Improving Noun Compounds Identification with Verb-Centered Dependency Relations

Ruiji FU[†], Wei ZHANG, Bing QIN, Ting LIU

Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001, China

Abstract

This paper presents a supervised approach for identifying Chinese noun compounds (NCs) in context. Due to the lack of morphological inflections, quite a few Chinese noun compounds contain verbs without any inflections. Therefore, it is important for NC identification to distinguish whether a verb is in an NC or not. In this paper, verb-centered syntactic dependency relations are used as features to improve NC identification. The results show that our approach outperforms the baseline, a supervised model based on Conditional Random Fields (CRF) without dependency relation features.

Keywords: Noun Compounds Identification; dependency relations; CRF

1. Introduction

Noun compounds (NCs) identification is a task of identifying NCs in context, which makes an important role in natural language processing (NLP). Downing (1977) [1] defines a noun compound as any consecutive sequence of nouns that functions as a noun. However, many words can be used as nouns in some context and as verbs or other part of speech in others without any word form inflections. Church (1988) [2] addresses that the ambiguities of NC boundary identification are usually caused by verbs. Due to the lack of morphological inflections, this phenomenon is more common in Chinese. For example, *信息抽取* (*Information Extraction*) is an NC which is composed of a noun *信息* (*information*) and a verb *抽取* (*extract, there is no inflection in Chinese.*). It makes Chinese NC identification more difficult. Therefore, it is helpful that distinguishing whether a verb is in an NC or not.

To distinguish whether a verb is in an NC or not, syntactic dependency relations can be used. Figure 1 gives two examples, (a) *中国驻俄罗斯大使馆* (*Chinese embassy in Russia*) is an NC, but (b) *北约打击利比亚* (*NATO strikes Libya*) is not. The two NCs both contain verbs, *驻* (*station*) in (a) and *打击* (*strike*) in (b). It is difficult to distinguish whether the verbs in an NC according to the lexical and POS features. But we can use dependency relations. In (a), *驻* (*station*) has an attribute relation from *大使馆* (*embassy*), which indicates it is within an NC. In (b), *打击* (*strike*) is the predicate, which indicates it is out of an NC.

In our experiments, we use the verb-centered dependency relations as features for a CRF model to identify NCs. We compare our method with the baseline – a CRF model with common features. The results show that these features can improve NC identification.

The contributions of this paper are: (1) we import verb-centered dependency relations to a supervised

[†] Corresponding author.
Email addresses: rjfu@ir.hit.edu.cn (Ruiji FU).

model and improve the performance of NC identification; (2) we propose two formulae to measure the closeness between verb-centered dependency relations and NCs.

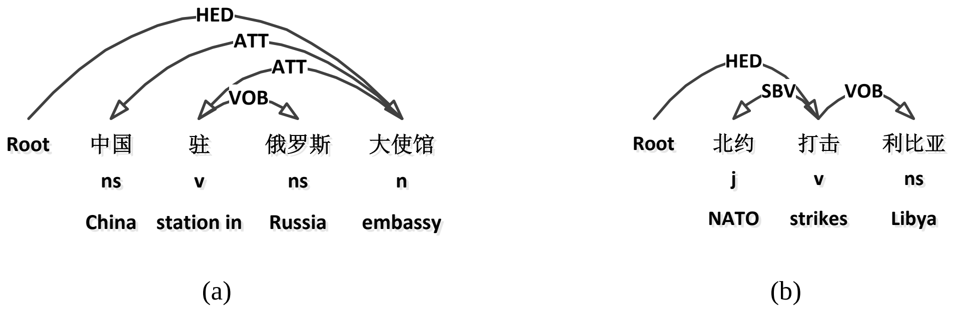

Fig.1 Examples of Verb-Centered Dependency Relations

This paper is organized as follows. Section 2 discusses the related work. Section 3 describes our approach in detail. Section 4 presents and discusses the results of our experiments. Finally, we present our conclusions and future work in section 5.

2. Related Work

In this section, we introduce some previous work about NC identification, including supervised and unsupervised methods.

Zhou et al. (2003) [3] build a maximum entropy model to recognize Chinese and English BaseNP. Sha and Pereira (2003) [4] employ a CRF model to identify BaseNPs. They use only lexical features and POS features. BaseNPs are similar with NCs, but more common. NC identification is more difficult than BaseNP identification because it suffers from data sparse problem more seriously. This paper tries to improve the performance of NC identification by adding dependency relation features.

Reiter and Frank (2010) [5] explore a corpus-based learning approach for identifying generic noun phrases, using selections of linguistically motivated features. The features include NP-level features, sentence-level features, syntactic features and semantic features. However, this work focuses on English NC. Chinese NC identification is more difficult because of the lack of morphological inflections and the uncertainty in word segmentation.

