cruse [2] further extends the distinction to several sub-subtypes, such as restitutive, interactive, satisfactive, counteractive as the subtypes of complementary, and so on. Others' work, may not strictly follow their definitions, almost all agree on the main subtypes of complementary, antonymy and converse (for example, [1, 8, 9]).

2.3 Chinese Oppositions

As to the studies of Chinese opposites, Liu [10] points out that there are three subtypes of opposite pairs: complementary, such as *dead: alive*; converse, such as *buy: sell*; and, directional opposition, such as *up: down*. Later, Liu and Zhou [11] add polar oppositeness, such as *cold: hot*, as one subtype of Chinese opposites. The definitions for both complementary and converse in their work are the same as these of English; while pairs of *up: down* and *cold: hot* are examples of antonyms in traditional categorizations.

From the above, we may see that there are various kinds of oppositeness in natural language. For this paper, we select three most clearly defined subtypes, that is, complementary, antonym (or gradable opposite) and converse, to be examined in the later tests.

3 Negation and Oppositeness

3.1 Negation in Classical Logics

In Classic Logics, for a certain domainα, for the pair of A and B, if:

$A = \overline{B}$, $B = \overline{A}$

and, $A \cup B = U$, $A \cap B = \phi$

then:

1) the negation of A goes to B, and the negation of B goes to A;
2) the double-negation of A goes back to A, and the double-negation of B goes back to B.

In this paper, the first statement is called the negation of A or B, and the second one is called the double-negation of A or B.

3.2 Negation in Language Using

The question on whether opposition in language functions the same as the logical opposition has raised a long time discussion (for example, [2, 7, 8]). In Classic Logics, for the contradictory pair of A and B, the negation of A goes to its opposite point and the double-negation of A goes back to A directly. Furthermore, Aristotle distinct contrary from contradictory in logics: "the negation of one predication entails its contradictory", like *true: false*, *red: not red*; at the same time, "the assertion of one predicate entails the denial of its contrary, but in which both contraries may be false", like *red: green*, *big: small* [5].

As Lyons [7] correctly points out, “[t]he distinction of contradictories and contraries corresponds to the distinction of ungradable and gradable lexemes within the class of opposites in a language, but it applies more widely; and the fact that gradable antonyms can generally be taken as contraries, rather than contradictories, is a consequence of gradability, not its cause.” The relations between the opposite pair members are similar to logical negation but not necessarily follow the logical negation and double-negation rules. Hence, the negation and double-negation tests, which are used for logical contradictories, only work on the case of complementary pairs in language opposition.

4 Negation and Double-Negation Tests

The approach adopted here is to the negation and double-negation to test three subtypes of oppositeness, in order to compare their behavior in natural language

using. The selected examples of the subtypes are the most cited ones in previous studies. And, to avoid the possible ambiguity caused by syntactic structures, we use the most direct and simple way of having negation and double-negation, that is, to have the negation sentences translated with negator of NOT and double-negation with NOT NOT.

Now let's see how the purely logic assumption works for Chinese pair members of complementary, antonym and converse relations.

4.1 Negation Tests

a). (complementary) 死：活 (si 3: huo 2, dead: alive)

So for each member of the pair, we can have statements like:

小明死了。 And, 小明活着。

Xiao-ming si le Xiao-ming huo zhe

Xiao-ming dead LE Xiao-ming alive ZHE

Xiao-ming (is) dead. Xiao-ming (is) alive.

For them, the negations are:

小明没有死。 ＝ 小明还活着。

Xiao-ming mei you si Xiao-ming hai huo zhe

Xiao-ming not dead Xiao-ming still alive ZHE

Xiao-ming (is) not dead. Xiao-ming (is) still alive.

小明没有活下来。 ＝小明死了。

Xiao-ming mei you huo xia lai Xiao-ming si le

Xiao-ming not alive down Xiao-ming dead LE

Xiao-ming (is) not alive. Xiao-ming (is) dead.

For the pair of *si: hou*, the negation of one goes to the other of the pair. In other words, when *Xiao-ming is dead* is negated, and then it should mean that *Xiao-ming is (still) alive*. Meanwhile, the negation of *Xiao-ming is alive* only implies the one that *Xiao-ming is dead*. However, in real language using, it is also natural to have the phrases like *ban-si-bu-huo* (half-dead-half-alive), which means still alive but exhaust of energy. The negation of *huo* in this phrase does not have the equal meaning of *si*, because in such case the complementary opposite of *si: hou* coerces to being gradable, which will be explained in section 4.3.

b). (antonym) 高兴：伤心 (gao 1 xing 4: shang 1 xin1 , happy: sad)

For statements like:

她高兴。And, 她伤心。

ta gao xing ta shang xin

she happy she sad

She is happy. She (is) sad.

We will have their negated statements like:

她不高兴。＝？她伤心。 Or, ＝? 她平静。
ta bu gao xing ta shang xin ta ping jing
she not happy she sad she quiet
She (is) not happy. She (is) sad. She (is) quite.

她不伤心。 ＝她高兴。 Or, = ? 她平静。
ta bu shang xin ta gao xing ta ping jing
she not sad she happy she quiet
She (is) not sad. = She (is) happy. She (is) quite.

The negation of *gao-xing* could be understood in different ways, that it, there are more than one utterances implied from the negated sentences of *She is not happy*, and vice versa.

c). (converse) 买：卖 (mai 3: mai 4, buy: sell)

For statement:

张三买了一辆车。
Zhang-san mai le yi liang che
Zhang-san buy LE one liang (measure word) car
Zhang-san brought a car.
Its negated sentence is:

张三没有买一辆车。 ≠ 张三卖了一辆车。
Zhang-san mei you mai yi liang che Zhang-san mai le yi liang che
Zhang-san not buy one liang car Zhang-san sell LE one Liang car
Zhang-san did not buy a car. Zhang-san sold a car.

Similarly, for statement of:

李四买了一辆车。
Li-si mai le yi liang che
Li-si buy LE one liang car
Li-si sold a car.

Here comes its negative sentence:

李四没有买一辆车。 ≠ 李四卖了一辆车。
Li-si mei you mai yi liang che Li-si mai le yi liang che
Li-si not buy one liang car Li-si sell LE one liang car
Li-si did not buy a car. Li-si sold a car.

Similarly with the pair of *gao-xing: shang-xin*, the negation of the member of this pair does not have only one understanding. So, again, the negation test fails in the pair of *mai: mai*.

4.2 Double-Negation Tests

a). (complementary) 死：活 (si 3: huo 2, dead: alive)

For the above original sentences, their double-negated sentences are:

小明不是没有死。		小明死了。
Xiao-ming bu shi mei you si		Xiao-ming si LE
Xiao-ming not not dead		Xiao-ming dead LE
Xiao-ming is not not dead.		Xiao-ming is dead.
小明不是没有活下来。	=	小明活下来了。
Xiao-ming bu shi mei you huo xia lai		Xiao-ming huo xia lai LE
Xiao-ming not not alive down		Xiao-ming alive down
Xiao-ming is not not alive.		Xiao-ming is alive.

The double-negations of both sentences go back to the original sentences, which means that the double-negation of the member equal to the meaning of the original members: NOT-NOT *dead* is the same to *dead*.

b). (antonym) 高兴：伤心 (gao 1 xing 4: shang 1 xin1 , happy: sad)

Still, for the above sentence pairs, their double-negated sentences like:

她不是不高兴。	＝?她很平静。Or, =?而是很不高兴。		Or, =?她高兴。
ta bu shi bu gao xing	ta hen ping jing	er shi hen bu gao xing	ta gao xing
she not not happy	she very quiet	but very unhappy	she happy
She is not not happy	She is quite.	But (she) is very unhappy	She (is) happy.

她不是不伤心。	＝?她很平静。Or, =?而是不很伤心。		Or, =?她伤心。
ta bu shi bu shang xin	ta hen ping jing	er shi bu hen shang xin	ta shang xin
she not not sad	she very quiet	but not very sad	she sad
She is not not sad	She is quite.	But (she) is not very sad	She (is) sad.

The double-negations to the pair member of *gao-xing: shang-xin* also has more than one meaning. Interesting we notice that, *NOT-NOT happy* can both mean *being happy*, or *being very unhappy*; on the other hand, *NOT-NOT sad* prefers to be read as *being sad* or *being not very sad*.

c). (converse) 买：卖 (mai 3: mai 4, buy: sell)

As above, we have the double-negation for the original statements like:

张三不是没有买一辆车。		张三买了一辆车。
Zhang-san bu shi mei you mai yi liang che		Zhang-san mai le yi liang che
Zhang-san not not buy one liang car		Zhang-san buy LE one liang car
Zhang-san did not not buy a car.		Zhang-san brought a car.
李四不是没有卖一辆车。	=	李四卖了一辆车。
Li-si bu shi mei you mai yi liang che		Li-si mai LE yi liang che

LI-si not not sell one liang car	Li-si sell LE one liang car
Li-si did not not sell a car.	Li-si sold a car.

When we double negate *buy* of this pair, the only possible reading is still *buy*, and the same for *sell*.

4.3 Binary and Negation

The negation and double-negation tests success on the complementary and converse pairs and, at the same time, fail to work on gradable antonym pairs such as *happy: sad*. The reason lies on the binarity of opposite pairs.

For some subtype of opposition, like complementary, the two members of the pair dichotomize the related domain; while some others , like the gradable antonym, the domain contains more than two members which may said to be held meaning contrasts, along certain scale or dimension. That is to say, for a normal circumstance, someone that is not *alive* should be *dead*, but the utterance of someone is not *happy* is not equal to that of someone is *sad*. So in the later example, other kinds of emotions can be used instead of being sad to contrast with being happy, or, equal to being not happy.

Even for the complementary pairs, it is necessary to identify the related domain this pair is applied to, in order to have the negation and double-negation rules work. A table or a chair, for example, is not applicable for the pair of *die: alive*. (cf. e.g., [7])

Also, in natural language using, it is not rare to have the complementary pairs, very often is one member of the pair, coerced into being gradable. Again, let's take *die: alive* for example.

Cruse [2] concludes that some complementary adjectives are not normally gradable, but points out that very often one member of a pair is more likely to be grading than the other, like: ?*very dead*, ?*moderately dead*, ?*deader than before*; but, *very alive*, *moderately alive*, *more alive than before* [2].

And, Murphy [9] agrees that "complementaries can sometimes be used as contraries, and contraries sometimes are used as complementaries", even for some gradable pairs, "denial of one is usually taken to be the assertion of the other" [9], like *Ari is not honest* normally entails *Ari is dishonest*, and *vice versa*. Actually, according to Cruse [2], he "solves this problem by maintaining that such words must have two senses, one in complementary opposition and the other in contrary opposition to its antonym." (cf. [2, 9])

Pairs like *buy: sell* offers another aspect of looking at the binarity of being an opposite pair. In the normal trading domain, there are, usually, only the contrast of buying and selling. Their meanings are converse because: for any *X* buys *Y* from *Z*, then it is true that, *Z* sells *Y* to *X*. However, the two-member-relation is not necessary enough to generate the implication like complementary pairs in the negation and double-negation tests. That is because, logically speaking, if *X* is not buying *Y*, then it is not to say *X* is selling *Y*---- it is also even possible for *X* to do nothing with *Y*. The two-member-relation between buy and sell is coinciding with the human tendency to

to categorize experience in terms of binary contrasts, as linguists (see, e.g., [1, 7, 9]) pointed out.

4.4 Summary

As we have seen above, the negations of complementary members *si* 3(die) and *huo* 2 (alive) assert the other one of pair, and the double-negations of each imply the assertions of themselves; for the antonym pair of *gao* 1 *xing* 4 (happy) and *shang* 1 *xin* 1(sad), the negation of does not necessary imply, and vice verse, and the double-negation of either equal to; in the case of *mai* 3(buy): *mai* 4(sell), which combine a converse pair, the negation of is not always, but the double-negation of still means, and vice versa.

The results can be generalized in the table below:

Table 1. Negation and double-negation results of three subtypes of opposition

	negation	double-negation
complementary	goes to its opposite	goes back to itself
antonym	not necessarily go to its opposite	not necessarily go back to itself
converse	not necessarily go to its opposite	goes back to itself

Hence in natural languages, like Chinese, a purely logical negation or double-negation test it fails to work in many cases. For example, the utterance of "the house is big" negates that of "the house is small", while the one of "the house is not big" does not necessarily imply that "the house is small". Also, the negation of "prefect" could be "flawed", but the negation of "flawed" is probably "flawless", rather than "prefect". The emotion words may be more impressing examples. When saying someone is "not happy" or "unhappy", the speaker either implies the possibility of being in other emotional states. But for complementary pairs, the domain where they apply to is divided into two parts which combine the whole domain but mutually exclusive, so the negation of one member always goes to the other and double-negation goes back.

5 Conclusion

In this paper, we adopt logical negation and double-negation tests to compare the behaviors of three main subtypes of Chinese oppositeness. The test result indicates that: the three types of oppositeness are different in the way their pair members contrast each other; for a complementary pair, both the negation and double-negation go to the supposed members of the pair; for an antonym pair, neither bidirectional-negation nor double-negation goes to the supposed members of the pair, for a converse pair, the negation of each member of the pair does not always go to the other, but the double-negation of them goes back to themselves. Our test reveals that

the contrast relations holding between different subtypes of opposites are different from each other, at least in the logic negation and double-negation entailments.

However, this paper only employs several most often cited opposite pairs for the tests. Therefore, in future work we should extend the tests to opposite pairs in a larger scale, with the help of online corpora, to re-examine the results of the paper. Also, the comparative studies of opposite negation tests between different natural languages, such as Chinese and English, are supposed to be useful.

Acknowledgments. The work is supported by a General Research Fund (GRF) sponsored by the Research Grants Council (Project no. 544011). The authors would like to thank the anonymous reviewers for their useful comments on earlier drafts of this paper.

References

1. Cann, R.: Sense Relations. In: Maienborn, et al. (eds.) Semantics, pp. 456–479. de Gruyter (2011)
2. Cruse, D.A.: Lexical semantics. Cambridge University Press, Cambridge (1986)
3. Cruse, D.A.: Meaning in language: an introduction to semantics and pragmatics, 2nd edn. Oxford University Press, Oxford (2004)
4. Cruse, D.A., Togia, P.: Towards a cognitive model of antonymy. Lexicology 1, 113–141 (1995)
5. Lehrer, A., Lehrer, K.: Antonymy. Linguistics and Philosophy 5, 483–501 (1982)
6. Lyons, J.: Introduction to theoretical linguistics. Cambridge U.P., Cambridge (1986)
7. Lyons, J.: Semantics. Cambridge University Press, Cambridge (1977)
8. Mettinger, A.: Aspects of Semantic Opposition in English. Clarendon Press, Oxford (1994)
9. Murphy, M.L.: Semantic Relations and the Lexicon: Antonyms, Synonyms and other Semantic Paradigms. Cambridge University Press, Cambridge (2011)
10. Liu, S.X.: Chinese Lexical Semantics. Commercial Publication, Beijing (1990) (in Chinese)
11. Liu, S.X., Zhou, J.: Synonyms and Antonyms. Commercial Publication, Beijing (1992) (in Chinese)

A Hanzi Radical Ontology Based Approach towards Teaching Chinese Characters

Jia-Fei Hong[1] and Chu-Ren Huang[2]

[1] National Taiwan Normal University, Taipei, Taiwan
jiafeihong@gmail.com
[2] The Hong Kong Polytechnic University, Hong Hum, Hong Kong
churen.huang@inet.polyu.edu.hk

Abstract. Given the current popularity of learning Chinese language globally, design and creation of Chinese teaching materials is gaining recognition and has great impact. The preparation of teaching materials for reading and writing is particularly challenging because of the use of Chinese characters in the writing system. The aim of this study is to adopt an ontology-based description of the knowledge system of Chinese characters and to propose a knowledge-system based approach to the teaching of Chinese writing. In addition, we integrate Generative Lexicon Theory by [1] as bases of the concepts of Chinese character for Chinese teaching. By adopting this approach, Chinese learners can recognize and write Chinese words and then understand their lexical senses. In this study, we take Chinese radical representing"艸 (cao3, grass)", and five sense faculties in ShuoWenJieZi ,"目 (mu4, eyes)","耳 (er3, ears)",'口 (kou3, mouth)","鼻 (bi2, nose)",and "舌 (she2, tongue)"which all belongs to "body part" class in SUMO concepts as our research objects and explore their possible applications in Chinese teaching.

Keywords: character, Ontology, Chinese teaching, lexical semantics.

1 Introduction

Given the current popularity of learning Chinese language globally, design and creation of Chinese teaching materials is gaining recognition and has great impact. In Chinese education, it's necessary to build a systematized Chinese teaching system for learning, writing, and understanding senses of Chinese for second language learning. Regarding Chinese teaching and Chinese character teaching, there are so many different theories now; however, their teaching approaches are so multifarious and miscellaneous to cause learning achievements are different seriously. Therefore, we would like to provide a teaching system of Chinese character based on scientific theory.

In this study, we would like to explore Hantology by semantic symbol Ontology and discuss concepts of Chinese radicals. And according to [2] study, we then observe concept derivations of Chinese radicals are similar to Generative Lexicon Theory [1].

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 745–755, 2013.

About section presentation in this study, firstly, we talk some Chinese character teaching and some related researches of Chinese component teaching. Then, we discuss Ontology, and Hantology construction and its applications. Moreover, we point out "radicals" aspect instead of "components" aspect for Chinese character teaching. Last but not least, we take "艸 (*cao3*, grass)" and "Five Sense Faculties" as conventionalized by radicals to study the important and applications of Chinese character teaching by Hantology.

2 Previous Studies: Chinese Character Teaching, Chinese Radical Teaching

In the case of Chinese character teaching, several scholars mentioned different viewpoints and thoughts which include basic strokes, stroke order, component and construction of Chinese characters [3]-[7]. However, they did not have a scientific knowledge to support Chinese character teaching.

In [4] study, the author mentioned that it's an important point to focus on characters which include construction, evolvement and components. In addition, [4] study also pointed out three critical teaching approaches in talking original, stem and component of Chinese characters.

Although in [8] study, the author mentioned the application of Chinese character teaching by Hantology, he concentrated on digitized system, explained applications of Chinese character teaching, emphasize importance of digital learning and provided large and related Chinese character information and language knowledge. But, he did not discuss that Chinese learners can take Chinese radical by Hantology to learn and comprehend Chinese lexical senses.

For this reason, we would like to propose Hantology study to discuss Chinese radicals from their lexical senses in order to know and learn Chinese characters and words by Chinese radicals and their divided senses.

3 Hanzi Radical Ontology

3.1 Hantology

The Semantic Symbol Ontology is a system expressing the relations of Hanzi and its meaning cluster. This ontology system extended the basic structure constructed [9], which maps the meanings of 540 radicals in ShuoWenJieZi [10] with IEEE SUMO [11]. We use the results from analyzing derivative concepts to express the Semantic Ontology for each radical. Our current working interface allows easy query of existing database as well as recording of new entries.

Base on the definitions in ShuoWenJieZi [10] and our analysis of meaning cluster of the characters derived from the same radical, we can posit the basic concept for

each radical. For example, the basic concept for "艸 (*cao3*)" is "grass", and basic concepts for "目 (*mu4*)", "耳 (*er3*)", "口 (*kou3*)", "鼻 (*bi2*)" and "舌(*she2*)" are "eye", "ear", "mouth", "nose", and "tongue" individually.

3.2 The Classification of Hanzi Semantic Symbols

According to the definition in ShuoWenJieZi [10], our structure classifies the relationship between deriving meaning cluster and the basic concept of a radical. We use Pustejovsky's "Quilia Structure" [1] as base and observe the analysis on the definitions in ShuoWenJieZi [10], and then classify the deriving concepts of Hanzi radicals into 7 categories, expanded from the original four qualia aspects of Formal, Constitutive, Agentive, and Telic:

(1) Formal: This category can be further divided into 5 small categories: "sense", "characteristic", "proper names", and "atypical". The "sense" categories can be further divided into 5 small categories: "vision", "hearing", "smelling", and "taste".
(2) Constitutive: This category can be further divided into 3 small categories: "part," "member," and "group".
(3) Telic: Concepts related to function or usage.
(4) Participant: Words are classified into this category when the definition in ShuoWenJieZi mentions the participant involved.
(5) Participating: According to different events, concepts are divided into 6 small categories: "action", "state", "purpose", "function", "tool", and "others".
(6) Descriptive: This category can be further divided into two categories: "active" and "state".
(7) Agentive: The relationship between the radical and its meaning cluster coming from production or giving birth are classified in to agentive.

4 Research Motivation and Goal: Radical Aspect Instead of Component Aspect

In Chinese teaching programs, it is very important in Chinese character teaching. Many scholars and academic organizations propose different Chinese teaching approaches and teaching materials. In [6] study, the author mentioned some problems, difficulties and even bottleneck. In addition, the author found some specific problems of Chinese characters as 1) character origin can not be observed from changes of character form; 2) it's not reasonable in classifications for components of Chinese characters; 3) it's confused easily in components; 4) phonogram words can not present their senses and sounds in the meantime; and 5) there is fewer relationships among original senses, extended senses and loan senses.

From Chinese character teaching aspect, previous studies mentioned radical teaching approach, but only focused on radicals and used radicals in characters reading and writing. However, learning Chinese words need to learn character and senses at the same time; otherwise learners can not employ Chinese words. So that, we purpose some information of Chinese radicals and compositions based on Hantology and derive some learning approaches of lexical senses comprehensively and effectively.

Therefore, we propose a teaching system of Chinese writing scientifically and point out Chinese teaching by Chinese radicals and their relationships of divided words in this study. Following different categories of Chinese radicals by Hantology, these divided words represent their lexical senses. In other words, the goal of this Chinese teaching study is to attempt to propose Chinese radicals aspect based on Hantology as teaching objective instead of component aspect. Consequently, Chinese learners can study Chinese radicals systematically and obtain advanced level in reading compression and writing.

5 Domian Ontologies of "艸 (*cao3*, grass)" and "Five Sense Faculties" as Conventionalized by Radical

Following Ontology and lexical wordnet system, Chinese is a special writing system which includes radicals and phonograms. About radical, it refers to component in general. ShuoWenJieZi [10], the oldest dictionary of Chinese, is organized according to the radical forms as semantic symbols. Some related studies discussed some issues of Chinese semantic symbol Ontology [12]-[13] and they mentioned that Chinese radicals in Chinese semantic symbol Ontology are the basic core. In addition, Chinese radicals and their related characters can form Chinese words. Therefore, [14] studied some researches about Chinese radicals by Hantology, semantic symbol Ontology and Generative Lexicon Theory [1].

In Chinese words, there are more concepts and divided words related to "艸 (*cao3*, grass)", it's because many Chinese use traditional Chinese medicines or herbal medicines and the common distinguishing feature in Chinese. So, we can find there are many divided words from Chinese radical "艸 (*cao3*, grass)".

As regards "Five Sense Faculties" domain Ontology conventionalized by radical, we discuss their concept derivations and their knowledge representation. About "Five Sense Faculties", it refers to "目 (*mu4*, eyes)", "耳 (*er3*, ears)", '口 (*kou3*, mouth)", "鼻 (*bi2*, nose)", and "舌 (*she2*, tongue)". Among these five words, their frequencies are all higher in Chinese society and culture individually. We use them frequently and commonly, it's because we need to listen by ears, need to see by eyes, need to eat and speak by mouths, need to smell by noses and need to taste by tongues.

It's so common to use divided words related Chinese radical "艸 (*cao3*, grass)" and "Five Sense Faculties" in our daily activities. Therefore, in this study, we follow

[2] and [14] to discuss some applications in Chinese character teaching. We would like to introduce Chinese radical "艸 (*cao3*, grass)" and "Five Sense Faculties" and then to discuss their applications in Chinese character teaching respectively.

5.1 Radical "艸 (*cao3*, grass)"

In [2], the Chinese Radicals Study, they mentioned that there are divided words and divided concept as names, parts, constitutive, descriptive, formal, usages, and telic of plants for radical "艸 (*cao3*, grass)". They are shown as below Fig.1. The ontology of our Chinese Radicals "艸 (*cao3*, grass)" follows Pustejovsky's Generative Lexicon Theory [1] and the Qualia Structure of Pustejovsky's Generative Lexicon Theory can explains word divisions. Following this theory, we can divide Chinese Radicals " 艸 (*cao3*, grass)" into 4 categories as formal, constitutive, telic, and agentive which mean plants, parts of plants, usages of plants and descriptions of plants. Moreover, in [2] study, they constructed a concept construction of Chinese radicals and demonstrated that this concept construction has the division ability as Pustejovsky's Generative Lexicon Theory similarly.

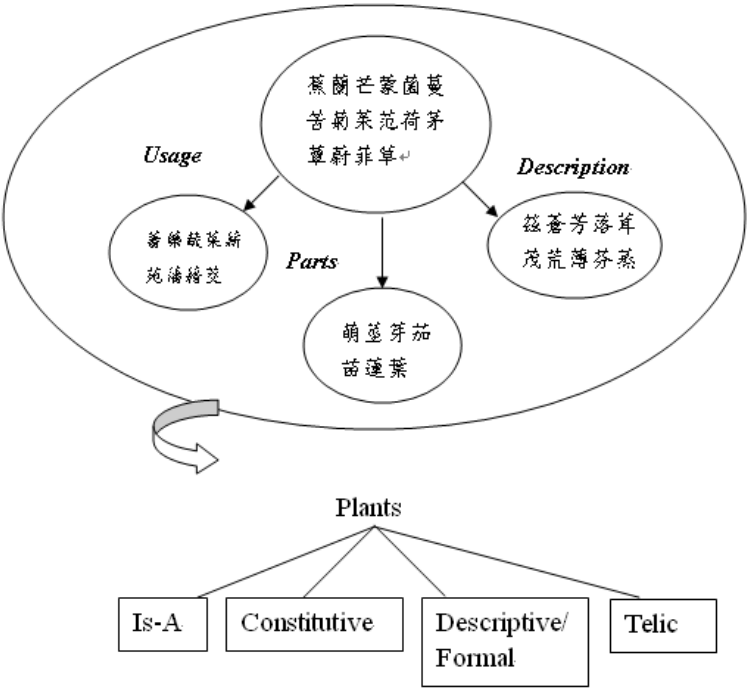


Fig. 1. Domain Ontology Conventionalized by Radical "艸 (*cao3*, grass)"

5.2 "Five Sense Faculties" Domain Ontology Conventionalized by Radical

According to [12] in "Five Sense Faculties" study, they pointed out the formal for concept construction and showed the relations, features, and ontologies for them. We would like to illustrate Chinese radicals of "eyes," "ears," 'mouth," "nose," and "tongue" and their related semantic symbols. We will show their divisions and divided concepts for Chinese radicals of "Five Sense Faculties" in Fig.2 to Fig.5.

Radical "目 (mu4, eyes)"

According to our analysis, the deriving concepts of " 目 (*mu4*, eyes)" on the category system includes "formal", "constitutive", "descriptive", "participating", and "telic". Among these five classes, the most active conceptual derivation can be classified as "vision activities" which is under "telic" category. The following is the concept deriving illustration of radical "目 (*mu4*, eyes)".

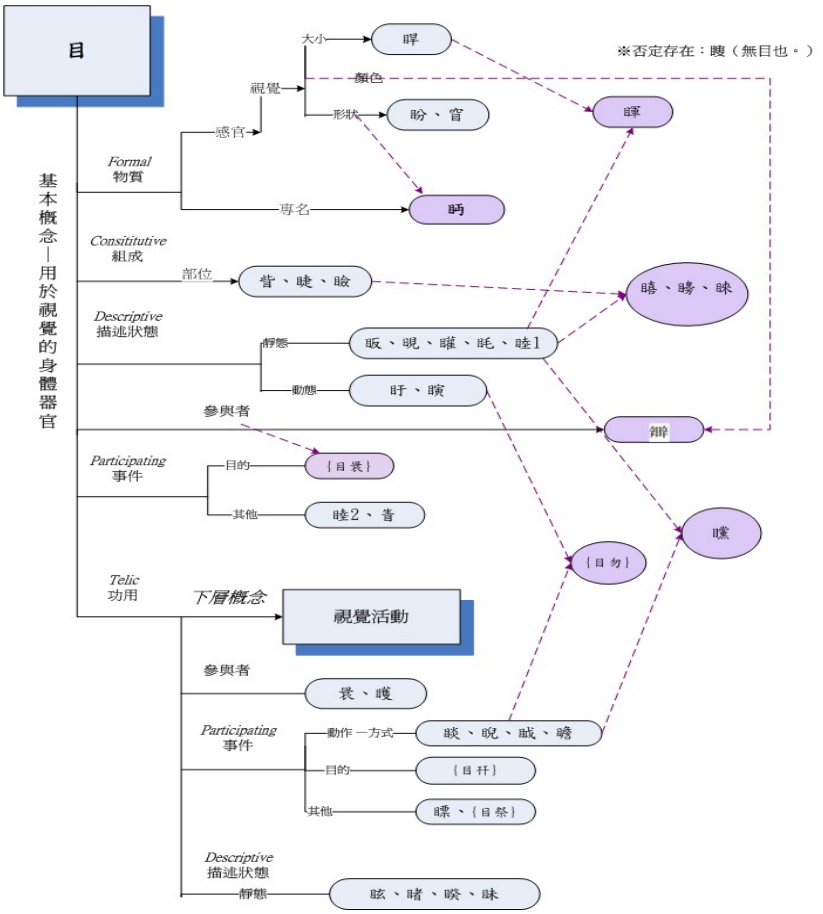


Fig. 2. Sense Faculties Domain Ontology Conventionalized by Radical "目 (*mu4*, eyes)"

Radical "耳 (*er3*, ears)"

According to the analysis of "耳 (*er3*, ears)", there are four deriving concepts as "formal", "descriptive", "participating", and "telic" for the related semantic symbols in the category system. We also can observe the deriving concept ---metaphor in this concept category system. Among these conceptual categories, the most conceptual derivations belong to "hearing activities" which is under "telic" category. Below Fig. 3 is shown the concept deriving illustration of radical "耳 (*er3*, ears)".

