ICIC Express Letters
Part B: Applications
Volume **7**, Number **1**, January **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **173–179**

RESEARCH ON CHINESE NOUN + NOUN COMPOUNDS SEMANTIC CLASSIFICATION AND AUTOMATIC INTERPRETATION

Min Gu[1], Yanhui Gu[1,2], Fang Xu[1], Bin Li[3], Bin Zhao[1,2] and Weiguang Qu[1,2]

[1]School of Computer Science and Technology
Nanjing Normal University
No. 1, Wenyuan Road, Xianlin, Nanjing 210023, P. R. China
{ gu; wgqu }@njnu.edu.cn

[2]Jiangsu Research Center of Information Security and Privacy Technology
Nanjing 210023, P. R. China

[3]School of Chinese Language and Literature
Nanjing Normal University
No. 122, Ninghai Road, Nanjing 210097, P. R. China

Received June 2015; accepted August 2015

Abstract. *The purpose of semantic classification is to determine the semantic relations of the compound. In this paper, we summarize the combination rules of different semantic relations of the Noun + Noun compound, and determine the semantic relations of the compounds based on the rules. In addition, we recover the implied verb between the nouns, apply the Xinhua corpus from 1991 to 2004 and the HowNet to acquire the paraphrase of Chinese Noun + Noun compounds. After continuously modifying and improving the program, the result of the paper can be applied in some other fields such as information retrieval, and machine translation.*
Keywords: Chinese Noun + Noun compounds, Semantic classification, Interpretation

1. **Introduction.** Chinese Noun + Noun compounds is a new phrase that contains two nouns which are constituted together directly, such as "guo2jia1jing1ji4" (the national economy). Ma [1] proposed that noun + noun compounds refers to that two nouns combine together that neither contains 'the' or 'and' nor '、' or '·'. To facilitate the writing and reading, this paper calls 'Noun + Noun compounds' 'N + N compounds' for short. In syntax analysis, N + N compounds can be analyzed into four types of relation: attributive-head relation, associative relation, parallel relation and subject-predicate relation [1]. Taking 'guo2jia1 jing1ji4' (the national economy) as an example, it is classified as attributive-head phrase. 'te4qu1 xiang1gang3' (Hong Kong SAR) belongs to associative relation and 'zhuan1jia1 xue2zhe3' (experts and scholars) is classified as parallel relation.

N + N compounds research focuses on the aspects of semantic comprehension, syntactic analysis, translation, automatic acquisition at present. As a result of N + N compound's strong generative capacity, the research on it comes up with new problems. If we cannot identify the semantic of N + N compounds, it will affect the semantic understanding of a text. Therefore, the study of Chinese N + N compounds has a positive effect on information retrieval, machine translation, and question answering system.

The research tasks on English N + N compounds mainly apply two strategies: the top-down strategy and the bottom-up strategy. Levin [2] proposed nine semantic relations (recoverably deletable predicates). The nine relations are 'CAUSE, HAVE, MAKE, USE, BE, FOR, IN, FROM, ABOUT'. Vanderwende [3] defined thirteen kinds of semantic relations, he used "wh-questions" (who, what, when, where, whose, how) to question

the N + N compound, so it can be classified. For example, 'bird' in 'bird sanctuary' is the answer for 'what for', so 'bird sanctuary' is classified as 'Purpose'. Lauer [4] uses eight prepositions (of, for, in, at, on, from, with, about) to define semantic relations, for example, "baby car" can be understood as "car for baby". Zhao et al. [5] defines four coarse-grained semantic relation 'Proto-Agent (PA), Proto-Patient (PP), the Range (RA) and Manner (MA)' with reference to the verb semantic role. He uses it to classify the relations of N + N compound. For example, "dong4wu4fen1lei4" (the classification of animals) can be understood as "dui4dong4wu4jin4xing2fen1lei4". "dui4" is the preposition used to describe the relationship between PA, so "dong4wu4fen1lei4" can be classified as PA. Preslav and Hearst [6] rewrites N + N compounds into a relative clause containing wildcards to classify the sematic relation. The methods that the researchers use to classify the sematic relation between Chinese N + N compounds are different. Li [7] analyzes five rules to judge attributive-head relation. The automatic interpretation of N + N compounds mainly uses two methods. Wang et al. [8] uses the method based on the verbs. She uses the bottom-up strategy to get the interpretation dynamically. However, the definition template she uses is single, and the interpretation is not exact. Wei and Yuan [9] sums up the different modes and templates of N + N compound. She finds out the verbs and then fills it into the template to get the interpretation of N + N compound. She uses 245 phrases to test [10-12]. The accuracy is 94.2%, but the coverage is only 63.7%. The cause of low coverage is mainly the lack of database information (including the semantic classes of nouns, verbs and templates).

