Automatic Acquisition of Chinese Novel Noun Compounds

Meng Wang[*§], Chu-Ren Huang[§+], Shiwen Yu[*], Weiwei Sun[$]

[*]Key Laboratory of Computational Linguistics, Peking University, China

[§]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

[+]Institute of Linguistics Academia Sinica

[$]Department of Computational Linguistics, Saarland University & DFKI

E-mail: wm@pku.edu.cn, churen.huang@inet.polyu.edu.hk, yusw@pku.edu.cn, wsun@coli.uni-saarland.de

Abstract

Automatic acquisition of novel compounds is notoriously difficult because most novel compounds have relatively low frequency in a corpus. The current study proposes a new method to deal with the novel compound acquisition challenge. We model this task as a two-class classification problem in which a candidate compound is either classified as a compound or a non-compound. A machine learning method using SVM, incorporating two types of linguistically motivated features: semantic features and character features, is applied to identify rare but valid noun compounds. We explore two kinds of training data: one is virtual training data which is obtained by three statistical scores, i.e. co-occurrence frequency, mutual information and dependent ratio, from the frequent compounds; the other is real training data which is randomly selected from the infrequent compounds. We conduct comparative experiments, and the experimental results show that even with limited direct evidence in the corpus for the novel compounds, we can make full use of the typical frequent compounds to help in the discovery of the novel compounds.

1. Introduction

Noun compounds are very common in most texts including press and technical materials, newswire and fictional prose. The properties of compounds have been extensively studied in linguistics literature especially in English. It is known that compounding as a concatenation of words is commonly used to pack meaning into a minimal amount of linguistic structure. Examples (a)-(d) are English noun compounds, where (a) and (b) are binary (composed of two words) compounds.

(a). stone fish
(b). party animal
(c). emergency bus fuel
(d). killer whale attack

Although noun compounds are common in English, they are by no means limited to one language. A substantial body of work has investigated the noun compounds in many languages, such as Chinese, Japanese, French, German and Italian. There is a variety of definitions of noun compounds employing different criteria (Quirk et al., 1985; Chomsky & Halle, 1991; Levi, 1978). A more restricted and functional definition proposed by Downing (1977) is adopted in this paper: A noun compound is a sequence of two or more nouns that functions as a single noun. See examples (e)-(g) for Chinese noun compounds following this definition.

(e). 水果 价格 (fruit price)
(f). 爱情 故事 (love story)
(g). 空气 质量 问题 (air quality issue)

One of the most significant properties of noun compounds is productiveness. New compounds are created from day to day, particularly in rapidly updating fields. The acquisition of noun compounds is very important for applications, such as machine translation. The translation quality of noun compounds will affect the performance of the whole system because most compounds are not compositional. An automatic and quantitative method for acquiring noun compounds is needed since novel compounds cannot be generated independently. In this paper, we focus on the automatic acquisition of novel (infrequent) Chinese noun compounds from the corpora, and restrict our attention on compounds formed by two consecutive nouns with the *modifier-head* relationship (see examples (e)-(f)).

Noun compounds acquisition is usually subsumed to the problem of identifying terms or collocations from the corpora, in which many statistical approaches are used. Novel compounds pose a great challenge to the acquisition task when statistical approaches are applied. It is well known that the basic assumption in the statistical approaches is that two lexically associated words tend to co-occur more often than expected on the basis of their individual occurrence frequencies (Church and Hanks 1989). This requires that candidate compounds will occur frequently in the corpus. Unfortunately, the novel compounds cannot meet such requirement and statistical tests do not necessarily give the correct prediction for them.

In this paper, we model this problem as a two-class classification problem in which a candidate compound is either classified as a compound or a non-compound. A SVM classifier is used for the classification. For comparison, two different methods are used to get the training data. First, we use three statistical scores to get the training data from the frequent candidate compounds. It is a kind of "virtual" data in a sense that training data and test data come from the different populations (i.e. one is the frequent, the other is the infrequent). Second, we randomly select the training data from the infrequent

candidate compounds. It can be viewed as the "real" data because the training data and test data come from the same population (i.e. the infrequent). We conduct comparative experiments, and the results show that the model using "virtual" data performs better than that of "real" data. This suggests the frequent noun compounds can provide useful information for the novel noun compounds acquisition when very little direct evidence is found in the corpus.