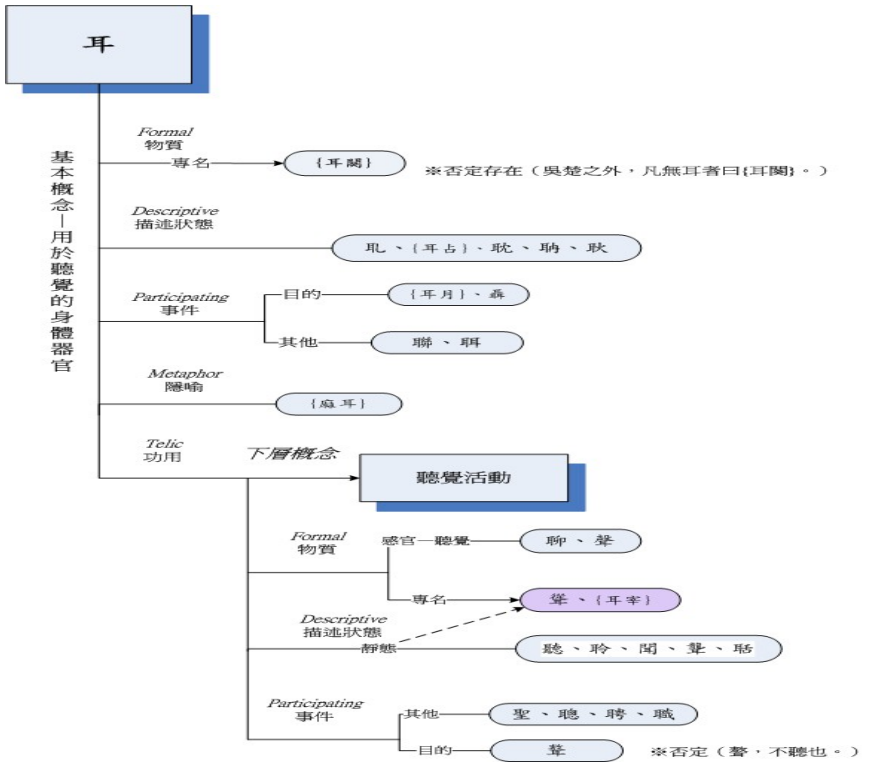


Fig. 3. Sense Faculties Domain Ontology Conventionalized by Radical "耳 (*er3*, ears)"

Radical "口 (*kou3*, mouth)"

About our analysis for the deriving concepts of "口 (*kou3*, mouth)"on the category system includes five classes as "formal", "constitutive", "descriptive", "participating", and "telic". In addition, the metaphor usages of conceptual derivations are also included in this system. The deriving concepts of "口 (*kou3*, mouth)"are similar to "目 (*mu4*, eyes)" and "耳 (*er3*, ears)". Their most related semantic symbols of the deriving concepts are from "telic".

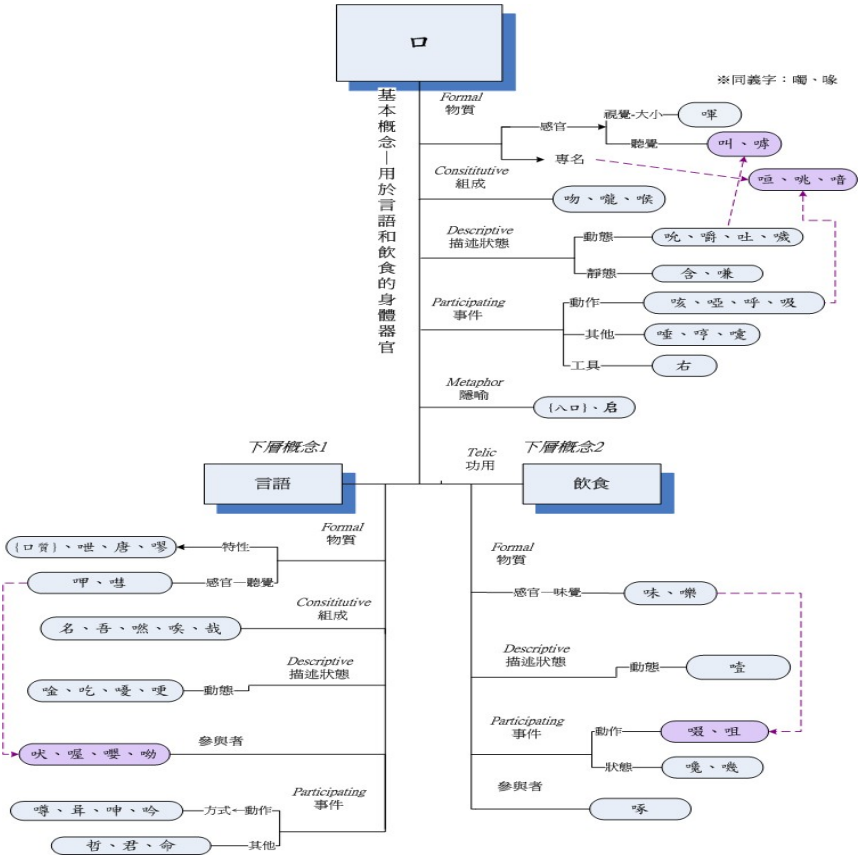


Fig. 4. Sense Faculties Domain Ontology Conventionalized by Radical "口 (*kou3*, mouth)"

Radical "鼻 (*bi2*, nose)"

About the radical "鼻 (*bi2*, nose)", the related semantic symbols are so few and only 4 words which their conceptual derivations belong to "participating". Therefore, either "breathing activities" or "smelling activities", they are all related to "function" of "鼻 (*bi2*, nose)".

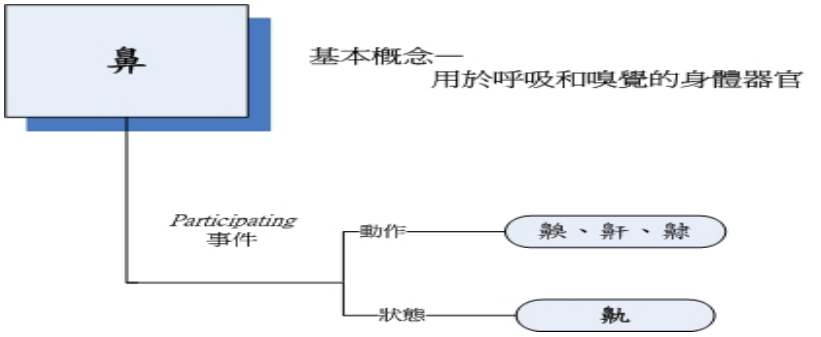


Fig. 5. Sense Faculties Domain Ontology Conventionalized by Radical "鼻 (*bi2*, nose)"

Radical "舌 (*she2*, tongue)"

There are only two words for the related semantic symbols of the radical "舌 (*she2*, tongue)" and therefore we can not construct the category system by Han-tology. However, according to ShuoWenJieZi, the explanation of "舌 (*she2*, tongue)" refers to "舌，在口所以言、別味者也". It means that tongue can distinguish different flavors. In fact, we can understand the function of the deriving concepts of "舌 (*she2*, tongue)". This viewpoint and conception is very similar to other four radicals of Chinese radicals of "Five Sense Faculties".

5.3 Applications in Chinese Character Teaching by Radical "艸 (*cao3*, grass)" and Radicals "Five Sense Faculties"

Through above analysis of Chinese Radicals, we can clearly understand the concept derivation and knowledge representation for Chinese Radicals "艸 (*cao3*, grass)" and "Five Sense Faculties" and understand their divided words and concept derivations based on the Qualia Structure of Pustejovsky's Generative Lexicon Theory [1]. In this section, we will discuss some teaching applications for Chinese Character teaching by following their divided words and concept derivations of Chinese Radicals "艸 (*cao3*, grass)" and "Five Sense Faculties".

Firstly, we analyze Chinese Radicals "艸 (*cao3*, grass)" by Pustejovsky's Generative Lexicon Theory [1] and we can observe some related concepts of plants for names, parts, descriptions and usages. For example, for plant parts, the divided words are "莖 (*jing1*, stem)", "芽 (*ya2*, bud)", "苗 (*miao2*, seedling)" and "葉 (*ye4*, leaf)" and so on. They are different parts for plants. Therefore, for our teaching applications, we can describe the stem of a tree: there are flowers, fruits and leaves above and connecting roots of a plant below, that is "莖 (*jing1*, stem)". About "芽 (*ya2*, bud)", it means that a partially opened flower or a swelling on a plant stem consisting of overlapping immature leaves or petals. About "葉 (*ye4*, leaf)", it means that the main organ of photosynthesis and transpiration in higher plants.

For the plant descriptions, these words are used in some states descriptions which are related plants. Moreover, these words can be described events or states in general. For example, "茂 (*mao4*, luxuriant)" means that produced or growing in extreme abundance or extendedly means that displaying luxury and furnishing gratification to the senses. As regards "芳 (*fang1*, fragrance)" and "芬 (*fen1*, fragrance)", they both refer to a distinctive odor that is pleasant and then extend to a pleasingly sweet olfactory property. With regard to Chinese teaching application in Chinese Radical "艸 (*cao3*, grass)", we can discuss Chinese Radical "艸 (*cao3*, grass)" and understand some senses of the related words by the Qualia Structure. We not only provide systemic and scientific teaching applications, we but also give a relational Chinese learning system for Chinese learners. For this reason, Chinese learners can understand and comprehend Chinese characters and words by their relationships in place of the mechanical memorizing approaches.

For example, "藥 (*yao4*, medicine)" is a common word but we can not learn it easily. It's because "藥 (*yao4*, medicine)" has two more senses. "藥 (*yao4*, medicine)" not only can represent "medicine", but also can represent "gunpowder".

These two senses both represent the related materials from "艸 (*cao3*, grass)" and both use the related materials from "艸 (*cao3*, grass)" to manufacture and reform. Therefore, according to Fig. 1, "藥 (*yao4*, medicine)" will be categorized into the "usage" category of divided words and concept derivations of Chinese Radicals "艸 (*cao3*, grass)".

In the same theory, we endeavor to discuss the teaching applications for Chinese Character teaching of "Five Sense Faculties". For Chinese Radical "目 (*mu4*, eyes)", "瞼 (*jian3*, eyelid)" is a related part of "目 (*mu4*, eyes)" and that is eyelids. "睦 (*mu4*, peaceful)" is descriptive of a static state of "目 (*mu4*, eyes)", and it means peaceful. "睦 (*mu4*, peaceful)" can be extended to describe events and their senses refer to kind, pleasant and kindly. For Chinese Radical "耳 (*er3*, ears)","耽 (*dan1*, ears appearance)" describes appearance states for ears and means that ears are big and downcast. "聽 (*ting1*, to listen)"、"聆 (*ling2*, to listen)"、"聾 (*long2*, deaf)" and" 聒 (*gua1*, noisy)" are all hearing activities of static states of the telic. Taking Chinese Radical "口 (*kou3*, mouth)" as an example, the dynamic states descriptions of "口 (*kou3*, mouth)" are "吮 (*shun3*, to suck)"、"嚼 (*jue2*, to chew)" and "吐 (*tu4*, to vomit)". On the contrary, the static state description of "口 (*kou3*, mouth)" is "含 (han2, to hold in the mouth)". About the constitutive of "口 (kou3, mouth)", they are "嚨 (long2, throat)" and "喉 (hou2, throat)", it's because "嚨 (*long2*, throat)" and "喉 (hou2, throat)" are components and compose the mouth. In the case of "鼻 (*bi2*, nose)", there are more descriptions for "smelling". For example, "鼾 (*han1*, to snore)" means someone breathes when he is in sleeping and "鼾 (*han1*, to snore)" belongs to the category of eventive activities. Lastly, although there are few Chinese words from Chinese Radical "舌 (*she2*, tongue)", the conceptions, comprehension and interpretations are very similar to other four radicals of Chinese radicals of "Five Sense Faculties".

Based on the divided words and concept derivations of Chinese Radicals, we would like to attempt to employ a systemic and scientific approach to analyze, interpret and study Chinese Character.

6 Conclusion

In this study, we use the Qualia Structure to construct excellent connections for lexical semantic relationships. We employ the Semantic Symbol Ontology --- Hantology to analyze concepts of Chinese radicals and follow the Qualia Structure of Pustejovsky's Generative Lexicon Theory [1] to discuss original senses, divided senses and extended senses of Chinese radicals. It's because we can take Chinese radicals in place of components in Chinese character teaching and build a scientific and systemic approach. Consequently, Chinese learners can learn Chinese characters systematically and conceptually and then they can make good use of reading and writing. The crucial excellence is Chinese learners don't need to memorize individual feature for each Chinese character, but they only understand and comprehend basic principle of Chinese radicals emphatically. Therefore, following our learning approach, they can accumulate comprehension of Chinese radicals.

It's so useful approach for Chinese characters teaching and learning to integrate the Qualia Structure of Pustejovsky's Generative Lexicon Theory [1] and Hantology [2]. That is to say Chinese radical symbol and divided senses in our approach and this approach is so different from alphabetic writing systems which there are no relationships in symbols and senses.

In a word, we can provide organized, systemic, basis and scientific teaching and learning systems of Chinese characters based on the Qualia Structure of Pustejovsky's Generative Lexicon Theory [1] and Hantology [2].

References

1. Pustejovsky, J.: The Generative Lexicon. The MIT Press (1995)
2. Chou, Y.-M., Huang, C.-R.: Hantology: conceptual system discovery based on orthographic convention. In: Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prevot, L. (eds.) Ontology and the Lexicon, pp. 122–143. Cambridge University Press, Cambridge (2010)
3. Wang, W.-Y., Cai, M. (eds.): Literacy and Writing Teaching. Language Course and Teaching Theory. Higher Education Publishing (2002) (in Chinese)
4. Huang, P.-R.: Theory and Practice in Chinese Teaching. Lexis Book, Taipei (2003) (in Chinese)
5. Yin, L.-Y.: On component part of Chinese Character teaching to foreigners. Journal of Yunnan Normal University 2(2) (2004) (in Chinese)
6. Chen, Y.-L.: Bottlenecks and strategies of the theoryical Chinese teaching. In: Proceeding of International Conference of Chinese Language Centers Operating Strategies and Teaching of the Twenty-First Century, pp. 193–198. Mandarin Training Center, National Taiwan Normal University, Taipei (2005) (in Chinese)
7. Chou, B.-X.: Improvement Strategies of Chinese Teaching from Transfer of Learning. Journal of Taipei Municipal University of Education 42(2), 1–22 (2011) (in Chinese)
8. Chou, Y.-M.: The Application of Hantology in Chinese Characters Teaching. Journal of Chinese Language Teaching 6(1), 91–112 (2009)
9. Chou, Y.-M.: Hantology: A Chinese Character-based Knowledge Framework and its Applications. Ph.D thesis, National Taiwan University (2005)
10. Xu, S.: ShuoWenJieZi (121)
11. Niles, I., Pease, A.: Linking Lexicons and Ontologies: Mapping WordNet to the SUMO Ontology. In: Proceedings of the IEEE International Conference on Information and Knowledge Engineering, Las Vegas, Nevada (June 2003)
12. Huang, C.-R.: Knowledge Representation with Hanzi: The relationship among characters, words, and senses. Presented at International Conference on Chinese Characters and Globalization, Taipei (January 2005)
13. Huang, C.-R., Chen, S.-Y., Chou, Y.-M.: Knowledge of the Language and the Language of Knowledge: An ontological study based on ShuoWenJieZi. In: Pan, W.-Y., Shen, Z.-W. (eds.) The Joy of Research: A Festschrift in Honor of Professor William S-Y Wang on His Seventy-Fifth Birthday, pp. 106–122. Shanghai Education Press (2010)
14. Huang, C.-R., Chen, S.-Y., Yang, Y.-C.: An Ontology based on Chinese Radicals:Concept Derivation and Knowledge Representationof the Five Sense Faculties. In: The 9th Chinese Lexical Semantics Workshop (CLSW 2008), pp. 95–109. National University of Singapore, Singapore (2008)

Discourse Coherence: Lexical Chain, Complex Network and Semantic Field

Mingyao Zhang[1,4], Hua Yang[1,2,*], Donghong Ji[2,3], Chong Teng[3], and Hongmiao Wu[4]

[1] College of Chinese Language and Literature, Wuhan University, Wuhan, 430072, China
myzhang@whu.edu.cn, yanghuastory@gmail.com
[2] School of Mathematics and Computer Science, Guizhou Normal University, Guiyang, 550001, China
[3] School of Computer, Wuhan University, Wuhan, 430072, China
donghongji_2000@yahoo.com, tchong616@126.com
[4] School of Foreign Languages and Literature, Wuhan University, Wuhan, 430072, China
hongmiao23@163.com

Abstract. Discourse coherence is one of the most essential and challenging topics in theoretical linguistics and computational linguistics. Complex network has found its initial application in discourse coherence analysis. We survey the corresponding work related to the complex network in discourse analysis and pinpoint the correlations among lexical chain, complex network and semantic field. We also demonstrate the prospect of applying complex network theory to discourse coherence analysis.

Keywords: discourse coherence, lexical chain, complex network, semantic field.

1 Introduction

Discourse coherence refers to the features of the discourse in which language constituents abide by grammatical rules, conform to each other in semantic from the outset to the end and ultimately achieve the communicative objectives [1]. The research of discourse coherence has been the focus of text linguistics [2]. Recently the computing model on discourse coherence has also received more and more attention in natural language processing (NLP) [3-4]. Theoretical linguists and computational linguists come to the consensus that discourse is not only the linear sequence of clauses and sentences, but a sophisticated structure. But no computable and accurate formalism has been available so far.

Complex network is an important model to describe complex system. Network consists of nodes denoting the elements of the system and edges denoting the interaction among elements. If the number of the nodes of the network reaches the degree so

* Corresponding author, E-mail: yanghuastory@263.net

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 756–765, 2013.

that its regularity, organization, and some statistical properties are influenced, the network is called complex network [6-7].

This paper gives an overview, analysis and prospect on the application of complex network in discourse coherence. Section 2 surveys the application of complex network in linguistics and NLP. Section 3 surveys the exploration of complex network in discourse coherence on the basis of [8]. Section 4 gives a detailed analysis on the related work by pointing out the substantial links among complex network, lexical chain and semantic field of the discourse.

2 Linguistic Complex Network

Language is a typical complex system after long-term evolution [9]. With the advent of complex network, more and more people began to apply it into different fields [10-13]. Many experts established network for language from the aspect of linguistics and cognition with the statistical and physical methods and acquired some useful conclusions which cannot be accessed from first-order statistics. The language network takes on the following forms: word co-occurrence network [10], word collocation network[10],[14-15], word dependency grammar network[16]. These networks' topology structures display the universality of complex networks and share the similar properties. In other words, different languages underlie relatively fixed regularities despite their miscellaneous morphology and syntax and network is a robust tool to mine these regularities.

3 Application of Complex Network in Discourse Coherence

Documents [8],[17] and[18] used similar approaches to assess the quality of discourse: complex network was used to represent document, acquire parameters which are positively correlated to manually evaluated score, and employ the parameters to evaluate text quality.

Reference [18] distinguished the professional-authors-written texts and high-school-student-written texts (in most cases, the former outperform the latter) by using network parameter. Document [17] distinguished the machine-generated summary and human-generated summary (in common sense that the latter outperforming the former) by using network parameter. [17] and [18] classified the quality of the texts into high-quality and low-quality. In contrast, [8] did a more precise work: coherence was set up as a standard to assess the text and precise correlation coefficient was computed. Though their objectives and data are different, they used the same network parameter, i.e. how to evaluate the discourse coherence, and [8] can be representative work to show how to evaluate discourse coherence with complex network parameters.

3.1 Acquiring Manual Score

40 texts of the same topic by high-schoolers were selected for coherence evaluation. All the texts have approximately the same length, with an average of 228 words. A

panel of five human judges, all of which are computational linguists, analyzed the texts using three criteria to mark them, namely (i) coherence and cohesion, (ii) adherence to standard writing conventions and (iii) theme adequacy/development, henceforth referred to as CC, SWC and TAD, respectively. However, Reference [8] did not give significant result concerning TAD. As for the human evaluation, the human scoring results are not evenly distributed.

3.2 Network Generation

Two kinds of network were acquired from each text: NET-A and NET-B. Within NET-A, words are the nodes, and the adjacency relations between two words are represented as the directed and weighted edges whose directions lead from the fore-occurring words to the latter and whose weight represents the frequency of the two adjacent words involved. Within NET-B, edge is defined as a word pointing to two words following it. These two networks represent Markov Model which memorized one or two past state.

3.3 Network Properties

Within complex network field, researchers have put forward parameters and models to assess the complexity of network and facilitate the understanding and prediction of these systematic behaviors [7], [19]. The network parameter and concepts adopted by [8] are significant from the following dimensions:

Node indegree, node outdegree, node degree: the degree of the node is defined as the number of edges which connect the nodes. In a directional graph, the indegree and the outdegree of a node are defined as the number of the edges point to or point out from the node respectively. Network node average degree is defined as the average degree of all the nodes. The concept of average indegree and outdegree is also involved in the directed graph.

Component: for a non-empty graph G, assume there exists at least one path for all pairs of nodes $v,w \in V$, this graph is called connected graph, and the connected subgraph is defined as the (connected) component.

Clustering coefficient: node clustering coefficient refers to the probability that the two adjacent nodes of a certain node are linked. For a node v, its local clustering coefficient is defined in formula (1), where ke is the real number of links among v's neighbors and k_v is the number of node v's neighbors. That is, node clustering coefficient is the ratio of the actual number edges among v's neighbors to the maximum edges can exist among its neighbors. Let $C_v = 0$ if $k_v = 0$ or $k_v = 1$.

$$C_v = \frac{k_e}{k_v \frac{k_v - 1}{2}} \tag{1}$$

Network clustering coefficient is defined as the average value of local clustering coefficient of all the nodes. As shown in formula (2), where N is the total number of network nodes. But directional graph is employed in [8], when calculating clustering

coefficient, the edges with opposite directions between the same pair of nodes are counted as two different edges.

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{2}$$

Components dynamics deviation (CDD) is a parameter used specifically in [8] to describe the dynamic growth process of a network. After one edge is added to the graph, the number of connected components is counted, thus generating the dynamic topology feature, which is a function of edge number, and also a function of discourse building process. For each text, network is initiated as a set of N nodes representing N words in the text, and 0 edges. Each word represents a component, regarding the text as the linear sequence of words. When the relation between two words is read, new edge is created or weight of existed edge is increased. Fig. 1 shows the variation tendency of number of components with the increase of edges for three texts A, B and C. The curve in each graph shows the variation tendency in which the number of the components varies with the addition of edges. The direct line is the reference line to represent the case that the variation process of component number is a uniform. The reference line is drawn because: in [17] expert-generated summary' CCD is always a line like this, but machine-generated summary's CCD deviates from the line. Human-generated summary tends to be a direct line while machine-generated summary tends to deviate from this line. In order to quantify the deviation of the curve from the direct line, $f_a(x)$ is defined as the function of the number of components and the number of word relation, $f_a(x)$ is the reference line, L is the number of all the words in the text, N is the number of nodes in the network. The algorithm is shown in formula (3). The CCD of text A, B and C is 0.014, 0.045 and 0.064 respectively.

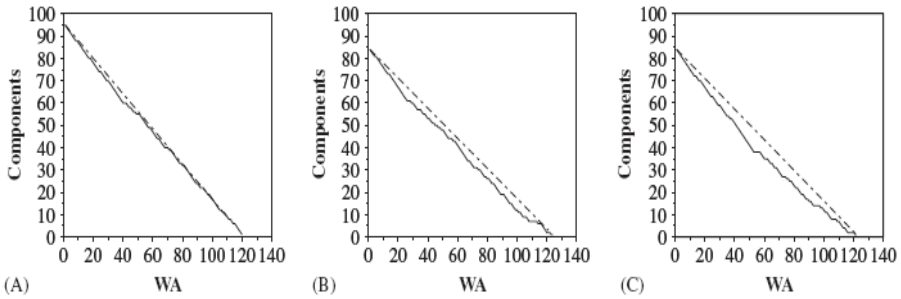


Fig. 1. Relation between WA and number of Components for Document A, B and C

3.4 The Relation between Network Parameter and Discourse Coherence

[8] attempts to relate the concept of complex network to manual score. Due to the disparateness of human scoring, 20 texts with lowest disparateness scores according to the standard are adopted. Three network parameters are found to be highly relevant to three types of score. In the following figures, the horizontal axis represents network

parameter, and the ordinate axis represents the human score, A and B represent NET-A and NET-B respectively. Both network parameter and human score are standardized as N(0,1) distribution. For each figure, corresponding reference line, Pearson correlation coefficient and p-value are obtained. [8] pointed out that the CC is the parameter most related to human score. And this relevance seems to capture the features of the text construction. NET-A is enough for capturing the features.

$$CDD = \frac{\sum_{x=1}^{L} |f_a(x) - f_s(x)| / N}{L} \tag{3}$$

3.4.1 Relation between Outdegree Distribution and Discourse Coherence

Fig. 2-A and Fig. 2-B show that human score declines with the increase of outdegree as far as three coherence standards are concerned. The most sensible are CC and SWC. The relevance between high Pearson correlation coefficient and low p-value is not coincidental. As for TAD, the curve also takes on declining tendency but not so noticeable. No noticeable difference between A and B. [8] accounts for this phenomenon this way: too high outdegree and clustering coefficient means the author introduces new concept too quickly.

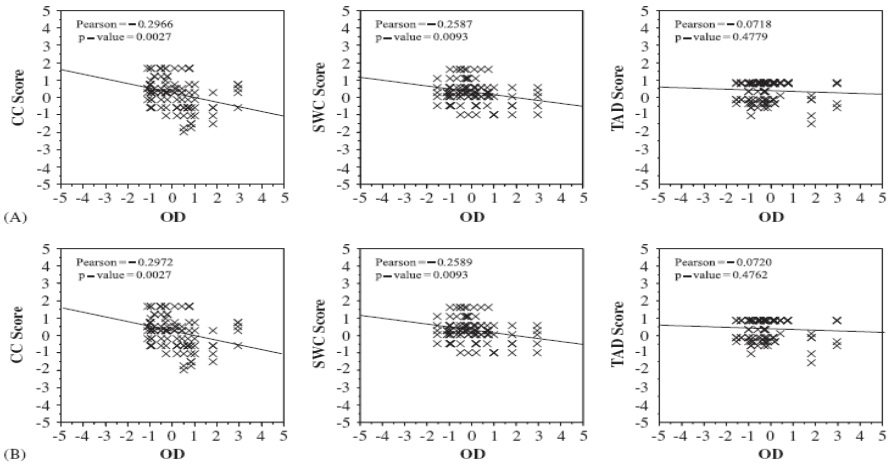


Fig. 2. The relevance between outdegree and coefficient

3.4.2 Relation between Clustering Coefficient and Discourse Coherence

Fig. 3-A and Fig. 3 - B show that the quality of the texts declines with the increase of CLC. The relevance on A is higher than that on B, especially CC and SWC. The high relevance is the most obvious one of all the parameters used. This phenomenon can be accounted for in terms of linguistics: the more crossed the concepts of the texts, the lower quality of the texts. That is, over-interconnection may result in poor quality of the text.