The contributions of this paper are as follows.

(1) We classify the semantic relations of N + N compound. In this paper, we find out the rules of the composition of N + N compound and use the rules to identify the four relations of N + N compound.

(2) We do the research of automatic interpretations of N + N compounds which belong to attributive-head relation. The method to realize automatic interpretation is based on interpretative verbs. Using verbs and templates, we can create the interpretation phrases.

(3) We develop a semantic classification and automatic interpretation platform for Chinese N + N compound. Using 1000 high frequency N + N compounds got from the Xinhua corpus from 1991 to 2004 to test the program, we can find out the shortcomings and improve it. The remainder of this paper is organized as follows. Section 2 describes the method of semantic classification; Section 3 describes the method of automatic interpretation; Section 4 introduces the experiments and results analysis, and we conclude the paper and present the future work in Section 5.

2. **Semantic Classification Methodology.**

We first find out the N + N compounds from the first half of 1998 People's Daily corpus, then calculate the word frequency and select out the top 300 N + N compounds ordered by the frequency. We mark the semantic relations between the nouns in N + N compounds artificially (including attributive-head relation, associative relation, parallel relation and subject-predicate relation) and inspect the definition of nouns in "The Semantic Knowledge-base of Contemporary Chinese", "HowNet" and "Tongyici Cilin (Extended Edition)". Then we sum up the rules of identifying the different semantic relations of N + N compounds and use the rules to write the program of semantic classification.

On the basis of previous relevant work, we summarize the judgement rules of four basic relations as follows.

Rule 1. *When the nouns in the N + N compounds belong to the same semantic class, the relation of the compound is parallel relation.*

Parallel relation refers to the two nouns of N1 + N2 compounds which are different things related to each other. For example, the two nouns in "xiao3mai4yu4mi3" (wheat

and corn) belong to "zhuang1jia1" (crop) in SKCC and "HowNet". The compound is classified as parallel relation.

Rule 2. *Time/Human + Time = Subject-predicate relation*

Subject-predicate relation refers to that two nouns of N1 + N2 compounds are subject and predicate of a sentence. N2 is the supplement of N1 which explains the specific content of N1. For example, "jin1tian1xing1qi1tian1" (Today is Sunday) is a subject-predicate phrase. The semantic class of "jin1tian1" (today) and "xing1qi1tian1" (Sunday) is time.

Rule 3. *Body component + Body component/Manual/Color/Waste/Natural objects = Subject-predicate relation*

For example, the semantic class of "leng3han4" (cold sweat) is "Waste", so the compound "hun2shen1leng3han4" (The body breaks into a cold sweat) is a subject-predicate phrase. However, the class "Color" is an adjective in "HowNet". So when using "HowNet" as the resource, "Color" needs to be ruled out.

Rule 4. *Name + Id/Career/Relationship/Personage = Associative relation*

The two nouns in associative relation phrase explain the same thing from their perspective view. For example, "Xi2jin4ping2zhu3xi2" (the chairman Jinping Xi) is an associative relation phrase. "Xi2jin4ping2" is a name and "zhu3xi2" (chairman) is a job. In "HowNet", the definition of "Xi2jin4ping2" is "human" and "propername". The definition of "zhu3xi2"(chairman) is "human, occupation, male and female".

Rule 5. *Identity/Personage + Name = Associative relation*

For example, "tie3ren2wang2jin4xi3" (the iron man Jinxi Wang) is an associative relation phrase. "tie3ren2" (iron man) is classified as personage and "wang2jin4xi3" is a name.

Rule 6. *The semantic class of two nouns has a relationship between up and down, which usually is noun and place names which can form associative relation phrase.*

For example, "te4qu1" (SAR) in "te4qu1xiang1gang3" (Hong Kong SAR) is a property of Hong Kong, so the compound belongs to associative relation.