Besides these supervised methods, some researchers also try unsupervised methods.

Liberman and Sproat (1992) [6] present an adjacency algorithm. Their proposal involves comparing the mutual information between the two pairs of adjacent words and bracketing together whichever pair exhibits the highest. Lauer (1995) [7] proposes a dependency model to identify the boundaries of NCs and compared with the adjacency model. The results show that the dependency model is more accurate than the adjacency model.

Zhang et al. (2000) [8] proposes a method based on mutual information and context dependency to extract Chinese compound words from a very large corpus. They select NC candidates according to mutual information and filtered them by context dependency.

3. Our Approach

In this section we describe our approach of Chinese NC identification. Firstly, the definition of Chinese NC is presented. Then, the CRF model is described. At last, we mainly introduce the verb-centered dependency relation features.

3.1. The Definition of Noun Compounds

Wang et al. (2010) [9] consider a NC is composed of some modifiers and a head. The head is usually the last word in an NC. Modifiers are before the head word. Following this notion and combining the definition of BaseNP (Zhao and Huang, 1999) [10], we defined a Chinese NC as follows:

NC = Modifiers + Head
Head → Noun | Nominalized verb | Alphanumeric string
Modifier → Noun | Verb | Adjective | Numeral | Quantifier | Alphanumeric string

According to this definition, *中国航空工业总公司* (*China Aviation Industry Corporation*), *CDMA 系统* (*CDMA system*), *中国人民解放军三〇二医院* (*Chinese People's Liberation Army 302 Hospital*), *经济发展* (*economic development*), and so on are all NCs.

3.2. Model Description

We treat NC identification as a sequence labeling problem and make use CRF model to solve it. The inference of CRF is that given an observable sequence $\vec{x}$, we want to find the most likely set of labels $\vec{y}$ for $\vec{x}$. The probability of $\vec{y}$ given $\vec{x}$ is calculated as follows (J.Lafferty et al., 2001) [11]:

$$p(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})}\prod_{j=1}^{n}\psi_j(\vec{x},\vec{y}) \tag{1}$$

$$Z(\vec{x}) = \sum_{\vec{y}'}\prod_{j=1}^{n}\psi_j(\vec{x},\vec{y}') \tag{2}$$

$$\psi_j(\vec{x},\vec{y}) = exp\left(\sum_{i=1}^{m}\lambda_i f_i(\mathrm{y}_{j-1},y_j,\vec{x},j)\right) \tag{3}$$

In formulae 2, 3 and 4, *j* denotes the index of the *j*th word in sequence $\vec{x}$. *n* denotes the length of $\vec{x}$. *m* denotes the number of the features.

Five tags B, I, E, S and O are imported to transform NC identification into a sequence labeling problem. The meanings of the tags are shown in Table 1.

Table 1 The Description of Tags

Tags	Description
B	Beginning of an NC
I	Inner of an NC
E	End of an NC
S	Single-word NC
O	Out of NCs

For example:

[美国 股票 市场] 经历 “ [黑色 星期四] ” ……

(*[US stock market] experienced ‘[Black Thursday]’ ...*)

is labeled as:

美国(US)/B 股票(stock)/I 市场(market)/E 经历(experienced)/O “/O 黑色(Black)/B 星期四(Thursday)/E ”/O ……

The input of the model is the results of Chinese segmentation and POS tagging. The features usually include lexical features and POS features.

3.3. Verb-Centered Dependency Relation Features

Since the ambiguities of NC boundary identification are usually caused by verbs, we add verb-centered syntactic dependency relations as features. We reduce these features into four values:

- **NV**: This word is not a verb;
- **VI**: This word is a verb and it can be in an NC according to its dependency relations;
- **VO**: This word is a verb and it cannot be in an NC according to its dependency relations;
- **VU**: This word is a verb and we cannot judge it can be in an NC or not according to its dependency relations;

Algorithm VerbDependency_judge(rawSentence, supValue, desupValue)

```
Input:    A raw sentence with segmentation and pos tags rawSentence，the threshold support value supValue,
            the threshold desupport value desupValue
Output:   A processed sentence with verb dependency conversion feature value proSentence
i = 0, wordNum = len(rawSentence);    // the number of word in rawSentence
while i < wordNum do,
    wordi = rawSentence.get(i);
    if wordi.pos = ‘v’ then
        wordj = rawSentence.get(wordi.reId); //get the words that wordi modifies
        rel = getRel(wordi, wordj); //get the type of wordi←wordj dependency relation
        voteValue = valueRelIn(rel); //calculate the value of wordi←wordj
        for k = 0 to k = wordNum-1 do,
            wordk = rawSentence.get(k);
            if wordk.reId = i then
                rel = getRel(wordk, wordi); //get the type of wordi→wordk dependency relation
                voteValue += valueRelOut(rel); //calculate the value of wordi→wordk
        if voteValue <= supValue and voteValue >=desupValue then,
            wordi.newfeature = “VU”;
        elif voteValue > supValue then,
            wordi.newfeature = “VI”;
        else
            wordi.newfeature = “VO”;
    else
            wordi.newfeature = “NV”;
    proSentence.append(wordi);
return proSentence
```