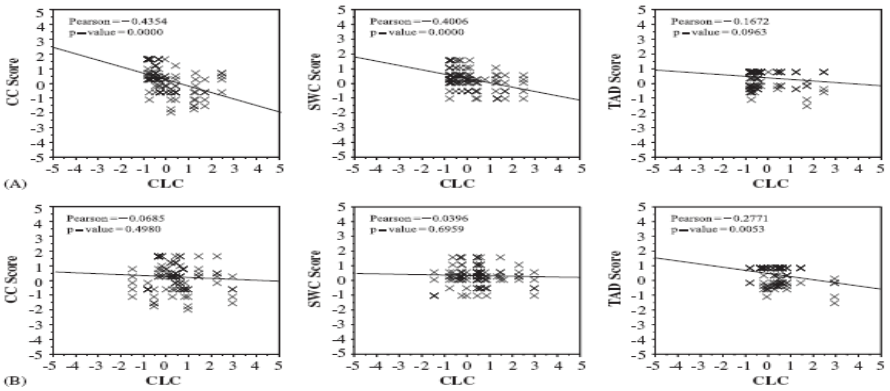


Fig. 3. The relationship between CLC and the scores of the text quality

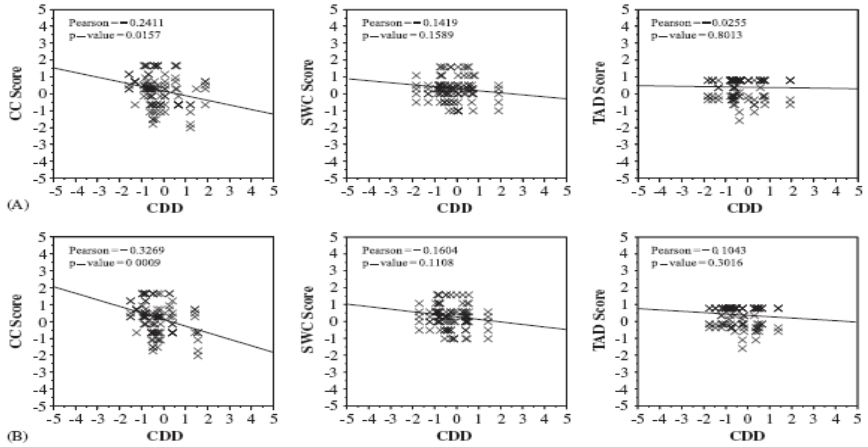


Fig. 4. Relation between CCD and text quality

3.4.3 Relation between CCD and Discourse Coherence

Fig. 4-A and Fig. 4-B show that text quality declines with the increase of CDD. There is no significant difference between A and B concerning the results. High CCD shows that the author is most likely to repeat some concepts in the writing process and consequently the text quality declines and leads to the rapid decline of the number of component. For example, the three texts in Fig. 1 have the scores 7.9 (A), 5.2 (B) and 3.7 (C) in terms of CC, and, 0.014 (A), 0.045 (B) and 0.064 (C) in terms of CCD.

3.5 Analysis and Explanation

As for the work in [8], we analyze it as follows:

As far as the building of network is concerned, [8] adopts word adjacency network. This process has some defects:

1. Problems concerning nodes. i) It is universally acknowledged that single word does not necessarily denote complete meaning. Lexics involves two layers of words:

aggregation and combination. For example, 全球变暖(quan qiu bian nuan, global warming) is not simply combined by 全球(quan qiu, global) and 变暖(bian nuan, warming). Instead, this term has become a fixed term which cannot be separated. ii) Using word as node ignores the fact of polysemy.

2. Problems concerning the edge. The adjacency of words doesn't necessarily represent the semantic relevance. One- or two-order Markov Model can capture the syntactic and semantic relation to some extent. However, they cannot capture all the relations as well as capture some error relations. We can find that NET-A captures more accurate results in terms of syntactic and semantic relations but loses more relations than NET-B. NET-B captures more complete syntactic and semantic relations while more false relations are captured compared with NET-A. In terms of NLP terminology, NET-A gets higher precision while NET-B gets higher recall. 1) About 87% of the syntactic relation occurs within the distance of 2 words [16], [22]. Word co-occurrence network lacks the precise definition of edge, so it cannot capture the relation between long-distance words in the text [23]. The percentage that word co-occurrence network captures non-syntactic dependency co-occurrence whose number is 2 in the window D is 50%, and 30% when the window is 1[16]. Syntactic dependency network overcomes the errors the co-occurence network may capture concerning the edges; however, this relation is acquired by human annotation instead of raw materials. On the other side, adjacent words tend to have grammatical relations. 87% of the syntactic relations occur between the words whose distance is less than or equals to 2. In this sense word co-occurrence network can be used to uncover the statistical properties of a language. In other words, word co-occurence network can be considered as the approximation of syntactic network. 2) As far as semantic relation is concerned, no methods using sememe analysis have been found in NLP field to analyze the semantic relations of a text. But NET-A and NET-B can be used to capture the words which have common semanteme to some extent.

In essence, lexical chain technique captures the lexical relation within the same semantic field, i.e. synonymy. In other words, NET-A and NET-B are incomplete and imprecise lexical chains. Lexical cohesion refers to the semantic relations among words [24]. [25] proposes the classification methods on the basis of dependency. Generally, lexical cohesion can be categorized as two kinds: reiteration and collocation. Reiteration includes identical repetition, non-identical repetition, hyponymy repetition, hyponymy repetition and synonymy repetition. Collocation falls into two categories: systematic semantic relations and non-systematic relations. Systematic semantic relations refer to antonymy, orderly membership, non-orderly membership and part-whole relation. Non-systematic relations refer to the words which occur within the same context. For example, non-systematic relations can be found in {garden, digging}, {post office, service, stamps, pay, leave}, and {car, lights, turning}. According to [24] and [26], the collocation in [25] is similar to words which belong to the same fields. Lexical cohesion exists not only between a pair of words, but among a chain of adjacent semantically-related words spanning across the topical unit in the text. These semantically-related words form the lexical chain within the text and co-occur within a fixed span. Lexical chains don't stop at the boundary of sentences. They may connect all the adjacent words or span across the entire text.

We may find that the network built on [8] captures the above-mentioned 5 kinds of lexical cohesion. Though the lexical cohesion it captures is incomplete and imprecise, network used in [8] captures relation between words in the same field.

4 The Prospect of Complex Network in Discourse Coherence

Since the network built in [8] is an incomplete and imprecise lexical chain, documents [24], [27-31] display the application prospect of network-based approaches in NLP. In addition, lexical chain can be used to locate the most important semantic units within the texts [24], it may find wide application in discourse coherence.

As mentioned before, the definition to the node of word co-occurrence network is not precise enough, and lexical cohesion the edge captures is not complete, [8]'s work can thus be extended. That is, to apply complex network into the analysis of discourse coherence. In order to achieve this goal, we need to settle the following problems: 1) to define node more accurately so that the node can be used to express complete semantic meaning. For example, the United States as a term should not be segmented as three single words, and such is the case with Chinese terms. 2) to define edge more accurate, i.e. syntactic relation or precise semantic relation. Semantic relation can be transformed into the partition of semantic field and the practice involves semanteme analysis. Though it has a long way to go on the task of semanteme analysis, it is still a practical approach in linguistics. Furthermore, the result of sememe analysis is formal, which, it is not totally impracticable for computation.

Moreover, lexical cohesion cannot be used to analyze discourse coherence abundantly. The essence of discourse coherence lies in the coherence of events, the logical relation among events. Therefore, we may start from lexical chain—the precise lexical complex network to derive event chain and logical chain to analyze discourse coherence.

Acknowledgement. This paper is supported by Natural Science Foundation Project (61070243, 61133012, 61070082, 61173062, 61202193), Major Project of Invitation for Bid of National Social Science Foundation (11&ZD189), Guizhou High-level Talent Research Project (TZJF-2010-048), Guizhou Normal University PhD Start-up Research Project (11904-05032110011), Governor Special Fund Grant of Guizhou Province for Prominent Science and Technology Talents (identification serial number "黔省专合字(2012)155号"), the Post-70s Scholars Academic Development Program of Wuhan University, and Autonomous Research Grant of Academy of Humanities and Social Sciences, Wuhan University(2012YB005).

References

1. Liu, C.D.: Text Linguistics for Teachers. Shanghai Foreign Language Education Press, Shanghai (1999)
2. Cheng, X.T.: Language Teaching Approaches based on Texts. Foreign Language Teaching (001), 8–16 (2005)
3. Feng, Z.W.: Formal Models of Natural Language Processing. Chinese University of Science and Technology Press, Hefei (2010)

4. Carlson, L., Marcu, D., Okurowski, M.E.: Building a discourse-tagged corpus in the framework of rhetorical structure theory, pp. 1–10. Association for Computational Linguistics (2001)
5. Guo, L., Xu, X.M.: Complex Network. Shanghai Science and Technology Education Press (2006)
6. Milgram, S.: The small world problem. Psychology Today 2(1), 60–67 (1967)
7. Newman, M.E.: The structure and function of complex networks. Arxiv preprint cond-mat/0303516 (2003)
8. Antiqueira, L., Nunes, M.G., Oliveira, J.O., et al.: Strong correlations between text quality and complex networks features. Physica A: Statistical Mechanics and its Applications 373, 811–820 (2007)
9. Steels, L.: Language as a Complex Adaptive System. In: Deb, K., Rudolph, G., Lutton, E., Merelo, J.J., Schoenauer, M., Schwefel, H.-P., Yao, X. (eds.) PPSN 2000. LNCS, vol. 1917, pp. 17–28. Springer, Heidelberg (2000)
10. Ferrer i Cancho, R., Sole, R.V.: The small world of human language. Proceedings of the Royal Society B: Biological Sciences 268(1482), 2261–2265 (2001)
11. Ferrer i Cancho, R.: The structure of syntactic dependency networks: Insights from recent advances in network theory. In: The Problems of Quantitative Linguistics, Ruta, Chernivtsi, pp. 60–75 (2005)
12. Sole, R.V., Murtra, B.C., Valverde, S., et al.: Language Networks: their structure, function and evolution. Trends in Cognitive Sciences (2006)
13. Mehler, A.: Large Text Networks as an Object of Corpus Linguistic Studies (2007)
14. Dorogovtsev, S.N., Mendes, J.F.: Language as an Evolving Word Web. Proceedings: Biological Sciences 268(1485), 2603–2606 (2001)
15. Heyer, G., Quasthoff, U., Wittig, T.: Text Mining: Wissensrohstoff Text Konzepte, Algorithmen, Ergebnisse. W3L-Verl. (2006)
16. Ferrer i Cancho, R., Solé, R.V., Köhler, R.: Patterns in syntactic dependency networks. Physical Review E 69, 51915 (2004)
17. Pardo, T.A., Antiqueira, L., Nunes, M.G., et al.: Using complex networks for language processing: The case of summary evaluation. In: Proceedings of the International Conference on Communications, Circuits and Systems (ICCCAS 2006), Special Session on Complex Networks, pp. 2678–2682 (2006)
18. Antiqueira, L., Maria, G.V., Oliveira, O.N., et al.: Complex networks in the assessment of text quality. Arxiv preprint physics/0504033 (2005)
19. Newman, M.E.: The Structure and Function of Complex Network. Arxiv preprint cond-mat/0303516 (2003)
20. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature (London) 393(6684), 440–442 (1998)
21. Ferrer i Cancho, R., Solé, R.V., Köhler, R.: Patterns in syntactic dependency networks. Physical Review E 69(5), 51915 (2004)
22. Ferrer i Cancho, R.: Euclidean distance between syntactically linked words. Proc. Natl. Acad. Sci. USA Phys. Rev. E. 70, 56135 (2003)
23. Chomsky, N.: Syntactic structures. Mouton, The Hague (1957)
24. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics 17(1), 21–48 (1991)
25. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman, London (1976)
26. Jia, Y.: Chinese Lexics. Peking University Press, Beijing (1992)
27. Okumura, M., Honda, T.: Word sense disambiguation and text segmentation based on lexical cohesion, pp. 755–761 (1994)

28. Hirst, G., Stonge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An Electronic Lexical Database, 305–332 (1998)
29. Hirst, G., Budanitsky, A.: Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering 11(01), 87–111 (2005)
30. Chen, Y., Liu, B., Wang, X.: Automatic text summarization based on textual cohesion. Journal of Electronics (China) 24(3), 338–346 (2007)
31. Ercan, G., Cicekli, I.: Using lexical chains for keyword extraction. Information Processing and Management 43(6), 1705–1714 (2007)

The Nature of Semantic Primitive and Its Role in Synset Construction

Li Feng[1], Yiqun Zhang[2], and Yaxuan Chen[3]

[1] College of International Exchange, Shenzhen University, Shenzhen 518060
lily_von@126.com
[2] School of Bioscience and Bioengineering,
South China University of Technology, Guangzhou 510006
drum.s@163.com
[3] School of Humanities and Social Sciences,
Nanjing University of Science and Technology, Nanjing 210094
435972595@qq.com

Abstract. Semantic primitive exists in lexical level, conceptual level and semantic level. This paper analyzes the meaning and usage of semantic primitive and explores their definitions and connotations in different linguistic levels. It mainly discusses the role of semantic primitive in synset construction. Through the analysis and discussion, the paper not only provides the objective standard for the synset construction, but also highlights the relations and distinctions among the members in synset. Meanwhile, it also contributes to the semantic formalized description.

Keywords: semantic primitive, synset construction, formalized description.

1 Introduction

In recent years, with the development of semantic research, the word primitive appears with an increasing frequency in semantic research field. Some concepts such as semantic primitive, primitive word, interpretative primitive and conceptual primitive have emerged in some research works. These different conceps sometimes refer to the same thing and sometimes mixed with other concept, particularly the concept "semantic primitive". Semantic primitive exists in lexical level, conceptual level and semantic level, with one name but different references. It makes itself much more difficult to be understood and applied. As semantic research has become a popular issue in recent years, the classification of semantics in language information processing and semantic auto labeling becomes more and more detailed. Under this circumstance, it is necessary to have an analysis on the nature of semantic primitive and the application of it in linguistics.

2 Semantic Primitive in Linguistics

The study on semantic primitive can be traced back to the 17th Century. Many philosophers and logicians, such as Leibniz, Descartes, and Arnaud had mentioned

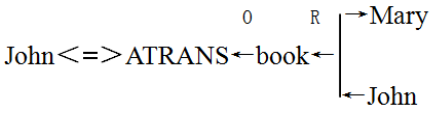
D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 766–773, 2013.

the necessity to study the semantic primitives. Some forerunners like Anna Wierzbicka, the representative of the Polish Semantic School, Ray Jackendoff, the scholar of cognitive semantics school, and Roger Schank, the American pioneer in the field of natural language processing all put forward the theories about semantic primitives.

According to Anna Wierzbicka, semantic primitive refers to those words that are easy enough to be understood without any explanation[1]. These simplest primitives can be used to explain some words with complicated meanings and they themselves cannot be further interpreted[2]. They compose a metalanguage together with other grammatical rules. The semantic primitives mentioned by Wierzbicka in fact are interpretative primitives, which refer to the primitives that can meet the demand of language dictionary defining[3]. Apart from interpretative primitives, primitive words also conclude communication primitive words and teaching primitive words in the division of functional level. To understand semantic primitives from the perspective of basic expressing, all these kinds of primitive words are the smallest units of words in expressing or interpreting.

Jackendoff holds the view that concepts are infinite, each of them is generated by finite "primitives" according to different formation rules[4]. The purpose of his research is to find out the primitives and the internal structure of a concept. Jackendoff started from lexical concepts to construct the conceptual structures. He believes that the concepts of lexical items can be decomposed into a group of components which are limited in their number. These components expressed a concept have their own internal structures. For example, the primitive of the verb "throw" is GO, and the internal argument structure of GO is GO ([object] [path])[5]. It is an understanding about the semantic primitives from the perspective of internal structure of concepts.

Roger Schank puts forward the theory of Conceptual Dependency which is used in the field of natural language processing. He claims that there exists a conceptual base in people's heads. Language undersdanding is the process of reflecting language statements into a conceptual base which exists in human's brain and the semantic structure of different sentences in different languages are the same. Therefore, Schank uses 11 primitive acts to describe all the acts and 12 primitive states to describe all the states of natural language[6]. Using these primitives, he tries to present semantic information hidden in sentences by analyzing the dependent relationship between primitives and other sentence components. For example, the sentence "John gives Mary a book" can be expressed with the CD theory like this:



The primitive act node ATRANS indicates the verbs like *"give" "send"* and *"buy"*, etc. In this expression, the node refers to transferring an abstract relationship of the verb *"give"*. The three arrows with an R indicate the dependent relationships of accepting or giving among *"Mary", "John"* and *"book"*. The arrow with an O refers to the dependent relationship between *"book"* and *"ATRANS"*, that is, *"book"* is the

goal of the basic act of *"ATRANS"*. It is an understanding about the semantic primitives from the perspective of concepts and natural language processing.

To sum up, all of the semantic primitives above in linguistics are used to interpret or describe meanings no matter what the content they are. They have a close relationship with lexical meanings in different language levels.

3 The Nature of Semantic Primitive in Different Levels

As mentioned above, semantic primitive in linguistics exists in three levels: conceptual level, lexical level and semantic level. These three levels are corresponding to three layers of human's cognitive concept structures. Just as Rosch pointed out in his category theory[7]: in the classification system of concrete objects, human's cognitive category can be divided into three layers according to different abstract degree: upper layer (1), basic layer (2) and subordinate layer (3). The higher layer of category has higher abstractness. Except for the most abstract one, each category is embodied by a higher level[8]. See the following diagram (take "tree" as an example):

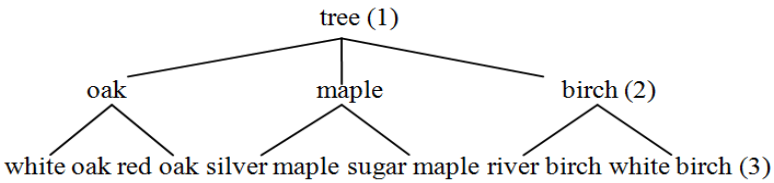


Fig. 1. "Tree" Diagram of Category Levels

In this diagram, the basic layer (2) relates to the start point of human's cognition, so it is called cognitive layer. It is the concrete, well-known gestalt knowledge unit based on people's life experience. The upper layer (1) generalized from the baisc layer has its general characteristics; the lower layer (3) is subordinate one, which includes unique characteristics of each member in this category.

Although semantic primitive is not a newly introduced concept in linguistics, there has not been a standard definition given to it so far. Previous studies show that semantic primitives are existing in different linguistic levels and their definitions are not the same because of various study purposes. However, the common feature of the semantic primitives can be seen from the different levels. It refers to the smallest unit with the most basic meaning in one particular research field compared with other linguistic units. Consequently, semantic primitive has different semantic content and refers to different thing because of different research purposes and fields in different levels.

In lexical level, the analysis on semantic primitives mainly focuses on words because of the study purpose. In this level, the basic words with simple meanings can be used to explain complicated words or meanings just as the semantic primitives

mentioned by Wierzbicka. To be more specific, this kind of words are interpretative primitive words which belong to expressive primitives. They are mainly used in lexical defining. In lexical system, these common basic words are the basic unit, which should be called "primitive words" as An Hualin named.

In the abstract conceptual level, the research field is general concepts. The semantic primitive in this level usually refers to a kind of semantic meaning that is highly generalized and has a feature of abstractness. Huang Zengyang used five primitives {v, g, u, z, r}, which refer to dynamic, stative, attribute, value, and effect respectively in his Hierarchical Network of Concepts Theory[9] to describe the features of abstract concepts. Roger Schank's research on primitives are also focus on abstract concepts. Linguistic concepts are meanings expressed by words or conventional linguistic units. Under this circumstance, the semantic primitives in concept level should be named "conceptual primitives" precisely. They are not concrete lexical words, but the abstract common characteristics of them.

The lower part under the lexical level is semantic level, which is mainly composed by semantic unit—sememes and semes. Sememe is the smallest unit that can be used freely in semantic system. It is usually represented by semantic components, a group of semes, which used to describe the meaning of a word. The semes can be divided into two parts: primary part that expresses the conceptual meaning of a word and secondary part that shows the concomitant meaning. The core semes that describe the conceptual meaning are distinguished from the other semes. This part is semantic primitives.

4 The Role of Semantic Primitive in the Snyset Construction

The semantic primitive in semantic level is closely associated with synset construction. The construction of synset mainly focuses on each sememe of polysemys with the purpose of describing the similarity and difference of a synonyms set (synset) in detail. The similarity of synonyms set is universal while the difference is individual. The determination of common semantic primitive not only provides the reference standard for the synonyms group building, but also highlights the differences. That makes the relationship and distinction among the members in the synset very clear. It also provides a convenient way in semantic formalized description.

4.1 Semantic Primitive and Seme

Semantic Primitive and Seme are two associative concepts with some differences as well. Semantic primitive in semantic level is extracted from the microscopic study on sememes. Sememes are composed by basic meaning and foil meaning[10], corresponding to conceptual meaning and concomitant meaning. Sememes are usually described by semes. The semes that used to describe the conceptual meaning are semantic primitives. So the view that semes are equated with semantic primitives is not exactly right. We can say that word's meaning is described by semes when a

sememe of a word is treated as a whole, because from the perspective of semantic analysis and description, semantic primitives and semes are both semantic components and the smallest semantic units for describing meanings. However, their references are different if sememe is divided into two parts. The different functions of the both are showed adequately in comparison between synonyms in synset. What semantic primitives describe is the conceptual meaning of a sememe, and semes describe the whole sememe. Therefore, the selections of semes are different. Semantic primitives used to describe conceptual meaning are core semes that are limited in their number; while the number of semes used in describing semantic meaning is not fixed and can be increased or decreased according to researchers' subjective divisions.

The difference between semantic primitives and semes can be also distinguished from the perspective of concepts and conceptual words. Concepts refer to contents contained by words. They can be independent from the words. Conceptual words are the combination of concepts and certain linguistic forms, which are usually considered as the common words. For example:

"爹diē", "爸爸bàba" and "父亲fùqīn" (dad and father) refer to the same concept with three different conceptual words in Chinese. Although they are different in forms, their reference and connotation are same: all of them refer to a male who reproduces children (the definition 'a male parent' is not accurate in dictionary) [11]. To describe it with core semes is like this: [reproductive + children + male]. These three core semes are the semantic primitives of conceptual connotation, and "爹diē", "爸爸bàba" and "父亲fùqīn" are all the variations of the one concept. These variations are known as concept anamorphoses which have concrete meanings. These concept anamorphoses have same semantic primitives and also have their own characteristics:

	Semantic Primitives	Distinguishers
爹diē	[+reproductive +children +male]	[+colloquial][+addressing] [+old addressing]
爸爸bàba	[+reproductive + children + male]	[+ colloquial] [+addressing]
父亲fùqīn	[+reproductive + children + male]	[-colloquial] [-addressing]
	Conceptual Meaning	Concomitant Meaning

It can be said that semantic primitives describe the never-changed connotation of a concept, and distinguishers (distinctive semes) display the differences between conceptual words with the same connotation.

To be more precise, the composition of conceptual connotation is:

[semantic primitive1+semantic primitive2+semantic primitive3+…];

the composition of conceptual words is [semantic primitives + distinguishers].

4.2 A Case Study in Snyset Construction and Formalized Description

The living language in daily communication does not always follow a fixed mode or structure. It is usually expressed by different variations. Synset construction means to collect word variations (synonyms) together on one concept connotation in order to

form synonyms set and their system (semantic web) accessible in computer information processing.

With the shared semantic primitives, different word groups can be constructed, such as synset of nouns, verbs and adjectives, etc. Nominal synonyms set (as the example of "爹diē", "爸爸bàba" and "父亲fùqīn") is a kind of word group that collects nouns with same concept but different forms into one group so as to have comparative study and analysis. It is the shared semantic primitives of a synset that provide the reference standard to collect the variations into the group. Through this detailed study and analysis, the nuances lay between the conceptual meaning and concomitant meaning can be clearly displayed, as shown above.

Synonyms set of verb is complicated in construction that mainly focuses on polysemys. A polysemy has more than one sememe. Synset construction has to collect the variations with the same concept based on each sememe. For example:

The verb "拿ná" (take) is a polysemy. It has several sememes: "取qǔ" (to take), "抓zhuā" (to grasp), "搬bān"(to carry), etc. Their common characteristic (conceptual connotation) is [+fingers+to fold]. Only the verbs with this semantic primitives can be collected into the synset "拿ná". After determining the semantic primitives, next step is to describe their differences with semes. The seme analysis should conclude the necessary collocation components of the sememes used in the sentence, including the agent (person), the patient (object), instrument (hand) and the location of the patient (in hand). To take "取qǔ", one sememe of the verb "拿ná" as an example,

"取qǔ" (take) has more components of its concept meaning. Besides the same one [+fingers+to fold] with "拿ná", the semantic primitives of "取qǔ"express like this:

[+agent+to fold fingers+patient+in hands]. We can construct the synset of "取qǔ" according to this standard.

The verb"取qǔ" in the sentence 他取了支笔 (He took a pen) can be replaced by" 摸mō", "掏tāo" and "抽chōu" (synonyms of take) and the meaning of the sentence does not change. The semantic compositions of each word are:

	Semantic Primitives	Distinguishers
取qǔ	[+agent+to fold fingers+patient+in hands]	[○]
摸mō②	[+agent+to fold fingers+patient+in hands]	[+sensing]
掏tāo①	[+agent+to fold fingers+patient+in hands]	[+from mouth of container]
抽chōu①	[+agent+to fold fingers+patient+in hands]	[+among others]

Compared to"取qǔ", the other variations are more concrete in their characteristics. "摸mō" is to sense and take the pen with hands; "掏tāo" is to take the pen from a mouth of container; "抽chōu" is to draw the pen among other things .

In this verb synset, the different variations of the concept are concrete reflections of semantic compositions on syntax. "取qǔ" is a typical syntactic realization, so the distinguishers is none, expresses as [○]; "掏tāo" and "抽chōu" are variations with certain patients, one is to take from the mouth of a container, and the other is to take among something; "摸mō" is a variation with the "sensing" movement of fingers.

The shared semantic primitives here, on the one hand, emphasize the same concept meaning of these verbs, set up the standard in building synset; on the other hand, they

highlight the differences of every member of the synset. What is more, the determination of semantic primitives makes it advantageous to describe the words meaning with function expressions, which are good for the semantic formalized description. The function expression is:

$$y = f(x) \tag{1}$$

To discover the differences among polysemys is to work out the independent variable x. The x here refers to distinguishers; the f represents semantic primitives; y stands for words in synset. For example:

A group of addressing for "father": {爹diē, 爸爸bàba ,父亲fùqīn}:

When f = a male who reproduces children

y_1爹 Diē = f (x: colloquial +old addressing)

y_2爸爸Bàba = f (x: colloquial)

y_3父亲Fùqīn= f (x: written)

A group of hand action of take: {取qǔ, 摸mō, 掏tāo, 抽chōu}:

When f = [+agent+to fold fingers+patient+in hands]

y_1摸mō②= f (x: sensing)

y_2掏tāo①= f (x: from mouth of a container)

y_3抽chōu①= f (x: among others)

5 Conclusion

This article reveals that the semantic primitive exists in different language levels, such as conceptual level, lexical level and semantic level because of different research purposes. But they have a common nature, which refers to the smallest, basic unit in a certain research field. The semantic primitive in semantic level plays an important role. It is the limited semes that express the concepual meaning. The determination of the semantic primitive means to set up a standard for synset construction. With this objective standard we know whether or not a word can be the member of a synset. Meanwhile, it makes the similarity and the difference of the lexical meaning clear in a synset. The case study of synset construction has a universal significance in model building and semantic formalized description.

References

1. Wierzbicka, A.: Semantic Primitives. Athen um, Frankfurt (1972)
2. Li, J.Y.: A Comparison of the Semantic Theories of Wierzbicka and Jackendoff from Perspective of Semantic Primitives. Foreign Language Education 27, 16–18 (2006) (in Chinese)
3. An, H.L.: The Research on Mordern Chinese Primitive Words. China Social Sciences Publishing House, Beijing (2005) (in Chinese)
4. Jackendoff, R.: Semantics and Cognition. MIT Press, Massachusetts (1983)

5. Cheng, Q.L.: A Survey of Jackendoff's Theory of Conceptual Semantics. Foreign Language Teaching and Research 2, 8–13 (1997) (in Chinese)
6. Schank, R.: Conceptual Dependency: A Theory of Natural Language Understanding. Cognitive Psychology 3, 552–631 (1972)
7. Rosch, E.: Cognitive Representations of Semantic Categories. Experimental Psychology 104, 192–233 (1975)
8. Lu, X.L.: The Relational Structure of the Semantic Network and Semantic Retrieval. J.Jiangsu Polytechnic University 10, 82–85 (2009) (in Chinese)
9. Huang, Z.Y.: The Fundamental Theorem and Mathematical Expression on Space of Language Concept. Ocean Press, Beijing (2004) (in Chinese)
10. Zhang, Z.Y., Zhang, Q.Y.: Lexical Semantics. Commercial Press, Beijing (2005) (in Chinese)
11. Mordern Chinese Dictionary, 5th edn. Commercial Press, Beijing (2005) (in Chinese)

The Text Deduction and Model Realization of the Lexical Meanings in Dictionaries Based on "Synset-Lexeme Anamorphosis" and "Basic Semantic Elements and Their Structures"

Guozheng Xiao[1] and Xinglong Wang[1,2,3]

[1] Center for Study of Language and Information, Wuhan University, Wuhan 430072
gzxiao@foxmail.com
[2] College of Chinese Language and Literature, Ludong University, Yantai 264025
[3] Key Laboratory of Language Resource Development and Application of Shandong Province, Yantai 264025
wangxinglong100@163.com

Abstract. This paper discusses the definition deduction and model realization of word meanings in dictionaries from the combined perspectives of "Synset-Lexeme Anamorphosis" and "Basic Semantic Elements and their Structure". The paper attempts to prove the reliability of the semantic items in dictionaries and the implementation of the model used to deduce some semantic items in *Modern Chinese Dictionary* (MCD), including the deduction of semantic items of some compound-word and phrases. Then, some semantic items deriving from the two theories are illustrated.

