Exploiting Lexicalized Statistical Patterns in Chinese Linguistic Analysis

Yu Zhao and Maosong Sun

Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
National Lab for Information Science and Technology,
Tsinghua University, Beijing 100084, China

Abstract. The web corpus has been used for linguistic analysis with the help of search engines. In this paper, we describe the concept of lexicalized patterns, which we exploit to obtain statistical information using the simple string matching strategy via search engines. We discuss the usage of lexicalized statistical patterns at three linguistic levels of Chinese analysis: lexical, syntactic and semantic. We develop a specialized search engine to get frequency counts for these patterns on SogouT[1] corpus. Experimental results show that lexicalized statistical patterns are effective on analyzing the cohesion of phrases, determining the phrasal category and discovering patient objects.

Keywords: Lexicalized statistical pattern, Chinese linguistic analysis, Web corpus, Natural language processing.

1 Introduction

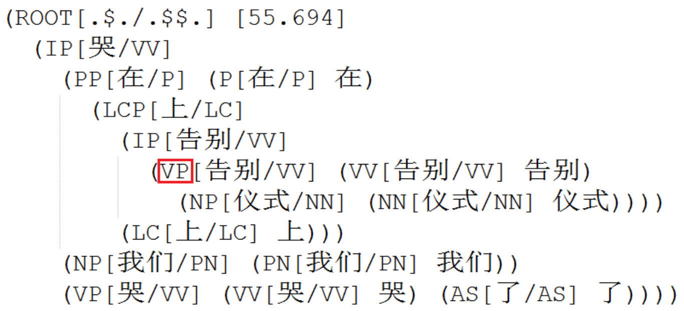
Most of current statistical natural language processing systems rely on large well-organized annotated corpus. For example, the state-of-the-art dependency parser uses Treebanks to extract POS-tag features [9]. Nevertheless, these corpora are highly time-consuming and labor-intensive to build and extend. Moreover, they are mostly in limited size. The main cause of error for many natural language processing task is the lack of related statistical information in the training set.

Let us consider the task of determining the phrasal category, which is one of the most significant issues in shallow parsing. In Chinese, a chunk that has two components: VP+NP, can possibly be a verbal phrase or a substantive phrase. For instance, 告别仪式 (farewell ceremony) and 告别朋友 (say goodbye to a friend) are both composed by VP+NP. while the former is a substantive phrase and the latter is a verbal phrase. However, the famous Stanford parser incorrectly categorizes 告别仪式 as a verbal phrase, as shown in Figure 1. Resolving this type of error requires information that is not present in Treebanks.

Therefore, a growing number of researchers have been realizing the potential of web-scale corpus to NLP tasks. The key advantage of web corpus lies in

[1] The 2008 version is available online at `http://www.sogou.com/labs/dl/t.html`

M. Sun et al. (Eds.): CCL and NLP-NABD 2013, LNAI 8202, pp. 238–246, 2013.



```
(ROOT[.$./.$$.] [55.694]
  (IP[哭/VV]
    (PP[在/P] (P[在/P] 在)
      (LCP[上/LC]
        (IP[告别/VV]
          (VP[告别/VV] (VV[告别/VV] 告别)
            (NP[仪式/NN] (NN[仪式/NN] 仪式))))
        (LC[上/LC] 上)))
    (NP[我们/PN] (PN[我们/PN] 我们))
    (VP[哭/VV] (VV[哭/VV] 哭) (AS[了/AS] 了))))
```

Fig. 1. An error occurred in the parse output of the Stanford parser. The correct tag of phrase “告别 仪式” (a farewell ceremony) should be NP instead of VP.

its massive scale, which contributes to the completeness of statistical information. The main challenge is that web corpus usually consists of huge amounts of unstructured plain-text documents. As far as we know, there is not a flexible approach for analyzing the deep structure of natural language automatically, so that the major methodology of utilizing web search engine is simply based on string-matching strategy.

In this paper, we propose to design valuable patterns for different NLP tasks and access word count statistics via the search engines. For example, we find that 告别了朋友 (have said goodbye to a friend) appears more frequently than 告别了仪式 (have said goodbye to a ceremony) by comparing the amount of results acquired from web search engine[2]. Furthermore, it indicates that for VP NP phrases, the occurrence of “VP+了+NP” structure is probably a discriminative cue to determine its phrasal category.