The rest of this paper is organized as follows. Section 2 shows the methodological data we experimented with. Section 3 introduces how we collect the two types of training data. Section 4 explains the experiments and the results. Section 5 concludes with a discussion of the results.

2. Methodological data

The corpus we used in the experiment is a half year collection of *People's Daily* in 1998 segmented and POS-tagged. We use a two-step procedure to extract noun sequences of length two as candidate compounds.

1. Look for the consecutive bigrams of nouns which are not preceded or succeeded by a noun in order to avoid selecting noun pairs in a larger compound (e.g., "空气 质量 问题", air quality issue);
2. Bigrams containing letters or digitals (e.g., "PC机原理", principle of PC) are filtered out.

This procedure yielded a total number of 127,701 tokens which consist of 48,283 distinct types of candidate noun compounds.

Co-occurrence Frequency	**>=5**	**>=2**	**>=1**	**=1**
No. of types	4,140	13,886	48,283	34,397

Table 1 Distribution of candidate compounds of length two

Table 1 shows a close inspection of the distribution of noun compounds of length two. Majority (more than 71%) of the candidate compounds occur only once, while those frequent compounds (co-occurrence frequency larger than 5) account for less than 10%. The distribution in our data is predicted by Zipf's law. Such distribution shows that it is inevitable to deal with the hapaxes for the acquisition of noun compounds.

3. Training data preparation

Statistical scores have been widely used for the acquisition of terms and collocations. We explore three statistical scores: Co-occurrence Frequency (CoocF), Mutual Information (MI) and Dependent Ratio (DR) to get the virtual training data from frequent candidate compounds.

3.1 Co-occurrence Frequency (CoocF)

High repetition of a sequence of words implies some relations between them. Previous work in terminology acquisition has shown that CoocF is a good indicator of the termhood of word sequence (Justeson & Katz, 1995). For noun compound acquisition, CoocF has also proved to work better than the other statistical scores such as mutual information in English (Lapata 2000). Table 2 shows the samples of candidate compounds with the highest CoocF.

Candidate compounds	**CoocF↓**	**MI**	**DR**
领导 干部 (leader cadre)	1106	5.7583	0.2867
人民 群众 (people)	911	4.6613	0.1572
国际 社会 (international society)	544	4.0037	0.0963
*国 关系 (country relation)	510	4.6064	0.1226
金融 机构 (financial institution)	415	5.1144	0.1939

Table2 Samples of candidates with the highest CoocF[1]

3.2 Mutual Information (MI)

Mutual Information, as a measure of word association (Church & Hanks, 1989), has been widely used in collocation and term extraction. We use mutual information to compare the probability of observing noun *n1* and *n2* together (joint probability) with the probabilities of observing *n1* and *n2* independently (chance). The bigram mutual information is defined as (1):

$$\mathrm{I}(n1,n2) = \log_2 \frac{P(n1,n2)}{P(n1)P(n2)} \quad (1)$$

Here, P(*n1*) and P(*n2*) are estimated by the number of occurrence of *n1* and *n2* divided by the size of the corpus *N*. P(*n1*,*n2*) is the number of times *n1* and *n2* co-occur divided by *N*.

Candidate compounds	**CoocF**	**MI↓**	**DR**
脊髓 灰质炎 (Poliomyelitis)	7	13.0646	0.8660
*厂规 厂纪 (factory regulations and disciplines)	9	12.5819	0.6667
鲨鱼 软骨素 (Shark Cartilage)	8	12.4866	0.7416
算术 平均数 (Arithmetic mean)	8	12.3996	0.5164
交响诗 大合唱 (chorus of symphonic poem)	6	12.1765	0.4000

Table3 Samples of candidates with the highest MI

3.3 Dependent Ratio (DR)

As mentioned before, compounding is a highly productive language phenomenon. In fact, there are some very productive modifiers or heads which can appear in different compounds. So we propose the dependent ratio to measure the dependency of the components in

