A Machine Learning Approach to Determine Semantic Dependency Structure in Chinese [*]

Jiajun Yan and **David B. Bracewell** and **Fuji Ren** and **Shingo Kuroiwa**
Department of Information Science and Intelligent Systems,
Faculty of Engineering, The University of Tokushima
Tokushima 770-8506, JAPAN

Abstract

In this paper, we attempt to automatically annotate the Penn Chinese Treebank with semantic dependency structure. Initially a small portion of the Penn Chinese Treebank was manually annotated with headword and semantic dependency relations. An initial investigation is then done using a Naive Bayesian Classifier and some handcrafted rules. The results show that the algorithms and proposed approach are effective at determining semantic dependency structure automatically. The Naive Bayesian Classifier makes a good baseline algorithm for future research.

Introduction

In natural language processing, semantic dependency structure is a practical approach to semantic representation, knowledge acquisition and machine translation. Text annotated with semantic dependency structure can make implicit knowledge in documents more explicit and thus the annotated documents will provide an easy way of processing knowledge extraction. In English, a lot of research has been done in semantic parsing using statistical and machine learning methods (Gildea & Jurafsky 2002) to semantically annotated corpora such as FrameNet (Johnson & Fillmore 2000) and the proposition Bank (Palmer, Gildea, & Kingsbury 2005) in recent years. So far much of the research has been focused on English due to the lack of semantically annotated resources in other languages.

For Chinese, automatic and manual annotation of semantic information, sememe variation, and validation of the corpus is underway. (Gan & Wong 2000) have annotated a subset of the Sinica balanced corpus with semantic dependency relations as defined in HowNet. (Li *et al.* 2003) reported that they annotated a 1,000,000-word-scale Chinese corpus with semantic dependency structure manually. However, corpora with semantic information are still scarce for Chinese NLP researchers due to the fact that such corpora, like the above mentioned, are rarely publically available.

After annotating the corpus with syntactic information, the issue becomes what kind of information will be needed and how to define the granularity of the word sememe and relations between words in the context. How to get the semantic information is also still a problem.To align or to specify the semantic structure is more difficult. (Yang & Li 2002) pioneered structural disambiguation at the same time of solving word sense disambiguation by using sememe co-occurrence information in sentences from a large corpus and transferring the information to restricted rules for sense disambiguation.

(Xue & Palmer 2003) (Xue & Palmer 2005) reported results on semantic role labeling for Chinese verbs using a pre-release version of the Chinese Proposition Bank. They reported that results on experiments using the handcrafted parses in the Penn Chinese Treebank were slightly higher than the results reported for the state-of-the-art semantic role labeling systems for English, even though the Chinese Proposition Bank is smaller in size. However, they were only focused on verbs and in this paper we look at all parts-of-speech. (Yan *et al.* 2005) reported a method for specifying semantic structure for NPs. First, they performed a shallow parse to extract all the possible NPs from the segmented data. Then they matched the syntactic structure of the information structure of HowNet to the possible NP, if an NP matched with more than one semantic structure, the word-similarity between the possible NP and the multiple candidate semantic structures would be calculated.

Auto-tagging Chinese corpora with semantic dependency structure is still a difficult problem. In this paper, our aim is to try to automatically annotate the semantic dependency structure for the Penn Chinese Treebank (Xia *et al.* 2000). Initially a small portion of the Penn Chinese Treebank was manually annotated with headwords and dependency relations. A Naive Bayesian Classifier with varying features was then adopted to learn the relations. Finally, a rule-based system was created based on features of Chinese to solve some problem patterns that were found in the Penn Chinese Treebank dealing with ambiguous structures.

[*] This research has been partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid for Scientific Research (B), 14380166, 17300065,Exploratory Research, 17656128, 2005 and the Outstanding Overseas Chinese Scholars Fund of the Chinese Academy of Sciences (No.2003-1-1).

The rest of this paper is organized as follows. In Section 2 this paper's approach of solving the problem will be examined. Section 3 reports on the experiments based on the manually annotated corpus. Finally, in section 4 conclusions are drawn and future work is discussed.

Proposed Approach

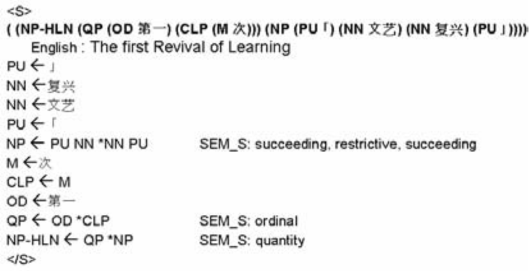
In this section we show the entire process of learning the relations for headword-modifier pairs from the Penn Chinese Treebank 5.0. First the annotation process will be examined. Then, the algorithms that were used will be discussed.