We denote structures like VP+了+NP as **templates**. The instantiated cases for the given template are defined as **lexicalized patterns**, such as 告别了仪式 and 告别了朋友. Our work is based on a reasonable hypothesis that the phrase frequency information collected from large corpora can reflect the correctness of phrase usage. If a lexicalized pattern frequently occurs in web-scale corpora, we would be confident that it is linguistically valid. In contrast, if a lexicalized pattern hardly appears, we would say that it is linguistically invalid. We denote those lexicalized patterns with frequency counts as **lexicalized statistical patterns**, examples of which are shown in Table 1.

Analysis of the Chinese language can be performed at different linguistic levels. In this paper, we discuss the usage of lexicalized statistical patterns on three levels of Chinese grammatical analysis. At the lexicalization level, we focus on analyzing the cohesion of compound noun phrases. At the syntactical level, we focus on determining the phrasal category. At the semantic level, we focus on discovering patient objects among the predicate-object phrases. We integrate templates and utilize valuable lexicalized statistical patterns at each level.

[2] We obtain 745,000 results for 告别了朋友, and 197 results for 告别了仪式 on Google.

Table 1. Frequency information of three lexicalized statistical patterns at different linguistic levels. Here, 兔子的尾巴 means *the tail* (尾巴) *of rabbits* (兔子) in English. 把食堂吃了 means *to eat* (吃) *the canteen* (食堂) in English while the original phrase 吃 食堂 means *eating at the canteen.*

	Lexical level	Syntactic level	Semantic level
Phrase	兔子(NP_1) 尾巴(NP_2)	告别(VP) 朋友(NP)	吃(VP) 食堂(NP)
Template	NP_1+的+NP_2	VP+了+NP	把+NP+VP+了
Pattern	兔子的尾巴	告别了朋友	把食堂吃了
Frequency	272	8460	0

Recent works have shown that exploiting web-scale corpus is an effective way to enhance the performance of NLP systems. Volk used frequencies counts of query patterns to resolve PP attachment ambiguities via web search engine [5]. Lapata and Keller used web counts to resolve preposition attachments and compound noun interpretation [3][4]. Bansal used Google n-grams to generate full range of syntactic attachments [1]. Moreover, contextual statistics collected from web have also significantly improved the quality of automatic thesaurus extraction [6]. Yuan designed a series of rule-based scoring method for word categorization [7][8]. It was proven to be effective for validating parts of speech of Chinese words and categorizing phrases. Nevertheless, the principle of judging whether a word or a phrase follows a rule is completely based on decisions of linguistic experts. It requires significant amount of manual effort, which to a great extent weakens the scalability of patterns.

In section 2, we introduce a specialized search engine for lexicalized statistical patterns to conveniently acquire frequency information from any plain-text datasets. In this work, we experimentally collected Chinese plain-text sentences from the SogouT corpus. In section 3, we present some useful templates for Chinese analysis at different linguistic levels. For each level, the results of our case study are demonstrated respectively. Finally, we provide conclusion in Section 4.

2 Search Engine for Lexicalized Patterns

A naive way to obtain frequencies from web-scale corpus is to directly query from traditional web search engine. For example, if we query the phrase “告别朋友” using Google, it will return about 37,200 results. However, there are three main disadvantages of traditional search engine:

- There are duplicate pages and spam sites in the web environment. It will make the frequency counts unreliable.
- Web search engines ignore stop words and punctuation in general. But under some conditions we need these features to guarantee the quality of results.
- The users have no way to perform complex type of queries, such as near query, wildcards query or slop query.

For the example mentioned above, results from traditional search engine contain noises like “这是一场无声的告别，朋友们再见!” (It’s a silent farewell. Goodbye friends!). The solution to this problem is to complement a period or a question mark behind the query, such as “告别朋友.” and “告别朋友?”. Moreover, we hope to permit the punctuation and query to be separated by no more than two words, like “告别朋友吧!” or “告别朋友了吗?”. To our knowledge, we are not aware of any web search engine that can deal with such kind of queries.

Hence, we manage to develop a flexible search engine to dig up lexicalized patterns more accurately. In fact, a similar product known as the Sketch Engine.[3] has been published by Lexical Computing Ltd. It can deal with complex types of word queries, but it is not capable of handling Chinese documents without word segmentation.