[1] The asterisk means it is not a valid compound. The symbol ↓ represents the descending order and ↑ represents the ascending order.

compounds. For a candidate compound *AB* in question (both *A* and *B* are nouns), the dependent ratio is defined to be

$$\mathrm{DR}(AB) = \sqrt{\mathrm{LD(B)} \times \mathrm{RD(A)}} \quad (2)$$

And the definition of LD and RD is:

$$LD(X) = \frac{\max_{w \in Left(X)} freq(wX)}{freq(X)} \quad (3)$$

$$RD(X) = \frac{\max_{w \in Right(X)} freq(Xw)}{freq(X)} \quad (4)$$

where freq(*X*) is the frequency of *X*, Left(*X*) and Right(*X*) represent a set of all the left adjacent words and right adjacent words of *X* respectively, and w represents an element in the set. Table 4 shows some samples of candidate compounds in ascending order by DR values.

Candidate compounds	DR↑	CoocF	MI
议长 先生 (Mr. Speaker)	0.0320	4	4.1814
全省 系统 (the system of the province)	0.0348	1	1.2392
记者 先生 (Mr. Reporter)	0.0385	1	-1.0812
总统 先生 (Mr. President)	0.0390	12	2.7772
计算机 系统 (computer system)	0.0392	15	4.3845

Table 4 Samples of candidates with the lowest DR

3.4 Training data collection

For each statistical score, we choose the top 700 candidate compounds according to its respective ranks. After which, we get three groups of candidate compounds. All of them are manually classified within their contexts: whether a candidate is a compound or not. We find that the three groups of candidates above actually have very little overlap, which means every statistical score can cover a certain type of compounds. We put these samples together to form the first part of "virtual" training data. However, the collected data is not a balanced one in the sense that they contain too much positive samples (valid compounds).

An appropriate score should assign higher values to valid compounds and lower values to non-compounds when it is used in the acquisition task. So we can get the negative samples (non-compounds) from those with lower statistical values. MI is used to select the negative samples as the second part of "virtual" training data. These two parts of data are combined to produce the final "virtual" training data which contains 2518 samples.

As mentioned before, the "virtual" training data are actually extracted from the frequent candidates, while the test data are novel compounds with CoocF equal to one. Obviously, the two data sets come from different populations and the distribution of this kind of "virtual" data is very different from the test data. It is necessary to prepare another kind of "real" training data which comes from the same population with the test data. So we randomly selected 2518 candidate compounds which occur only once in the corpus. It is the final "real" training data.

4. Experiments using SVM

4.1 Features

Semantic features and character features are shown to work effectively in the acquisition task (Wang & Huang, 2010). Semantic features capture meaning regularity in the compounding process. For example, *木头 家具 (wood furniture)*, *金属 盆子 (metal basin)* and *玻璃 瓶子 (glass bottle)* can be viewed as the results of semantic combination of substance and artifact. A way to obtain such information is to use the concepts which the nouns represent in a taxonomy. The Chinese Semantic Dictionary (CSD) is used to provide such semantic knowledge. CSD (Wang et al., 2003) is a large machine-readable dictionary containing a large amount of semantic information such as semantic hierarchy and collocation features for 37675 nouns.

Both unigram and bigram semantic features are extracted: the unigram features include each noun's semantic category; and the bigram feature is the combination of the two nouns' semantic categories. If the noun does not appear in the dictionary, we use "NULL" as the feature.

Character feature is another important type of feature which comes from two aspects: one is the word itself, and the other derives from the word formation. In Chinese, the nouns which denote the same kind of things are always *modifier-head* construction, and most of them have the same headword. For example, *松树 (pine tree)*, *柳树 (willow tree)* and *桃树 (peach tree)* are different kinds of trees, and they have the same headword *树 (tree)*. Using headwords can avoid data sparseness in a sense that it actually acts as a backward method. Both unigram and bigram character features are used for the experiments.

4.2 Experiments

We randomly choose 570 un-supported candidate compounds which are attested only once in the corpus as test data. The two types of features are used in the LibSVM package tool. In the experiments, we use the linear kernel and other default parameters. Table 5 reports the classification performance of two kinds of training data using different features.

