Chapter 27
A Study on the Semantic Interpretation of Chinese Noun Compounds

Meng Wang and Lulu Wang

Abstract Noun compound interpretation is determined by implicit semantic relations encoded by constituent nouns. In this chapter, we will present preliminary research on the interpretation of Chinese noun compounds (NCs) using two different strategies. For the abstract relation strategy, we proposed a novel taxonomy of Chinese noun compounds and a method for interpreting Chinese NCs based on word similarity. For the verbal paraphrasing strategy, we proposed the simple dynamic approach of using paraphrasing verbs, which not only provided possible interpretations of noun compounds but also captured the subtle semantic differences of similar NCs. Both strategies can be applied in other fields, such as question answering, information retrieval, and lexicography.

Keywords Chinese noun compounds · Interpretation · Semantics

27.1 Introduction

A noun compound (NC) is a sequence of two or more nouns that functions as a single noun (Downing 1977). The use of NCs is very frequent in English-written text, including press and technical materials, newswires, and fictional prose. In other languages, such as Chinese, NCs are also abundant in texts since the compounding of nouns is the most common way of naming new things. The syntax and semantics of noun compounds has remained an active research field in linguistics, which includes the broader research of multiword expressions (MWEs). As a well-established subtask of language understanding, the interpretation of noun compounds involves uncovering the underlying semantic relations encoded by

M. Wang (✉)
School of Humanities, Jiangnan University, Wuxi, China

L. Wang
Department of Linguistics, Communication University of China, Beijing, China
e-mail: lulu.wang@cuc.edu.cn

C.-R. Huang et al. (eds.), *Chinese Language Resources*, Text, Speech and Language Technology 49, https://doi.org/10.1007/978-3-031-38913-9_27

constituent nouns. For example, 爱情故事 *aiqing gushi* "love story" can be illustrated as 讲述爱情的故事 *jiangshu aiqing de gushi* "a story that tells about love" and 别墅女人 *bieshu nvren* "villa woman" means 住在别墅的女人 *zhuzai bieshu de nvren* "a woman living in a villa." Understanding the semantic relations between noun compounds is helpful for many tasks, such as machine translation, information retrieval, and question answering, among others.

In this chapter, we will focus on the semantic interpretation of Chinese noun compounds. The remainder is organized as follows: Section 27.2 will describe related work, while Sect. 27.3 will present a novel taxonomy of Chinese noun compounds based on the transparency of the compounds. In Sect. 27.4, a method for predicting the semantic relations of novel NCs based on word similarity will be introduced. Section 27.5 will illustrate how to interpret noun compounds using verbal paraphrasing, while Sect. 27.6 will offer the conclusion and future work.

27.2 Previous Studies

In theoretical linguistics, there are contradictory views regarding the semantic interpretation of NCs. Most linguists describe the semantics of noun compounds via a set of abstract relations, as represented in the work of Levi (1978), who presented nine recoverable deletable predicates (RDPs)—be, cause, have, make, use, about, for, from, and in—that are universal and primitive in generating noun compounds, and Warren (1978), who proposed a four-level hierarchical taxonomy derived from the Brown Corpus. Following this tradition, some scholars in the computational field have focused on the taxonomies of noun compounds. Ó Séaghdha (2007) proposed six semantic relations—BE, HAVE, IN, ACTOR, INST(-RUMENT), and ABOUT—and each relation was subdivided into subcategories. For example, HAVE is subdivided into the possession, condition-experiencer, property-object, part-whole, and group-member subcategories. Tratz and Hovy (2010) presented a large, fine-grained taxonomy of 43 noun compound relations, which were notably tested by Amazon's Mechanical Turk service. However, there is still no consensus as to which set of relations binds nouns in a noun compound.

Overall, the semantic relations proposed by different scholars have ranged from general to more specific, with the general ones aiming for broad-coverage analysis of unrestricted text and the specific ones aiming for specialized applications in some domains. In this line of research, the semantic interpretation of NCs is viewed as a multiclass classification problem, where the predefined semantic relations are the categories to be assigned. However, the approach of abstract relations is problematic in several ways. As Nakov and Hearst (2013) pointed out, it is unclear which relation inventory is best, as relations capture only part of the semantics and multiple relations are possible. For example, Wei (2012) assumed that 中国电影 *zhongguo dianying* "Chinese movies" is classified into the categories of LOCATION and CONTENT.

