ICIC Express Letters
Part B: Applications
Volume **7**, Number **6**, June **2016**

ICIC International ©2016 ISSN 2185-2766
pp. 1215–1221

AUTOMATIC INTERPRETATION OF CHINESE NOUN COMPOUNDS BASED ON WORD SIMILARITY

Meng Wang[1], Lulu Wang[2], Na Tian[1] and Bin Li[3,4]

[1]School of Humanity
Jiangnan University
No. 1800, Lihu Avenue, Wuxi 214122, P. R. China
wangmengly@163.com

[2]School of Literature
Communication University of China
No. 1, Dingfuzhuang East Street, Chaoyang District, Beijing 100024, P. R. China
lulu.wang@cuc.edu.cn

[3]Research Center of Language and Informatics
School of Chinese Language and Literature
Nanjing Normal University
No. 122, Ninghai Road, Nanjing 210097, P. R. China
libin.njnu@gmail.com

[4]Department of Computer Science
Brandeis University
No. 415, South Street, Waltham, MA 02453, US

Received November 2015; accepted February 2016

Abstract. *Noun compound interpretation is to make the compressed semantic relation between the nouns explicit. In this paper, we present a method for interpreting Chinese two-word noun compounds automatically based on word similarity. The experimental results show that our method can provide reasonable interpretations for novel NCs, and word similarity is useful information in solving the interpreting problem.*

Keywords: Chinese noun compounds, Interpretation, Semantic relation, Word similarity

1. **Introduction.** A noun compound (NC) is a sequence of two or more nouns (e.g., diamond ring, love story) that syntactically behaves as a single noun. NCs occur very frequently in English written text, including technical materials, newswire and fictional prose [1,2]. In Chinese, NCs are also abundant in text since compounding of nouns is a common way of naming new things. Research on the syntax and semantics of noun compounds belongs to the broader field of Multi-Word Expression (MWE). The interpretation of NC is to determine the semantic relationships between adjacent nouns. For example, "love story" can be interpreted as "a story that tells about love", and "diamond ring" means "ring inlaid with diamond". Understanding relations between noun compounds is an important problem within a wide variety of natural language processing (NLP) applications, such as machine translation, information retrieval and question answering, among others.

In this paper, we focus on the interpretation of Chinese two-word NCs and present an automatic method for predicting the semantic relations to the novel NCs based on word similarity. The remainder of the paper is organized as follows. Section 2 describes the motivation of this research. Section 3 gives a brief introduction to the taxonomy of the Chinese NC relations. Section 4 presents the word similarity measures. Section

5 introduces the proposed approach. Section 6 presents an evaluation of this approach, while Section 7 offers conclusion and future work.

2. **Motivation.** In the interpretation of NCs, earlier work uses hand-coded rules which require large human efforts [3]. Recent work has investigated methods for interpreting NCs automatically. Following this line of research, semantics of NCs can be represented as abstract relations drawn from a small closed set. Thus, the interpretation can be treated as a classification problem. In this paper, we present a method using word similarity to predict the semantic relations of novel NCs. Given an NC in the test data, we compute the similarities between the correspondence nouns in the training data to acquire the semantic relation.

For example, we have a test NC "nong2cun1 shi4chang3 (rural market)" and two training NCs "shou3du1 yi1yuan4 (capital hospital)" and "huang2jin1 shi4chang3 (gold market)". Figure 1 shows the correspondences between them, where S_{ij} is a measure of noun-noun similarity in the training and test data. Table 1 lists the word similarities which are computed by HowNet. In this case, S_{11} is the similarity between "shou3du1 (capital)" and "nong2cun1 (rural)".

The similarity of the NC pair can be derived by the product of the individual similarities. Note that "nong2cun1 shi4chang3 (rural market)" is *market located in rural* (LOCATION), "huang2jin1 shi4chang3 (gold market)" is *market that sells gold* (PATIENT), and "shou3du1 yi1yuan4 (capital hospital)" is *hospital located in the capital* (LOCATION). Although "shi4chang3 (market)" in the test NC also occurs in the training exemplar, the semantic relation is different. By comparing the similarity of both constituents of the training NCs, we can draw the conclusion that "nong2cun1 shi4chang3 (rural market)" is more closely related to "shou3du1 yi1yuan4 (capital hospital)". Then the semantic relation of "nong2cun1 shi4chang3 (rural market)" is labeled as LOCATION.