Keywords: Synset-Lexeme Anamorphosis, basic semantic element (BSE), structure of the basic semantic elements (SBSE), definition model, semantic item, deduction.

1 Drawback in Traditional Lexicographical Theories and Methodologies

Lexicography serves as the important basis for word meaning and natural language processing (NPL), as well as the key object for studies in dictionary compilation and semantic theories. Ideal and accurate definition is expected to secure meaning computation and identification.

Noted Chinese dictionary defining models or approaches fall into three categories: 1) the six "dictionary defining models" devised by Fu[1], who depicts the word class of acts and behaviors with the definitions in *Modern Chinese Dictionary* (MCD) as the basis; 2) the meta language defining model with like field and manner designed by Li[2], inspired by theories in semantics, grammar and meta language; and 3) the three defining models for nouns established by Pan[3], i.e., the model with central elements at the front, the model with central elements in the beginning, and the model with central elements in the middle.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 774–783, 2013.

These models and approaches have their disadvantages. Firstly, the defining models are more theoretically-oriented than practically-feasible, rendering it hard for the realization of semantic items. Secondly, most of them lack comprehensive, scientific or systematic theories as their supporting cornerstone. Thirdly, the deduction of defining models is seldom workable, and the arrangement of semantic items is weak in logic. Lastly, they fail to make the best of existing dictionary defining frameworks and defining sources.

As a modified or improved move, this paper attempts to testify the feasibility of the defining models and the realization of the semantic items. A template is actually stipulated as to how to innovate and improve the models in modern lexicographic definition. The advantages of the defining model designed in this paper will be detailed elsewhere.

2 Theories of "Synset-Lexeme Anamorphosis" (SLA) and "Basic Semantic Elements and Structure of Basic Semantic Elements" (BSE/SBSE)

The SLA theory and BSE/SBSE theory are fabricated, after years of study on related semantic theories and techniques, by Xiao. The former is a theory in which words are aggregated on the basis of word senses. A lexeme consists of the sememe and grapheme of a word. The grapheme is the form in which a word is articulated and written. One sememe may correspond to several graphemes; in other words, these graphemes express the same sememe. Therefore, these graphemes stay together as a synonym group, which we term as "Synset." A Synset is thus a set of words which share the same sememe. [4] The basic semantic elements (BSE's) in the latter BSE/SBSE theory refer to the most fundamental and necessary conceptual ingredients that make up a word sense. Different BSE's of a word or Synset aggregate in certain structure or hierarchy, constituting the SBSE, or structure of BSE's, for the meaning of the word or Synset. BSE's are divided into core BSE's, generic BSE's, main BSE's and pragmatic BSE's. [5] If the core or basic word in a Synset is called the "essence word", then the SBSE of the core word is called "essence SBSE". Likewise, the other words in the Synset establish what is called "anamorphosis (or variant) SBSE's". The relations of such concepts are shown in Fig. 1.:

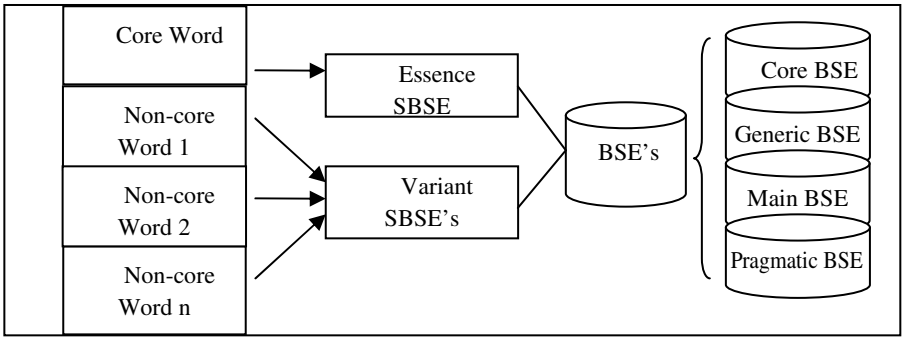


Fig. 1. Basic semantic elements and structure of basic semantic elements in a Synset

3 Systematic Deduction of Semantic Items in Modern Chinese Dictionary (MCD)

3.1 Deduction of Semantic Items from the Perspective of BSE/SBSE

The various semantic items of a dictionary entry or word are usually closely related. If we attempt to deduce, with the guidelines given in the BSE/SBSE theory, the relationships between the items under the Chinese word【用】(*YONG*, Use), we are near to a new template that could be used to separate and establish the semantic items of other words entered in MCD. The above-mentioned "core word", however, will be modified as "core meaning", which usually occurs as the first semantic item under an entry. Hereinafter, the contents contained within [] are the BSE's of a sense, ∧ means conjunction ("and"), ∨ means disjunction ("or"), ∅ means the zero form or absence of such BSE's, j means the trace for meaning transfer, () and {} are for better hierarchy, and the arrow shows the route or approach of the deducing process. If "① 使用" is deemed as the essence SBSE, then items ②, ③ and ④ could be seen as the variant SBSE's of the first essence structure. They are deduced as follows:

【用, YONG】

① 动(*dong*, v.) **使用(*shiyong*, to use)**: ~具(*yongju*, utensil, thing to use)｜~力(*yongli*, be forceful, use force or strength)｜~兵(*yongbing*, resort to arms, use military forces)｜公~(*gongyong*, common, for everyone to use)｜大材小~(*dacai xiao yong*, waste one's talent on a petty job, use talented people for trivial tasks)｜~笔写字(*yong bi xiezi*, write in pen or pencil, use a pen or pencil to write)。

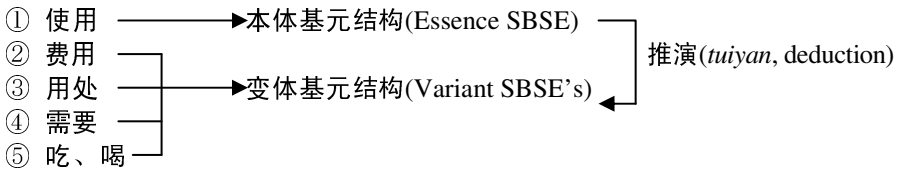
② 名(*ming*, n.) **费用(*feiyong*, fare or expenditure)**: ~项(*yongxiang*, budget, possible expenditure)｜家~(*jiangyong*, family expenses, home expenditure)。

③ 名(*ming*, n.) **用处(*yongchu*, use, usefulness)**: 功~(*gongyong*, function, use)｜多少总会有点~(*duoshao zonghui youdian yong*, be more or less useful, have usefulness anyway)。

④ 动(*dong*, v.) **需要(*xuyao*, need to)**(多用于否定式(*duo yongyu foudingshi*, often in negation)): 天还很亮，不~开灯(*Tian hai mei liang*, *buyong kaideng*, You don't need to turn on the light for it's still daytime)｜东西都准备好了，您不~操心了(*Dongxi dou zhunbei hao le*, *nin buyong caoxin le*, You don't need to bother as everything is ready)。

⑤ 动(*dong*, v.) **吃、喝(*chi/he*, eat/drink)**(含恭敬意(*han gongjing yi*, respectfully)): ~饭(*yong fan*, have or eat one's meal, help oneself to food)｜请~茶(Q*ing yong cha*, Please have or drink tea)。

⑥ <书>连(<*shu*, lit.>) **因此; 因(*yinci/yin*,** hence; therefore)(多用于书信(*duo yongyu shuxin*, often in letter)): ~特函达(*yong te han da*, hence this letter)。⑦名(*ming*, n.) **姓(*Xing*, Surname)**。



①*使用：[主体]∧[支配] ∧ [某事物] ∧ [服务] ∧ [某目的] ∧ [行为]*
[消费者]∧[支配]∧[货币]∧[服务]∧[交易]∧[∅]∧[行为]
[消费者]∧[支配]∧[货币]∧[服务]∧[交易]∧[花销]∧[行为]
[消费者]∧[支配]∧[货币]∧[服务]∧[交易]∧[花销]∧[∅]
②费用：**[消费者]∧[支配]∧[货币]∧[服务]∧[交易]∧[花销]**
①*Use:[Subject]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Act]*
[**Consumer**]∧[Dominate]∧[**Currency**]∧[Serve]∧[**Deal**]∧[∅]∧[Act]
[Consumer]∧[Dominate]∧[Currency]∧[Serve]∧[Deal]∧[**Expense**]∧[**Act**]
[Consumer]∧[Dominate]∧[Currency]∧[Serve]∧[Deal]∧[Expense]∧[∅]
②Expenditure**:[Consumer]∧[Dominate]∧[Money]∧[Serve]∧[Deal]∧[Expense]**
①*使用：[主体] ∧ [支配] ∧ [某事物] ∧ [服务] ∧ [某目的] ∧ [行为] ∧ [∅]*
{[人]∧[动物]}∧[支配]∧[某事物]∧[服务]∧[某目的]∧[行为]∧[方面]
[∅]∧{[人]∧[动物]}∧[支配]∧[某事物]∧[服务]∧[某目的]∧[∅]∧[方面]
[某事物j]∧[被]∧{[人]∧[动物]}∧[支配]∧[∅j]∧[服务]∧[某目的]∧[方面]
③用处：**[某事物]∧[被]∧{[人]∨[动物]}∧[支配]∧[服务]∧[某目的]∧[方面]**
① *Use:[Subject]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Act]∧[∅]*
{[**Human**]∧[**Animal**]}∧[Dom]∧[Th]∧[Serve]∧[Aim]∧[**Act**]∧[**Someway**]
[∅]∧{[Hum]∧[Ani]}∧[Dom]∧[Th]∧[Serve]∧[Aim]∧[∅]∧[Someway]
[**Thj**]∧[**Be**]∧{[Hum]∧[Ani]}∧[Dominated]∧[∅j]∧[Serve]∧[Aim]∧
[Someway]
③Usefulness:**[Th]∧[Be]∧{[Hum]∨[Ani]}∧[Domed]∧[Serve]∧[Aim]∧**
[Someway]

①*使用：[主体]∧[支配]∧[某事物]∧[服务]∧[某目的]∧[∅]∧[行为]*
[主体]∧[支配]∧[某事物]∧[服务]∧[某目的]∧[需求]∧[行为]
[主体]∧[支配]∧[某事物]∧[服务]∧[某目的]∧[需求]∧[∅]
④需要：**[主体]∧[支配]∧[某事物]∧[服务]∧[某目的]∧[需求]**
① *Use: [Subject]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[∅]∧[Act]*
[Sub]∧[Dom]∧[Th]∧[Serve]∧[Aim]∧[**Need**]∧[**Act**]
[Sub]∧[Dom]∧[Th]∧[Serve]∧[Aim]∧[**Need**]∧[∅]
④ **Need to:[Sub]∧[Dom]∧[Th]∧[Serve]∧[Aim]∧[Need]**

①*使用：[主体]∧[支配]∧[某事物]∧[服务]∧[某目的]∧[行为]*
⑤吃喝：{[人]∨[动物]}∧[支配]∧[食物]∧[服务]∧[身体]∧[行为]
① *Use: [Subject]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Act]*
⑤ **Eat/Drink:** {[**Hum**]∨[**Ani**]}∧**[Dom]**∧[**Food**]∧**[Serve]**∧[**Health**]∧**[Act]**

Observed from the perspective of BSE and SBSE, the structures of BSE for semantic items ② to ④ almost remain constant with that for ①, except for the few variable elements in the structure of semantic item ①.

3.2 Deduction of Semantic Items from the Perspective of BSE/SBSE

Under the entry of 【用】(*YONG*, Use) in MCD are 30 double-syllable items (two Chinese characters united as one word/phrase), which constitute a Synset with the BSE's and SBSE of 【用】 as the basis. The senses of 【用】 form the essence SBSE, while those of other words/phrases establish the variant SBSE. We attempt to improve, driven by the fore-said two theories, present-day lexicographic models and items, an effort that is expected to make dictionary definitions more systematic, logical and reader-friendly.

The BSE's and SBSE of the entry 【用】(*YONG*, Use) are generalized, after examination of large quantities of language data and reference to dictionary definitions, as follows:

[主体]∧[支配]∧[某事物]∧[服务]∧[某目的]∧[行为]
[ZT]∧[ZP]∧[MSW]∧[FW]∧[MMD]∧[XW]
（Abbreviations of their Chinese *pinyin* initials are equally effective.）
[Subject]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Act] (Translated)

Besides, the above SBSE can be regarded as the essence structure for the Synset of 【用】(*YONG*, Use), rendering the SBSE for other words or senses attainable through similar deduction. Discussions as such are made from the aspects of core BSE's, generic BSE's, main BSE's and pragmatic BSE's. Only one example for each type is provided.

3.2.1 Core BSE's: [支配] ([ZP], [Dominate]) and [服务] ([FW], [Serve])

It is evident from analysis that the various SBSE's for the items under the entry of 【用】(*YONG*, Use) share the core BSE's of *[支配](Dominate)* and *[服务](Serve)*. In other words, these two BSE's stand as the essential reasons for the aggregation of the Synset members. Core BSE deduction means the SBSE of some word meanings can only be attainable through substitution with other core BSE's, and on the basis of the essence SBSE. The deduction of this type of BSE's are inevitably related to other core BSE's like [ZP] (Subject) and [FW] (Serve).

(1) Deduction of Core BSE [ZP] ([Dominate])**:**
【用】(*YONG*, Use)
[Subject]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Act]
(This line to be omitted hereinafter)
【用命】(*YONGMING*, Obey)
[Inferior]∧[Dominate]∧[Command]∧[Serve]∧[Superior]∧[Act] Deduction I
[Inferior]∧[Execute]∧[Command]∧[Serve]∧[Superior]∧[Act] Deduction II

（2）Deduction of Core BSE [FW] ([Serve]):

【用刑】 (*YONGXING*, Punish)

[Punisher]∧[Dominate]∧[Instrument]∧[Serve]∧[Punishee]∧[Act]

Deduction I

[Punisher]∧[Dominate]∧[Instrument]∧[Punish]∧[Punishee]∧[Act]

Deduction II

3.2.2 Generic BSE: [行为]([XW], [Behavior])

Most SBSE's for the items under the entry of 【用】(*YONG*, Use) share the core BSE of *[行为]* (*[Act]*). In other words, [Act] stands as the essential reason for the aggregation of the Synset members.

(1) Main BSE→Generic BSE

The original generic BSE will disappear, under the effect of the interaction of other BSE's, and leave a vacancy for its absence. The original main BSE(’s) will then be displaced and fill the vacancy, thus forming a new generic BSE.

- **Type [ZP] ([Dominate]):**

【用户】(*YONGHU*, User)

[Human]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Act] Deduction I

[∅j]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Act]∧[Human] Deduction II

[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Human] Deduction III

- **Type [MSW] ([Thing]):**

【用人】(*YONGREN*, Servant)

[∅]∧[Human]∧[Dominate]∧[**Another**]∧[Serve]∧[Aim]∧[**Act**]

Deduction I

[**Be**]∧[Human]∧[Dominated]∧[**∅j**]∧[Serve]∧[Aim]∧[∅]∧[**Another j**]

Deduction II

[Be]∧[Human]∧[Dominated]∧[Serve]∧[Aim]∧[Another]

Deduction III

(2) Pragmatic BSE→Generic BSE

The original generic BSE will disappear, under the effect of the interaction of other BSE's, and leave a vacancy for its absence. A new pragmatic BSE may occur under the effect of the compound-word senses. It will then fill the vacancy, thus forming a new generic BSE.

【用场】(*YONGTU*, Usefulness)

[∅]∧[Human]∧[Dominate]∧[**Thing**]∧[Serve]∧[Aim]∧[Act]∧[∅] Deduction I

[**Thingj**]∧[**Be**]∧[Hum]∧[Dominated]∧[**∅j**]∧[Serve]∧[Aim]∧[∅]∧[**Someway**]

Deduction II

[Thing]∧[**Be**]∧[Hum]∧[Dominated]∧[Serve]∧[Aim]∧**[Someway]** Deduction III

3.2.3 Main BSE's: [主体] ([ZT], [Subject]), [某事物] ([MSW], [Thing]), and [某目的] ([MMD], [Aim])

The SBSE's for the Synset members under the entry of 【用】(*YONG*, Use) in MCD share the three BSE's of *[Subject]*, *[Thing]* and *[Aim]*. A few entry members possess

only the three elements, while a larger part have the hyponymy variants from these three. It is because of the existence of the main BSE's that the Synset, with 【用】 (*YONG*) as the essence, has become varied and rich.

(1) BSE Variant 1: [ZT]([Subject]), [MSW] ([Thing]), and [MMD] (Aim)

The SBSE's of some word senses can usually be acquired by replacing *[Subject]*, *[Thing]* or *[Aim]* with their hyponyms.

【用兵】(*YONGBING*, Use arms)

[**将领**]∧[Dominate]∧[**士兵**]∧[Serve]∧**[战争**]∧[Act]

When the elements of *[Subject]*, *[Thing]* and *[Aim]* are replaced with *[*将领*]* (*[General]*), *[*士兵*]* (*[Soldiers]*) and *[*战争*]* (*[War]*), respectively, the structure for 【用兵】 (*YONGBING*) is thus formed.

(2) BSE Variant 2: [ZT] ([Subject]) and [MSW] ([Thing])

- **Direct Replacement**

Direct replacement of BSE's means that the structure of some senses can be directly developed by replacing *[ZT]* (*[Subject]*) and *[MSW]* (*[Thing]*), usually with their hyponyms.

【用人】(*YONGREN*, Employ)

[**上级**]∧[Dominate]∧[**下级**]∧[Serve]∧[Aim]∧[Act]

When *[ZT]* (*[Subject]*) and *[MSW]* (*[Thing]*) are replaced with *[*上级*]* (*[Superior]*) and *[*下级*]* (*[Inferior]*), respectively, the structure for 【用人】(*YONGREN*, Employ) is thus established.

- **Indirect Replacement**

Indirect replacement of BSE's means that the structure of some senses cannot be directly developed. But rather, the structure shall be deduced for two or more than two steps, and be connected with the main element *[MSW]* (*[Thing]*).

【用事1】(YONGSHI 1, Be in (supreme) power)

[Human]∧[Dominate]∧[**事务**]∧[Serve]∧[Aim]∧[Act] Deduction I

[Human]∧[Dominate]∧[权利]∧[Serve]∧[Aim]∧[Act] Deduction II

The structure for 【用事1】(YONGSHI 1, Be in (supreme) power) can only establish itself when *[Thing]* is replaced with *[*事务*]* (*[Affair]*), which in turn is replaced with *[*权利*]* (*[Power]*).

(3) BSE Variant 3: [ZT] ([Subject])

【用处】(YONGCHU, Use(s))

[∅]∧{[人]∨[动物]}∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[**Act**]∧[∅] Deduction I

[**Be**]∧{[人]∨[动物]}∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[∅]∧

[**Someway**]

Deduction II

[**Thing j**]∧[**Be**]∧{[人]∨[动物]}∧[Dominated]∧[**∅j**]∧[Serve]∧[Aim]∧ [**Someway**] Deduction III

[Thing]∧[Be]∧{[人]∨[动 物]}∧[Dominated]∧[Serve]∧[Aim]∧**[Someway]** Deduction IV

The structure for 【用处】(YONGCHU, Use(s)) can only be developed by replacing *[Subject]* with its hyponyms *{[人]∨[动物]}* (*[Human]* or *[Animal]*), and by deducing otherwise.

3.2.4 Pragmatic BSE's: [X]

[X] refers to an open, uncertain pragmatic element. It is based on the essence structure for BSE's, and aimed at distinguishing the several variant structures. Therefore, the number of [X] is controllable.

Pragmatic elements are effective in two ways. On the one hand, they help to fine tune and clarify the particular features of some word senses, an effect the deduction with only core elements, generic elements and main elements is unable to achieve. On the other, they serve to tell apart some items that are very close in meaning and are only subtly different from each other.

(1) Direct Addition

Direct addition of BSE refers to the practice in which the structures for some word meanings can be attained by replacement with main BSE's, and then by direct addition of some new pragmatic elements.

- **Core-BSE Oriented**

In this type of element addition, a new pragmatic BSE is added on the basis of core element(s), or to modify the core element(s).

【用事2】(*YONGSHI* 2, Act on impulse)

[Human]∧[**草率**]∧[Dominate]∧[Affair]∧[Serve]∧[Aim]∧[Act]

The newly-added *[草率]* (*[Impulsively]*) is oriented to the core element *[Dominate]*.

- **Main-BSE Oriented**

In this type of element addition, a new pragmatic BSE is added on the basis of main element(s), or to modify the main element(s).

【用劲】(*YONGJIN*, Exert one's strength)

{[Human]∨[Animal]}∧[Dominate]∧[**大量**]∧[**力气**]∧[Serve]∧[Aim]∧[Act]

The newly-added *[大量]* (*[Much]*) is oriented to the main element *[力气]* (*[Strength]*).

- **All-BSE's Oriented**

In this type of element addition, a new pragmatic BSE is added on the basis of all the elements, or to modify the whole set of BSE's.

【用餐】(*YONGCAN*, Have one's meal, Help oneself to food or drink)

{[Human]∨[Animal]}∧[Dominate]∧[Food]∧[Serve]∧[Health]∧[Act]∧

[**正式**]

The newly added *[正式]* (*[Formally]*) is oriented to all the BSE's preceding it.

（2）Deducing Addition

Deducing addition of BSE refers to the practice in which the structures for some word meanings cannot be attained by further adding—after replacement of main element(s)—a new pragmatic element, but can be attained—after replacement—by other means like replacing, adding, deleting or even displacing some other elements.

- **Main-BSE Oriented**

【用功2】(YONGGONG 2, Earnest, Dedicated)

[Human]∧[Dominate]∧[∅]∧{[时 间]∨[精 力]}∧[Serve]∧[Aim]∧[**Act**]∧[∅] Deduction I

[Human]∧[Dominate]∧[**大量**]∧{[时间]∨[精力]}∧[Serve]∧[Aim]∧[∅]∧

[**特点**]

Deduction II

[Human]∧[Dominate]∧[大 量]∧{[时 间]∨[精 力]}∧[Serve]∧[Aim]∧[特 点] Deduction III

The newly-added *[*大量*]* (*[Much]*) is oriented to the main elements *{[*时间*]*∨*[*精力*]}* (*[Time]* or *[Energy]*). The added *[*特点*]* (*[Quality]*) in the end denotes this entry is an adjective (or adverb).

- **All-BSE's Oriented**

【用费】(*YONGFEI*, Use money, Expend)

[消费者]∧[Dominate]∧[货币]∧[Serve]∧[交易]∧[**Act**]∧[花销]∧[∅] Deduction I

[消费者]∧[Dominate]∧[货币]∧[Serve]∧[交易]∧[∅]∧[花销]∧[**普通**] Deduction II

[消费者]∧[Dominate]∧[货币]∧[Serve]∧[交易]∧[花销]∧[普通]

Deduction III

The newly-added element *[*普通*]* (*[Neutrally]*) is oriented to all the basic elements including *[*消费者*]* (*[Consumer]*), *[*货币*]* (*[Currency]*), *[*交易*]* (*[Deal]*), and *[*花销*]* (*[Expense]*).

- **Core-BSE Replaced**

In this type of element replacement, the structure for some word meanings are attained by deleting the generic elements within the essence SBSE and then adding some other generic BSE('s).

【用法】(*YONGFA*, Usage)

[Human]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[Act]∧[∅] Deduction I

[Human]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[∅]∧[**方法**] Deduction II

[Human]∧[Dominate]∧[Thing]∧[Serve]∧[Aim]∧[**方法**] Deduction III

After deleting the element *[Act]*, the newly-added generic element *[*方法*]* (Means) constitutes, along with others, the structure for the entry 【用法】(*YONGFA*, Usage)

4 Transition from SBSE and Text Deduction to Defining Practice

So far, the structure of BSE has set up the defining framework for dictionary entries. As BSE's are the necessary conceptual ingredients for entries to be defined, these basic semantic elements shall be aggregated—in defining practice—in an effective manner with meta language. A few examples are given below[1]:

[1] For better readability, some defining texts could be further improved.

【用】(*YONG*, Use) 主体支配某事物服务某目的的行为(Act of Subject Dominating a Thing to Serve an Aim)

【用典】(*YONGDIAN*, Use literary quotations) 人利用典故服务写作和言论的行为(Act of Human Quoting a Sentence/Story to Serve Writing and Speech)

【用兵】(*YONGBING*, Use arms) 将领支配士兵服务战争的行为 (Act of General Dominating Arms to Serve War)

【用工】(*YONGGONG*, Use labor force) 管理者支配工人服务厂企等的行为 (Act of Management Dominating Staff to Serve Production)

【用语】(*YONGYU*, Use language) 说话者支配言语服务交际的行为 (Act of Speaker Dominating Language to Serve Communication)

【用费】(*YONGFEI*, Use money, Expend) 消费者支配货币进行交易的花销，较普通(Act of Consumer Dominating Currency to Serve Purchase)

Similar main BSE's shall be put together to better compare and distinguish different senses. Alphabetic order is not necessarily preferable in this manner.

5 Conclusion

Based on the theories of "SLA" and "BSE/SBSE", the essence SBSE of an entry's senses is depicted, the relations between and regularity of such senses generalized, and the original dictionary-borne semantic items deduced in a systematic manner. The variant SBSE's of all other semantic senses under the entry, except for the very first core sense, are subsequently described. In turn, a whole new template of semantic defining version is formulated.

References

1. Fu, H.Q.: Analysis and Description of Meanings. Language Publishing House, Beijing (1996) (in Chinese)
2. Li, B.J.: Introduction to Semantics and Grammar. China Book Store, Beijing (2007) (in Chinese)
3. Pan, X.L.: Definition Model Selection for Nouns in Dictionaries. Lexicographical Studies. 26–36 (2011)
4. Xiao, G.Z.: The Description of the Lexical Meaning Structure of the Verb "da" and the Construction of its Synset—A Study on "Synset-Lexeme Anamorphosis" Shared by Human and Computer. In: Xiao, G.Z., Ji, D.H., Sun, M.S. (eds.) ICCC 2007. Publishing House of Electronics Industry, Beijing (2007)
5. Xiao, G.Z., Xiao, S., Guo, T.T.: The Mapping from Space of Concept Primitives to Space of Word Meaning Primitives. Journal of East China Normal University (Philosophy and Social Sciences), 139–143 (2011)

Semantic Labeling of Chinese Serial Verb Sentences Based on Feature Structure

Bo Chen[1,3], Hongmiao Wu[2], Chen Lv[3], Hua Yang[3], and Donghong Ji[3]

[1] Dept of Language & Literature, Hubei University of Art & Science, Xiangyang, China
cb9928@gmail.com
[2] School of Foreign Languages & Literature, Wuhan University, Wuhan, China
Hongmiao23@163.com
[3] Computer School, Wuhan University, Wuhan, China
lvchen1989@gmail.com, yanghuastory@263.net,
donghongji_2000@yahoo.com.cn

Abstract. Parsing Chinese serial verb sentences is a key issue in NLP. Many controversies arise from serial verb sentences. This paper puts forward a novel model "the Feature Structure theory" to resolve the semantic labeling of Chinese serial verb sentences. We analyze the difficulties in annotating these sentences, and compare Feature Structure with traditional dependency structure. Feature Structure represents more semantic information and more semantic relations. Feature Graph is a recursive undirected graph, allows nesting and multiple correlations.

Keywords: Chinese serial verb sentences, Feature Structure, semantic labeling, dependency structure, graph.

1 Introduction

In natural language processing, semantic parsing is one of the most challenging topics in the modern fields of computational linguistics, as well as one of the main bottlenecks of large-scale applications of language information technology today[1-3]. As a typical Chinese special sentence pattern, Chinese serial verb sentence has two and more verbs, which arises many problems in semantic parsing. This article studies a novel model for Chinese semantic representation based on Feature Structure, and analyzes Chinese serial verb sentence using Feature Structure. Comparing traditional dependency structure, we achieve a better result of semantic parsing based on Feature Structure.