Considering these drawbacks, other researchers have used verbal paraphrasing to interpret noun compounds (Girju et al. 2005; Nakov and Hearst 2006; Nakov 2008;

Table 27.1 Levi's (1978) transparency scale for noun compounds

	Types	Examples
a	Transparent	*Orange peel*
b	Partly opaque	*Grammar school*
c	Exocentric	*Ladybird*
d	Partly idiomatic	*Flea market*
e	Completely idiomatic	*Honeymoon*

Ó Séaghdha 2008). Finn (1980) interpreted "salt water" with "dissolved in." Butnariu and Veale (2008) summarized eight relational possibilities, for example, "headache pill" might be paraphrased as "headache-inducing pill," "headache prevention pill," "pill for treating headaches," "pill that causes headaches," "pill that is prescribed for headaches," and "pill that prevents headaches." With these verbs, the paraphrases are more specific than that of the abstract relations. Following this view, the SemEval 2010 task 9 "Noun Compound Interpretation Using Paraphrasing Verbs and Prepositions" and SemEval 2013 task 4 "Free Paraphrases of Noun Compounds" both intended to promote a paraphrase-based approach to this problem.

Accordingly, there are two ways to interpret noun compounds in Chinese. Theoretically, there have been some achievements in the analysis of semantic relations, while very little work on the automatic semantic interpretation of Chinese NCs has been done. Zhao et al. (2007) focused on a subset of Chinese NCs in which the head word is a verb nominalization, such as 血液循环 *xueye xunhuan* "blood circulation," and four coarse-grained semantic roles were proposed for the classification of noun modifiers in compound nominalization. Our study took a static approach in which the interpretation was viewed as a classification problem. As for the second line of research, Wang (2010) and Wang et al. (2014) adopted a bottom-up strategy to capture the verbs of noun compounds and provided four types of paraphrase patterns. As Wei (2012) pointed out, these four types are not specific enough to give proper interpretations. Instead, Wei (2012) classified the noun compounds into eight major types and 346 subcategories, which proved to be fine-grained.

27.3 Taxonomy of Chinese Noun Compounds

Whether using abstract relations or verbal paraphrasing, there are still some noun compounds that are not interpretable. We hypothesized that this is due to the lack of consideration of the decomposable possibilities and the semantic transparency of noun compounds. Taking the noun compound 夫妻肺片 *fuqi feipian* "pork lungs in chili sauce" as an example, it is not decomposable; that is, the meaning of the compound is not simply the combination of the literal meanings of the parts. Levi (1978) proposed a transparency scale for noun compounds, as shown in Table 27.1.

In Table 27.1, Levi (1978) summarized five types of noun compounds based on semantic transparency, each type showing a different interpretation pattern of the

Table 27.2 Basic types of noun compounds

	Transparency scale	Examples
a	Transparent	机组人员
		jizu renyuan
		"crew member"
b	Partly opaque	钻石戒指
		zuanshi jiezhi
		"diamond ring"
c	Partly idiomatic	试管婴儿
		shiguan yinger
		"test tube baby"
d	Completely idiomatic	夫妻肺片
		fuqi feipian
		"the spouse pork lung"

noun compounds. For example, "orange peel" is simply the combination of "orange" and "peel." However, "grammar school" cannot be combined literally because there is a hidden verb in this compound, as in "grammar teaching school." In contrast, the other types cannot be combined literally or be interpreted by hidden verbs. For instance, "ladybird" is not a kind of bird but a kind of bug, "Coccinellidae,"[1] and "honeymoon" has nothing to do with "honey" or "moon" but instead refers to the vacation that brides and grooms take to celebrate their marriage. The type "partly idiomatic" is special because it is partly idiomatic that verbs are not easy to recover. For example, it is not acceptable to say "flea selling market" for the market selling small commodities.

In light of Levi's (1978) transparence scale and Nunberg et al.'s (1994) claims on idioms, we collected 428 noun-noun compounds (N1-N2) and classified them into the following four categories shown in Table 27.2.

As Table 27.2 shows, the first three types are decomposable at the syntagmatic level, but the last one is non-decomposable. Initially, we decided that non-decomposable idioms should be analyzed as a whole unit both syntactically and semantically, and since the other types were decomposable, they could be divided into N1 and N2. However, the semantic relations of these types are different in terms of semantic transparency. Therefore, we proposed a novel taxonomy of Chinese noun compounds based on semantic transparency. Table 27.3 summarizes 11 subcategories of noun compounds based on their semantic relations.

To interpret the noun compounds in Table 27.3, we created different interpretation patterns with different conditions. Category 1 corresponds to the noun compounds of type a in Table 27.2, which can be interpreted as the literal meanings of the parts, for example, 机组人员 *jizu renyuan* "crew members" in the paraphrased 属于机组的人员 *shuyu jizu de renyuan* "the members that belong to the crew."