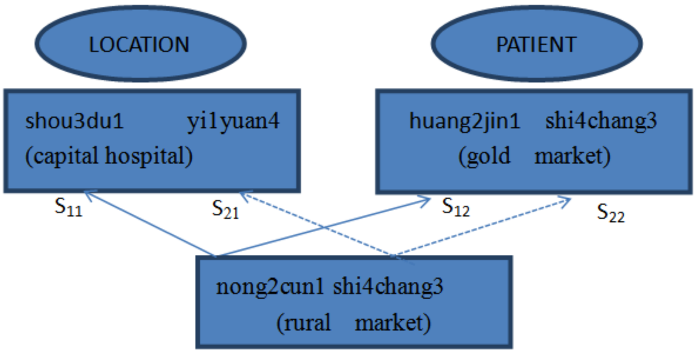


FIGURE 1. Similarity between test NC and training NCs

TABLE 1. Similarities for the component nouns based on HowNet

Similarity	Training word	Test word	Values
S_{11}	shou3du1 (capital)	nong2cun1 (rural)	0.550000
S_{21}	yi1yuan4 (hospital)	shi4chang3 (market)	0.483182
S_{12}	huang2jin1 (gold)	nong2cun1 (rural)	0.024242
S_{22}	shi4chang3 (market)	shi4chang3 (market)	1

3. **Taxonomy of Chinese Noun Compounds Relations.** To elaborate the interpretation of noun compounds, previous studies describe the semantics of noun compounds in two ways: one is to define the abstract relations, such as BE, HAVE, IN, ACTOR, INST, ABOUT [2]. The other is to use verbal paraphrases to interpret the noun compounds.

For instance, "salt water" could be interpreted with "*dissolved in*" [3]. And "headache pill" might be paraphrased as "headache-inducing pill" or "headache prevention pill" [4]. In dealing with Chinese noun compounds, Wang et al. suggest four types of paraphrase patterns of Chinese noun compounds based on the paraphrased verbs [5]. However, these four types are not specific enough to give proper interpretations [6]. Instead, Wei classifies the noun compounds into 8 major types and 346 subcategories, based on the semantic types of the parts [6]. However, some of these subcategories can be merged, and some noun compounds belong to more than one subcategory. Moreover, some noun compounds are not interpretable so that we could not find the hidden verbs. We hypothesize that this is due to the lack of considerations of the decomposable possibilities and the semantic transparency of noun compounds.

Idioms are classified into decomposable idioms and non-decomposable one [7]. The former is combinational and the other is idiosyncratic. For example, the noun compound "fu1qi1 fei4pian4 (pork lungs in chili sauce)" is not decomposable, that is, the meaning of the compound is not simply the combinations of the literal meanings of the parts. Based on the transparency scale, Levi classifies the noun compounds into five subtypes: transparent, partly opaque, exocentric, partly idiomatic and completely idiomatic [2]. For example, "orange peel" is transparent as it is simply the combination of the parts of "orange" and "peel". And "grammar school" is partly opaque because it cannot be combined literally, that is because a hidden verb should be revealed to illustrate this compound, which is "school that teaches grammar".

Enlightened by these ideas, we present a novel taxonomy of Chinese noun compounds [8].

Table 2. Basic types of noun compounds

Type	Transparency scale	Examples
decomposable	transparent	机组 人员 (ji1zu3 ren2yuan2) (crew members)
	partly opaque	钻石 戒指 (zuan4shi2 jie4zhi3) (diamond ring)
	partly idiomatic	试管 婴儿 (shi4guan3 ying1er2) (test tube baby)
non-decomposable	completely idiomatic	夫妻 肺片 (fu1qi1 fei4pian4) (pork lungs in chili sauce)

As Table 2 shows, the first three types are decomposable, while the last one is non-decomposable. For the first three types, only the first two could reveal the hidden verbs. For example, "zuan4shi2 (diamond)" and "jie4zhi3 (ring)" imply the verb of "xiang1qian4 (inlaid)", but "shi4guan3 (test tube)" and "ying1er2 (baby)" cannot be combined literally, that is because "test tube" denotes as *in vitro (glass) fertilization*. They are not simply the combinations of the literal meanings of the components, but involve a process of metaphors or metonyms, which enhance the difficulty in revealing the hidden verbs.

In order to reveal the hidden verbs, Wei first adopts the idea of qualia roles by Pustejovsky into the interpretation of Chinese noun compounds and discovers the semantic relations within [6,11]. We believe that there is a clear correspondent relationship between the semantic relations and the qualia roles of the head noun. To illustrate, we summarize this correspondence in Table 3.