Rule 7. *When the two nouns belong to different semantic classes, the N + N compound belongs to attributive-head relation.*

Attributive-head relation indicates that the N + N compounds are modifier and center. For example, "zhuan1jia1xiao3zu3" (the group of experts) is an attributive-head phrase. "zhuan1jia1" (experts) and "xiao3zu3" (group) belong to different semantic classes.

Rule 8. *Inferior + Superior = Attributive-head relation*

For example, "shu1fa3" (calligraphy) belongs to "yi4shu4" (art), so "shu1fa3yi4shu4" is an attributive-head phrase.

Rule 9. *Whole + Part = Attributive-head relation*

N1 is the whole object and N2 is part of N1. For example, because Nanjing is part of Jiangsu, "jiang1su1nan2jing1" is an attributive-head phrase.

Rule 10. *When the number of nouns in "Tongyici Cilin" is the same, the N + N compound belongs to parallel relation.*

The number of "zhuan1jia1" (experts) and "xue2zhe3" (scholars) is "Al02B01", so "zhuan1jia1xue2zhe3" is a parallel-relation phrase.

Rule 11. *When top four number of the two nouns in "Tongyici Cilin" is the same and the nouns are headword, the N + N compound is a parallel-relation phrase.*

For example, the number of "yu3" (rain) is "Bf01A01" and "xue3" (snow) is "Bf01B01". So "yu3xue3" (rain and snow) is a parallel-relation phrase.

Rule 12. *If the numbers of the two nouns are only different in the last two numbers and N2 is a headword, the N1 + N2 compound is an attributive-head phrase.*

In "Tongyici Cilin", the number of "ke1ji4" (science) is "Dk03A12" and the number of "zhi1shi1" (knowledge) is "Dk02C01". "ke1ji4zhi1shi1" (the knowledge of science) is an attributive-head phrase.

3. Automatic Interpretation Methodology.

3.1. Methodology of automatic interpretation based on verbs.

Because attributive-head relation plays an important role in semantic relations, we choose attributive-head N + N compounds as our research object. The interpretation of N + N compounds is to find out the interpretative verbs and fill into the interpretive templates to generate the interpretations. While the recall rate of using single template is only 68.6%, so we choose the interpreting template which Wei and Yuan [9] summarized in her paper to achieve better effect of interpretation.

3.2. Formation and filtration of interpretative phrases.

We get the semantic class of the nouns in the N + N compounds and search the database we mentioned above to find out the template which corresponds to the semantic class. Then, we insert the interpretative verbs into the template to form the interpretative phrases. The purpose of filtering the interpretative phrases is to remove the unreasonable phrases and improve the accuracy of interpretation. We send the phrases to the search engine Baidu and Biying to find out the number of the results. The specific steps are as follows:

(1) *Transcode the phrases;*
(2) *According to the search engine URL interface, get the corresponding URL;*
(3) *Use the web spider to get the corresponding web page information;*
(4) *Use the regular type matching to collect the number of the results.*

Then we sort the number of the results and select the top 3 as the interpretative phrases of the N + N compounds.

Taking "li4shi3lao3shi1" as an example, the results form Baidu and Biying are as Table 1.

Table 1. Results of search engines

N + N compounds	Baidu	Biying
li4shi3 lao3shi1	Zuo4wei2+li4shi3+de1+lao3shi1 36100000	jiao1+li4shi3+de1+lao3shi1 4160000
	jiao1+li4shi3+de1+lao3shi1 35500000	wan2cheng2+li4shi3+de1+lao3shi1 2110000
	wan2cheng2+li4shi3+de1+lao3shi1 35300000	Zuo4wei2+li4shi3+de1+lao3shi1 2080000
	zun1zhong4+li4shi3+de1+lao3shi1 34000000	zun1zhong4+li4shi3+de1+lao3shi1 839000

("jiao1+li4shi3+de1+lao3shi1" (teacher who teaches history) is the interpretative phrase. "35500000" is the number of search results.)

As we can see from Table 1, "jiao1+li4shi3+de1+lao3shi1" (the teacher who teaches history) should be the correct output, but due to retrieving in the search engine, the search query is broken up, so the result is not accurate. Therefore, we use the advanced search skills (put the search query in double quotes) to ensure that the search query will not be broken up. The modified query results are in Table 2.

According to the search engine, the best interpretation is "jiao1+li4shi3+de1+ lao3-shi1".