[1] Here, "lady" refers to the "Virgin Mary"; see more at http://www.hkhk.edu.ee/nature/ladybird_legends.html

Table 27.3 Semantic relations of NCs

	Semantic relations	Interpretation patterns	Examples
1	Possessive	N2 belongs to N1	机组人员
			jizu renyuan
			"crew member"
2	Property	N2's property is N1	股份制企业
			gufenzhi qiye
			"joint stock company"
3	Locative	N2 is located in N1	印尼火山
			yinni huoshan
			"Indonesia volcano"
4	Time	N2 is made in N1	清代家具
			qingdai jiaju
			"Qing Dynasty furniture"
5	Content	N2 is about N1	爱情故事
			aiqing gushi
			"love story"
6	Material	N2 is made of N1	钻石戒指
			zuanshi jiezhi
			"iamond ring"
7	Patient	V-N1-N2	围棋高手
			weiqi gaoshou
			"Chess master"
8	Actor	N1-V-N2	教委文件
			jiaowei wenjian
			"the board of education document"
9	Cause	N1 causes N2	考试焦虑
			kaoshi jiaolv
			"exam anxiety"
10	Partly idiomatic	Metaphoric or metonymic meaning of N1	试管婴儿
			shiguan yinger
			"test tube baby"
11	Idiomatic	Idiomatic meaning of N1-N2	夫妻肺片
			fuqi feipian
			"pork lungs in chili sauce"

In categories 2 to 5, these four types correspond to both type a and type b, since the meaning of the compounds can be interpreted by the fixed pattern of the components and can also be predicted by hidden verbs. For instance, the paraphrased verb of the compound 雅典奥运会 *yadian aoyunhui* "Athens Olympics" could be 举办 *juban* "to hold," and thus the paraphrased sentence would be 在雅典举办的奥运会 *zai yadian juban de aoyunhui* "The Olympic Games that were held in Athens." As for 爱情故事 *aiqing gushi* "love story," it could be paraphrased as 关于爱情的

故事 *guanyu aiqing de gushi* "the story about love" and 讲述爱情的故事 *jiangshu aiqing de gushi* "the story telling about love."

Moreover, categories 6 to 9 correspond to type b, in which the hidden verb must be revealed. In this group, the qualia roles of the head noun are different for each type. For example, the qualia role in category 6 is AGE because "material" usually relates to the MAKE relation, and the relation of "patient" in category 7 relates more with TELIC roles,[2] which are interpreted as the function of N1. For example, 围棋高手 *weiqi gaoshou* "chess master" could be paraphrased as 下围棋的高手 *xia weiqi de gaoshou* "the masters of playing chess." Here, 下 *xia* "to play" is the TELIC role of 围棋 *weiqi* "chess."

The last two categories correspond to type c and the non-decomposable idioms separately. Noun compounds in category 10 should be interpreted as having a metaphoric meaning, and thus they cannot be interpreted by hidden verbs. Taking 试管婴儿 *shiguan yinger* "test tube babies" as an example, the compound cannot be illustrated using expressions like 在试管里 孕育的婴儿 *zai shiguan li yunyu de yinger* "the babies that are fertilized in test tubes." The word 试管 *shiguan* "test tubes" has the metonymic meaning of 试管孕育技术 *shiguan yunyu jishu* "in glass fertilization." Therefore, the metaphoric meaning of the compound needs to infer 用试管技术孕育的婴儿 *yong shiguan jishu yunyu de yinger* "the babies that are fertilized by the technique of using test tubes." For these types of idioms, they are not decomposable at all and should be treated as a whole unit. For example, 夫妻肺片 *fuqi feipian* "pork lungs in chili sauce" refers only to the name of the dish.

27.4 Interpretation Based on Word Similarity

Kim and Baldwin (2005) introduced a method for interpreting novel English noun compounds with semantic relations using WordNet: Similarity. Based on the taxonomy above, we proposed a method using word similarity to predict the semantic relations of novel Chinese NCs. Given an NC in the testing data, we calculated the similarities between the correspondence nouns in the training data to acquire the semantic relation, which was our first strategy.

27.4.1 Word Similarity Measures

HowNet-based similarity. HowNet is a commonsense knowledge base of interconceptual relations and inter-attribute relations of concepts as connoted in lexicons of Chinese and their English equivalents (Dong and Dong 2005). As a knowledge base, the knowledge structured by HowNet is represented by a graph

[2] Pustejovsky (1995) proposed four qualia roles of nouns: formal, constitutive, agentive, and telic.

Fig. 27.1 Definition of the Chinese word 学校 "school" in HowNet

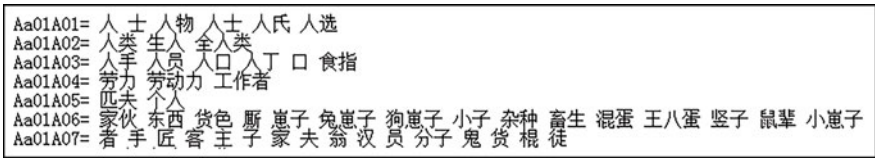

Fig. 27.2 Examples in *Cilin*

rather than a tree, and it is devoted to demonstrating the general and specific properties of concepts. For every word sense c_i (i.e., concept), its definition is composed of a set of sememes and corresponding relations. For instance, the Chinese word 学校 "school" is defined as follows in Fig. 27.1.