To interpret these noun compounds, we summarize various interpretation patterns with qualia roles of N1 or N2. For example, "wei2qi2 gao1shou3 (chess master)" could be paraphrased as "the masters of playing chess" where "to play" is the TELIC role of

Table 3. The semantic relations of noun compounds

Semantic relations	Qualia roles	Interpretation patterns	Examples
possessive	constitutive	N2 is ‘belonged to’ N1	机组 人员 (ji1zu3 ren2yuan2) (crew members)
property	formal	N2’s property is N1	股份制 企业 (gu3fen4zhi4 qi3ye4) (joint stock enterprise)
locative	formal/agentive	N2 is located in N1	印尼 火山 (yin4ni2 huo3shan1) (Indonesia volcano)
time	formal/agentive	N2 is made in N1	清代 家具 (qing1dai4 jia1ju4) (Qing dynasty furniture)
material	constitutive/agentive	N2 is made of N1	钻石 戒指 (zuan4shi2 jie4zhi3) (diamond ring)
patient	telic	V-N1-N2	围棋 高手 (wei2qi2 gao1shou3) (chess master)
actor	agentive	N1-V-N2	教委 文件 (jiao4wei3 wen2jian4) (the document issued by the board of education)
content	constitutive/telic	N2 is about N1	爱情 故事 (ai4qing2 gu4shi4) (love story)
cause	agentive	N1 causes N2	考试 焦虑 (kao3shi4 jiao1lv4) (tests anxiety)
partly-idiomatic	–	metaphoric or metonymic meaning of N1+*de*+N2	试管 婴儿 (shi4guan3 ying1er2) (test tube baby)
idiomatic	–	idiom	夫妻 肺片 (fu1qi1 fei4pian4) (pork lungs in chili sauce)

“chess”. Also, “ai4qing2 gu4shi4 (love story)” could be paraphrased as “the story about love” where the constitutive role of “gu4shi4 (story)” is “ai4qing2 (love)”.

4. Word Similarity.

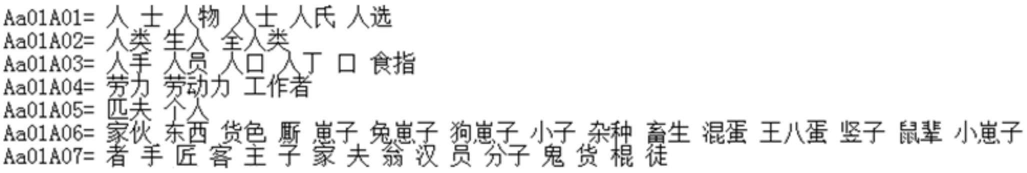
4.1. HowNet-based similarity.

HowNet is a common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. As a knowledge base, the knowledge structured by HowNet is a graph rather than a tree. It is devoted to demonstrate the general and specific properties of concepts. For every word sense c_i (i.e., a concept), its definition is composed by a set of sememes and the corresponding relations. For instance, the Chinese word “xue2xiao4 (school)” is defined as follows:

NO. = 0.95550
W_C = 学校
G_C = N
W_E = school
G_E = N
DEF = InstitutePlace| 场所,@teach| 教,@study| 学,education| 教育

HowNet allows the users to measure the semantic similarity and relatedness between a pair of two concepts based on the overlap of sememes. In this paper, we adopt the similarity measure provided by Liu to achieve the similarity of two nouns [9].

4.2. **Cilin-based similarity.** Cilin is a Chinese thesaurus defining and describing "concepts" and revealing their relations by Synset. The semantic category of words (i.e., concepts) is encoded by a 5-layer tree. Figure 2 gives some examples in Cilin.



Aa01A01= 人 士 人物 人士 人氏 人选
Aa01A02= 人类 生人 全人类
Aa01A03= 人手 人员 人口 人丁 口 食指
Aa01A04= 劳力 劳动力 工作者
Aa01A05= 匹夫 个人
Aa01A06= 家伙 东西 货色 厮 崽子 兔崽子 狗崽子 小子 杂种 畜生 混蛋 王八蛋 竖子 鼠辈 小崽子
Aa01A07= 者 手 匠 客 主 子 家 夫 翁 汉 员 分子 鬼 货 棍 徒

FIGURE 2. Examples in Cilin

The similarity of two words is measured by the distance in the tree. Formally, it is defined as:

$$Sim_{cilin}(w_1, w_2) = 1 - \frac{pathlen(w_1, w_2)}{pathlen(w_1, Root) + pathlen(w_2, Root)} \tag{1}$$

where $pathlen(w_1, w_2)$ is the minimum path length of (w_1, w_2) to their common parent node and $Root$ represents the root of the tree [10].