Fig.2 The Algorithm of Verb-Centered Dependency Relation Features Value Judgement

We compute the scores of all kinds of verb-centered dependency relations by formulae 4 and 5.

$$\text{valueRelIn}(rel) = \frac{\#\text{ of }\overrightarrow{rel}\text{ in NCs}}{\sum_{\vec{r}}\#\text{ of }\vec{r}\text{ in NCs}} - \frac{\#\text{ of }\overrightarrow{rel}\text{ out of NCs}}{\sum_{\vec{r}}\#\text{ of }\vec{r}\text{ out of NCs}} \tag{4}$$

$$\text{valueRelOut}(rel) = \frac{\#\text{ of }\overleftarrow{rel}\text{ in NCs}}{\sum_{\overleftarrow{r}}\#\text{ of }\overleftarrow{r}\text{ in NCs}} - \frac{\#\text{ of }\overleftarrow{rel}\text{ out of NCs}}{\sum_{\overleftarrow{r}}\#\text{ of }\overleftarrow{r}\text{ out of NCs}} \tag{5}$$

Here, $\overrightarrow{rel}$ denotes the relationship from another word to a verb. $\overleftarrow{rel}$ denotes the relationship from a verb to another word. We use the two values valueRelIn(rel) and valueRelOut(rel) to measure the closeness between the verb-centered relation *rel* and NCs. If a relationship usually appears in NCs rather than out of NCs, the value is positive, such as attribute relationship (ATT). Conversely, the value is negative, such as head relationship (HED).

We consider all of the relationship of a verb and other words and sum all the values to get the score of the verb. Then we compare the score with two thresholds *supValue* and *desupValue*. If the score is greater than *supValue*, the feature value is set as VI. If the score is lesser than *desupValue*, the feature value is set as VO. Else, the feature value is set as VU. In the experiments, we set the two thresholds *supValue* and *desupValue* zero experientially.

4. Experiments

In our experiments, we train CRF models on a corpus and test the effect of the verb-centered dependency relation features. The experiment results are analysed in detail.

4.1. The Data Set and Experimental Setup

The training data is generated from HIT-SCIR Dependency Treebank[1] by some heuristic rules. The dependency treebank is annotated manually with 14 kinds of dependency relations, which contains 60,000 sentences from the People's Daily. The training data contains 90,203 NCs.

For evaluation, we manually label 1,000 sentences from the People's Daily as the test data in the same domain (general news domain) and 1,000 sentences from finance domain as the test data in a different domain. The test data in the general news domain contains 1,805 NCs. The test data in the finance domain contains 3,192 NCs. From these amounts, we can infer that there is difference between the distributions of the two test data sets.

HIT-SCIR Language Technology Platform (LTP)[2] is used to do the Chinese segmentation, POS tagging and dependency parsing. The results of LTP are the input of the NC identification model.

The CRF model of Sha and Pereira (2003) [4] is employed as the baseline. The features include lexical features and POS features. The size of the feature window is 5. We add verb-centered dependency relation features to the baseline and build a new model. The baseline and our approach are compared on the two test data sets respectively.

[1] http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=147&MMN_position=52:48

[2] http://ir.hit.edu.cn/demo/ltp/

4.2. Experimental Results and Analysis

We calculate the $\mathrm{valueRelIn}(rel)$ and $\mathrm{valueRelOut}(rel)$ according to formulae 4 and 5 on the training data. The results are shown in Table 2. We can see that the relation ATT gets the greatest positive values, and HED gets the least negative value. The results are reasonable. ATT relations often exist in NCs. But if a verb has HED relation, it is the predicate of the whole sentence and might not be in an NC. The values can reflect the closeness between the relations and NCs. It is accordant with the analysis in section 3.3. We use these values to compute the verb-centered dependency relation feature values.