HowNet allows users to measure the semantic similarity and relatedness between a pair of two concepts based on the overlapping of sememes. In our study, we adopted a similarity measure provided by Liu and Li (2002) to achieve the similarity of two nouns.

***Cilin*-based similarity.** *Cilin* is a Chinese thesaurus that defines and describes "concepts" and reveals their relations using Synset. The semantic category of words (i.e., concepts) is encoded by a five-layer tree, as shown in Fig. 27.2.

The similarity of two words in *Cilin* is measured by the distance in the tree. Formally, it is defined using Formula (27.1):

$$\mathrm{sim}_{\mathrm{cilin}}(w_1, w_2) = 1 - \frac{\mathrm{pathlen}(w_1, w_2)}{\mathrm{pathlen}(w_1, \mathrm{Root}) + \mathrm{pathlen}(w_2, \mathrm{Root})} \tag{27.1}$$

where $\mathrm{pathlen}(w_1, w_2)$ is the minimum path length of (w_1, w_2) to their common parent node, and Root represents the root of the tree.

27.4.2 Method

The similarity between NCs (t_1, t_2) and (n_1, n_2) was calculated by the similarities of the component nouns. Formally, the similarity of each NC pair was defined using Formula (27.2):

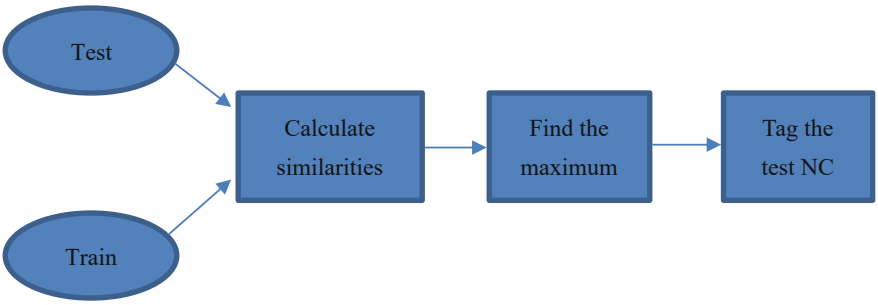

Fig. 27.3 The procedure of our method

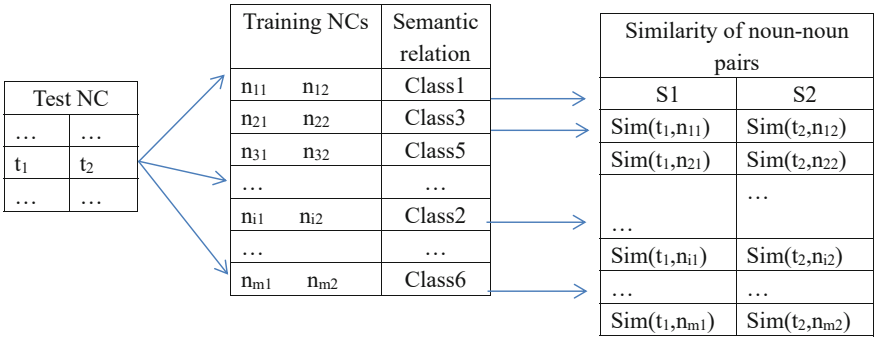

Fig. 27.4 Detailed similarities between the test NC and training NCs

$$\mathrm{Sim}((t_1, t_2)\ (n_1, n_2)) = \frac{(\alpha \mathrm{S1} + \mathrm{S1}) \times ((1-\alpha)\mathrm{S2} + \mathrm{S2})}{2} \tag{27.2}$$

where S1 is the modifier similarity (i.e., Sim(t_1, n_1)) and S2 is the head similarity (i.e., Sim(t_2, n_2)), while $\alpha \in [0, 1]$ is the weighting factor that balances the contributions of the modifier and head.

For each test NC, we calculated the similarities of all NCs in the training data. Then, we chose the NC in the training data that had the highest similarity and labeled it the test NC according to the sematic relation associated with that training data. Formally, the semantic relation of the test NC (t_1, t_2) was determined using Formula (27.3):

$$\begin{aligned} &\mathrm{Relation}\ (t_1, t_2) = \mathrm{Relation}\ (n_{i1}, n_{i2}), \text{where } i \\ &= \underbrace{\mathrm{argmax}}_{i}\ \mathrm{Sim}\ ((t_1, t_2), (n_{i1}, n_{i2})) \end{aligned} \tag{27.3}$$

Figure 27.3 shows the complete procedure of our method, while Fig. 27.4 illustrates in detail how we calculated the similarities between a test NC (t_1, t_2) and the NCs in the training data.