1.1 Characteristics of Chinese Serial Verb Sentence

Chinese serial verb sentences can be described as: "Subject +Verb Phrase_1 +Verb Phrase_2 +Verb Phrase_n". Its characteristics include that two or more serial Verb Phrases compose the predicate of the sentence, they share a common subject. The structures of these verb

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 784–790, 2013.

phrases are compact. There aren't syntactic relations between them, such as: subject-predicate, modifier-head, predicate-object, predicate-complement, etc. In terms of time sequence or logic sequence, these verb phrases are arranged in order [4-6].

Example 1:

我 开 车 去 车站 接 他
/wo/ /kai/ /che / /qu / /chezhan / /jie / /ta /.
I drive car go station pick up him
I drive a car to go to station to pick up him.

In Example 1, there are three verb phrases: "开车" (means drive a car), "去车站" (means go to station), "接他" (means pick up him). They are arranged by time sequence.

The semantic relations of a Chinese serial verb sentence include two parts: the semantic relations between the subject and these verb phrases, the semantic relations among these verb phrases.

1.2 Difficulties in Parsing Chinese Serial Verb Sentences

Currently, traditional dependency structure is the main semantic analysis method to parse Chinese language [7-8]. When we build Chinese labeling corpuses, most of us will choose dependency structure. However, when we use it to parse Chinese serial verb sentence, there are many difficulties.

In accordance with traditional dependency rule, only one verb can be the head of a sentence, all of the other words depend on it [1], [8]. As for Chinese serial verb sentence, there are at least two verbs, and maybe more. It is hard to ensure which one is the head.

Chinese information processing provides a compromise that the first verb of Chinese serial verb sentence is regarded as the head; the following verbs depended on it [9-10]. Fig.1 is the dependency tree of Example 1.

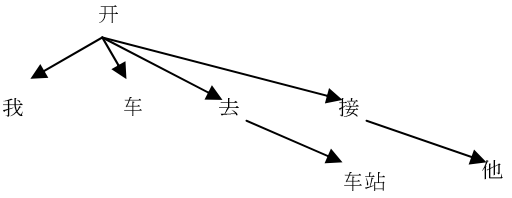


Fig. 1. Dependency tree of Example 1

According to semantic relatedness and semantic cognition, we analyze Example 1. There are 6 word pairs with semantic relation at least, such as:

(我, 开), (我, 去), (我, 接), (开, 车), (去, 车站), (接, 他).

Besides, Example 1 also contains other semantic information, we can ask questions to find it:

"where do we pick him up?"
"How do we pick him up?"
"How can we go there?"

The answers are “车站” (means station), “用车接” (means by car), “开车” (means drive).

Using traditional dependency structure to parse Chinese serial verb sentence will lost much semantic information, and will bring problems for the following Chinese processing.

2 Feature Structure Theory

The final purpose of semantic parsing in Machine Translation is to find the semantic relations in a sentence [11]. We focus on the representation of semantic relatedness and relatedness classification.

Example 2:

从	广州	飞	飞	到	武汉
/cong/	/Guangzhou/	/fei/	/fei/	/dao/	/Wuhan/
From	Guangzhou	fly	fly	to	Wuhan
Fly from Guangzhou			fly	to	Wuhan

In Example 2, there are semantic relatedness between “飞” (means fly) and “广州”(means Guangzhou), between “飞”(means fly) and “武汉”(means Wuhan). If we apply relatedness classifications to Example 2, it can be described as:

飞-从(the beginning)-广州
飞-到(the destination)-武汉

The two triples can be expressed as a set of triples of an Entity, a Feature and a Value. We name it “Feature Triple” of the phrase or sentence structure [12].

Feature Triple: [Entity, Feature, Value]

A Feature Triple can represent a group of semantic relatedness. Example 2 can be described as following:

[飞, 从, 广州]
[飞, 到, 武汉]

Fig. 2 shows the graph of Feature Structure. Formally, a triple can be represented as two nodes and the edge connecting them. The nodes stand for words, the edge stands for feature. The node serves as the owner of the feature, while the other nodes as value.

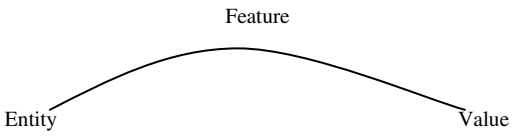


Fig. 2. Feature Structure

Example 3:

他 说 他 是 大学 教师
/ta/ /shuo/ /ta/ /shi/ /daxue/ /jiaoshi/
He say he is university teacher
He says he is an university professor.

Example 3 can be described as following:

[说, ,他]

[说, ,他是大学教师]

[是, ,教师]

[是, , 他]

[教师, ,大学]

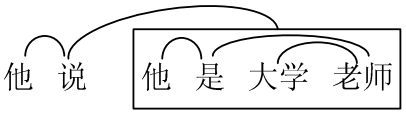


Fig. 3. Feature graph of Example 3

Fig. 3 shows that Example 3 has 5 feature triples. “他 (means he)” is the entity and one value of “说 (means say)”. “他是大学教师 (means he is a university professor)” is the other value of “说 (means say)”. Here “他是大学教师 (means he is a university professor)” is treated as an entirety to produce semantic relations with “说 (means say)”. And the node “他是大学教师 (means he is a university professor)” itself is a Feature Structure. “是 (means is)” is the entity, “大学教师 (means a university professor)” is its value, “他(means he)” is its another value. Besides, the node “大学教师 (means a university professor)” itself is also a Feature Structure. “教师 (means professor)” is the entity; “大学 (means university)” is its value.

Fig. 4 shows the Feature Graph. Formally, Feature Structure can be seen as a recursive undirected graph, which means that the node itself can be a graph [1], [8]. Feature Structure allows nesting and multiple correlations.

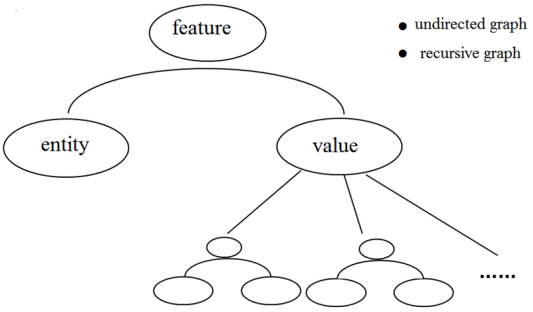


Fig. 4. Feature Structure Graph

3 Parsing Chinese Serial Verb Sentences with Feature Structure

Example 4:

我 推开 门 走 出去.
/wo/ /tui kai/ /men/ /zou/ /chuqu/
I push door walk outside
I open the door and go outside.

Example 4 is a typical Chinese Serial Verb Sentence. The first verb phrase is "推开门" (means open the door), the second verb phrase is "走出去" (means go outside), the two phrases are arranged in time order. Its feature graph is Fig.5:

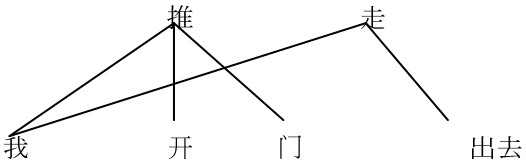


Fig. 5. Feature Graph of Example 4

Example 5：

我 买 了 碗 面 吃.
/wo/ /mai/ /le/ /wan/ /mian/ /chi/
I buy bowl noodle eat
I bought a bowl of noodle to eat.

In Example 5, the first verb phrase is "买了碗面" (means bought a bowl of noodle), "面" (means noodle) is the object of the verb "买" (means buy). The second verb phrase doesn't have object, only have a verb "吃" (means eat). The characteristics of Example 5 are the object of the first verb "买" (means buy)—— "面" (means noodle) is the patient of the second verb "吃" (means eat). The two verb phrases have semantic relatedness. Its feature graph is Fig.6[1]:

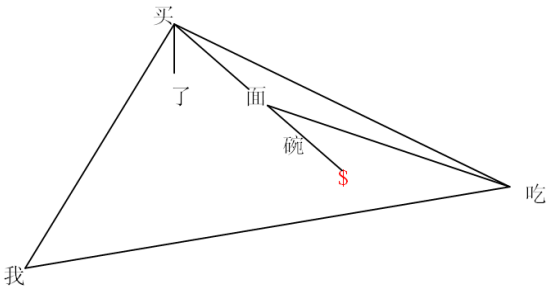


Fig. 6. Feature Graph of Example 5

[1] The characteristics of Example 5 is that the number "一"(means one) in the phrase "一碗面" is omitted. we use"$"to describe the phenomena. It is the characteristics of Chinese language.

4 Comparing Feature Structure with Dependency Structure

We use two methods of Dependency Structure and Feature Structure to parse Example 5. Table 1 is the Dependency Tree and Feature Structure graph of Example 5, and Table 2 is the specific parsing results.

Table 1. Dependency Tree and Feature Graph of Example 5

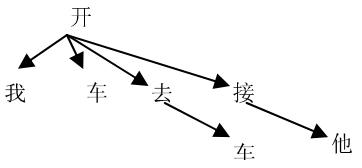
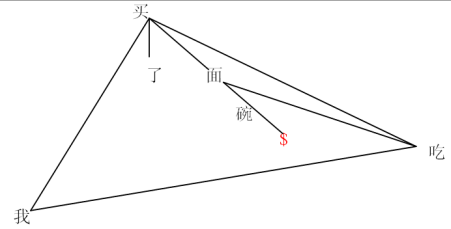
Dependency Tree	Feature Structure graph
	

Table 2.

Semantic relations between words		The consequence of Feature Structure	The consequence of Dependency Structure
exist	Not exist		
我,买		+	+
买,了		+	+
买,面		+	+
碗,面		+	+
吃,面		+	-
我,吃		+	-
	买,吃	-	+

In Table 2, Dependency Tree can not represent the semantic relations between “我” (means I) and “吃” (means eat), “吃” (means eat) and “面” (means noodle). However, it represents the semantic relation between “买” (means buy) and “吃” (means eat) which does not exist. Dependency Tree omitted two semantic relations, and labeled a semantic relation that does not exist.

As for Chinese Serial Verb Sentences, Feature Structure graph can represent more semantic relations and more semantic information.

Firstly, Dependency structure can only describe the semantic relation between the subject and the first predicate verb. Feature Structure can describe completely the semantic relations between the subject and all predicate verbs.

Secondly, Dependency structure can not describe the semantic relation between the object of one predicate verb and another predicate verb. Feature Structure can describe it.

Thirdly, in Chinese Serial Verb Sentences, two predicate verbs may have semantic relations or not. Dependency Tree cannot discriminate the situation. No matter whether the actual semantic relation does exist or not, Dependency Tree will order the first

verb as the head of the sentence. Feature Structure can represent the semantic relations in terms of the actual situation.

5 Conclusion and Future Work

According to the semantic representation, we put forward a mechanism "Feature Structure". It is used to represent Chinese phrases and sentences. Feature Structure can be represented as a recursive undirected graph, and allows nesting and multiple correlations.

It is an attempt to use Feature Structure to label Chinese Language. Now we have built the basic concepts and description frameworks of Feature Structure, and built a large–scale Chinese semantic resource with 30,000 sentences. As for the applications, our research can be used directly to relation extraction, event extraction, automatic question & answering as well as the syntactic parsing in machine translation.

Acknowledgment. The work is funded by the following projects: the Major Projects of Chinese National Social Science Foundation No.11&ZD189, National Natural Science Foundation of China No.61202193,61133012,61173062,61070082,61070083, and 61070243.

References

1. Chen, B., Ji, D.H.: Chinese Semantic Parsing Based on Dependency Graph and Feature. In: 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, Shenyang, China, pp. 1731–1734 (2011)
2. Nivre, J., Scholz, M.: Deterministic dependency parsing of English text. In: Proceedings of the 20th International Conference on Computational Linguistics, pp. 64–70 (2004)
3. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 Shared Task on Dependency Parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL, pp. 915–932 (2007)
4. Li, L.D.: Modern Chinese Sentence Patterns, pp. 302–307. The Commercial Press (1986)
5. Xing, X.: on the characristics of Chinese Serial Verb Sentences. Journal of Xinjiang University, 116–122 (1987)
6. Yang, Y.R.: The semantic relation of Chinese Serial Verb Sentences and Chinese Pivotal Sentences. Journal of Southwest China Normal University, 96–100 (1992)
7. Buchholz, S., Marsi, E.: CoNLL-X Shared Task on Multilingual Dependency Parsing. In: Proc. of the 10th CoNLL-X, NY, pp. 149–164 (2006)
8. Chen, B., Ji, D.H., Lv, C.: Semantic Labeling of Chinese Subject-Predicate Predicate Sentence Based on Feature Structure. Journal of Chinese Information Processing, 24–32 (2012)
9. Zhou, M., Huang, C.N.: Approach to the Chinese Dependency Formalism For the Tagging of Corpus. Journal of Chinese Information Processing, 35–53 (1994)
10. Zhou, Q.: Annotation Scheme for Chinese Treebank. Journal of Chinese Information Processing, 1–8 (2004)
11. Feng, Z.W.: Machine Translation. Translation and Publishing Corporation, 412–434 (1998)
12. Chen, B.: Building a Chinese Semantic Resource Based on Feature Structure. Doctoral Dissertation. Wuhan University, China (2011)

Study on Predicate-Only Lexical Items in Mandarin Chinese

Xiaojuan Ma[1] and Zhanhao Jiang[2]

[1] Department of Chinese Language and Character, Hubei University, Wuhan, 430062
maxiaojuan@whu.edu.cn
[2] School of English studies, Xi'an International Studies University, Xi'an, 710128

Abstract. The predicate-only lexical items are subsidiary in Chinese lexicon in terms of parts of speech. From some relatively closed corpus, we choose and classify such lexical items out of Verb Group, Adjective Group and Descriptive Word Group. We also classify the adhesive verbs and conclude that the adhesion to other syntactic components is the main reason why these verbs function only as the predicate in a sentence.

Keywords: single notional words, predicate-only lexical items, adhesion, semantic feature.

1 Introduction

Predicate, as the counterpart of subject in a sentence, belongs to the first level of a sentence together with subject, with the predicate as the research focus. In mandarin Chinese, predicate is indispensable to a sentence while a subject may not be. However, this article shall focus its attention on predicate-only lexical items(referred to as Pre-Only hereafter).

On the basis of the concept "non-predicate", Lü Shu-xiang once advanced the term "predicate-only adjectives" [1]which, as the name suggests, indicate that such adjectives can serve only as predicate in a sentence. The current article will expand this notion and coin the term-"Pre-Only", lexical items that can function grammatically as predicate in a sentence. By "grammatically" we mean that such lexical items can serve as one of the syntactical components in a narrow sense. Literature review shows that Pre-Only can be found among verbs themselves(Verb Group), adjectives(excluding attributive words)(Adjective Group) and descriptive words(Descriptive Word Group), which implies that Pre-Only are special in terms of parts of speech.

2 Related Work

There are Pre-Only in both the Verb Group and the Adjective Group in mandarin Chinese. For example, in the former, we have倒 dao "pour in", 急于 jiyu "hurry to", 加以 jiayi "make"; In the latter, we have adjectives such as难 nan "difficult", 容易

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 791–799, 2013.

rongyi “easy”, 多duo “many”, 少shao “little”, 对dui “right” 错cuo “wrong”, which were once listed by Lü Shuxiang. According to YuanYulin, in Chinese, some verbs are tightly combined with objects, which do not allow the objects to live without it. Examples are 属于 shuyu “belong to”, 成为 chengwe i“to become”, 不及buji “less than”, 不如 buru “rather than” 姓xing “name”, 是 shi “is”, 等于dengyu “equal to”, 具有 juyou“have”, 号称 haocheng “to be known as” . The object of these words cannot be moved to the head of sentence as a topic. Yin Shichao [2] focus on "adhesive verb" in detail. In his article Study on the Adhesive Verb, in which he assumes that "adhesive verb" refers to those which aren’t self-sufficient in syntactic sense and which alone cannot be responses to question sentences, and must have the co-occurrence of other corresponding verbs in a sentence. In Adhesive Verb Theory we have such sub-categorizations as object-adhesive verbs, adverbial-adhesive verbs, complement-adhesive verbs, predicate-adhesive verbs [3-5] and others .

Qi and Wang delineated "Predicate-only Adjectives"—adjectives which can’t be attribute in terms of syntactic components. Qi and Wang also divided them into two groups: in group one adjectives can’t serve as attribute in any case while in group two, adjectives can be the center of a modifier-head construction, which in turn as a whole serve as the attributive adjective [6]. However, this article will pay particular attention to the nuanced and subtle disparities between "Predicate-only adjectives" and the "Predicate-only lexical items" we’ll cover, although they both contain "predicate-only" in their names. The “predicate-only” in the title of this paper is defined only from the perspective of lexical items’ function of being a predicate, while the "predicate-only adjective" is in contrast to the non-predicate adjective, which means that such definition excludes adjective’s attributive function. Besides being predicate, adjectives have other syntactic functions such as subject, object, adverbial and complement. Consequently, in this paper, we will pay close attention to these adjectives in the next section when we choose "Pre-Only”.

According to Mo Pengling, Shan Qing [7], the mean frequency of verbs as predicates is 76.7%. Although the scholars give us the frequency of verbs as predicate from the statistical perspective, few scholars have further probed into the corresponding relation between verb and predicate. As a result, we first should exclude those verbs that serve syntactically as subject or object. Then we narrow the scope of Pre-Only by excluding verbs that function as attributive and adverbial in a syntactical sense. And the remaining would be the predicate-only verbs we’ll handle in the article. Data in this research are elicited from two sources: The Word Class Handbook by YuanYulin [8] (referred to as "Handbook" hereafter) and The Interpretation of Dictionary of Grammar Information in Modern Chinese (referred to as "Interpretation" hereafter) by YuShiwen [9]. We select the two corpora mainly because they have processed and counted grammatical information of words to some extent and we can use them for reference and comparison.

Through the statistics and selection, we get about 120 Pre-Only from "Handbook", where the number of verbs is 80, adjective, 20, descriptive words[1], 21 , and words which have more than two parts of speech,7. There are verbs such as 活像 huoxiang

[1] Descriptive Words section is independent in "Handbook".

“ like”, 急于jiyü “hurry to”, 加以jiayi “in addition”, 酷似 kusi “like”, 懒得lande “not feel like”; we have adjectives such as "碍口aikou “be too embarrassing to mention”, 省事shengshi “simplify matters”, 识货shihuo “ be able to tell good from bad”, 衰败shuaibai “decline”,费工feigong “require a lot of labor”; here are descriptive words like "皑皑aiai “pure white”, 遍野bianye “all over the plains”, 超群chaoqun “preeminent”, 连绵lianmian “continuous”, 扑鼻pubi “to assail the nostrils”, 至上 zhishang “supreme”. When it comes to Pre-Only in “Interpretation”, we select them out of Verb Library, Adjective Library and Descriptive Words Library. In Verb Library alone, there are 40 special groupings, some of which can help us to determine which groupings belong to Pre-Only. The labels for such groupings include "After the Noun[2], Single Subject[3], Single Predicate[4], Single Object[5], Single Adverbial[6] and Single Complement[7]". These labels can help us to exclude verbs which may work syntactically as attribute, subject, object, adverbial and complement and the remaining are the Pre-Only we need. In the same token, we select Pre-Only out of Adjective Group and Descriptive Word Group and we get 424 Pre-Only, among them there are 360 as verbs, 63, as adjectives, and 1, as descriptive word. Predicate-only verbs occupies 16.8% among 2147 verbs, predicate-only adjectives, 4.3% among all 1473 adjectives, and predicate-only descriptive words account for 0.49% among the 203 descriptive words.

We test all the Pre-Only we got from the above-mentioned data sources in the PKU corpus (CCL). We mentioned that there might be some problems when we selected data from "Handbook" and "Interpretation". For example, we should observe whether the predicate-only verbs from the "Handbook" can serve syntactically as complement or subject, or object in the corpus, and then analyze their frequency of such usage. As for adjective group and descriptive group, we focus our attention mainly on those that serve as subject or object, and excluded them if they do. Because of detailed grammatical information, we can directly get the result we needed according to the labels of groupings in "Interpretation", while as for Adjective Group and Descriptive Word Group, we need to test them in the corpus to see whether they function as subject or object. According to this procedure we observed the specific usage of the above cases in the corpus and excluded over 5% of the cases of other syntactic functions other than predicate. Special attention should be paid to the fact that not all syntactic functions of the lexical items with the help of corpus and we have to resort to the circumstantial evidence. A good case in point is the usage of 相等xiangdeng “equal”. Corpus findings show that more than 95% of its usage is predicate with few other syntactic functions. However, the description of this word in Collocation Dictionary of Notional Words in Modern Chinese shows us that it can serve as object like "认为renwei

[2] This label indicates that lexical items with this label can only be put after nouns.
[3] This label indicates that those words stand as subject themselves.
[4] This label indicates that those words stand as predicate themselves.
[5] This label indicates that those words stand as object themselves.
[6] This label indicates that those words stand as adverbial themselves.
[7] This label indicates that those words stand as complement themselves.

“think ~[8] ", and complement like "分配得fenpeide “distribute ~". Therefore, we will not include this word in the Pre-Only. On the basis of the above work, the preliminary data was selected: in "Handbook" there are 110 Pre-Only, among which there are 67 verbs, 17 adjectives, 18 descriptive words, and 8 words that have two parts of speech; in "Interpretation" there are 389 Pre-Only, among which there are 360 verbs, 28 adjectives, and 1 descriptive word. Up to now, the Pre-Only we selected are from three groups, i.e., Verb Group, Adjective Group and Descriptive Word Group. Due to limited space, this paper focuses on the predicate-only verbs alone. As we mentioned above, Pre-Only account for 16.8% of all verbs in "Interpretation", which shows from one aspect that Pre-Only aren’t the typical members in verb group. Statistics of these Pre-Only tell us that they have such co-existent features as freedom and adhesion in syntactic sense. So the following section will concentrate on their similarities and differences.

Above-mentioned observation and selection of Pre-Only demonstrate that they are monosyllabic or disyllabic verbs and they have one feature in common: it is difficult for them to be "nominal". Meanwhile, they are abstract in semantic meaning, which is often found in the inherent features of adhesive verbs. Although some scholars believe that it is difficult for the adhesive verbs to function as free syntactic components, they do demonstrate the features of prototype verb, which is worth studying as the sub-category of verb.

The adhesion of the adhesive verbs

A legion of previous studies on free verbs have shifted our research interest in the adhesion of adhesive verbs in this paper, with our special emphasis on the fact that although these adhesive verbs can function as predicate only with other relevant syntactic components, we still regard them as Pre-Only for their verb features depend on the other syntactic components. In addition, the more adhesive of the grammatical components are, the more limited syntactic function these verbs have, namely, it is natural for these verbs to exhibit adhesion.

Near null syntactic parameters of prototype verbs

Some lexical items like "留念liunian “to keep as a souvenir” should work with other verbs to be predicate or the core of predicate. So should some adhesive verbs. In addition, more than 95% of the adhesive verbs cannot be modified by the adverb of degree "很hen “very”, which is the biggest difference between verbs and adjectives and marks the parts of speech of some lexical items. More than 90% of the adhesive verbs don’t have any overlapping forms of themselves. Even if some have overlapping forms of themselves, there is only one form. A good case in point is "担待dandai “forgive”. In spoken Chinese we can say “please ~ ~", but we can’t say “担担待待”*.

Without self-sufficient semantic meaning

Yin Shichao said that "the stronger verb’s adhesion is, the weaker its grammatical function is and the more abstract its semantic meaning is; the higher the freedom degree of the verb is, the richer its grammatical function is and the more specific its

[8] ~ stands for the lexical item as the example mentioned above(here相等*xiangdeng* “equal”).

semantic meaning is." On the basis of Fu's "mode of meaning analysis of action words" [10], we will analyze the following words with "D+E" mode, where D denotes action and E, relations between objects or relations between matters and E must be closely related with D. There is a consistent feature in the paraphrase patterns of these verbs: the doer does not appear. The decoding of some lexical items' semantic meaning needs to be compensated by syntactic measures sometimes. As we mentioned above, free verbs are free while adhesive verbs exist only by relying on other relevant syntactic components. However, by observation of their semantic and syntactic meaning we can see they are also "quasi-adhesive", namely, these words either require the concurrence of subject, or appear in the context where the doer is implied. Here are some specific cases:

爱好1 aihao1 " have a keen interest in something".verb, integral 70, degree of membership 0.7, belongs to not so typical verb group.

摆脱baituo "to get rid of (pinned, bound, difficult, bad conditions etc.)".verb, integral 90, degree of membership 0.9, belongs to the typical verb group.

包括 baokuo "contain (or enumeration of each part, or emphasize a part)".

More researches have been on the internal differences of free verbs. Therefore we focus on the internal differences of adhesive verbs. We discussed the main internal similarities of adhesive verb above, and now we will find their internal differences. The most obvious difference we can get is that every adhesive verb has specific adhesive composition, which has already been mentioned by Zhou Haifeng. Adhesive verbs can be classified into five sub-types in theory: subject-adhesive verbs, object-adhesive verbs, predicate-adhesive verbs, adverb-adhesive verbs, complement-adhesive verbs. Three articles have been written on these topics. But the subject-adhesive verbs and predicate-adhesive verbs are less studied. Therefore we will have detailed discussion of several types of adhesive verbs. As the specific components of adhesion are different, adhesive verbs are divided into:

Subject-adhesive verbs

Subject-adhesive verbs can not exist independently as the predicate, but they can only when they adhere to subject. Such verbs are mostly the predicate verbs of fixed phrases in our investigation. And nearly all are the intransitive verbs. Here are a few subject-adhesive verbs and their collocations.

遍野 bianye: all over the field, meaning a multitude of. 尸横shiheng corpse lies｜饿殍e'fu bodies of the starved｜哀鸿aihong victims ～

不息buxi: no stop. 奔流pour benliu｜生命life shengming｜奋斗struggle fendou～

超群Chaoqun: more than a general. 技艺skill jiyi｜武艺martial arts wuyi｜实力actual strength shili～

Object-adhesive verbs

Object-adhesive verbs are ones which can't serve independently as predicate, but they can when they work together with their objects in a sentence. Object-adhesive verbs are the most frequent ones in the group of adhesive verbs. They also have an obvious feature different from other adhesive verbs-they have formal markers, which can help us to

judge whether the verbs are object-adhesive ones or not. Now on the basis of previous researches and our investigation we will discuss this type of verbs in detail.

Unmarked Object-adhesive verbs

Unmarked object-adhesive verbs do not have so obvious regulations as marked object-adhesive verbs. Semantically, they can be classified as:

Imperative-verbs:让3 rang3 “let”,使1shi1 “cause”, 叫2 $jiao_2$ “order”, 致使zhishi “cause”,迫使poshi “force”;

Link-verbs:是3 shi3 “is”;

call-verbs:叫1 jiao1 “call”, 统称tongcheng “collectively called ”, 通称tongcheng “be generally called”, 称1cheng1 “say”, 称呼chenghu “call”;

change-verbs:变1 bian1 “change”, 呈现 chengxian “emerge”;

psychological and intention-verbs: 失望shiwang “lose hope”, 乐意leyi “be wiling to”, 热衷rezhong “be fond of”, 想3 xiang3 “want to” ,希望 xiwang “enthusiastic hope”; and

allow- verbs: 纵容zongrong “connive”

Marked Object-adhesive verbs:

Marked object-adhesive verbs can be categorized as:

"于yü(to)"-verbs: 敢于ganyü “dare to”,急于 jiyü “anxious to”, 等于dengyü “equal to”,忙于mangyü “busy to”;

"以yi “with,by”-verbs: 加以jiayi “in addition”, 予以yüyi “give”, 致以zhiyi “present”,给以geiyi “be given to”;

得de “particle”-verbs:懒得lande “not feel like”,博得bode “win”,获得huode “acquire”;

为wei “by”-verbs:成为chengwei “become”,作为zuowei “as”;

成cheng“into”-verbs: 酿成niangcheng “lead to”,变成biancheng “become”;

出chu “out”-verbs:露出luchu “show”,指出zhichu “pointed out”;

给gei “give”-verbs:留给liigei “leave”,献给xian’gei “dedicate to”; and

似si “like”-verbs: 酷似kusi “ be exactly like”, 貌似maosi “seemingly”;

According to YinShichao, there are 16 adhesive verb formations, some of which are related with object-adhesive verbs, i.e., "X-以yi, X-于yü, X –得de, X-不得bude,X-为wei, X-做zuo, X-似si, X-着zhe and X-如ru". It still remains a question whether such word formation can form words. But the degree of lexicalization of such words is different, and the formation method is also different. Some of them develop from two morphemes of two different syntactic structure such as 于yü “to”-verbs; Some disyllabic verbs come into being from its second verb by voiding its semantic meaning and weakening its syllable. A case in point is 成cheng “into”-verbs. Among these marked object-adhesive verbs, the highest degree of lexicalization and the most productivity belongs to the "V+于yü" object-adhesive verb group, which is of dual tone and often appear as common words. "于yü(to)"-verbs belong to the typical cases of developing from two morphemes of two different syntactic structure. "于yü(to)" verbs develop from changing its back attached components to a front attached com-

ponents while their semantic meaning, sound and syntactic function all have changed. However, from the perspective of its degree of adhering to the object in the sentence, the function of ancient and modern "于yü" verbs has changed little, that is to say, "于yü" verbs have a strong command of co-occurrence of their objects that come after the verbs.