Corpus annotation

First random sentences were selected from the Treebank and manually annotated. They were annotated with headword and dependency relation information. In the end there were 3,639 semantic dependency relations from 116 sentences consisting of 3,510 words. Almost the entire dependency relation tag set reported by (Li *et al.* 2003) was used. It consists of 59 semantic relations, 9 syntactic relations and 2 special relations.

In Chinese, punctuation has an important role in the sentence. In the Penn Treebank, the punctuations are annotated. So for the relation between punctuations and other constituents, we annotated them mainly with the relation of "succeeding."

In the semantic dependency grammar, the headword of a sentence represents the main meaning for the entire sentence and the headword of a constituent represents the main meaning of the constituent. In a compound constituent, the headword inherits the headword of the head sub-constituent, and headwords of other sub-constituents are dependent on that headword. The word that was able to best represent the meaning of the constituent was chosen as the headword. Figure 1 gives an example of an annotated sentence, "*" denotes the headword. Figure 2 shows the conversion from a parse tree to a semantic dependency tree.



```
<S>
( (NP-HLN (QP (OD 第一) (CLP (M 次))) (NP (PU 「) (NN 文艺) (NN 复兴) (PU 」))))
    English : The first Revival of Learning
PU ← 」
NN ←复兴
NN ←文艺
PU ← 「
NP ← PU NN *NN PU          SEM_S: succeeding, restrictive, succeeding
M ←次
CLP ← M
OD ←第一
QP ← OD *CLP               SEM_S: ordinal
NP-HLN ← QP *NP            SEM_S: quantity
</S>
```

Figure 1: Manually annotating the corpus with headword and semantic dependency relation

When annotating the headword, some non-proper annotations in the original bracketed data of the Penn Chinese Treebank were found in the raw data. These annotations were too shallowly parsed. In some of these non-proper annotations, the modifier was at the same height in the parse tree as the word that should become the headword for the parse tree. Figure 3 shows some examples of these non-proper annotations. The tree structure of the original sentence for the second example is shown in Figure 4(a).

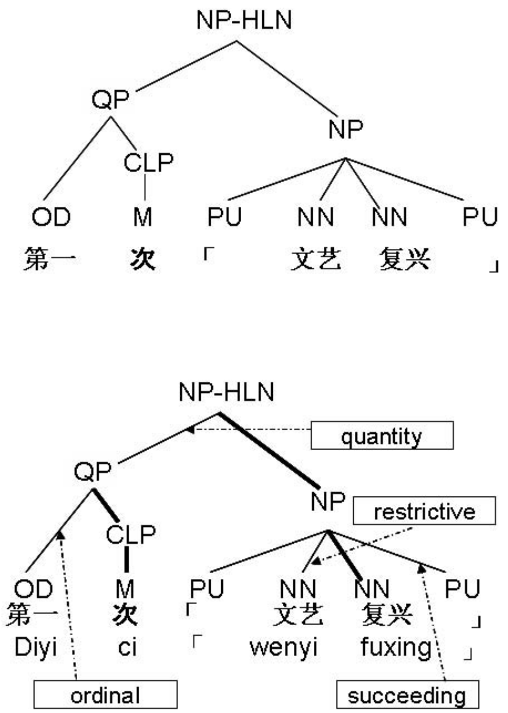
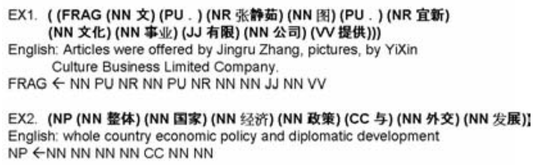


Figure 2: From parse tree to semantic dependency tree



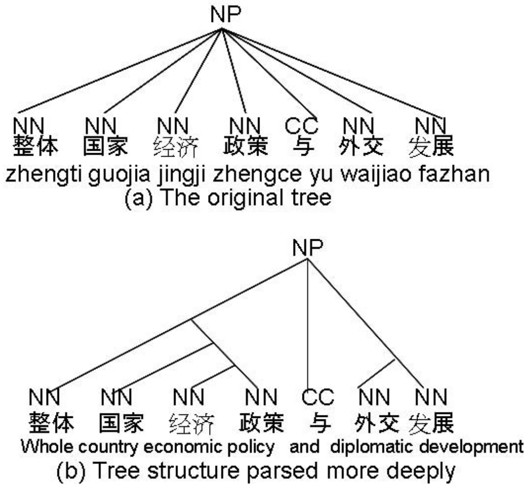
```
EX1.  ( (FRAG (NN 文) (PU . ) (NR 张静茹) (NN 图) (PU . ) (NR 宜新)
       (NN 文化) (NN 事业) (JJ 有限) (NN 公司) (VV 提供)))
English: Articles were offered by Jingru Zhang, pictures, by YiXin
        Culture Business Limited Company.
FRAG ← NN PU NR NN PU NR NN NN JJ NN VV

EX2.  (NP (NN 整体) (NN 国家) (NN 经济) (NN 政策) (CC 与) (NN 外交) (NN 发展))
English: whole country economic policy and diplomatic development
NP ←NN NN NN NN CC NN NN
```

Figure 3: Examples of shallow parsing

The sentence was left ambiguous. If there had been a deeper parse then the resulting parse tree would most likely look that in Figure 4 (b) and selecting the headword and relations would be more straightforward. However, as it is in Figure 4(a) it is difficult to decide which word is the headword and what kind of relation is proper.