As can be seen, the test NC is associated with a total number of "m" similarities, where "m" is the number of NCs in the training data. Then, the semantic relation of the test NC was determined by the training instance with the highest similarity.

27.4.3 Experiments and Evaluation

We retrieved two-word Chinese NCs from the *People's Daily* of 1998 and 2000, which were segmented and POS tagged (Yu et al. 2002). After excluding proper nouns and coordinate constructions, we obtained 1483 NCs for our experiment. The semantic relations of all the NCs were judged by two annotators who had majored in linguistics. Overall, we used 978 NCs for the training data and 505 NCs for the testing data.

We experimented with the two similarity methods introduced above, assuming that the contribution of the head and modifier noun was equal ($\alpha = 0.5$). Table 27.4 shows the experimental results. Note that the HowNet and *Cilin* similarities were based on dictionary-based methods. Thus, if the test word did not appear in HowNet or *Cilin*, our method could not tag the test NC (i.e., unlabeled data) because of the lack of similarities. The performances of HowNet and *Cilin* similarity were very close, and they each classified 35% of the NCs correctly.

Table 27.5 lists some test NCs and the most similar NCs found in the training data. As can be seen, our method provided reasonable interpretations, which is very useful in understanding novel NCs. For instance, if a reader did not know the meaning of the novel NC 网络医生 *wangluo yisheng* "network doctor," our method provided NCs such as 出租车司机 *chuzuche siji* "taxi driver," which were easy to understand. Our method could also help a reader to predict the semantic relation of two nouns. Taking 布料玩具 *buliao wanju* "cloth toy" and 黄金首饰 *huangjin shoushi* "gold treasury" as an example, they both shared the same semantic relation of "material," and thus their similarity was very high, so with our method, a reader could learn the semantic relation of the former and the unfamiliar relation of the latter, as well as the more frequently used relation.

27.5 Interpretation Using Verbal Paraphrasing

In linguistic theories, it has been proven that verbs play an important role in the process of noun compound derivation. In this section, we will present a simple and unsupervised approach for characterizing the semantic relations held in two-word

Table 27.4 Accuracy based on HowNet and *Cilin* similarity

Similarity measure	Unlabeled	# Correct (accuracy)
HowNet	25	174 (34.46%)
Cilin	16	178 (35.25%)

Table 27.5 The most similar NCs based on the two similarity measures

	The most similar NCs in the training data	
Test NCs	HowNet similarity	*Cilin* similarity
残疾儿童	白内障患者	白内障患者
canji ertong	*baineizhang huanzhe*	*baineizhang huanzhe*
"disabled children"	"cataract patient"	"cataract patient"
玻璃茶几	水晶花瓶	钻石戒指
boli chaji	*shuijing huaping*	*zuanshi jiezhi*
"glass table"	"crystal vase"	"diamond ring"
网络医生	因特网用户	出租车司机
wangluo yisheng	*yintewang yonghu*	*chuzuche siji*
"network doctor"	"Internet user"	"taxi driver"
蔬菜收入	水果价格	水果价格
shucai shouru	*shuiguo jiage*	*shuiguo jiage*
"vegetable income"	"fruit price"	"fruit price"
大学校长	中学教师	政府领导
daxue xiaozhang	*zhongxue jiaoshi*	*zhengfu lingdao*
"university president"	"middle school teacher"	"government leader"
布料玩具	黄金首饰	冰秋千
buliao wanju	*huangjin shoushi*	*bing qiuqian*
"cloth toy"	"gold treasury"	"ice swing"

Chinese NCs. What is especially novel about this approach is that NCs are interpreted in terms of verbal phrases, rather than by a set of concrete verbs. This is a richer and more flexible paraphrasing model in the sense that one semantic relation can be expressed by different verbal phrases.

27.5.1 Acquisition of Paraphrasing Verbs

In English, popular approaches to the acquisition of paraphrasing verbs have searched for snippets that have both nouns as endpoints as well as collected verbs from intervening materials. For example, Nakov and Hearst (2006) used the phrase "noun2 THAT * noun1" for Google queries and extracted verbs between THAT and noun1 from the returned pages. However, there are neither inflections nor clear form markers in Chinese, such as the complementizers that indicate relative clauses, which is why it is difficult to acquire Chinese verbs using explicit clues.