Adverbial-adhesive verbs

Adverbial-adhesive verbs can't serve independently as the predicate, but they can only when they adhere to adverbials such as "看待kandai "look upon", 着想 zhuoxiang "take into consideration", 看B3 kan B3[9] "look at", which means that they can work as the predicate center only when they adhere to adverbials. Here are the examples:

我 是 这样 看待 这件 事情 的。
Wo shi zheyang kandai zha jian shiqing de. "That's the way I look at it."
*我 是 看待 这件事情的
* Wo shi kandai zhe jian shiqing de.
我 这样 做 也 是 为 你 着想。
Wo zheyang zuo ye shi wei ni zhuoxiang. "I do it for you."
*我 这样 做 也 是 着想。
* Wo zheyang zuo ye shi zhuoxiang.
你怎么看这件事情？
Ni zenme kan zhejian shiqing? "How do you see this matter?"
* Ni kan zhejian shiqing?

Complement-adhesive verbs

Complement-adhesive verbs can not serve independently as the predicate, but they can only when they adhere to complements. According to Zhou Haifeng, complement-adhesive verbs generally only adhere to one complement and can not repeat themselves in form. They cannot take dynamic auxiliary like着, zhe auxiliary word;了, le auxiliary word; 过, guo auxiliary word either, and they can be modified by adverbials when they adhere to their complements.

Predicate-adhesive verbs

Predicate-adhesive verbs are non-free adhesive verbs which can't serve independently as the predicate.They must be in front of the core of predicate verb. It seems to us that predicate-adhesive verbs mainly consist of auxiliary verbs and typical predicate-adhesive verbs.

Auxiliary verbs

The auxiliary verbs are typically the verbs which stand before the verbs in a sentence. Their syntactic function is limited. A lot of researches regard the structure of auxiliary verbs with predicate verb as adverbial-center structure while some regard it as conjunction-predicate structure. Since auxiliary verbs belong to the category of verbs, we tend to regard the auxiliary verbs as predicate-adhesive verbs, behind which can only appear

[9] B3 stands for one paraphrase of "*kan*",which infers "take care"and can be used in imperative sentence.

verbal components. The structure of auxiliary verbs with predicate verb should be viewed as the conjunction-predicate structure. In our opinion the analysis of adverbial-center structure shows us that the auxiliary verbs in front of the main verb have changed to be one subsidiary part of other main verbs. Typical predicate-adhesive verbs.

While there are some controversies about the position of auxiliary verbs as predicate-adhesive verbs, the following words belong to the typical predicate-adhesive verb group:强制qiangzhi "force", 强行qiangxing "force", 抽空choukong "manage to", 挺身tingshen "rise up", 酌情zhuoqing "take into consideration the circumstances" and so on.

这件事情需要强制执行。

Zhe jian shiqing xüyao qiangzhi zhixing. "The thing needs to be enforced".

他强行改了自己的密码。

Ta qiangxing gai le ziji de mima. "He changed password himself by force"

你抽空去一趟超市吧。

Ni choukpng qü yitang chaoshi ba. "Would you manage to find your time to go the supermarket?"

如果你遇到事情，他还是会挺身而出的。

Ruguo ni yüdao shiqing , ta haishi hui tingshe er chu de. "If you encounter bad things, he would rise up bravely."

请酌情处理一下这件事情。

Qing zhuo qing chuli yixia zhejian shiqing. "Please handle this matter at your discretion."

We think that compared to auxiliary verbs, the typical predicate-adhesive verbs mentioned above are inclined to stand as adverbial, and cannot be analyzed as conjunction-predicate structure. Therefore we do not classify such verbs into Pre-Only. Considering form-verbs are also verbs before another verb while they dominate the predicate verb behind, they still belong to Pre-Only as we delineate.

3 Conclusion

This article is designed to study the Pre-Only, a relatively closed set of words and make a statistical analysis of them selected out of Verb Group, Adjective Group and Descriptive Word Group. With different adhesive components of Pre-Only as the yardstick, categorization is also made of them. What's more, analysis is made of their syntactic and semantic features. The research of this paper is more of statistic research and classification of Pre-Only with little on the discussion of their mechanism which thereby deserves our further and closer attention in the future.

References

1. Lü, S.X.: Problems in Analyzing Chinese Grammar. The Commercial Press, Beijing (1979)
2. Yin, S.C.: Discussion On Adhesive Verbs. Chinese Language 6, 401–410 (1991)

3. Yang, X.P.: On Chinese Morphemes. Nanjing University Publishing Press, Nanjing (2003)
4. Lin, H.A.: A preliminary Study on Adverbial-adhesive Verbs. Linguistic Researches 3, 21–24 (1996)
5. Zhou, H.F.: A preliminary Study on Complement-adhesive Verbs. Chinese Language Learning 3, 26–27 (2000)
6. Qi, H.Y.(ed.): Functional Comparison of Adjective Phrases and Adjectives. Chinese Language Learning 2, 1–9 (2001)
7. Mo, P.L.(ed.): The Statistical Analysis on Three Classes of Notional Words of Syntactic Function. Journal of Nanjing Normal University 2, 55–63 (1985)
8. Yuan, Y.L. (ed.): The Word Class Handbook. Beijing Language and Culture University Press, Beijing (2009)
9. Yu, S.W. (ed.): The Interpretation of Dictionary of Grammar Information in Modern Chinese. Qinghua University Publishing Press, Beijing (2003)
10. Fu, H.Q.: Description and Analysis of Chinese Lexical Meaning. Foreign Language Teaching and Research Press, Beijing (2006)

A Chinese-English Comparative Study on Non-conventional Verb-Object Collocations in Chinese

Qiong Wu

College of Chinese Language and Literature, Wuhan University, Wuhan 430072
cindy61361@hotmail.com

Abstract. In Chinese language, "verb + non-conventional object" is a very specific and complex structure. Discussions about categories of such kinds of objects have not reached a consensus. This paper employs the "stereotypical relation" to sort out and defines seven kinds of "non-conventional object". It also attempts to explain the syntactic functions of these objects based on a Chinese-English comparative analysis. A hypothesis is eventually devised of how the non-conventional collocations come into being.

Keywords: semantic relation, non-conventional objects, Chinese-English comparative analysis, syntactic function.

1 Introduction

Chinese “Verb+Non-conventional Object” is a special structure because the object in the structure is not a grammatical object. Thus, it is difficult to infer the meaning from the literal. How to categorize these non-conventional objects is also a problem for linguists, because there is no authoritative categorization in Chinese language hitherto. For these reasons above, this paper employs the "stereotypical relation"[1] to sort out and defines seven kinds of "non-conventional object", and demonstrates that, Chinese non-conventional objects may function as adverbial modifier or subject.

2 Definition of Chinese Non-conventional Object

There is a characteristic in Chinese Non-conventional object---simple surface framework, but complex interrelationship [2]. Because of such a characteristic, there are so many categorizations of Chinese objects. This paper first defines 7 Chinese non-conventional objects from the perspective of stereotypical relation in Hanyu Dongci Yongfa Cidian.

There are 14 kinds of objects in Chinese. We first eliminate heterozygous object, because this kind of object is often considered as idioms. For example: 喝西北风 (he xibeifeng, have nothing to eat), 跑龙套(pao longtao, to play a small role), 哭鼻子 (ku bizi, snivel), 开小差儿 (kai xiaochar, slope away) etc.. Another kind of object

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 800–808, 2013.

eliminated is homologous object. By Meng,C.[3], there are two qualifications in homologous object: a). the object does not have new meanings; b). the object can only have meaning when it accompanies with a verb. By analyzing corpus we collected, we found this kind of V+O collocation can be considered as V+O separable word. For example:谈话(tan hua, talk),跳舞(tiao wu, dance),and吹气(chui qi, blow).

There are still 12 kinds of objects left which we need to consider about. We will employ the stereotypical relation to categorize these objects again.

Xu, S.H.[1]demonstrates the stereotypical relationship from the perspective of cognition. He argues that stereotypical relations are the relations inherent in things themselves, and they could, through the cognitive refraction, be turned into a cognitive instrument in the process of language use. On the basis of this idea, stereotypical relations are the projection of the laws of objects in reality. By the cognitive projection, this relation can be used as the style and media for human's comprehension of the world, and also can be used to express human emotions. Besides these, this relation can also be used as the foundation of human comprehension. Take two words for example. When people say 'drink', they can associate fluid with this verb; and when people say 'eat', they definitely associate food with it, not other things, like gas. Hence, this stereotypical relation not only related to human's daily life, but also related to the semantics of the verb. In Chinese, conventional object has a direct collocation with the verb (吃饭 chi fan, chi is a verb, and fan is its conventional object), while non-conventional object does not have such a collocation.

Xing, F.Y. [4] also discusses the definition of Chinese non-conventional objects. He holds the view that the word "conventional" reflects the relationship between verb and object which could be accepted by all human beings. Xing also use the stereotypical relationship to define the objects. He uses triangle relationship (Fig.1) to explain the characteristics of Chinese non-conventional object.

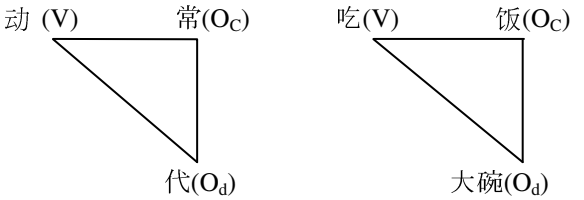


Fig. 1. Matrix of Triangle Relationship

He argues verb and objects in non-conventional V+O collocations can construct a triangle, the verb and the object cab build a relationship with the help of the conventional object. In his work, Xing named the non-conventional object 代体宾语（daiti binyu, vicarious object）.Though the name is different, the nature of the object is the same. From this point of view, we can see that the most important qualification for defining non-conventional object is stereotypical relation.

Fan, X. [5] also defines objects from the view of stereotypical relation. In his work, he categorized 6 kinds of conventional objects and 5 kinds of non-conventional objects. We agree with Fan's definition, because this definition is logical and

comprehensive. But there are still two other kinds of objects which cannot match with Fan's categorization, one is reason object, like养病(yang bing , get over illness) ; and another is goal object, like考研究生(kao yanjiusheng, take part in the entrance exams for postgraduate schools).We prefer to define these as non-conventional object for the following two reasons: a). the object is not the patient of the verb; b). With the help of conventional object, the verb and the object can build a triangle relationship.

Table 1. Definition of Chinese non-conventional objects (by Fan,X. 2006)

Non-conventional objects in Chinese	Examples
1. **Instrument object**	写毛笔 (xie maobi , write with brush) 吃大碗 (chi dawan , eat with large bowl)
2. **Manner object**	唱A调 (chang A diao , sing in A major) 写仿宋体 (xie fangsongti ,write in Fang song)
3. **Location object**	睡大床 (shui dachuang , sleep in a big bed) 吃食堂 (chi shitang , eat in cateen)
4. **Time object**	熬通宵 (ao tongxiao , sit up all night) 休息星期天(xiuxi xingqitian , rest on sunday)
5. **Agent object**	晒太阳 (shai taiyang , sun oneself) 红着脸 (hong zhe lian , blush)

Hence we get 7 kinds of non-conventional objects in Chinese: Instrument object, Manner object, Location object, Time object, Agent object, reason object and goal object. What we curious about next is why Xing uses 代体宾语(daiti binyu, vicarious object) to call non-conventional object. In Chinese, 代 means replacement, or means in instead of something. What is 代体really means in Chinese non-conventional object? What is 代体 represents for? Hence, in the next section, we will mainly discuss the following questions:

What is the function of non-conventional object in Chinese? And what "daiti" means in non-conventional object?

3 Grammatical Function of Non-conventional Object in Chinese

3.1 Non-conventional Object Study Based on Chinese Language

We listed 6 kinds of non-conventional objects in Chinese below1. On the left-hand, we listed their original forms, and on the right-hand side, we transformed the collocations on the basis of Parallel Principles of Transformation Theory [6]. We add some words or phrases to the original form without change their original meanings.

[1] We will talk about agent object later.

Table 2. Parallel Transformation of Chinese Non-conventional Objects

Object Types	Examples	Transform to
Manner Object	存 活 期 cun huoqi save money in a current account	[以]活期 [的 方 式] 存（钱） [yi] huoqi [de fangshi] cun(qian)
	寄 特 快 ji tekuai send a letter by express	[以]特 快 [的 方 式] 寄（信） [yi] tekuai [de fangshi] ji xin
	写 黑 体 xie heiti write characters in bold	[以]黑体 [的 方 式] 写（字） [yi] heiti de fangshi xie zi
Location Object	吃 麦当劳 Chi McDonald's Eat in McDonald's	[在] 麦当劳 吃 [zai] McDonald's chi
	睡 沙 发 Shui shafa Sleep on the sofa	[在]沙 发 [上] 睡 [zai] shafa [shang] shui
	吃 馆 子 Chi guanzi Eat in restaurant	[在]馆 子 吃 [zai] guanzi chi
Instrument Object	吃 大 碗 Chi dawan Eat with large bowl	[用] 大 碗 吃 [yong] dawan chi
	看 放 大 镜 Kan fangdajing See with magnifying glass	[用] 放 大 镜 看 [yong] fangdajing kan
	写 铅 笔 Xie qianbi Write with pen	[用] 铅 笔 写 [yong] qianbi xie
Reason Object	养 病 Yang bing Get over illness	[因为] 病 [所以] 养 [yinwei]bing [suoyi] yang
	逃 荒 Tao huang Run away from a famine	[因为] 荒 [所以] 逃 [yinwei] huang [suoyi] tao
Time Object	大 干 红 五 月 Da gan hong wuyue Work on red may	[在] 红 五 月 大 干 [zai] hong wuyue da gan
	休 息 星 期 天 Xiuxi xingqitian Rest on sunday	[在] 星 期 天 休 息 [zai] xingqitian xiuxi

Table 2. (*Continued*)

Goal Object	等 朋 友 Deng pengyou Wait for friends	[为 见 到] 朋 友[而] 等 候 [wei jiandao]pengyou er denghou
	考 研究生 Kao yanjiusheng take part in the entrance exams for postgraduateschools	[为 成 为] 研 究 生 [而]考试 [wei chengwei] yanjiusheng er kaoshi
	筹 备 展 览 会 Choubei zhanlanhui Prepare for the exhibition	[为] 展 览 会 [而] 筹 备 [wei] zhanlanhui [er] choubei

Through the above transform analysis we found that all the above categories of non-conventional collocations can add some verb-object reference as a word or phrase in front of the verb. We then look at these added ingredients, whether it is "以……的方式"(yi……de fangshi), "为 …… 而 ……"(wei.....er…...), or" 用 ……"(yong.....),or" 在 ……"(zai.....)etc., they are prepositions, or prepositional phrases, which transformed in association with non-conventional objects as adverbial modifiers in a sentence.

Through the above analysis we can see, non-conventional object features can include time, location, tool, reason, instrument etc., and can also add components to form a "adverbial + verb" structure. From a semantic perspective, non-conventional object = verb + adverbial + verb. From the functional point of view, "verb + non-conventional object" embodied in the object is in the mix "adverbial + verb" to an adverbial function. It can be reflected from the English expression also.

3.2 Chinese-English Comparative Study on Non-conventional Object

In the previous section, we talked about the characteristics of Chinese non-conventional VN structure and the grammatical function of its object. In this section, we will translate the structure into English, and try to see if there is any similarity between these two

Table 3. Chinese Non-conventional object in English PP Form

Object Types	Examples		English PP Form
Manner object	存活期 寄平信	cun huoqi ji pingxin	keep money in a current account send the letter by surface mails
Location object	睡沙发 吃全聚德	shui shafa chi Quan Jude	sleep on the sofa eat at Quan Jude
Instrument object	看放大镜 写铅笔	kan fangd jing xie qianbi	look at things with a magnifying glass write with a pencil
Goal object	考研究生	kao yanjiusheng	take part in the entrance exams for post- graduate schools
Reason object	养病 逃荒	yang bing tao huang	get over illness run away from a famine
Time object	休息星期天 大干红五月	xiuxi xingqitian da gan hongwuyue	rest on Sunday work in Red May

languages. Chinese 吃食堂(chi shitang)、吃大碗 (chi dawan)、写毛笔 (xie mao-bi)and etc. can be expressed by V+PP form in English. For example (see Table 3):

We found that after translation into English, the original object which located after the verb becomes adverbial constituents in English. And also in English, there are some similar collocations, such as the following:

- to sail the sea
- to jump the hedge
- to walk the street
- to sit a boat

Like Chinese, these structures are also cannot transformed into passive forms, *the sea is sailed./*the hedge is jumped. And it is also not appropriate to take sea, hedge as the object. There are some phrases like the examples below in English:

- to come a walk = to come [for] a walk
- to swim a river = to swim [across] a river
- to jump a hedge = to jump [over] a hedge

Although the noun which located after the verb seems to be a direct object, but its essence more strongly represent as adverbial, therefore, the meanings are closer to an adverbial. The noun after the verb has a strong nature of the adverbial. Because object and adverbial adjunct is constituted with the verb phrase, therefore, it is not easy to distinguish them from, and even may. Both are closely linked, there is “only one step away”. The function of objects in non-conventional V+O collocations is to supplement and illustrate verbs. We find that, whether in English or in Chinese, objects in most of the V+O non-conventional collocations are nouns, these nouns can function as adverbial modifiers. Therefore, comprehensive analysis can give an answer in the previous question: objects in non-conventional collocations can have an adverbial function.

4 An Explanation for Chinese Agent Object

There is still another kind of object in Chinese verb usage dictionary that we do not mention, it is called agent object. In linguistics, a grammatical agent is the cause or initiator of an event [7]. Wei, H.[8] made a statistical analysis of 179 Chinese verbs, and found that only 35 verbs can collocate with agent object. That is to say, this kind of verbs is not a few. And most of the verbs are intransitive verbs, the agent object can be put in front of the verb without marks. And, after changing the location, the meaning of the collocation does not change. For example:

– 开始新的一年	Kaishi xinde yinian	Begins a new year
– 新的一年开始	Xinde yinian kaishi	A new year begins

– 来客人	Lai keren	Comes the guest
– 客人来	Keren lai	The guest comes

Not only in Chinese, there also exists agent object in English. But the difference is, there must be markedness when the verb changes its location with the agent object. For example:

— The teacher comes. Here comes the teacher.
— The tree stands. There stands the tree.

Here the word “here” and “there” are the markers. They must appear in the sentences, if not, the sentences are not grammatical. The reason for the difference is, Chinese is a language of parataxis, while English is a language of hypotaxis. Chinese speakers understand the sentence by meaning, while English speakers understand the sentence by form. Hence, the agent object can be located in front of the verb without marked in Chinese, but English cannot. If we want to exchange the location of verb and agent object in English, marker is needed. But both in Chinese and English, agent objects play a role of subject in the collocations, that is, the objects are subjective. Agent object is also a kind of 代体宾语(daiti binyu, vicarious object),what agent object represents is the subject. Thus, agent object plays a role of subject in V+O non-conventional collocation.

We hold the view that agent object is a special kind of object, it is different from other kinds of objects. Further research is needed in the future. We here just give the initiation in order to attract more attention from the academic.

5 The Cause of Chinese Non-conventional Object

For the cause of the non-conventional object, some Chinese scholars hold the view that it is because of the economic principles of language. We agree with the view, but we don’t think it is enough, we believe there is still some reasons behind this view, for example, cognitive mechanism. Thus, in this section, we try to explain the reasons from the cognitive perspective.

As Xu, S.H. [1] demonstrated, the leading tendencies of human cognitive activity is to optimize thinking. Under the restriction and the guidance of it, language performance is also sought to optimize. The main way of optimization is to offer more information with lesser words or phrases. It is not necessary to mention the details when people can get it; it is not necessary to create a new form when the existed form can illustrate the meaning. Human can understand the world with the implicit statement, or with the stereotypical relationships. In this way, it gradually formed a mechanism that to use stereotypical relationship implied implicit expression.

In a conversation, the speaker uses stereotypical relationship to make their words simple, and the listener uses stereotypical relationship to enrich the information and understand it. As in Chinese non-conventional V+O collocations, the grammatical object is well known, or is defaulted, so both the speaker and the listener can understand it without speak it out. Oppositely, non-conventional object in V+O collocations always offer the information about time, location, agent, goal and etc. which the listener cannot comprehend them if the speaker does not make it explicit. So the speaker must speak out in order to make the listener understand.

In the end, we want to make a hypothesis based on the stereotypical relationship to illustrate the forming process of Chinese non-conventional object. We conclude, the original form of Chinese non-conventional V+O collocation is" PP + V + Conventional Object", and after a certain phase, people get familiar with these conventional collocation, or in other words, people construct a common cognition of the world, a stereotypical relationship is built. So when we say write, we can connect the verb with the conventional object--the character by the stereotypical relationship. Because of the optimal principle-- It is not necessary to mention the details when people can get it, so it is not necessary to speak out the conventional object when it appears with the non-conventional object at the same time. So people can use"PP+V" instead of" PP + V + Conventional Object" .Again, as the optimization principle illustrates, that it is not necessary to create a new form when the existed form can illustrate the meaning. As people are familiar with the V+O form, and it is more succinct than PP+V. V+O (O is non-conventional object) replaced PP+V form, therefore appears the Chinese non-conventional V+O collocation.

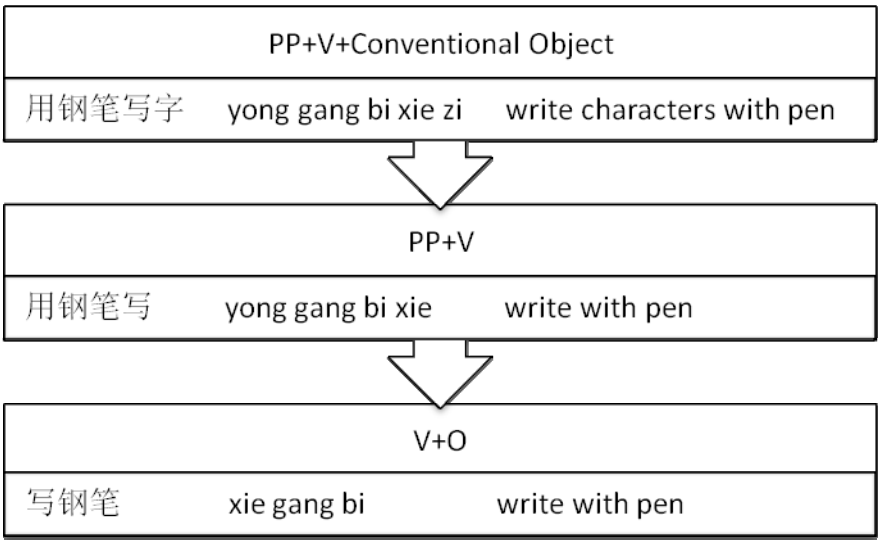


Fig. 2. The forming process of Chinese non-conventional object

6 Conclusion

This paper first defined 7 kinds of non-conventional object in Chinese, then talked about their function by Chinese-English comparative analysis study and found out the non-conventional object can be seen as an adverbial or a subject in a sentence. In the end, this paper explained the cause of Chinese non-conventional object from the perspective of cognitive optimization and made a hypothesis of the forming process of non-conventional object. Further research is needed to focus on the characteristic of agent object and the differences between Chinese and English non-conventional objects.

References

1. Xu, S.H.: Stereotypical Relation: The Study of Syntactic Construction. Journal of Foreign Languages 144, 8–16 (2003)
2. Chu, Z.X.: Chinese Agent-Object Clause and Means of Distinction of Agent and Patient in SVO language— Also on Features of Chinese high-performance of syntax. Chinese and Ethnic Languages in the Perspective of Typology Academic Forum (2010)
3. Meng, C.: Hanyu Dongci Yongfa Cidian. The Commercial Press, Beijing (1999)
4. Xing, F.Y.: Chinese Grammar. Northeast Normal University Press, Changchun (1996)
5. Fan, X.: An Introspection to the Study of Chinese Object. Chinese Language Learning 3, 3–13 (2006)
6. Zhu, D.X.: The Principles of parallel. Studies of Chinese Language 191, 81–87 (1986)
7. Kroeger, P.: Analyzing Grammar: An Introduction. Cambridge University Press, Cambridge (2005)
8. Wei, H.: A Study on the Conditions of Common Verbs with Objects for Chinese Acquisition. People's Press, Beijing (2009)

A Chinese Sentence Segmentation Approach Based on Comma

Shengqin Xu, Fang Kong, Peifeng Li, and Qiaoming Zhu

Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu, 215006
School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, 215006
{20104227033,kongfang,pfli,qmzhu}@suda.edu.cn

Abstract. Chinese sentence segmentation is considered to be a very fundamental step in natural language processing. A successful solution for sentence boundary detection is a key step in the subsequent NLP tasks, such as parsing and machine translation, etc. In this paper, we consider comma as a sign-of-the-sentence boundary, and then divide it into two major types, i.e., the true (EOS) and the pseudo (Non-EOS). Finally, a system framework of Chinese sentence segmentation based on two-layer classifiers is presented and implemented. The experimental results on Chinese Treebank 6.0. Results show that our model achieve the F-measure of 90.7% overall, which improves by 1.5%.

Keywords: Chinese Sentence Segmentation, Maximum Entropy Models, Comma, Two-layer classifiers.

1 Introduction

Sentence segmentation is an important initial processing step for many natural language processing applications, such as part-of-speech (POS) tagging, machine translation, and syntactic parsing, etc. All these tasks require their input text to be alienated into sentences for further processing. However, segmenting a text into sentences is not a trivial task, since the end-of-sentence punctuation marks are ambiguous. Comparatively speaking, English sentence segmentation is an easily handled problem. Sentence boundaries can be depended on period, exclamation mark and question mark. Although period, decimal point and ellipsis in English use the same symbol (dot), which may lead to confusion, it can be resolved fairly easily in a machine-learning framework [1].

Generally, Chinese language also uses period, question mark, and exclamation mark to indicate sentence boundaries. Sentence boundaries can be detected without doubt where these punctuation marks exist. Unlike English comma, comma in Chinese plays the same role as period in English in certain context. Take Run-on Sentence for example, two or more independent clauses (i.e., complete sentences) are joined with comma without an accompanying coordinating conjunction. The comma in Run-on Sentence can be considered as sentence boundary. An example is given in (a), which one Chinese sentence is translated into two English sentences.

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 809–817, 2013.

(a) Chinese: 亚洲是一个有着悠久历史和重要地位的大洲，[1]她是人类文明的摇篮之一，[2]对人类文明的进步和科学文化的发展作出过辉煌的贡献.

Pinyin: yàzhōu shì yígè yǒuzhe yōujiǔ lìshǐ hé zhòngyào dìwèi de dàzhōu， [1] tā shì rélèi wénmíng de yáolán zhīyī，[2] duì rénlèi wénmíng de jìnbù hé kēxué wénhuà de fāzhǎn zuòchūguò huīhuáng de gòngxiàn.

English: Asia is a vast continent with a long history and increasing strategic importance. [1] As one of the cradles of human civilization, [2] it has mad brilliant contributions to human progress and scientific and cultural development.

In this paper, we consider Chinese sentence segmentation as a comma disambiguation problem. Comma is classified into two types. One is the kind of comma that can be detected as the end of sentence (EOS, such as [1] in Example (a)), while the other cannot be considered as the end of sentence (Non-EOS, such as [2] in Example (a)). Sentences that can be split on commas are generally loosely coordinated structures that are syntactically and semantically complete on their own, and they do not have a close syntactic relation with another. As illustrated in Figure 1, the comma in this sentence meets the above-mentioned condition, so it's regarded as an EOS comma. However, some commas cannot mark the sentence boundary as the example showed in Figure 2, where a LCP is separated from the rest of the sentence with a comma. We believe that the sentence boundary detection task to disambiguate commas, if successfully solved, simplifies downstream tasks such as syntactic parsing, semantic role labeling and machine translation[2-4].