Part of the problem is that it is a fragment and not a sentence as shown in Figure 3. However, in Chinese much information can be gained from fragments and semantic relations can and should be assigned. Fragments, though, are more difficult than regular sentences to assign headwords


NP
NN NN NN NN CC NN NN
整体 国家 经济 政策 与 外交 发展
zhengti guojia jingji zhengce yu waijiao fazhan
NP
NN NN NN NN CC NN NN
整体 国家 经济 政策 与 外交 发展
Whole country economic policy and diplomatic development

(a) The original tree

(b) Tree structure parsed more deeply

Figure 4: Tree structure of the original data and improved one

and relations to. A later section will show how to propose a method for disambiguation.

Algorithms

After we manually annotated part of the corpus with headwords and assigned semantic dependency relations, we created programs to build multiple training and test sets. Two algorithms were used, a Naive Bayesian Classifier and a baseline. Both of the algorithms are capable of doing multi-category classification and thus can be straightforwardly applied to the problem at hand. In addition, as this is an initial investigation simpler algorithms were tested to see the feasibility of machine learning techniques for this problem. We hope to be able to use this initial work as a baseline for future research. The features that were looked at as well as more information about the two algorithms will be explained in the following subsections.

Feature Selection The features (Xue & Palmer 2005) used for their semantic role labeling for Chinese verbs consist of the following.

- Position
- Path
- Head word and its part-of-speech
- Predicate
- Subcat frame
- Phrase type
- First and last word of the constituent in focus
- Phrase type of the sibling to the left
- Syntactic frame
- Combination features

In contrast to their feature list, in this paper, only a subset was used. Also, because of the size of the corpus a smaller feature should help improve results. Since the headword and its modifier are the most important indicator of the semantic dependency relation, it will be the basis for the chosen characteristics. The 5 chosen features are as follows.

- Headword
- Modifier
- Headword part-of-speech
- Modifier part-of-speech
- Context

The context feature is the modifier part-of-speech(POS)es that are between the headword and the modifier of interest. In addition to these features a small rule set was used. The rule set and the reason for it will be discussed in detail in a later section.

Naive Bayesian Classifier (NBC) The Naive Bayesian Classifier is widely used in machine learning due to its efficiency and its ability to combine evidence from a large number of features (Manning & Schutze 1999). The combinations of features that were used are listed below.

- Headword POS and modifier POS
- Headword and its POS and modifier and its POS
- Headword POS, modifier POS, and context
- Headword and its POS, modifier and its POS, and context

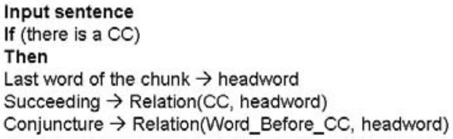
For example in Figure 2, in the phrase “wenyi fuxing,” the part-of-speech features are “NN NN,” the word features are ”wenyi fuxing,” and the context feature is “[]” meaning empty. For an example of context, in Figure 4, if “fazhan” is taken as the headword and a relation is being assigned between “zhengce” and “fazhan,” the context feature would be “[CC NN].”

Baseline For a baseline algorithm, the most probably relation was used. This algorithm simply assigns the most probably relation seen in the training data to every headword modifier pair seen in the testing data.

Rule-Based Correction

To resolve the problem patterns in the Penn Chinese Treebank, rule-based correction is performed. First, in order to test if a rule-based method is feasible some preference rules were created by hand and added to the system. The rules were made according to the features of some of the problem sentences. For the phrase in Figure 4, the rule in figure 5 was created. However, creating rules by hand is tedious and time consuming and as such only 4 rules were created.

The rules are to mostly fix problems with NP. A rule set of about 35 to 40 rules would probably be enough to fix these problems. It is, perhaps, more desirable to automatically create or extract these rules.



```
Input sentence
If (there is a CC)
Then
Last word of the chunk → headword
Succeeding → Relation(CC, headword)
Conjuncture → Relation(Word_Before_CC, headword)
```

Figure 5: Sample rule

Experiments

A 10-Fold-Cross-Validated test was adopted. First to test if the Naive Bayesian classifier and the baseline had a chance at being effective a closed test was performed using some of the features. Table 1 shows the results for the closed test, for brevity the Naive Bayesian Classifier is listed as NBC. The Accuracy is simply the number of correctly guess relations divided by the total number of relations in the testing data set. Please note that while not shown, the recall was always 100%.