5. **Approach.** The similarity between NCs (t_1, t_2) and (n_1, n_2) is calculated by the similarities of the component nouns. Formally, the similarity of the NC pair is defined as:

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

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

For each test NC, we calculate the similarities with all NCs in the training data. Then we choose the NC in the training data which has the highest similarity, and label the test NC according to the sematic relation associated with that training data. Formally, the semantic relation of test NC (t_1, t_2) is determined by :

$$Relation(t_1, t_2) = Relation(n_{i1}, n_{i2}) \tag{3}$$

where

$$i = \underbrace{\arg\max}_{i} Sim((t_1, t_2), (n_{i1}, n_{i2}))$$

Figure 3 shows the complete procedure of our method. Figure 4 illustrates how to compute the similarities between a test NC (t_1, t_2) and the NCs in the training data in detail. As can be seen, a 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 is determined by the training instance with the highest similarity.

6. Experiments and Evaluation.

6.1. **Data collection.** We retrieved Chinese two-word NCs from the People's Daily of 1998 and 2000 which are segmented and POS tagged. After excluding proper nouns and coordinate constructions, we finally get 1483 NCs for our experiment. The semantic relations of all the NCs are judged by two annotators who major in linguistics. Overall, we use 978 NCs for the training data and 505 NCs for the test data.

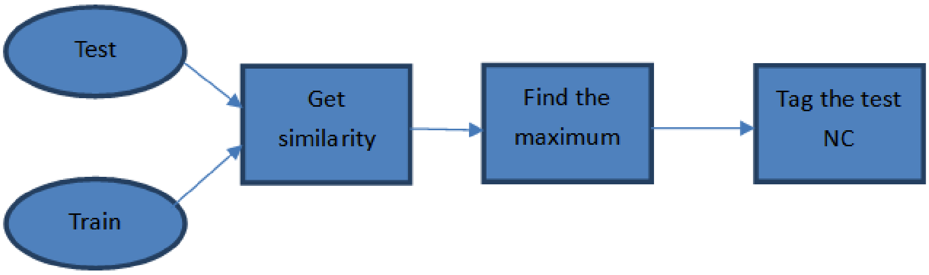
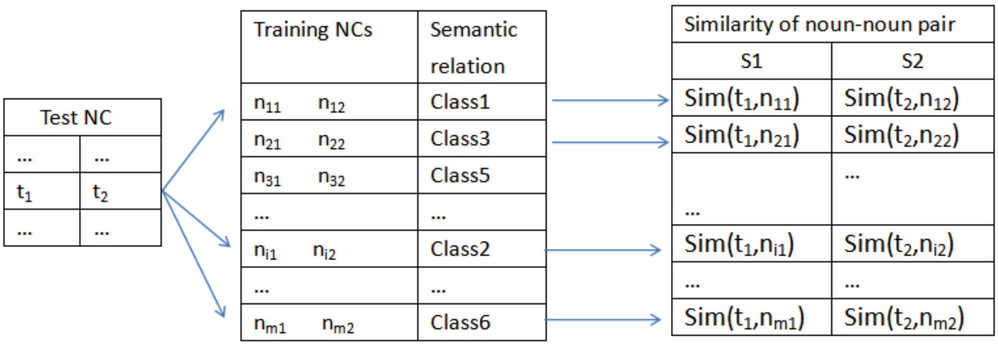


FIGURE 3. The procedure of the method



Test NC	
…	…
t_1	t_2
…	…

Training NCs		Semantic relation
n_{11}	n_{12}	Class1
n_{21}	n_{22}	Class3
n_{31}	n_{32}	Class5
…		…
n_{i1}	n_{i2}	Class2
…		…
n_{m1}	n_{m2}	Class6

Similarity of noun-noun pair	
S1	S2
$\text{Sim}(t_1,n_{11})$	$\text{Sim}(t_2,n_{12})$
$\text{Sim}(t_1,n_{21})$	$\text{Sim}(t_2,n_{22})$
…	…
$\text{Sim}(t_1,n_{i1})$	$\text{Sim}(t_2,n_{i2})$
…	…
$\text{Sim}(t_1,n_{m1})$	$\text{Sim}(t_2,n_{m2})$

FIGURE 4. Detailed similarities between the test NC and training NCs

6.2. **Experimental results.** We experiment with two similarity methods introduced in Section 4, assuming that the contribution of the head and modifier noun is equal ($\alpha = 0.5$). Table 4 shows the experimental results. Note that the similarities based on HowNet or Cilin both belong to dictionary-based methods. Thus, if the test word does not appear in HowNet or Cilin, our method cannot tag the test NC (i.e., unlabeled data) because of the lack of similarities. The performances based on HowNet and Cilin similarity are very close, and they can classify 35% NCs correctly.