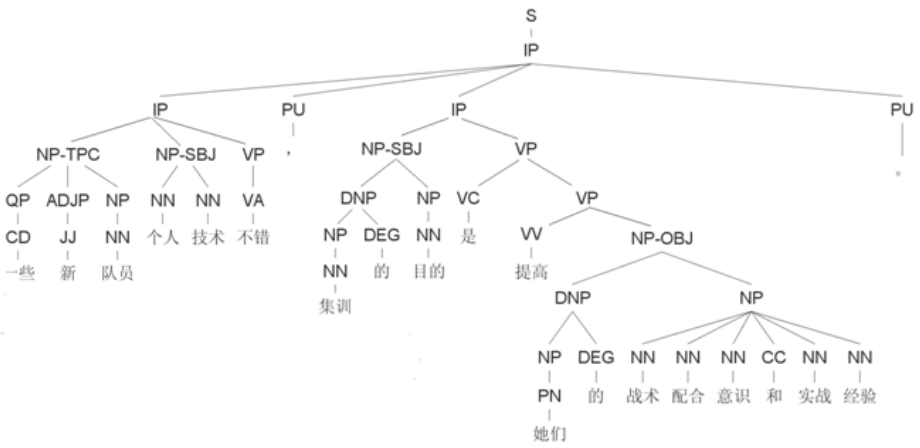


Fig. 1. Sentence-boundary denoting comma

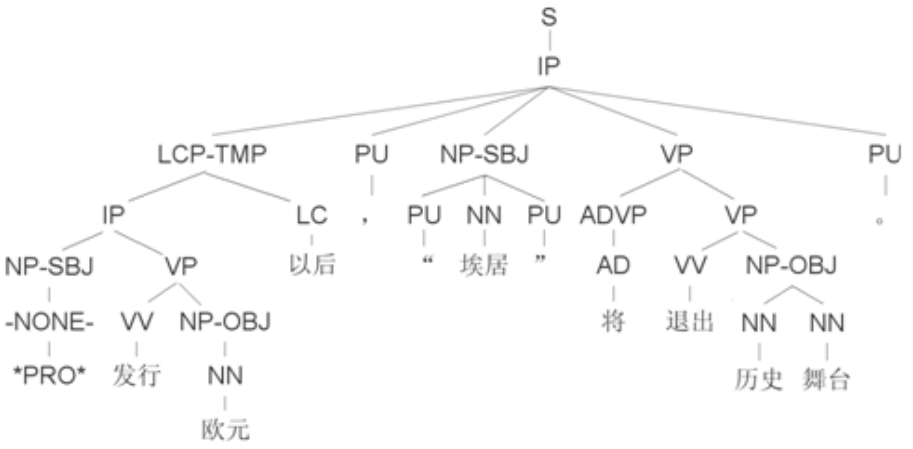


Fig. 2. Non-sentence boundary denoting comma

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 describes the framework of Chinese segmentation. Section 4 discusses the features used in the framework. Section 5 reports our results. Section 6 concludes our paper.

2 Related Work

There are two approaches generally used to detect English sentence boundary. The first approach uses manually built rules which are usually encoded in terms of regular expression grammars. For instance, the Alembic workbench[5] contains a sentence splitting module which employs over 100 regular expressions rules written in Flex. The second approach employs machine learning techniques such as Maximum Entropy (ME) model[6], neural networks[7], etc.

Chinese also uses punctuation to indicate sentence boundaries. The difference is that the Chinese comma also functions similarly as the English period in some context and signals the boundary of a sentence. Chinese comma has also been studied in the context of syntactic parsing for long sentences[8-9], where the study of comma is seen as part of a “divide-and-conquer” strategy to syntactic parsing. Long sentences are split into shorter sentence segments on commas before they are parsed, and the syntactic parses for the shorter sentence segments are then assembled into the syntactic parse for the original sentence.

Xue et al. divided Chinese comma into two types: EOS and Non-EOS, and then devised a heuristic algorithm to label each comma with either EOS or Non-EOS [10]. After careful observation of each kind of commas, they selected appropriate features to train the ME classifier, and their experimental results achieved F-measure of close to 90% overall.

In this paper, we present a Two-layer Maximum Entropy Model (MMEM) system to distinguish commas. In the first layer classifier, we use Xue’s approach to reproduce their system and use it as the baseline of our system. Then we add six new

features to train the second layer classifier. Finally, the experimental results show that our approach can significantly improve the performance.

3 The Framework of Chinese Segmentation System

The framework of our system is shown in the figure 3. The whole system can be seen as a MMEM, which consists of four parts as follows：

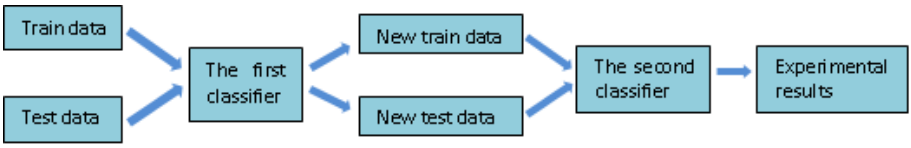


Fig. 3. The framework of Chinese sentence segmentation

- Train the first ME classifier

First of all, we do the preprocessing such as word segmentation with the CTB raw text, and then get the syntax tree of each sentence using the Berkeley parser. After that, we obtain the features of each comma and create the training file. In the end, we train the first classifier using the training file.

- Use the first ME classifier to classify the two kind of commas

After using the similar approach as the above step to obtain test file, we submit the file to the first classifier and then obtain the results.

- Obtain new features and train the second ME classifier

We process the training file and then obtain the new features to train the second classifier. The difference between the training data and test data of the second classifier is that when trying to obtain the test file we are supposed to use the results of the first classifier. Procedures to obtain test file are showed as follows: 1)define a threshold for the results of the first classifier; 2)find commas in test file which confidences are above the threshold; finally, we get the relative feature of test commas and then get final results.

- Contrast the results of the second ME classifier with the corpus and get the final results

4 Feature Selection

The ME model is an extremely flexible technique for linguistic modeling, since it can use a virtually unrestricted and rich feature set in the framework of a probability model. So we can concentrate on choosing features instead of using them. Many scholars have discovered that the most obvious influences on the performance of the system are the feature they chose rather than learning algorithm.

4.1 Features of the first classifier

Since turning comma disambiguation into a binary classification problem, we consider the comma and its context when we choose the features which are defined as follows:

- Lexical features: words before and after the comma
- POS features: POS of words before and after the comma
- Syntactic information: Syntactic information of the sentence which contains the comma

According to the feature set, we define the features of our system as shown in Table 1. The features of the first classifier come from Xue's system, which most of them are syntactic feature. So we could estimate that the accuracy of syntactic parsing will affect the experimental results.

Table 1. Features used in the first classifier

features in the first classifier	PreWord&POS	words and its POS before the comma
	FolWord&POS	words and its POS after the comma
	LeftSibling	The phrase label of left sibling of the comma in syntactic parse tree
	RightSibling	The phrase label of right sibling of the comma in syntactic parse tree
	Left&RightSibling	The phrase label of left and right sibling in the syntactic parse tree
	A&L&RSibling	The conjunction of the ancestors, the phrase label of left and right sibling
	HasCS	Whether there is a subordinating conjunction(e.g., "if", "because") to the left of the comma
	ISCoordIP	Whether the parent of the comma is a coordinating IP construction
	ISTopChild	Whether the comma is a top-level child, defined as the child of the root node of the syntactic tree
	ISTopCoordIP	Whether the parent of the comma is atop-level coordinating IP construction
	PunctuaMark	The punctuation mark template for this sentence

4.2 Features of the Second Classifier

After carefully observation of the first classifier experimental results, we found that parts of Non-EOS commas which are falsely classified into EOS commas have the same characteristic. That is, they are always in a sentence which clauses have same syntactic structure. And in fact they are Non-EOS.

(b) Chinese: 二十多年来，[1]中国经深体制改革不断深化，[2]综合国力明显增强，[3]对外经贸合作日益扩大。

Pinyin: èrshí duō niánlái, [1] zhōngguó jīngjì tǐzhì gǎigé bùduàn shēnhuà, [2] zōnghé guólì míngxiǎn zēngqiáng, [3] duìwài jīngmào hézuò rìyì kuòdà 。
English: Over the past two decades and more, [1] China has deepened the reform of its economic system, [2] visibly increased its overall national strength, [3] and steadily expanded its foreign economic cooperation and trade.

Since clauses separated by the comma in this sentence have their own subject-verb-object structure and fit most of features of the first classifier, it is more inclined to be classified into EOS (such as [2] and [3] in Example (b)). This phenomenon have been seriously affected the experimental results of the first classifier, so we introduce the CommaProportion feature in the second classifier based on the results of the first classifier. CommaProportion, which is used to describe the comma's tendency to be an EOS or Non-EOS, has three values: 1) Negative when the number of Non-EOS is larger than those of EOS, 2) Same when the number is same as those of EOS and 3) Positive when the number of Non-EOS is smaller than those of EOS.

Table 2. Features used in the second classifier

features in the second classifier	PTwoWord&POS	The second words and its POS before the comma
	FTwoWord&POS	The second words and its POS after the comma
	LeftSiblingHasPP	Whether there is a phrase label (e.g., "PP", "ADVP" and "LCP") to the left of the comma
	WordSeman	Semantic information of two words before and after the comma
	CommaProportion	Comma proportion between EOS commas and Non-EOS commas in the sentence
	LeftSiblingHasIP	Whether there is a coordinating IP construction to the left of the comma

In addition, we observe that the comma is more inclined to be classified to be a Non-EOS when the clause before comma has phrase structure with a POS of 'PP', 'ADVP' or 'LCP'(such as [1] in Example (b)). So we introduce the feature LeftSiblingHasPP used to describe this situation. The results show that this feature can improve the performance of the framework. Other features are described in the Table 1.

5 Results and Discussions

To compare with the baseline, we used Xue's corpus of comma (a subset of the Chinese TreeBank (CTB) 6.0) to reproduce their system and established our system. The CTB file IDs used in our experiments are listed in Table 2. We conducted our experiments with a ME classifier trained with the Mallet package[11].

Table 3. Data set division (1,510 commas in the test set)

Data	Training set	Test set
CTB	41-325,400-454,500-554,590-596, 600-885,900,1001-1078, 1100-1151	1-40, 901-931

Firstly, we compare the performance of the first classifier when it uses automatic syntactic trees or golden-standard ones. In Exp1, the features in the training /test data are derived from the automatic syntactic trees. In Exp2, the features in the training data are derived from the golden-standard syntactic trees while the features in the test data are derived from automatic syntactic trees. Table 4 shows the experimental results, which can conclude that the performance of the syntactic parser does affect the experimental results.

Table 4. Results in the first classifier

Performance	Exp1: Automatic syntactic trees			Exp2: Golden-standard syntactic trees		
(%)	P	R	F1	P	R	F1
Overall	89.2	89.2	89.2	99.3	98.4	98.9
EOS	64.7	76.4	70.1	96.7	96.0	96.3
Non-EOS	95.1	91.7	93.4	99.8	98.9	99.4

The results of our experiments are presented in Table 5. The second classifier achieved a modest improvement over the baseline. The F-measure score of the second classifier is 90.7% while the precision and recall for EOS commas are 81.0% and 80.4% respectively and the F-measure score is 80.1%. For Non-EOS commas, the precision and recall are 93.2% and 92.4% respectively, with the F-measure score being 92.7%.

Table 5. Experimental results of the baseline and the two classifier

	The first classifier (Xue)			The second classifier		
(%)	P	R	F1	P	R	F1
Overall	89.2	89.2	89.2	91.2	90.4	90.7
EOS	64.7	76.4	70.1	81.0	80.4	80.1
Non-EOS	95.1	91.7	93.4	93.2	92.4	92.7

Table 6 shows the contribution of individual feature groups. The numbers reflect the F-measure when each feature group is taken out of the model. All the features have made a contribution to the overall. We have done lots of experiments to find out a threshold which could ensure the second classifier has better performance, and make the threshold to be 0.65 at last.

From the experimental results, we can find that the most effective feature is the CommaProportion feature. However, since the proportion of EOS to Non-EOS is 1:5 in the whole corpus, there is some EOS commas are falsely classified into Non-EOS commas. Therefore, the F-measure of EOS commas has been affected after introducing this feature.

As the features of the first classifier required high accuracy of the syntactic tree, this paper introduces some other features instead of syntactic features. Table 6 shows the contribution of individual feature groups, which could conclude that features of the second classifier have made a contribution to our framework.

Table 6. Feature effectiveness

	Features	overall	F1(EOS)	F1(Non-EOS)
features in the first classifier	all	89.2	70.1	93.4
	-PreviousWord&POS	-0.1	+1.6	-0.5
	-FollowingWord&POS	-0.1	+1.6	-0.5
	-LeftSibling	-0.1	+1.6	-0.5
	-RightSibling	-0.3	0	-0.4
	Ancestor&L&RSibling	-0.2	+1.6	-0.6
	-ISCoordIP	-0.1	+1.6	-0.5
	-ISTopCoordIP	-0.8	-10.4	+1
features in the second classifier	all	90.7	80.1	92.7
	-PTwoWord&POS	-1.8	-5.4	-1
	-FTwoWord&POS	-1.8	-6.6	-0.8
	-LeftSiblingHasPP	-1.9	-6.6	-0.7
	-WordSeman	-1.5	-6.2	-0.5
	-CommaProportion	-1.8	-5.4	-1.1
	LeftSiblingHasIP	-2.0	-7.4	-0.8

6 Conclusion

The sentence is a standard textual unit in natural language processing applications. As a sequence of string, sentence boundaries must be detected in raw text before we do the lexical analysis and syntactic parsing. The accuracy of sentence boundaries will affect the downstream work in NLP. We trained a statistical model using data from Xue's system and reported our results. Our model achieves a classification F-measure of close to 91% overall, which improves by 1.5 percent.

Hidden Markov Models (HMMs) are a powerful probabilistic tool for modeling sequential data and have been applied with success to many text-related tasks, such as shallow paring. In these case, the obssservations are usually modified as multinomial distributions over a discrete dictionary and the HMM parameters are set to maximize the likelihood of the observations[12-13]. In future work, we will explore HMMs in Chinese sentence segmentation and try to improve the performance of our system.

References

1. Jeffrey, C.R., Adwait, R.: A Maximum Entropy Approach to Identifying Sentence Boundaries. In: Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP), pp. 803–806 (1997)
2. Junhui, L., Guodong, Z., Qiaoming, Z., Peide, Q.: Syntactic Parsing with Hierarchical Modeling. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 561–566. Springer, Heidelberg (2008)
3. Qiaoming, Z., Junhui, L., Hongling, W., Guodong, Z.: A Unified Framework for Scope Learning via Simplified Shallow Semantic Parsing. In: Proceedings of the 2010 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 714–724 (2010)

4. Junhui, L., Guodong, Z., Hongling, W., Qiaoming, Z.: Learning the Scope of Negation via Shallow Semantic Parsing. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pp. 671–679 (2010)
5. John, A., John, B., David, D., Lynette, H., Patricia, R., Marc, V.: MITRE: Description of the Alembic system used for MUC-6. In: The Proceedings of the Sixth Message Understanding Conference (MUC-6), pp. 141–155 (1995)
6. Neha, A., Kelley, H.F., Max, S.: Sentence boundary detection using a MaxEnt classifier
7. David, D.P., Marti, A.H.: Adaptive sentence boundary disambiguation. In: The Proceeding of the 1994 Conference on Applied Natural Language Processing (ALNP), pp. 241–267 (1994)
8. Meixun, J., Miyoung, K., Dong, K., Jong, L.: Segmentation of Chinese Long Sentences Using Commas. In: Proceedings of the SIGHANN Workshop on Chinese Language Processing (2004)
9. Xing, L., Chengqing, Z., Rile, H.: A Hierarchical Parsing Approach with Punctuation Processing for Long Sentence Sentences. In: Proceedings of the Second International Joint Conference on Natural Language Processing: Companion Volume Including Posters/Demos and Tutorial Abstracts
10. Nianwen, X., Yaqin, Y.: Chinese sentence segmentation as comma classification. In: The Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 631–635 (2010)
11. Andrew, K.M.: Mallet: A machine learning for language toolkit (2004), `http://mallet.cs.umass.edu`
12. Zhou, G.: Direct modeling of output context dependence in discriminative Hidden Markov Model. Pattern Recognition Letters, 545–553 (2005)
13. Zhou, G.: Discriminative hidden Markov modeling with long state dependence using a kNN ensemble. In: Proceedings of the 20rd International Conference on Computational Linguistics (COLING), pp. 22–28 (2004)

An Ontology-Based Approach for Topic-Based Interpretation Training

Man Feng

School of Foreign Languages, Zhongnan University of Economics and Law, Wuhan, China
omaggief@yahoo.com.cn

Abstract. Ontology, used to systematically model the domain knowledge, facilitates not only computation but also cognitive processing of human brain. Whereas, interpretation process involves similar brain activities, an ontology-based training method is proposed for topic-based interpretation training (TBIT) to meet its two major teaching objectives: Both subject knowledge and LSP (Lan-guage for Special Purpose) knowledge are to be acquired by learners. Ontology can help interpreters to clarify unfamiliar concepts, acquire knowledge and trans-form it into long term memory (LTM). An ontology-based training can improve interprter's information processing ability, anticipation ability, and short term memory capacity so that a high efficiency of prior knowledge operation can be re-alized and better quality of rendering can be produced.

Keywords: ontology-based approach, topic-based interpretation training.

1 Introduction

Ontology is originally a philosophical term, describing the nature and universals of being by categorizing things and phenomenon, depicting their properties and relations. The ontology in this paper is the scientific ontology, and was defined by Studer as "the formal explicit specification of a shared conceptualization" [1]. It can be seen as a knowledge network to model a specific field with a set of definitions and categorized concepts.

Having been rigorously formalized, ontology can be used to systematically analyse the domain knowledge, realize human-machine and human-human information sharing, knowledge consultancy, knowledge disambiguation and knowledge reuse within a specific domain and/or between different domains[2]. In general, ontology largely serves in knowledge management, information search, long-distance learning, e-administration and e-commerce, etc.

Its functions of knowledge management, inference of knowledge and facilitation of automation share great similarity with the language transformation process through instantaneous knowledge retrieval and information processing in interpretation. This paper tries to explore and propose an ontology-based training method for TBIT.

There are various types of ontologies in computer science according to Professor FENG Zhiwei: common ontology, domain ontology, language ontology and formal

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 818–828, 2013.

ontology[3]. Among different types of ontology, domain ontology and language ontology are more relevant to TBIT.

2 Cognitive Difficulties Encountered by Interpreters in TBIT

TBIT, a most popular training course for interpretation learners, is given much weight in the overall training program of interpreters as it combines linguistic knowledge, subject knowledge and interpreting skills into the exercise. China's TBIT are facing unique difficulties from western programs: Trainees in China are rarely equivalent bilinguals due to their lack of bilingual family background and of enough foreign culture exposure. Their acquired foreign language level is still far from equivalent to their mother language level, therefore, subject knowledge acquisition and LSP (Language for special purpose) competence improvement become the two teaching focus. The traditional training consists of 3 modules: foundation building (preparation of glossary and subject knowledge), brainstorming (glossary exchange and delivery of impromptu speech), and interpreting practice incorporating interpreting skills with special subject knowledge. However, the practice module could be extremely difficult when the topic is unfamiliar to the instructees, even with the help of a glossary: unfamiliar topic increases the information processing load of the interpreter; the newly-obtained knowledge and vocabulary may stay asleep at the bottom of interpreter's LTM and cannot be elicited efficiently through the prior knowledge operation during the instantaneous bilingual transformation. Those cognitive difficulties encountered by interpreters in the training are where ontology could be introduced as a resolution.

3 Applying Ontology to TBIT

3.1 Applying Ontology to Acquire Subject Knowledge and LSP

In interpretation studies, the importance of knowledge structure of interpreters has been well emphasized. Professor Selescovitch explained that the synthesis of background knowledge and senses formed a conceptualized mental representation that is of vital importance for interpreter's comprehension and rendering[4]. Professor Gile [5], representative of the cognitive school, and Professor Zhong Weihe[6], a Chinese expert on interpretation studies all pointed out the similar significance.

3.1.1 Using Ontology to Collect Terminologies and Concepts

The acquisition of subject knowledge relies not only on intperter's daily accumulation but also on a good preparation before interpreting task. The normal practice for the preparation is that interpreters will collect by themselves or, if lucky, be provided with a disor-ganized or at most alphabetically sorted glossary, which often does not cover all the aspects. Ontology serves well in collecting major concepts in a systematic way. Do-main ontology itself is the abstraction of domain knowledge. There exist all kinds of domain-specific ontologies, such as domain-specific ontology of communication, transportation, Chinese medicine, archeology, business administration, etc.,

while at the same time more and more domain ontologies have being built up. Even if a rele-vant ontology is not available, a classified thesaurus can be referred to at least. The collected ontology or classified thesaurus can help interpreters to obtain important concepts within a domain within a short time period. For example, if we are preparing a meeting on weapons of mass-destruction, from SUMo, we can filter out a compre-hensive list of concepts about weapons of mass destruction (excerpt from search result of SUMO) [7]:

...

```
(subclass BacterialAgent ToxicOrganism)
(subclass BacterialAgent Bacterium)
(biochemicalAgentDelivery BacterialAgent Breathing)
(biochemicalAgentDelivery BacterialAgent Touching)
(documentation BacterialAgent EnglishLanguage
"&%BiologicalAgents that are instances
of &%Bacterium.")
(=>
   (instance ?BACTERIUM Bacterium)
   (exists (?NUMBER)
      (and
         (width ?BACTERIUM (MeasureFn ?NUMBER Meter))
         (greaterThanOrEqualTo ?NUMBER 0.000001)
         (lessThanOrEqualTo ?NUMBER 0.000002))))
```

...

3.1.2 Using Ontology to Comprehend and Clarify Concepts

Due to the lack of relevant knowledge, even with a bilingual glossary, interpreter may get concepts con-fused. “Ontologies are often equated with taxonomic hierarchies of classes, with class definitions, and the subsumption relation.”[8] Referring to the categorization, definition, and property descriptions provided by an ontology, a concept can be well clarified by figuring out whether its relations with other terms are sub-class, instance, attribute, cause, effect or relation of other kinds. For example, if trainees are not clear about “inflation”, “WordNet” that they refer to will offer explanation as below (ex-cerpt from search result of WordNet) [9] in several ways like giving hyponym, hypernym, antonomy and etc:

- S: (n) inflation, rising prices (a general and progressive increase in prices) "in inflation everything gets more valuable except money"
  - direct hyponym / full hyponym
- S: (n) cost-pull inflation (inflation caused by an increase in the costs of production)
- S: (n) demand-pull inflation (inflation caused by an increase in demand or in the supply of money)
- S: (n) reflation (inflation of currency after a period of deflation; restore the system to a previous state)

- S: (n) stagflation (a period of slow economic growth and high unemployment (stagnation) while prices rise (inflation))
  - direct hypernym / inherited hypernym / sister term
  - antonym

W: (n) deflation [Opposed to: inflation] (a contraction of economic activity resulting in a decline of prices)

W: (n) disinflation [Opposed to: inflation] (a reduction of prices intended to improve the balance of payments)

[Excerpt from search result of WordNet]

The clarification of concepts can be easier with the visualization of an ontology. With the help of an engineer, by filtering certain properties, we can get different geometrical representation of the requested ontology. For example, from "Figure 1" [10], a section of a visualized ontology simply representing the taxonomy, interpreters can clarify the concepts easily: the applicative programming, automatic programming, concurrent programming, sequential programming, object-oriented programming, logic programming, and visual programming actually are different programming techniques, whereas distributed programming and parallel programming are two types of concurrent programming.

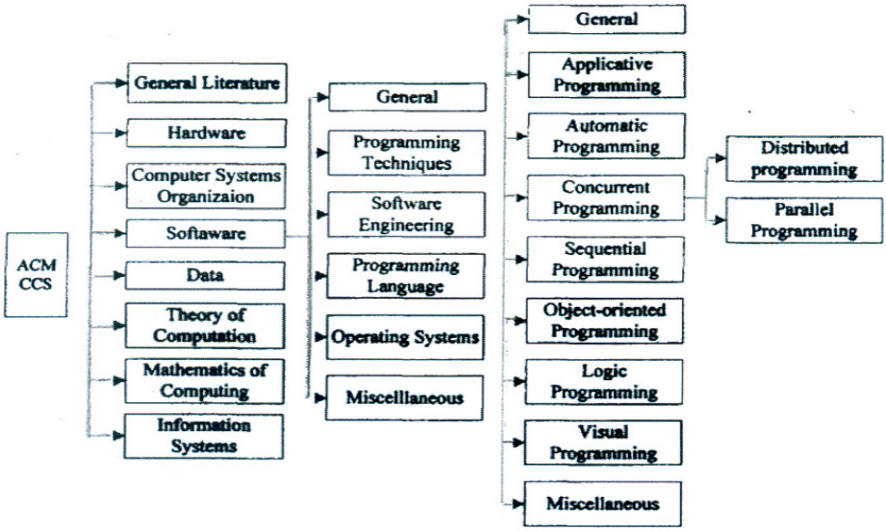


Fig. 1. Structure of ACM CCS (Source: CNKI)

3.1.3 Using Ontology to Assess Interpreter's Preparation before Task

With the help of ontology, an overall view of a domain knowledge, an interpreter can check its own knowledge acquisition and efficiently find out their weak points in the com-prehension of a certain topic. For instance, if the term "distributed programming" is vague to trainees, they know where to make efforts to retrieve the useful

information for further study. They can search the corpora with keywords representing any adja-cent concepts to “distributed programming” and get language material to practice with. Also, trainees can check with each other for their comprehension of terminology and concepts with each other in both working language by referring to an ontology. Certainly this work can also be done by instructors.

3.2 Adapting Ontology to Automatic Processing of Assorted Chunks

Interpreters have to work with very short time limit, a powerful STM or working memory can help interpreters identify, retain and retrieve information with minimum note-taking thus time will be spared for interpreter to comprehend and interpret the source message. According to German psychologist Ebbinghous, STM capacity is very limited, averaged as 7±2 information items, no matter the items are letters, figures, irrelevant words or chunks [11].

To expand the STM capacity, Wood proposed a model of speech fluency based on automatic processing and retrieval of prefabricated chunks [12]. This model explained that formulaic language units are fundamental to fluent language production, as they allow language production to occur while bypassing controlled processing and the constraints of short-term memory capacity. Also multiword lexical units combine language form and functions, grammar and meaning, and vocabulary together in a way that language points have been integrated and withdrawn when language is planned and produced, as a result, retrieval time to language points has been shortened.

The above point inspires us to adapt ontology for our own use: Each concept of the ontology can be represented by multiword lexical chunks expressing concrete meaning with a specific relation to the term, if many chunks has been stored in interpreter’s LTM along the ontological knowledge network, they will help interpreters render a speech without thinking over how to put language points together by grammatical rules but simplify the process into one step. As the prefabricated chunks have been stored, errors of grammar and wrong usage of words could be much reduced so the rendering is more accurate and idiomatic. Therefore, if interpreters can build up their own work sheets of ready-made chunks to represent the domain ontology in both languages, it’s quite sure that the STM capacity and working efficiency will be augmented with the increased accessibility to and the retrieval of those chunks. For in-stance, when presenting principles of insurance, the learners can convert this knowledge into 4 sub-classes, under each subclass different chunks are given to mark their relations with this sub-class, whether it’s definition/explanation, instance, consequence of violation, comparison or other kinds. If the concepts under a single topic are quite a few, this topic can be signaled out. Detailed illustration is given as below:

Principles of Insurance

1) Principle of Insurable interest: financial or other interest insured against an in-sured risk (explanation/definition).

2) Principle of Proximate cause: direct cause (explanation /definition); loss aroused by indirect causes will not get indemnified (consequence of violation of this principle).

3) Principle of Indemnity: compensation for loss or damage to indemnify the insured (explanation/ definition); life insurance, cargo insurance (Instance).

4) Utmost good faith: disclosure of relevant information regarding the insured risk (object); breach of good faith / principle of good faith / duty/ contract/ confidence / security (opposite, alias).

Instructors can require instructees to practice those chunks to a level of automatic processing, a feasible target for Chinese learner at the level of chunks. Later on, training materials containing relevant information will be provided to check the processing of those chunks in a specific context. A comparison study is suggested here to compare the STM capacity between Group A only preparing glossary in the traditional way and Group B adapting ontological knowledge into required worksheet on the same topic.

3.3 Applying Ontology to Information Processing Exercise in TBIT

Interpretation is a process of information processing, consisting of listening, monitoring, storing, retrieving, comprehending through decoding, and encoding of received information in target language. Dr. Liu Minghua mentioned in a lecture that "more-skilled interpreters differ from less-skilled interpreters in their information processing being more semantic-based, being more selective in what to interpret, being more efficient at lexical processing, having a better grasp of text structure, being more selective in listening, and having a more enhanced self-awareness of the task." [13] Interpreter trainee shall learn to utilize their knowledge structure to timely grasp the overall structure of an utterance instantaneously and construct a clear logic network between key concepts of a discourse, and process the essential information with appropriate attention split. In terms of information processing, ontology can build up a solid background knowledge schema, providing more contextual information for the comprehension of language signs, thus back up the mental process of understanding.