In this paper, we propose to make use of Apache Solr[4], which is an open source enterprise search platform. Solr provides document indexing APIs and support term proximity with slop factors. We implement a query builder to support formalized complex queries based on Solr. The unified regular expression of query is described as follow:

$$\mathrm{W(RW)^*(RE)?} \tag{1}$$

where W represents a candidate word list, R represents a range of wildcard gaps, and E represents a set of punctuation at the end of sentences. For example, one possible query to describe the pattern 告别朋友 can be written as:

$$\{\text{告别}\}<0\text{-}1>\{\text{朋友}\}<0\text{-}2>E \tag{2}$$

where “告别朋友吧!” and “告别朋友了吗?” both match this query.

In order to get the Chinese Web text corpus, we extract over 2.1 billion sentences from SogouT web page dataset, removing unnecessary html tags, hyperlinks, scripts and independent anchor texts. We take the number of matched sentences as the word count result for corresponding query pattern. To make sure the word count of our sketch search engine is reliable, we eliminate duplicates and short sentences (no longer than 2). We eventually index 729,008,561 unique sentences with a cost of 165.2 GB free disk space. The lexicalized search engine can handle all quires in the format described in (1), with an average of 10 seconds query response time. Three examples of word count statistics are demonstrated in Table 1.

3 Linguistical Analysis for Chinese Phrases

In this section, we analyze Chinese phrases at three linguistic levels: lexical, syntactic and semantic. The basic idea is to integrate handcrafted templates and then automatically acquire phrase counts via the search engine. For each level, we present the result of our case study on SogouT corpus respectively.

[3] http://www.sketchengine.co.uk/

[4] http://lucene.apache.org/solr/

3.1 Lexical Level

We focus on analyzing the cohesion of a compound noun phrase at lexical level. Phrases with low cohesion should not be included into vocabulary. For instance, 兔子尾巴 (rabbit tail) has low cohesion, while 圆桌会议 (round table) has high cohesion. Mutual information is commonly used to measure the cohesion of a phrase, but the boundary value between high and low cohesion is often fuzzy. For example, if we use SogouT to get frequency counts, the point-wise mutual information of 兔子 (rabbit) and 尾巴 (tail) is：

$$\begin{aligned} mi(\text{兔子}, \text{尾巴}) &= \log_2 \frac{N \cdot \text{Count}(\text{兔子尾巴})}{\text{Count}(\text{兔子}) \cdot \text{Count}(\text{尾巴})} \\ &= \log_2 \frac{729008561 \cdot 562}{171325 \cdot 193158} \\ &= 3.63 \end{aligned} \tag{3}$$

The result indicates that the mutual information of 兔子尾巴 is not relatively low. It is not enough to prove that this phrase has low cohesion.

In linguistic perspective, an important clue of high cohesion phrase is that the component words can hardly be separated by particles. Hence, we construct a template “NP_1+的+NP_2” to determine whether a compound noun phrase “NP_1 NP_2” has low cohesion. The statistic of lexicalized patterns are provided in the following:

Count(兔子尾巴)= 562	Count(兔子的尾巴)= 272
Count(圆桌会议)= 6895	Count(圆桌的会议)= 8

The result shows that 兔子的尾巴 (the tail of rabbit) and 兔子尾巴 has a similar amount of occurrences, while 圆桌的会议 (the conference of round table) is significantly less frequent than 圆桌会议. In fact, 圆桌的会议 hardly appears in SogouT corpus. It indicates that the occurrence ratio of lexicalized patterns of NP_1+的+NP_2 is more effective than mutual information.

3.2 Syntactical Level

A typical problem at syntactic level is to determine the phrasal category of “VP NP” phrases. For example, 告别仪式 is a substantive phrase, while 告别朋友 is a verbal phrase. These two phrases share the same verb 告别 (say goodbye to). We can give many other examples, such as 修理公司 (repair company) and 修理自行车 (repair the bicycle), 学习小组 (study group) and 学习英语 (study English), 购买需求 (purchasing demand) and 购买基金 (purchase funds), etc.