Semantic relations between words should be expressed through certain syntactic forms and structures. The semantic relations held between nouns, and verbs are directly expressed by "Verb-Object" and "Subject-Verb" structures, in which the noun acts as the subject or object of the verb. For example, the Verb-Object structure 切割钻石 *qiege zuanshi* "cut the diamond" shows that 钻石 *zuanshi* "diamond" is a solid substance that can be cut. Thus, we aimed to acquire concept-related verbs for

the nouns using the two grammatical relations above. It was determined that a large-scale corpus with phrase-structure annotation was necessary for this task. However, such resources in Chinese are limited, resulting in a lack of coverage of the acquired verbs. Therefore, we adopted a backward strategy that extracted the verbs from specific grammatical relations (i.e., Subject-Verb and Verb-Object) in terms of collocation using Chinese Word Sketch (CWS).

Chinese Word Sketch. CWS[3] is a combination of the Chinese Gigaword Corpus and the corpus management tool in Sketch Engine (Kilgarriff et al. 2004; Huang et al. 2005). The Chinese Gigaword Corpus (second edition) is a comprehensive archive of newswire text data in Chinese containing about 1.4 billion Chinese characters. All the texts have been segmented and POS tagged automatically. We included all the data in our study. The main functionality of Sketch Engine includes KWIC displays, co-occurrence statistics, grammatical relations, and word sketches, which provide grammatical descriptions of a word in terms of corpus collocations. For nouns, the grammatical description includes nine relations: "A_Modifier/N_Modifier/Modifies," "Subject_of," "Object_of," "And/Or," and "Possession/Possessor." All the collocations were formalized as triples of Rel; Word1; Word2, where Rel is a relation, Word1 is a keyword of a query, and Word2 is the collocation involved with respect to the relation in question.

We used a two-step procedure to acquire the verbs that were related to the compound "n1 n2." First, we collected the collocations with "Subject_of" and "Object_of" relations using n1 and n2 as the keywords of the queries, respectively. We chose only the top 200 words with the highest salience for each relation. Thus, we obtained two sets of collocating verbs denoted as VerbSet1 and VerbSet2 for n1 and n2. Then, we found the intersection of VerbSet1 and VerbSet2, which provided the final paraphrasing verbs. Table 27.6 shows an example of the procedure.

We used this method for 电影公司 *dianying gongsi* "film company" and 啤酒公司 *pijiu gongsi* "beer company," which have the same head. The paraphrasing verbs are shown in Table 27.7, which shows that the two similar compounds have very few common verbs. Fine-grained semantic distinctions were captured with our approach.

27.5.2 Generating Verbal Paraphrases

Yuan (1995) proposed four typical Chinese syntactic patterns for the recovery of the implied predicates, as shown in Table 27.8. In our approach, we used those patterns to generate verbal paraphrases for a compound based on the acquired paraphrasing verbs. We obtained the verbal paraphrases to the maximum using those patterns; however, many of them did not make sense. Next, we filtered out the inappropriate paraphrases via search engines.

[3] http://wordsketch.ling.sinica.edu.tw/

Table 27.6 Verbs acquired for the noun compound 钻石戒指 *zuanshi jiezhi* "diamond ring"

VerbSet1	镶有 *xiangyou* "embed with"
钻石	盛产 *shengchan* "be rich in"
zuanshi	镶 *xiang* "embed"
"diamond"	镶满 *xiangman* "be studded with"
	走私 *zousi* "smuggle"
	镶嵌 *xiangqian* "inlay"
	打磨 *damo* "polish"
VerbSet2	戴上 *daishang* "wear"
戒指	戴 *dai* "wear"
jiezhi	定情 *dingqing* "promise"
"ring"	戴有 *daiyou* "wear"
	试戴 *shidai* "try on"
	抢走 *qiangzou* "snatch"
	交换 *jiaohuan* "exchange"
	镶 *xiang* "encrust"
	镶嵌 *xiangqian* "encrust"
Intersection	拥有 *yongyou* "own"
钻石戒指	获得 *huode* "get"
zuanshi jiezhi	购买 *goumai* "buy"
"diamond ring"	镶嵌 *xiangqian* "encrust"
	戴 *dai* "wear"
	抢走 *qiangzou* "snatch"
	镶 *xiang* "encrust"
	包括 *baokuo* "contain"

Table 27.7 Examples of paraphrasing verbs for 电影公司 *dianying gongsi* "film company" and 啤酒公司 *pijiu gongsi* "beer company"

Noun compounds	Paraphrasing verbs				
电影公司	发行	制作	投资	进出口	服务
dianying gongsi	*faxing*	*zhizuo*	*touzi*	*jinchukou*	*fuwu*
"film company"	"distribute"	"produce"	"invest"	"import and export"	"serve"
啤酒公司	销售	制造	经销	代理	经营
pijiu gongsi	*xiaoshou*	*zhizao*	*jingxiao*	*daili*	*jingying*
"beer company"	"sell"	"make"	"distribute"	"import and distribute"	"manage"