3.3.1 Applying Ontology to Logic Training in TBIT

During the initial stage of training, memory training and information processing exercise are often a prerequisite. When listening to a speech, trainees are required to categorize major concepts or keywords to represent the main idea of a discourse and depict the logic links between them. The "30 relations" described by the "frame ontology theory" [14] can be simplified and introduced to trainees so that they have a rough idea how to describe various relations between concepts.

Vice versa, terms and concepts can be withdrawn from an ontology so that trainees can make a discourse by organizing them with logic links. For example, we can randomly list out "seed production", "transplanting techniques", "diagnosis" and "pest control" from "Figure 2"[15] and ask trainees to make a discourse with them with or without changing their appearing sequence. As a discourse is based on logically linked concepts, interpreter's information processing ability can be practiced through working with logic and their minds.

If we go deeper with the ontology and know more specific information, we will be able to understand the following sentence which contains highly specialized knowledge and terminologies. An instructor can cut it in pieces, disarrange them for trainees to reformulate the message by listening to those disordered pieces.

"This is the most successful method used in hybrid seed production of tetraploid cottons wherein 40 to 50% or more seed setting is obtained. The method involves removal of corolla along with anther sheath by giving shallow cut at the base of the bud with thumb nail and removing corolla and anther column in one jerk twisting action. Care should be taken to ensure that the white cover membrane of the ovary is not damaged or removed during this operation as this affects the boll setting."[16]

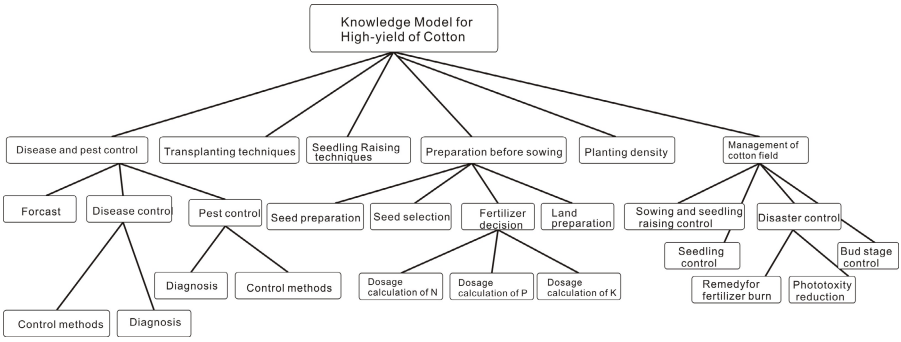


Fig. 2. Visualization of task ontology (Source: translated version from CNKI)

3.3.2 Applying Ontology to Anticipation Exercise in TBIT

Anticipation is often used as a coping strategy to reduce the load of short term memory (STM) and lessen the difficulty in listening and comprehension so as to obtain some information in advance and spare more energy to optimize the rendering[17]. Fred Van Besien found in his research that "extralinguistic information like general and situational knowledge, and information obtained in the course of translation, seems to play the most important part in the interpreter's hypothesizing of the speaker's utterances. Purely linguistic knowledge plays only a minor part"[18].

Ontological knowledge of an interpreter can help with the extralingual anticipation to a great extent. With the clear knowledge network of ontology, interpreter can easily build up concept knots between different concepts, thus facilitates the logical reasoning and inference along the extension of a concept. Therefore it's suitable to incorporate ontology to the training of anticipation skills. Taking "Figure 1" again as an ex-ample, since under the subclass of "parallel programming", attributes of it are explained (contents are omitted in the Figure1), then it's not difficult to interpret the following sentence with anticipation initiated at hearing the comparison between parallel and sequential programs: "Parallel computer programs are more difficult to write than sequential ones, because concurrency introduces several new classes of potential software bugs, of which race conditions are the most common. Communication and synchronization between the different subtasks are typically some of the greatest obstacles to getting good parallel program performance" [19].

Anticipation exercise can also be done with "Figure 2", instructor presents the topic of the conference as "Hybrid Seed Production in Cotton" while demanding instructees to predict the possible subtopics, or plays a video and stops in the middle to require students to anticipate the next following sentences. With the help of an ontology, anticipation exercise turns out to be possible and easier.

3.3.3 Applying Ontology for Emergency-Response Exercise in TBIT

Interpreters may get trapped because of internal and external factors while listening to speakers and reproducing the received message in target language. The interconnected rela-tions between concepts demonstrated by ontology have covered many aspects such as synonyms, antonyms, hypernyms, hyponyms, and meronyms which may offer alternatives of similar expressions for interpreters to adapt to difficult situations flexibly. For example, when visiting a factory the manager inquired about "瓦斯继电器 (wa si ji dian qi, gas relay)", you didn't know the English for "瓦斯(wa si, gas)", however you may recall that "gas" are more often used for explosive air, therefore you may try to interpret "瓦斯继电器(wa si ji dian qi, gas relay)" as "gas relay" accompanying with an explanation that you are not sure about the terminology, not knowing you're coincidentally correct. These kinds of coping strategies are quite effective.

It's obvious that frequent retrieval to synonyms, antonyms, hyperyms, hyponyms, and meronyms of a terminology shall be encouraged because they supply flexible alternatives for emergency-responses. Exercise can be carried out for difficult terminologies with the help of an ontology. We take difficult terms out of an ontology and ask trainees to provide alternative expressions, such as "a cave that stored Buddhist literature or a library cave of monks" for "藏经洞 (cang jing dong, sutra cave)", "a disease of neuromuscular junction" or "neuromuscular disorder" for "重症肌无力（zhong zheng ji wu li, Myasthenia gravis）", etc. For this kind of exercise, we can refer to some ontologies like WordNet, HowNet, etc.

4 Establishing an Ontology-Based Approach for TBIT

TBIT is a systematic process to to meet instructional goals effectively. From the above discussions on how the usage of ontology can be adopted for TBIT, an ontology-based approach is established for different training modules of TBIT: the pre-class module, in-class module and post-class module.

For the pre-class module, during the preparation stage, given a specific topic, trainees shall collect information and refer to the existing domain ontologies, better in both language versions; then convert them into a visualized worksheets with formulaic chunks to represent the ontological knowledge of that topic. If there does not exist an relevant ontology, trainees shall be able to build up their non-formalized ontology with a classified thesaurus within that domain. In this way terminology and concepts in this domain can be reached by learners to maximum extent and on a systematic basis. At the beginning of the class, learners shall be able to draw a knowledge map relevant to the demanded topic. They shall do self-assessment on how well the domain knowledge has been acquired in both languages at all levels from terminologies

to concepts to different view points and even academic analysis. Trainees can also test themselves how well they can do automatic processing of formulaic chunks and how far they can follow a certain topic after referring to domain ontology. At last, trainees are encouraged to share and exchange information to expand the scope of the adapted ontologies.

For the in-class module, the adapted ontological knowledge represented by assorted chunks can be referred to by trainees to do an impromptu speech on whatever they choose from within, in either working language. The speech could be anything related to the therein concept, either a brief introduction, an explanation, or extended arguments, which serves the purpose of assessing whether trainees can express themselves in a professional manner. Trainees can also use others' worksheets of assorted chunks to check themselves whether they can make sense of those chunks and make a discourse by building up logic links between them. In this way, public speech competence has been trained, the language knowledge and subject knowledge newly stored in the LTM can be activated and reinforced.

Logic exercise, anticipation exercise and emergency-response exercise can be carried out in the class with the adapted ontology, either in the training of a single skill or of comprehensive skills. For example, synonyms and antonyms, hyponyms and hypernyms are required to be given in exercises for high-frequency words or difficult terms; anticipation exercise can be done within a certain context; emergency-response training for jargons and etc.

Instructors, being aware of learners' weak points in mastering an ontology, can target at training of the weak parts by inputting the relevant high-frequency word/concept to obtain relevant materials so that the teaching content are easily adapted according to learner's different levels. With the accumulation of language knowledge and subject knowledge, both instructors and instructees can select more difficult materials to integrate into interpreting skills training, thus make a comprehensive training to improve the output quality of interpretation of specialized topics.

For the post-class module, learners are encouraged to exchange their worksheets of assorted chunks to represent specific domain ontology through blog, bbs or other online interaction tools. Continued tape hours can be done based on the weak points of the mastering of an ontological knowledge.

As for the different stages of TBIT, In China, it is normally divided into 2 stages: the initial stage and the following stage. At the initial stage, basic training of memory and information processing is a prerequisite. Ontology can be integrated into STM and LTM training, mind training, information processing exercise as aforementioned. In the following stage, it can be incorporated into the foundation building, professional language assessment, single interpreting skill training and comprehensive interpreting skills training, etc.

5 Conclusion

An ontology-based approach for TBIT assists to systemize the language knowledge and subject knowledge required by interpreters, making it easier for them to store,

retrieve, process and utilize both knowledge. Its adoption facilitates interpreters to improve their information processing, STM capacity, LTM capacity and activation, anticipation and flexibility in interpretation. All those effects will take interpreters closer to their training target of quasi-automatic or semi automatic interpretation.

However, this approach is facing many challenges. The interpreters shall widely collect materials for specific domain ontology, of which the accessibility and availability has been very much limited and the referred ontologies shall be post edited for the interpreters' own use. Although with the development and availability of ontology tools, individuals can develop ontologies and make them public, however, their quality and suitability is rather doubting to meet various educational purposes. Sometimes interpreters have to build up their own bilingual ontological knowledge from sources of textbooks and domain experts, which shall involve a team and can be time consuming. It's highly recommended that the translation industry shall work together to build up or convert different domain ontologies for our own use. At last, this paper discusses the ontology-based approach for TBIT only in terms of training methods, how to adopt it to achieve efficient prior knowledge operations of an interpreter and how to measure the benefits of using ontologies are not mentioned, nor does it demonstrate how well the application of ontology can improve the result of TBIT. All of these aspects shall be further discussed and explored with empirical studies.

Acknowledgement. This work is sponsored by the Research Foundation of Zhongnan University of Economics and Law (Grant No. 31541110201).

References

1. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge Engineering: Principles and Methods. J. Data & Knowledge Engineering 1, 161–197 (1998)
2. Hu, D.: Alloconcepts and Formalized Description. China Social Science Press, Beijing (2011)
3. Feng, Z.W.: Chinese Translation for Ontology: Ontology (Bentilun) and Knowledge Ontology (Zhishi Benti), http://www.survivor99.com/pscience/2006-9/ontology_fzw.html
4. Selescovitch, D.: Interpreting for International Conferences. Pen and Booth, Washington (1978)
5. Gile, D.: Basic Concepts and Models for Interpreter and Translator Training. John Benjamins Publishing Company, Amsterdam (1995)
6. Zhong, W.H.: Knowledge Requirements for Interpreters and Their Implication to Interpreting Course Designing. J. Chinese Translators Journal 4, 63–65 (2003)
7. Suggested Upper Merged Ontology, http://www.ontologyportal.org
8. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. J. Knowledge Acquisition 2, 199–220 (1993)
9. WordNet, http://wordnetweb.princeton.edu
10. Ou, Y.Y.: Ontology-based Adaptive E-learning Modeling in Semantic Learning Web. Dissertation for Doctoral Degree, p. 68. Zhejiang University (2007)
11. An, X.K.: On Memory of Interpreters. J. Chinese Science & Technology Translators Journal 4, 21–23 (2004)

12. Wood, D.: Formulaic Language in Acquisition and Production: Implications for Teaching. J. TESL Canada Journal 1, 1–15 (2002)
13. Liu, M.H.: The Making of a Skilled Interpreter: What we know about expertise development in interpreting, http://www.miis.edu/academics/programs/gstile/found-in-translation-series/past-lectures
14. Theory Frame Ontology, http://www.ksl.stanford.edu/knowledge-sharing/ontologies/html/frame-ontology/index.html
15. Wei, Y.Y.: Research of Ontology-based Agricultural Knowledge Modeling and Reasoning. Dissertation for Docter's Degree, pp. 107. University of Science and Technology of China (2011)
16. Hybrid Seed Production in Cotton, http://www.cicr.org.in/pdf/hybrid_seed_production.pdf
17. Gile, D.: Basic Concepts and Models for Interpreter and Translator Training. John Benjamins Publishing Company, Amsterdam (1995)
18. Fred, V.B.: Anticipation in Simultaneous Interpretation. J. Meta: Translators' Journal 2, 250–259 (1999)
19. Parallel Computing, http://en.wikipedia.org/wiki/Parallel_computing

The Construction of Music Domain Ontology

Li Yang[1] and Jinglian Gao[2]

[1] School of Art, Hubei University of Science and Technology, Xianning, 437000
[2] Guangdong Guobi Technology Co., Ltd, Guangzhou, 510620
keanhu@sina.com, gao@guobi.com

Abstract. The construction of music domain ontology is an urgent task for the automatic processing of the knowledge on music and musicology. This paper presents the process of constructing music domain ontology by the Seven-step method in detail. By defining the classes, the hierarchy of classes, the relation between classes and the properties of classes, an ontology model of music domain is established.

Keywords: domain ontology, music and musicology, the seven-step method.

1 Introduction

With the rapid development of the information technology and the increasing popularization of the Internet, the informationization degree of music and musicology is rapidly increased. Inexhaustible musical information keeps rushing into the Internet carried by massive texts, sounds, videos, images and databases. These resources have greatly enriched the storage modes and the transmission way of musical knowledge, and are quite useful for musicology information processing, music teaching and digital music processing.

However, because the musical information resources are described with different system of terms or thesaurus by different software, it is quite difficult to share and reuse knowledge among different systems. Furthermore, as to the intelligent web information retrieval, because so many data are unstructured or semi-structured, it cannot be guaranteed that the intelligent agent makes effective access and retrieval of the heterogeneous information in the web. Thus the utilization efficiency of the musical information resources is sharply reduced. [1]

A music domain ontology may give a satisfactory solution for these puzzles. This paper discusses the method of constructing music domain ontology, defines the basic elements of music domain ontology and finally gives a model of music domain ontology.

2 Method of Constructing Domain Ontology

Domain ontology is a professional ontology which provides the thesaurus of the concepts and the relation between these concepts on some certain domain.[2] Music

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 829–835, 2013.

domain ontology is such an ontology on music and musicology. Many disciplines have already possessed their own domain ontologies, but for music field, this research just starts.

In order to construct a domain ontology, we should firstly capture enough related dominical knowledge, provide a common understanding to these knowledge, establish the acknowledged thesaurus in this domain, and make a explicit definition to the thesaurus and all the relationship between terms.

Generally speaking we have several common methods for constructing a domain ontology: the TOVE method, the IDEF5 method, the Skeleton method, the KACTUS method, the SEN-SUS method, the Meth ontology method and the Seven-step method. [3]Among them the Seven-step method is regarded as the most mature one.[4] It was developed by the School of Medicine in Stanford University, which includes many aspects such as area analysis, ontology combination and concept expansion.[5]

In our research we adopt the Seven-step method, since it is the comparatively perfect one which is widely used and provides specific details on operation and related technical support.

3 Steps of Constructing Music Domain Ontology

The constructing of music domain ontology is in fact a process of the conceptualizing and formalizing the musical knowledge. Following the general principles of ontology construction and the Seven-step method, [6] we carry out our project.

3.1 Determine the Domain and Scope of the Ontology

We define our work the domain ontology on music and musicology. It is a conceptualized and formalized base of all musical knowledge. In our program the whole project is composed of eight sub-ontologies: Musical Activity, Musical Work, Musical Participant, Musical Activity Location, Musical Skills and Arts, Musicological Theory, Music Equipment and Digital Music Technique.

Although these sub-ontologies are independent relatively, they are not isolated in whole, but connected to each other by varied relations. For example, a music activity may involve the musical work, the participants, the location, the equipment and the technique. These involved objects are actually attributes of a Musical Activity, and the involved sub-ontologies are bases for the value of these attribute. The relation may be diagramed as Figure 1.

3.2 Select and Reuse Existing Resources

Although we didn't find any existing music domain ontology reusable, we still have some traditional resources in printed or electronic forms for reference. They are: The Music Volume of the Encyclopedia of China (published by Encyclopedia of China Publishing House), The Concise Oxford Dictionary of Music (published by People's Music Publishing House) and The English-Chinese Dictionary of Electronic Music (published by People's Music Publishing House).

Besides, the abundant online resources, such as the Wikipedia (http://zh.wikipedia.org), the China Music Network, (http://www.yyjy.com/yybk/), the Baidu Encyclopedia (http://baike.baidu. com), etc. are also quite useful for our work.

All these resources are the important foundation for us to obtain the domain thesaurus and to define the structure of concepts.

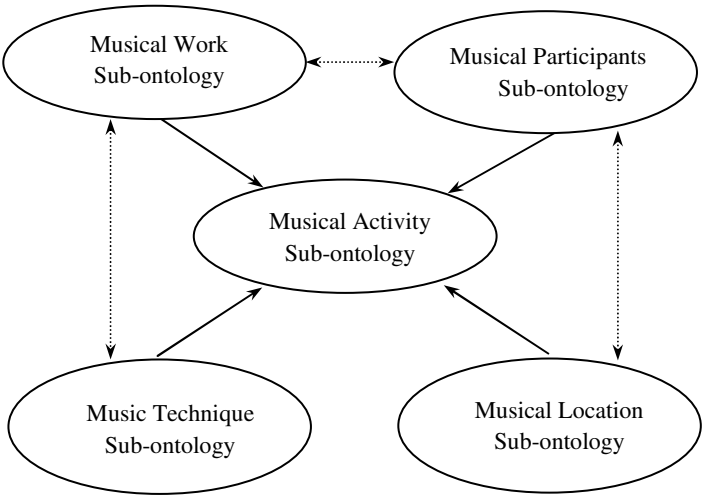


Fig. 1. Relation between sub-ontologies

3.3 Enumerate the Important Concepts in the Ontology

This is the work of recognizing the important terms in this domain so as to establish the ontology structure model. So it is important to get a comprehensive list of words without worrying about confusion.

Although there are many experimental precedents of automatic and semi-automatic constructing domain ontology, we still would like to carry out our projects by manual way so as to construct a professional high quality music domain ontology as possible as we can. Besides, without an accessible descriptor list in music domain, the automatic or semi-automatic construction cannot be satisfactory.

In the first stage of the project, we have extracted 3866 important concepts from the resources above. These concepts are accurate in meaning, they are not disjoint mutually, and they can cover almost all the knowledge in this sub-domain.

3.4 Define the Classes and the Class Hierarchy

According to Uschold and Gruninger, [7] there are three possible approaches in developing the class hierarchy:

A Top-Down development process starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts. A Bottom-Up development process starts with the definition of the most specific classes, the leaves of the hierarchy, with subsequent grouping of these classes into more general concepts. A Combination development process is a combination of the top-down and

bottom-up approaches: we define the more salient concepts firstly, and then generalize and specialize them appropriately.

Considering the stability of our model, we adopt the third one in our project to avoid the duplication of works and the inconsistency of knowledge.

It is the key to define the core classes by this method. Here we define the sets of core classes for each sub-ontology:

Musical Activity: {Music Creation, Music Performance, Music Competition, Concert, Music Instruction}

Musical Participant: {Composer, Librettist, Player}

Musical Works: {Vocal Music, Instrumental Music, Lyrics}

Musical Activity Location: {Opera House, Concert Hall, Odeum, Auditorium}

Musical Skills and Arts: {Singing Skills, Singing Posture, Singing Style, Playing Techniques}

Musicological Theory: {Basic Music Theory, Temperament, Music Criticism, Music Appreciation, Macro Musicology}

Music Equipment: {Human Voice, Musical Instrument}

Digital Music Technique: {Digital Music Hardware, Digital Music Software, Digital Music Interface, Digital Music Standard, Digital Music Format}

The hierarchal structure model for these classes is diagramed as figure 2.

3.5 Define the Relations between Classes

The relations between classes fall into three types:

Synonymy: If class A has the same meaning with class B, then they are in synonymy relation. For example, {Eroica} has the same referent as {Symphony No. 3 in E-flat major}.

Hyponymy: If class A is a class B or a kind of class B, then they are in hyponymy relation. For example, {Ludwig Van Beethoven} is a {composer}; {flute} is a kind of {aerophone}.

Meronymy: If class A is a part of class B, then they are in meronymy relation. For example, {string} is a part of {violin}, {soprano} is a part of {chorus}.

Correlativity: It means class A and class B are in some kind of correlativity relation. For example, {A Faust Symphony} is revised from {The Faust Symphony for Two Pianos}, {guitar} is developed from {lute}.

The class hierarchy model is diagramed as figure 2.

3.6 Define Slots: The Properties of Classes

Slots are the descriptors used for describing the properties of classes and instances. The hierarchical structure is just a frame for the concepts system which can not reflect the varied relations between concepts, therefore we need to define the properties of classes.

Here are some examples we designed for the classes in our music domain ontology:

Aerophone: {alias, original tone, clef, diapason, constitutes, material, characteristic, typical application}

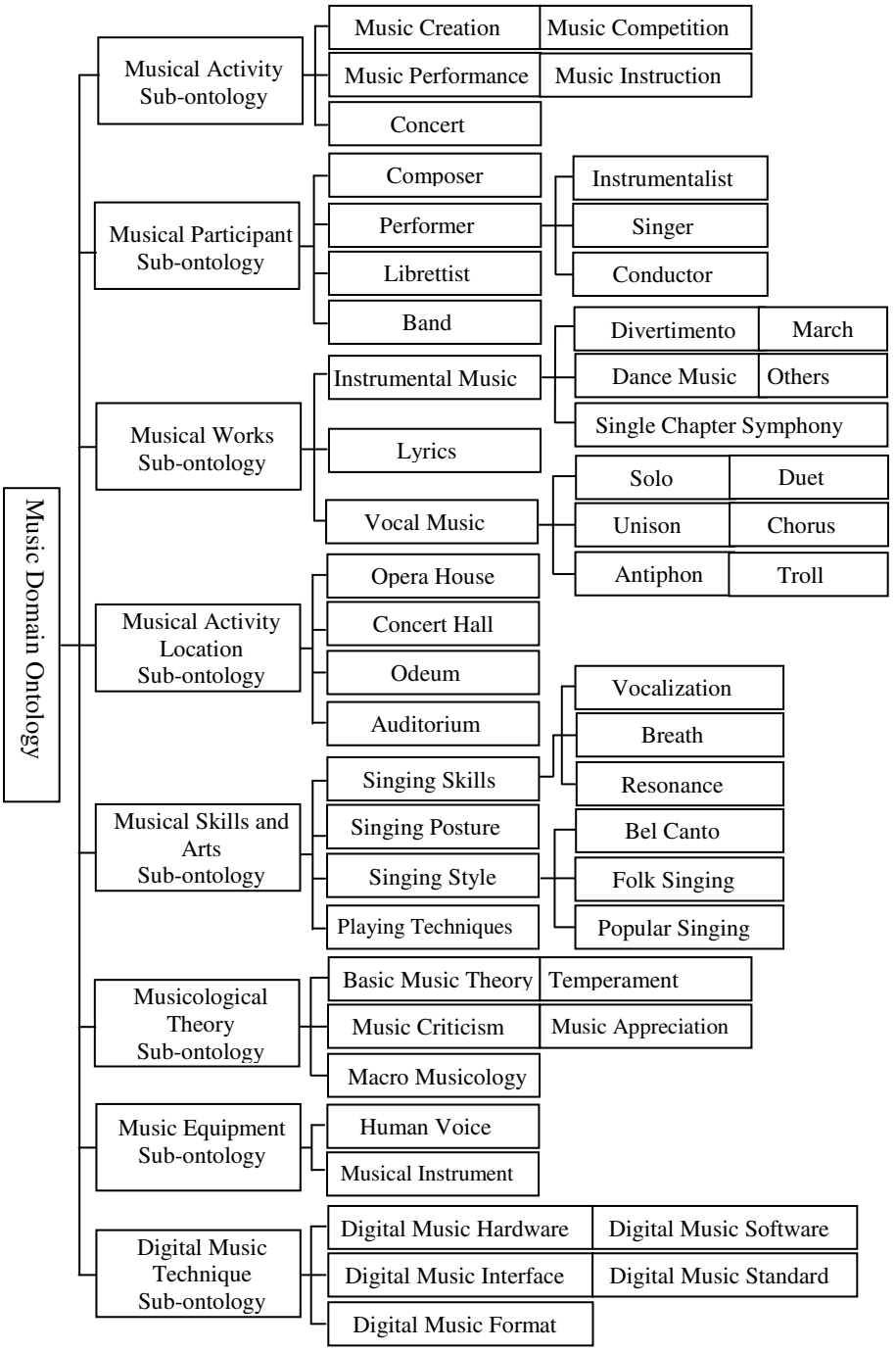


Fig. 2. The Concepts Structure Model of Music Domain Ontology

Composer: {full name, nationality, date of birth, date of death, place of birth, status, style, magnum opus}

Band: {place, age, organization, style, magnum opus, awards}

3.7 Define the Facets of the Slots

The slots in ontology may have different facets describing the value type, allowed values, the number of the values (cardinality), and other features of the values the slot can take.

For example, the value of an "alias" slot (as in "the alias of an instrument") is one string. That is, "alias" is a slot whose value type is "String". A slot "typical application" (as in "the typical application of aerophone") can have multiple values and the values are instances of the class "instrumental music". That is, "typical application" is a slot with value type "Instance".

3.8 Create Instances

With all the definition above, all the classes can be instantiation. Here are some examples of instance taken from our ontology.

eg. 1 Clarinet

```
CLASS NAME: {clarinet }
INHERITED ATTRIBUTE: {all the attributes inherited from
its superior class "Aerophone"}
ALIAS : {black pipe, speaker in the orchestra}
ORIGINIAL TONE:{bB}
CLEF:{treble clef}
DIAPASON:{from e to g3}
CONSTITUTES:{whistle head, small tube, main tube, bell
mouth, mechanical voice key system}
MATERIAL:{hard rubber, ABS plastic, phenolic resin, ebo-
ny, rosewood, mahogany, pmma}
CHARICTERISTIC:{loud and clear treble, limpid and grace-
ful median, vigorous and rich bass}
TYPICAL APPLICATION:{Clarinet Concerto by Mozart,  Rhap-
sody in Blue by George Gershwin}
```

eg. 2 Mozart

```
FULL NAME: {W.A.Wolfgang Amadeus Mozart}
NATIONALITY: {Austria}
DATE OF BIRTH: {1756-01-27}
DATE OF DEATH: {1791-12-05}
STATUS: {founder of piano concerto, the representative of
```

```
Viennese classical school}
STYLE: {Vienna classical style}
MAGNUM OPUS: {The Marriage of Figaro (opera), Don Juan
(opera), The Magic Flute (opera), Symphony No. 39 in E-
flat major, Symphony No. 40 in G minor, Symphony No. 41
"Jupiter" in C major, Violin Concerto No. 4 in D major}
```

3.9 Encode the Ontology

The last work is to encode the ontology. Up to now we have many ontology modeling tool such as Pro-tégé, Ontolingua, OntoSaurus ,WebODE and On-toEdit. Our final project is compiled by Protégé.

4 Conclusion and Future Work

This paper proposes a manual method for the formalized representing of the knowledge on music musicology on the theory of domain ontology. The method and the steps of constructing the music domain ontology are introduced, and the concept structure model is also established by defining the core class, class hierarchy, and the slots.

This is just the first stage of our whole projects. To establish the complete music domain ontology, we still need to do much more work, such as to optimize the classes and the slots, to define axioms, to ensure the consistency of the ontology. These issues will be further studied in our near future work.

Acknowledgments. This work is Sponsored by the Teaching Research Foundation of Xianning College, (Grant No. JO9130) and the Cooperation Project in Industry, Education and Research of Guangdong Province and MOE, P.R.C.(Grant No. 2010B090400170).

References

1. Gao, Y., Cao, C.G., Sui, Y.F.: Musical Domain-Specific Ontology Building and Analysis. Computer Science 31(1), 103–107 (2004)
2. Wang, F.S., Hou, L.W., Jiang, F.: Study on Establishing Domain Ontology. Information Science 23(2), 241–244 (2005)
3. Zhang, L., Huang, Y.C.: Establishment of Ontology on Crops Cultivation Domain. Journal of Library and Information Sciences in Agriculture 21(1), 68–72 (2009)
4. Li, J.: An Applied Study of Ontological Theory in Information Retrieval System (Benti Lilun Zai Wenxianjiansuo Xitongzhong De Yingyong). National Library of China Publishing House, Beijing (2005)
5. Lin, Z.F.: A Survey on the Theoretic Research of Ontological Conceptual Model Constructing. Information Research 5(139), 30–33 (2009)
6. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880 (2001)
7. Uschold, M., Gruninger, M.: Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 11(2), 93–136 (1996)

Author Index