Yuan (2010) distinguished five structural categories of compound phrases by designing handcrafted rules [8]. Inspired by these rules, we construct a set of templates for “VP NP” phrases via frequently-used auxiliaries and conjunctions in Chinese. For simplicity, we only take two-word compound phrase into account. We enumerate 14 templates in Table 2.

Table 2. Templates for determining the category of VP NP (do sth.) phrases

	Template	Translation	Example phrase
1	VP+了/着/过+NP	have done sth.	告别了朋友
2	不+VP+NP	do not do sth.	不告别朋友
3	VP+完/掉+NP	after doing sth.	告别完朋友
4	所+VP+的+NP	sth. done	所告别的朋友
5	把+NP+VP	to do sth.	把朋友告别
6	被+VP+的+NP	sth. that are done	被告别的朋友
7	为/拿+NP+VP	do for sth.	为朋友告别
8	VP+一+(量词)+NP	do a sth.	告别一个朋友
9	放着+NP+不+VP	sth. but not done	放着朋友不告别
10	VP+什么+NP	what sth. to do	告别什么朋友
11	NP+越+VP+越	the more one do sth.	朋友越告别越
12	连+NP+都/也+VP	don't even do sth.	连朋友都不告别
13	忙着+VP+NP	be busy doing sth.	忙着告别朋友
14	为什么+VP+NP	why one do sth.	为什么告别朋友

Given a "VP NP" phrase p, a lexicalized pattern $l(p,t)$ can be generated for each template t. If obtain the frequency counts via the search engine for lexicalized patterns, the phrase category of p is determined by the relative frequency $F(p)$, which is derived as follow:

$$F(p) = \frac{\sum_{t \in T} \mathrm{Count}(l(p,t))}{\mathrm{Count}(p)} \tag{4}$$

where $\mathrm{Count}(l(p,t))$ represents the frequency count of each lexicalized pattern and $\mathrm{Count}(p)$ represents the total count of phrase p. As we can see, the higher value of $F(p)$, the more likely the phrase p is to be verbal phrase. A simple validation algorithm is to set a lower threshold θ for verbal phrases, where θ may vary for different corpus.

To approximate the value of θ, we perform a case study on SogouT corpus. We collect 20 example phrases, half of which are verbal. The frequency results are demonstrated in Table 3. We find that the average phrase frequency of verbal phrases is only 60% higher than that of substantive phrases, while the average relative frequency $F(p)$ of verbal phrases is 52.3 times higher than that of substantive phrases. It indicates that the lexicalized patterns are effective to distinguish the two categories of phrases. Since the minimum $F(p)$ score among verbal phrases is 0.049 and the maximum $F(p)$ among substantive phrases is 0.006, the appropriate threshold θ for SogouT corpus can be set within the range of $(0.006, 0.049)$.

Table 3. Frequency results of lexicalized patterns for verbal and substantive phrases. The translations of all these phrases are given in Appendix.

Category	Phrase	Phrase frequency	Pattern frequency	$F(p)$
Verbal	告别朋友	116	99	0.853
	购买基金	9508	597	0.063
	修理自行车	752	40	0.053
	学习英语	27763	1361	0.049
	治疗病人	1453	334	0.230
	救济灾民	897	13	0.014
	判罚点球	5870	1373	0.234
	维护秩序	3972	173	0.044
	攻击敌人	9337	357	0.038
	更新系统	1504	147	0.098
	Average	**5617.2**	**449.4**	**0.080**
Substantive	告别仪式	8460	1	0.0001
	购买需求	4750	5	0.0001
	修理公司	1167	7	0.005
	学习小组	5079	23	0.005
	治疗手段	11273	8	0.0001
	救济中心	148	0	0.000
	判罚尺度	1222	1	0.0001
	维护工具	1107	0	0.000
	攻击计划	704	2	0.003
	更新日期	832	5	0.006
	Average	**3474.2**	**5.2**	**0.0015**

3.3 Semantic Level

At the semantic level, we focus on discovering the patient object, which is a nominal phrase that acts as the recipient of the action stated by a verbal phrase. It is a significant semantic relation between nominal and verbal phrases. For example, 晚饭 (dinner) is the patient of verb 吃 (eat) in the phrase 吃晚饭 (having dinner), while 食堂 (canteen) is not the patient of 吃 (eat) in the phrase 吃食堂 (eating at the canteen).