Training Data	Features	Accuracy (%)
virtual	Character	80.35
	Character+Semantic	80.70
real	Character	78.77
	Character+Semantic	78.07

Table 5 Performances of virtual and real training data using different features

4.3 Analysis

As can be seen, given the same amount of training data, the model using "virtual" data performed better than the "real" data. This result suggests that the "virtual" training data obtained by the statistical scores are more "informative" than the randomly selected data.

Data sparseness is the biggest factor which hinders the performance of classification. The words in the "virtual" training data coming from the frequent candidates are often used to form the new compounds. While for the real training data coming from the infrequent candidates, the data sparseness problem is aggravated. It is partly the reason why the "virtual" data outperforms the "real" data. To some extent, this kind of sampling method is similar to the active learning which always selects the most informative samples as the training data instead of selecting randomly.

5. Conclusion and Discussion

Novel compounds acquisition is a very challenging task since many statistical scores cannot be applied reliably on them. In this paper, we model this task as a two-class classification problem which decides whether a candidate is compound or not. We use two types of training data which come from different populations. One is the "virtual" training data obtained by three statistical scores, the other is "real" training data which are randomly selected from the infrequent candidate compounds. The experimental results show that the model using "virtual" data has a better performance than that of "real" data. This result shows that frequent compounds are very useful for novel compound acquisition.

Even with limited direct evidence in the corpus for the novel compounds, we can make full use of the statistical scores to get the typical frequent compounds to help discovering the infrequent ones.

In the future, we will explore other statistical scores to enlarge the training data. In addition, we will investigate how to combine these data to form a balanced training data for the classification.

References

Chen,Yirong, Qin Lu, Wenjie Li, Zhifang Sui, Luning Ji. (2006). A Study on Terminology Extraction based on Classified Corpora. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa. Italy.

Chomsky, N. and Halle, M. (1991). *The Sound Pattern of English*. 2nd Edition. MIT Press, Cambridge, MA.

Church, K. and Hanks P. (1989). Word Association Norms, Mutual Information and lexicography. *In Proceedings of the 27th Annual Meeting of the Association for computational Linguistics*, Vancouver, Canada.

Downing, Pamela. (1977). On the Creation and Use of English Compound Nouns. *Language*,53, No 4, 810—842.

Fazly, Afsaneh and Paul Cook and Suzanne Stevenson. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*. Volume 35 (1): 61-103.

Justeson, John S and Slava M Katz. (1995). Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* (1), 9-27.

Lapata, Maria and Alex Lascarides. (2003). Detecting Novel Compounds: The Role of Distributional Evidence. *In Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics*.

Lapata, Maria. (2000). *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. PhD Thesis. University of Edinburgh.

Lauer, Mark. (1996). *Designing Statistical Language Learners: Experiments on Compound Nouns*. Ph.D. thesis, Macquarie University. Australia.

Lauer, Mark. (1995). Corpus statistics meet the noun compound: some empirical results, *33rd annual meeting of the Association for Computational Linguistics.*

Levin, Judith. (1978). *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Manning, Christopher D., and Hinrich Schutze. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Miller, George A, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, New York, NY.

Su, Keh-Yih, Ming-Wen Wu and Jing-Shin Chang. (1994). A corpus-based approach to automatic compound extraction, *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, Las Cruces, New Mexico

Wang, Hui, Weidong Zhan and Shiwen Yu.(2003). The Specification of The Semantic Knowledge-base of Contemporary Chinese, *Journal of Chinese language and computing*, Vol. 13(No. 2)：159-176.

Wang, Meng and Chu-Ren Huang. (2010). Acquisition of Chinese Novel Noun Compounds. *The Second International Conference on Global Interoperability for Language Resources*, HongKong.

Yu, Shiwen, Huiming Duan and Xuefeng Zhu et al. (2002). The Basic Processing of Contemporary Chinese Corpus at Peking University SPECIFICATION. *Journal of Chinese Information Processing*, 16(5):49-64,(6):58-65.

Zipf, George Kingsley. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge (Mass.).