Table 2 The Values of Verb-Centered Dependency Relation

Relation Tags	**Description**	**ValueRelIn**	**ValueRelOut**
SBV	subject-verb	-0.008	0.015
VOB	verb-object	-0.105	-0.036
IOB	indirect-object	0	-0.001
FOB	fronting-object	0.003	0.082
DBL	double	0	-0.010
ATT	attribute	**0.722**	**0.184**
ADV	adverbial	-0.078	-0.116
CMP	complement	-0.040	-0.025
COO	coordinate	-0.204	-0.043
POB	preposition-object	-0.019	-0.006
LAD	left adjunct	-0.001	0.015
RAD	right adjunct	-0.004	-0.029
IS	independent structure	0	0
HED	head	**-0.266**	0

Table 3 compares the baseline and our approach on the test data in different domains. As the table shows, we get a raise (0.83%) of recall and a little drop (0.32%) of precision on the general news test data set by adding the verb-centered dependency relation features. And the F1-score gets a 0.27% raise. On the finance test data set, the raise of F1-score is greater (0.59%). The results show that our new features can help to improve Chinese NC identification.

Through the error analysis, we find that our verb-centered dependency relation features tend to help to recall more NCs, which might be missed by the baseline. However, the contribution for NC boundary identification is not very strong.

Table 3 The Results of NC identification

Domains	**Features**	**Results**		
		P	**R**	**F1**
General News	Baseline	**89.82%**	86.54%	88.15%
	Baseline+dep	89.50%	**87.37%**	**88.42%**
Finance	Baseline	**79.47%**	70.11%	74.50%
	Baseline+dep	78.17%	**72.24%**	**75.09%**

From Table 3, we can also see that the performance of models on general news domain is much better than the performance on finance domain. The reason is that the distribution of the training data and the distribution of the general news domain test data are similar. While, the distribution of the finance domain test data is quite different from the training data.

To measure the coverage of verb-centered dependency relation features, we count the ratio of the NCs that contain verbs from the test data sets. As shown in Table 4, there are 15.01% of NCs containing verbs in the general news domain, and there are 39.63% of NCs containing verbs in the finance domain. NCs in finance domain contain more verbs. For example, some familiar finance NCs *控股/v 股东/n* (*controlling shareholder*), *上市/v 公司/n* (*quoted company*), *商品房/n 施工/v 面积/n* (*commodity house floor space under construction*), and so on all contain verbs. This is why finance domain data is affected much more by the new features as Table 3 shows. And greater ratio of NCs containing verbs also leads the worse performance in finance domain.

Table 4 The Ratio of NCs Containing Verbs

Data Set	**# of all NCs**	**# of NCs containing verbs**	**Ratio**
General News	1,805	271	15.01%
Finance	3,192	1,265	39.63%

5. Conclusion and Future Work

In this paper, we proposed a CRF model to identify Chinese NCs and made use of verb-centered dependency relation features to improve the performance of the model. We compute the scores of all kinds of verb-centered dependency relations from the training data. Then the features are reduced into four values according to the scores and imported to the model. The results showed that the verb-centered dependency relation features can help to improve NC identification on both the same domain and cross-domain test data sets.

In the future work, we will try to improve NC identification by using huge human knowledge bases from the web. Moreover, we will try to generate more training data from other resources automatically, such as parallel corpora or web resources.

References

[1] Pamela Downing. On the Creation and Use of English Compound Nouns. *Language*, 53(4) , pages 810-842, 1977.

[2] Kenneth Ward Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, pages 136-143, 1988.

[3] Yaqian Zhou, Yikun Guo, Xuanjing Huang and Lide Wu. Chinese and English BaseNP Recognition Based on a Maximum Entropy Model. *Journal of Computer Research and Development*. 40(3), pages 440-446, 2003.

[4] Fei Sha and Fernando Pereira. Shallow Parsing with Conditional Random Fields. In *Proceedings of the Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL),* pages134-141, 2003.

[5] Nils Reiter and Anette Frank. Identifying Generic Noun Phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, Uppsala, Sweden, pages 40-49, 2010.

[6] Liberman, M. and Sproat, R. The Stress and Structure of Modified Noun Phrases in English. In Sag, I. and Szabolcsi, A. (eds.), Lexical Matters - csli Lecture Notes No.24, University of Chicago Press. pages 131–81, 1992.

[7] Mark Lauer. Designing Statistical Language Learners: Experiments on Noun Compounds. PhD thesis, Macquarie University, pages 25-51, 1995.

[8] Jian Zhang, Jianfeng Gao and Ming Zhou. Extraction of Chinese Compound Words: An Experimental Study on a Very Large Corpus. In *Proceedings of the Second Workshop on Chinese Language Processing*, Hong Kong, pages 132-139, 2000.

[9] Meng Wang, Churen Huang, Shiwen Yu and Bin Li. Chinese Noun Compound Interpretation Based on Paraphrasing Verbs. *Journal Of Chinese Information Processing*, 24(6), pages 3-9, 2010.

[10] Jun Zhao and Changning Huang. The Model For Chinese BaseNP Structure Analysis. *Chinese Journal of Computers,* 22(2), pages 141-146, 1999.

[11] John Lafferty, Andrew McCallum and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, pages 282-289, 2001.