4. Experiments and Result Analysis.

In this paper, the experimental data we use is the N + N compounds we get from the Xihua corpus from 1991 to 2004. In the corpus, the words which are followed by the tag "/n" are nouns. In the experiment we use regular expression matching to get the N + N compounds and calculate the frequency of the

Table 2. Improved results

N + N	Baidu	Biying
li4shi3 lao3shi1	jiao1+li4shi3+de1+lao3shi1 257000 zun1zhong4+li4shi3+de1+lao3shi1 9 zuo4wei2+li4shi3+de1+lao3shi1 3	jiao1+li4shi3+de1+lao3shi1 129 zuo4wei2+li4shi3+de1+lao3shi1 47 wan2cheng2+li4shi3+de1+lao3shi1 1

compounds. According to the order of the word frequency from big to small, we pick out the top 1000 N + N compounds.

4.1. **Results and analysis of semantic classification.** The method based on the rules are shown in Table 3. Test corpus is the 1000 high-frequency N + N compounds.

Table 3. Results of classification

Semantic relations	Results of classification	Hand-classified classification
attributive-head relation	955	990
associative relation	2	2
subject-predicate relation	2	2
parallel relation	5	4
unknown relation	36	2

The accuracy of attributive-head relation is 96.465%. The recall rate is 95.5% and the F-value is 95.98%. The specific results are as Table 4.

Table 4. Specific results of classification (The error results are in boldface)

Associative relation	mao2ze2dong1zhu3xi2() ba1jin1ye2ye2()
parallel relation	zhuan1jia1xue2zhe3() yin2pai2tong2pai2() you2dian4tong1xin4ye4() **jin3biao1sai4ban4jue2sai4()** yu3xue3()
subject-predicate relation	jin1tian1xing1qi1tian1() jin1tian1chun1jie2()

From the data analysis we can draw conclusions.

(1) In the experiment based on "HowNet", we get 36 unknown N + N compounds. The main reason of the result is:

① The part of speech is not the same. For example, "lv4se4tong1dao4" (green channel) is marked as noun in the corpus, but green is an adjective in the HowNet;

② The nouns are not included in "HowNet". For example, "gan3ran3zhe3" (infected person) is not included in "HowNet", so the compound "ai4zi1bing4gan3ran3zhe3" (the person who is infected with HIV) cannot be classified.

(2) Because the names we can recognize are very little, the compounds in the form of "name + noun" cannot be classified correctly.

(3) The rules of classification are not exactly accurate. We need to modify the rules. For example, "jin3biao1sai4ban4jue2sai4" is an attributive-head phrase. However, we classify it as a parallel phrase.

(4) Some of the N + N compounds we get from the corpus are not compounds. For example, "ge4/n ren2/n" is extracted in "10/m ge4/n ren2/n jin4ru4/v shi4jie4/n

qian2/f ba1/m ming2/ag". However, "ge4/n ren2/n" (personal) is an adjective. Because of the pos tagging, it is extracted. So we should take the context into consideration when we extract the compounds.

4.2. **Results and analysis of automatic interpretation.** The test data is 300 attributive-head compounds which are selected from the result of semantic classification. The accuracy of taking the top phrase as the interpretative phrase is only 72.67%. So we choose the top three phrases as the interpretative phrase. The results are shown in Table 5.

Table 5. Results of automatic interpretation under different metrics

Number of compounds	Interpreted number	Correct interpreted number	Accuracy	Recall	F-value
300	299	271	90.63%	90.33%	90.48%

As we can see from the table, the accuracy of taking the top three as the phrases is higher than the former. The main reason of it is that the results of the two search engines are different, the real correct phrases may be filtered by the program. For example, the interpretative phrases of "hui4yi4dai4biao3" (the representative of the conference) are as follows.

Table 6. Result of automatic interpretation under different datasets

N + N compounds	Baidu		Biying	
hui4yi4 dai4biao3	Zhao4kai1hui4yi4de1dai4biao3	77	Wei2hui4yi4de1dai4biao3	90
	Zuo4wei2hui4yi4de1dai4biao3	20	Shi4hui4yi4de1dai4biao3	7
	Ju3xing2hui4yi4de1dai4biao3	9	Zuo4wei2hui4yi4de1dai4biao3	6

We can see from Table 6 that the more accurate interpretation is "shi4hui4yi4de1dai-4biao3" (is the representative of the conference) or "zuo4wei2hui4yi4de1dai4biao3" (as representative of the conference). However, due to the search engines, the interpretations the system ordered in the first place are "zhao4kai1hui4yi4de1dai4biao3" and "wei2hui4yi4de1dai4-biao3". The interpretation is not accurate. So we choose the top three phrases as the interpretative phrase.