Given a predicate-object phrase "VP NP", the task of identifying the patient object usually rely on the judgement of linguists. In this paper, we propose to utilize a set of templates for this task, such as "把+NP+VP+了", "所+VP+的+NP" and "被+VP+的+NP". If the nominal phrase occur to be the patient object, the lexicalized pattern for this template tends to occur frequently in web-scale corpus. For example, 把晚饭吃了 (to have dinner) appears 63 times in SogouT, while 把食堂吃了 (to eat the canteen) has never appeared. Here, we present the frequencies of 8 examples in SogouT to verify the reliability of our template.

Patient objects:

Count(吃晚饭)= 16608	Count(把晚饭吃了)= 63
Count (炒菜) = 31473	Count (把菜炒了) = 18
Count(洗衣服)= 50989	Count(把衣服洗了)= 353
Count(割阑尾)= 157	Count(把阑尾割了)= 4

Non-patient objects:

Count(吃食堂)= 1798	Count(把食堂吃了)= 0
Count(去北京)= 6895	Count(把北京去了)= 0
Count(想主意)= 919	Count(把主意想了)= 0
Count(筹经费)= 4377	Count(把经费筹了)= 8

where the translations of these phrases are given in Appendix.

As we can see, for patient object, the frequency of lexicalized pattern is proportional to the frequency of the VP+NP phrase. For non-patient object, the frequency of lexicalized pattern is almost always zero. It indicates that our template is effective for discovering patient objects among VP+NP phrases.

4 Conclusion

We exploit lexicalized statistical patterns collected from the web corpus at three linguistic levels of Chinese analysis. Results of our case study indicates that these patterns are effective on analyzing the cohesion of phrases, determining the phrasal category and discovering patient objects.

Our automatic categorization of phrasal category contributes to reduce the workloads of linguistics [7][8]. The search engine for lexicalized patterns can also be used for verifying the effectiveness of batched and hand-crafted linguistic rules. In the future we aim at integrating the lexicalized statistical patterns as feature templates to enhance the precision of Chinese parser.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 61133012) and the National High Tech. Development Program of China (863 Program) (Grant No. 2012AA011102).

References

1. Bansal, M., Klein, D.: Web-scale features for full-scale parsing. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (2011)
2. Curran, J.R., Moens, M.: Scaling context space. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA (2002)
3. Keller, F., Lapata, M., Ourioupina, O.: Using the web to overcome data sparseness. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia (2002)

4. Lapata, M., Keller, F.: The Web as a baseline: Evaluating the performance of unsupervised Web-based models for a range of NLP tasks. In: Proceedings of HLT-NAACL (2004)
5. Volk, M.: Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In: Proceedings of the Corpus Linguistics 2001 Conference, Lancaster, UK, pp. 601–606 (2001)
6. Yates, A., Schoenmackers, S., Etzioni, O.: Detecting parser errors using web-based semantic filters. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2006)
7. Yuan, Y.: A Cognitive Investigation and Fuzzy Classification of Word-class in Mandarin Chinese. Shanghai Educational Publising House (2009)
8. Yuan, Y.: 汉语词类划分手册, Beijing Language and Culture University Press (2010)
9. Zhang, Y., Clark, S.: Syntactic processing using the generalized perceptron and beam search. Computational Linguistics 37(1), 105–151 (2011)

Appendix

Here we present the English glosses of some Chinese phrases in this paper to make sure one has a clearer understanding of their meanings.

告别朋友	to farewell a friend	告别仪式	the farewell ceremony
购买基金	to purchase funds	购买需求	the purchasing need
修理自行车	to repair the bicycle	修理公司	the repair company
学习英语	to study English	学习小组	the study group
治疗病人	to cure patients	治疗手段	the treatment
救济灾民	to relieve the victims	救济中心	the relief center
判罚点球	to give a penalty	判罚尺度	the principle of decision
维护秩序	to maintain order	维护工具	the maintenance tool
攻击敌人	to attack the enemy	攻击计划	the attacking plan
更新系统	to update the system	更新日期	the update date
吃晚饭	have dinner	炒菜	cook dishes
洗衣服	wash clothes	割阑尾	cut the appendix
吃食堂	eat at the canteen	去北京	go to Beijing
想主意	think of ideas	筹经费	raise money