Table 27.8 Patterns used to generate verbal paraphrases

No.	Pattern
P1	n1 + v + 的 *de* + n2
P2	n1 + v + n2
P3	n2 + v + n1
P4	v + n1 + 的 *de* + n2

Table 27.9 Top five verbal paraphrases ranked by Baidu and Google

	钻石 戒指 *zuanshi jiezhi* “diamond ring”	
Rank	Baidu	Google
1	带钻石的戒指 743	买钻石的戒指 *452,000*
	dai zuanshi de jiezhi	*mai zuanshi de jiezhi*
	“ring with diamond”	*“buy diamond ring”*
2	钻石镶嵌戒指 627	钻石镶嵌的戒指 362,000
	zuanshi xiangqian jiezhi	zuanshi xiangqian de jiezhi
	“diamond inlaid ring”	“ring embedded with diamond”
3	买钻石的戒指 *249*	镶嵌钻石的戒指 325,000
	mai zuanshi de jiezhi	xiangqian zuanshi de jiezhi
	“buy diamond ring”	“ring with inlaid diamond”
4	镶钻石的戒指 197	镶钻石的戒指 203,000
	xiang zuanshi de jiezhi	xiang zuanshi de jiezhi
	“ring embedded with diamond”	“ring embedded with diamond”
5	钻石镶戒指 173	没有钻石的戒指 *132,000*
	zuanshi xiang jiezhi	*meiyou zuanshi de jiezhi*
	“diamond inlaid ring”	*“ring without diamond”*

27.5.3 Filtering Verbal Paraphrases

The goal of this process aimed to remove the noise (i.e., inappropriate paraphrases) and retain the most reasonable verbal paraphrases by assigning a higher rank to them. For this purpose, we validated these paraphrases by finding evidence in a large corpus. The greater the evidence, the more appropriate a given paraphrase should be.

The notion of “Web as a corpus” has been widely accepted by researchers. Keller and Lapata (2003) applied web counts to a wide variety of NLP tasks involving syntax and semantics and demonstrated that realistic NLP tasks can benefit from web counts. In our approach, we viewed all the candidate paraphrases as queries, and all queries were submitted to the search engines and performed as exact matches. Thus, we obtained the web counts of the paraphrases. For each noun compound, the paraphrases were ranked by descending order of web counts. The paraphrases with a higher ranking were considered more reasonable than those with a lower ranking.

Baidu (www.baidu.com) and Google (www.google.com) were the most popular search engines for our Chinese search. We conducted experiments based on the web counts obtained from the two search engines, respectively. The number of hits from Baidu and Google was not identical, which resulted in some differences in the ranking. Table 27.9 shows the top five paraphrases for 钻石戒指 *zuanshi jiezhi* “diamond ring” based on Baidu and Google, respectively (incorrect phrases are in italics).

Table 27.10 Accuracy based on Google and Baidu

Google	Top *n*	1	3	5	10
	Accuracy (%)	68.79	90.28	93.35	96.41
Baidu	Top *n*	1	3	5	10
	Accuracy (%)	71.09	89.25	92.83	96.67

27.5.4 Experiments and Evaluation

We randomly selected 391 Chinese noun compounds from the newswire corpus *People's Daily* to test our approach. For each compound, the top 10 candidate paraphrases were collected. All the paraphrases were judged by three human subjects.[4] They were asked to make binary judgments (yes or no) for each paraphrase, that is, whether the paraphrases expressed a meaning similar to that of the compound. If more than two subjects labeled the paraphrase yes, it was viewed as correct. We defined the accuracy of the compounds using Formula (27.4):

$$\text{Accuracy} = \frac{\text{the number of compounds with correct interpretation}}{\text{the total number of compounds}} \times 100\% \tag{27.4}$$

Table 27.10 shows the different accuracy rates, where "*n*" equals 1, 3, 5, and 10. As shown, the performances based on Google and Baidu were very similar. Thus, our method provided correct interpretations for almost 70% of the compounds when only the topmost paraphrase was given, and accuracy increased with the number of candidate paraphrases.

27.6 Conclusion

In this chapter, we presented our preliminary research on the interpretation of Chinese noun compounds using two different strategies. For the abstract relation strategy, we proposed a novel taxonomy of Chinese noun compounds based on the transparency of the compounds. Then, we proposed a method for interpreting Chinese NCs based on word similarity. Our experimental results showed that word similarity provided useful information in solving interpretation problems. In the future, we plan to use corpus-based similarity methods such as word2vec to solve the out-of-vocabulary (OOV) problem. Moreover, the voting strategy can be used to determine the semantic relations of the test NCs since we chose only those NCs with the highest similarity.