In addition, the N + N compounds which has a noun to act as an adjective is not entirely accurate. For example, "zhong1xin1di4wei4" (central position), the word "zhong1xin1" (center) is a noun but acts as an adjective. The interpretation of the compound is wrong. So we cannot find the correct verbs for the interpretation in this situation.

5. **Conclusions and Future Work.** After summarizing the rules of different semantic relations, we try to classify the semantic relations of different compounds. In addition, we do automatic interpretation on the classification results. We search the verbs which are corresponding to the nouns in the compounds and search the database for the template. Then we put them together and send them to the search engines. According to the search results, we choose the top three as the interpretative phrases to export.

However, some issues need to be further improved.

(1) The roles of semantic classification are not exactly accurate. We need to summarize more compositive rules of different N + N compounds which are different in semantic relations to improve the rules.

(2) The verbs for interpretation in the automatic interpretation are not exactly accurate. We need to modify the verb collocation repository. The existing verbs are selected from

the Xihua corpus from 1991 to 2004. We only take the single noun into consideration. In the future, we need to filter the inaccurate verbs and take the context into consideration to get the verbs which is corresponding to the noun compounds.

(3) In the automatic interpretation, we use Baidu and Biying as search engines. However, the results of them may be different sometimes. So we also need to consider the choice of search engines to achieve the best interpretation.

(4) When we cannot find the noun in "HowNet", we use "Tongyici Cilin" to find synonyms. In this paper, the method of finding synonyms is direct match. However, we still cannot find the synonyms sometimes. So we consider computing the similarity of words to get the most similar words. In that case, we can improve the recall rate of semantic classification and automatic interpretation.

(5) Considering the ambiguity in N + N compounds, we need to summary the rules of removing ambiguity. Based on the analysis of huge corpus, we will find a suitable method to solve the problem.

REFERENCES

[1] H. Ma, The semantic study of noun + noun compounds, *Journal of Xinyang Teachers College*, 1999.

[2] J. Levin, *The Syntax and Semantics of Complex Nominals*, Academic Press, New York, 1978.

[3] L. Vanderwende, Algorithm for automatic interpretation of noun sequences, *The 15th International Conference on Computational Linguistics*, 1994.

[4] M. Lauer, *Designing Statistical Language Learners: Experiments on Compound Nouns*, Ph.D. Thesis, Macquarie University, Australia, 1995.

[5] J. Zhao, H. Liu and R. Lu, Semantic labeling of compound nominalization in Chinese, *Proc. of the Workshop on a Broader Perspective on Multiword Expressions*, Prague, pp.73-80, 2007.

[6] N. Preslav and M. Hearst, Using verbs to characterize noun-noun relations, *Proc. of the 12th International Conference on Artificial Intelligence: Methodology, Systems and Applications*, pp.233-244, 2006.

[7] S. Li, Study on determining the semantic relations in N1+(de)+N2 modifier-head construction in modern Chinese, *Journal of Chongqing Technology and Business University (Social Science Edition)*, 2009.

[8] M. Wang, C. Huang, S. Yu and B. Li, Chinese noun compound interpretation based on paraphrasing verbs, *Journal of Chinese Information Processing*, 2010.

[9] X. Wei and Y. Yuan, Towards a rule-based approach to automatic interpretation of Chinese noun compounds, *Journal of Chinese Information Processing*, 2014.

[10] J. Tan, Semantic relations between nouns and noun modifiers and their roles in dictionary definition, *Studies of the Chinese Language*, 2010.

[11] H. Wang, W. Zhan and S. Yu, Structure and application of the semantic knowledge-base of modern Chinese, *Applied Linguistics*, 2006.

[12] Z. Dong and Q. Dong, Construction of a knowledge system and its impact on Chinese research, *Contemporary Linguistics*, 2001.

[13] R. Girju, S. Szpakowicz, P. Nakov, P. Turney, V. Nastase and D. Yuret, SemEval 2007 task 04: Classification of semantic relations between nominals, *Proc. of Sem. Eval.*, Prague, Czech Republic, pp.13-18, 2007.

[14] S. Stan, B. Francis, N. Preslav and K. S. Nam, On the semantics of noun compounds, *Natural Language Engineering*, 2013.