TABLE 4. Accuracy based on HowNet and Cilin similarity

Similarity method	Unlabeled	#Correct (Accuracy)
HowNet	25	174 (34.46%)
Cilin	16	178 (35.25%)

TABLE 5. The most similar NCs based on two similar measures

Test NCs	The most similar NCs in the training data	
	based on HowNet similarity	based on Cilin similarity
残疾 儿童 (disabled children)	白内障 患者 (cataract patient)	白内障 患者 (cataract patient)
玻璃 茶几 (glass table)	水晶 花瓶 (crystal vase)	钻石 戒指 (diamond ring)
网络 医生 (network doctor)	因特网 用户 (Internet user)	出租车 司机 (Taxi driver)
蔬菜 收入 (vegetable income)	水果 价格 (fruit price)	水果 价格 (fruit price)
大学 校长 (university president)	中学 教师 (middle school teacher)	政府 领导 (government leader)

Table 5 lists some test NCs and the most similar NC found in the training data. As can be seen, our method can provide reasonable interpretation which is very useful in understanding a novel NC. For instance, if the reader does not know the meaning of a novel NC "network doctor", our method can provide some NCs such as "Taxi driver" which are easy to understand. It will help the reader to predict the semantic relation of the two nouns.

7. **Conclusion and Future Work.** We present a method for interpreting Chinese NCs based on word similarity. Experimental results show word similarity can provide useful information in solving the interpreting problems. In the future, we plan to use some corpus-based similarity methods such as word2vec to solve the OOV problem. What is more, the voting strategy can be used in determining the semantic relation of the test NCs since we only choose the NC with the highest similarity.

Acknowledgment. This work is supported by National Natural Science Foundation of China (61300152, 61300156 and 61402419).

REFERENCES

[1] P. Downing, On the creation and use of English compound nouns, *Language*, vol.53, no.4, pp.810-842, 1977.

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

[3] T. Finn, *The Semantic Interpretation of Compound Nominals*, Ph.D. Thesis, University of Illinois, Urbana, 1980.

[4] C. Butnariu and T. Veale, A concept-centered approach to noun-compound interpretation, *Proc. of the 22nd International Conference on Computational Linguistics*, pp.81-88, 2008.

[5] M. Wang, C. Huang, S. Yu and S. Kang, Chinese noun compound interpretation using verbal paraphrases, *ICIC Express Letters, Part B: Applications*, vol.5, no.5, pp.1377-1382, 2014.

[6] X. Wei, *Research on Chinese Noun Compound Interpretation for Semantic-Query*, Master Thesis, Peking University, 2012.

[7] G. Nunberg, I. A. Sag and T. Wasow, Idioms, *Language*, vol.70, no.3, pp.491-538, 1994.

[8] L. Wang and M. Wang, A study on the taxonomy of Chinese noun compounds, *The 16th Chinese Lexical Semantic Workshop*, 2015.

[9] Q. Liu and S. Li, Word similarity computing based on HowNet, *International Journal of Computational Linguistics & Chinese Language Processing*, vol.7, no.2, pp.59-76, 2002.

[10] Y. Jia, H. Zan, M. Fan, S. Yu and Z. Wang, Word relevance computation for noun-noun metaphor recognition, *Chinese Lexical Semantics*, pp.251-259, 2014.

[11] J. Pustejovsky, *The Generative Lexicon*, The MIT Press, 1995.

[12] M. Lapata, The disambiguation of nominalizations, *Comput. Linguist.*, vol.28, pp.357-388, 2002.

[13] D. Ó. Séaghdha, Designing and evaluating a semantic annotation scheme for compound nouns, *Proc. of the 4th Corpus Linguistics Conference*, 2007.

[14] S. Tratz and E. Hovy, A taxonomy, dataset, and classifier for automatic noun compound interpretation, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pp.678-687, 2010.

[15] N. S. Kim and T. Baldwin, Automatic interpretation of noun compounds using WordNet similarity, *The 2nd International Joint Conference on Natural Language Processing*, pp.945-956, 2005.

[16] Z. Dong and Q. Dong, *HowNet*, http://www.keenage.com.

[17] L. Yao, Z. Sui, Q. Zhao, Y. Hu and R. Wang, On automatic construction of medical ontology concept's description architecture, *International Journal of Innovative Computing, Information and Control*, vol.8, no.5(B), pp.3601-3616, 2012.

[18] Z. Sui, W. Kang and Y. Tian, Synchronously extracting instances and attributes for the concepts from the Web, *International Journal of Knowledge and Language Processing*, vol.3, no.3, pp.1-17, 2012.