[4] One of the subjects was a Ph.D. student in linguistics, and the other two had master's degrees in computational linguistics.

For the verbal paraphrasing strategy, we proposed the simple dynamic approach of using paraphrasing verbs, which could be useful in many NLP tasks. This approach not only provided possible interpretations of noun compounds but also captured interesting fine-grained semantic differences of similar noun compounds. In the future, we plan to acquire more verbs using web data, such as the Google 5-gram web index. We also plan to expand the paraphrasing patterns. Finally, we are also very interested in applying the methods proposed here to information retrieval.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 61300152, No. 61300156). We would like to thank the anonymous reviewers for their comments.

References

Butnariu, Cristina, and Tony Veale. 2008. A concept-centered approach to noun-compound interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 81–88. Manchester, United Kingdom.

Dong, Zhendong, and Qiang Dong. 2005. HowNet. Available at http://www.keenage.com. Accessed:10 Oct. 2015

Downing, Pamela. 1977. On the creation and use of English compound nouns. *Language* 53(4): 810–842.

Finn, T. 1980. The semantic interpretation of compound nominals. Ph.D. dissertation. University of Illinois, Urbana.

Girju, Roxana, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Journal of Computer Speech and Language—Special Issue on Multiword Expressions* 4(19):479–496.

Huang, Chu-ren, Adam Kilgarriff, Yiching Wu, Chih-Ming Chui, Simon Smith, Pavel Rychly, Ming-Hong Bai, and Keh-Jiann Chen. 2005. Chinese sketch engine and the extraction of grammatical collocations. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. Jeju Island, Korea.

Keller, Frank, and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29(3):459–484.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of Euralex*, 105–116. Lorient, France.

Kim, Su Nam, and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, 945–956. Jeju Island, Korea.

Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

Liu, Qun, and Sujian Li. 2002. Word similarity computing based on HowNet. *International Journal of Computational Linguistics & Chinese Language Processing* 7(2):59–76.

Nakov, Preslav. 2008. Noun compound interpretation using paraphrasing verbs: Feasibility study. In *LNAI* (Vol. 5253). In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems and Applications (AIMSA 2008)*, 103–117 Varna, Bulgaria.

Nakov, Preslav, and Marti A. Hearst. 2006. Using verbs to characterize noun-noun relations. In *LNCS* (Vol 4183). In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems and Applications (AIMSA 2006)*, 233–244. Varna, Bulgaria.

Nakov, Preslav, and Marti A. Hearst. 2013. Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Transactions on Audio, Speech, and Language Processing* 10(3):Article 13.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language* 70(3):491–538.

Ó Séaghdha, Diarmuid. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the 4th Corpus Linguistics Conference*. University of Birmingham, United Kingdom.

Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, MA: The MIT Press.

Ó Séaghdha, Diarmuid. 2008. Learning compound noun semantics. Ph.D. dissertation. University of Cambridge.

Tratz, Stephen, and Hovy, Eduard. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 678–687. Uppsala, Sweden.

Wang, Meng 王萌. 2010. Linguistic knowledge acquisition of noun for the construction of probabilistic lexical knowledge-based 面向概率型词汇知识库建设的名词语言知识获取. Ph.D. dissertation. Peking University, China.

Wang, Meng, Chu-ren Huang, Shiwen Yu, and Shiyong Kang. 2014. Chinese noun compound interpretation using verbal paraphrases. *ICIC Express Letters, Part B: Applications* 5(5): 1377–1382.

Warren, Beatrice. 1978. Semantic patterns of noun-noun compounds. Ph.D. thesis. Acta Universitatis Gothoburgensis, Sweden.

Wei, Xue. 魏雪. 2012. Research on Chinese noun compound interpretation for semantic-query. 面向语义搜索的汉语名名组合的自动释义研究. Master's thesis. Peking University, China.

Yu, Shiwen, Huiming Duan, Xuefeng Zhu, and Bin Sun 俞士汶, 段慧明, 朱学锋, 孙斌. 2002. The basic processing of contemporary Chinese corpus at Peking University SPECIFICATION 北京大学现代汉语语料库基本加工规范. *Journal of Chinese Information Processing 中文信息学报 16(5):49–64.*

Yuan, Yulin 袁毓林. 1995. Implying predicate and its syntactic implementation 谓词隐含及其句法后果—"的"字结构的称代规则和"的"的语法、语义功能 *Studies of the Chinese Language 中国语文 4:241–255.*

Zhao, Jinglei, Hui Liu, and Ruzhan Lu. 2007. Semantic labeling of compound nominalization in Chinese. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, 73–80. Prague, Czech Republic.