From Table 1 it can be seen that using part-of-speech and words the closed test results are very high[1]. This is a good indication that if the training data sufficiently describes the entire set that using these two features should result in a good accuracy. The next test was an open test. Table 2 shows the results of the open test.

Algorithm	Avg. Accuracy
Baseline	4.8% (±2.2%)
NBC (s)	70.45% (±3.02%)
NBC (s + w)	96.82% (±0.47%)

Table 1: Closed Test Results

As can be seen from Table 2 the best results came from the Naive Bayesian Classifier using POS, words, and context. In fact it can be seen that the addition of context helped improve the results in every test. This means that the context information provides useful information in the classification. If the manually annotated corpus were larger then the training set would be larger and this should result in better average accuracy. Since fragments were not omitted the system's accuracy was lower than it would be with just complete sentences.

Algorithm	Avg. Accuracy
Baseline	4.8% (±2.5%)
NBC (s)	67.19% (±3.09%)
NBC (s + w)	69.63% (±1.83%)
NBC (s + c)	71.22% (±5.03%)
NBC (s + w + c)	73.11% (±3.45%)

Table 2: Open Test Results

For the problem patterns in the original Treebank a rule-based correction approach was used. For experimentation, four hand crafted rules were looked at to judge the feasibility of such an approach. Table 3, shows the results of adding these rules to three of the different classifier setups. This slight improvement indicates that an approach that first uses a probabilistic model to assign relations and then uses rules to correct mistakes may be an efficient one. This approach would be similar to the one Brill's tagger uses (Brill 1992).

Algorithm	Avg. Accuracy
NBC (s+ c + r)	71.56% (±5.02%)
NBC (s + w + c + r)	74.05% (±3.55%)

Table 3: Open Test Results with Handcrafted Rules

Conclusion and Future Work

We see the principal results of our work to be the following, we presented, to our knowledge, the first method of automatically annotating semantic dependency relations for the Penn Chinese Treebank. Secondly, experiments of automatically annotating semantic dependency relations were carried out. The results indicate that Naive Bayesian Classifier is significantly more effective for annotating semantic dependency structure automatically than the baseline. We showed that the headwords provide useful knowledge for deciding semantic dependency relations. In this study, we also designed correction rules for the problem patterns of the Penn Chinese Treebank. Currently there are too few rules to see a significant improvement, but even with just four rules an small improvement of about 1% was seen.

Although we automatically annotated the sentences with semantic dependency structure successfully, much further work is still needed. The test set we used was made manually and thus was very small. We will aim at enlarging the size of the annotated corpus by using the algorithms in this paper to first assign a relation and then manually correcting the errors. After a larger annotated corpus is created we can use other machine learning algorithms. In particular we would like to examine the use of Support Vector Machines and Maximum Entropy. In addition the larger annotated corpus may improve the Naive Bayesian classifier's results due to a larger training data set.

In addition we will look at using more advanced genetic algorithms or transformation-based learning for automatically acquiring rules for problem patterns. In the end, perhaps, a hybrid system that first uses some probabilistic approach to assign relations and then uses a rule based system to correct errors will be the best.

References

Brill, E. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of 3rd Applied Natural Language Processing*, 152–155.

[1] In the tables "s" means part-of-speech, "w" means words, "c" means context, and "r" means rules.

Gan, K. W., and Wong, P. W. 2000. Annotating information structures in chinese texts using hownet. In *Proceedings of Second Chinese Language Processing Workshop*.

Gildea, D., and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistic* 28:496–530.

Johnson, C. R., and Fillmore, C. J. 2000. The framenet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, 56–62.

Li, M.; Li, J.; Dong, Z.; Wang, Z.; and Lu, D. 2003. Building a large chinese corpus annotated with semantic dependency. In *Proceedings of the Second SIGHAN Workshop*.

Manning, C. D., and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal* 31(1).

Xia, F.; Palmer, M.; Xue, N.; Okurowski, M. E.; Kovarik, J.; Chiou, F.-D.; Huang, S.; Kroch, T.; and Marcus, M. 2000. Developing guidelines and ensuring consistency for chinese text annotation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*.

Xue, N., and Palmer, M. 2003. Annotating propositions in the penn chinese treebank. In *Proceedings of the Second Sighan Workshop*.

Xue, N., and Palmer, M. 2005. Automatic semantic role labeling for chinese verbs. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*.

Yan, J.; Jiang, P.; Kuroiwa, S.; and Ren, F. 2005. Semantic analysis using compound rules (in japanese). In *the 11th Language Processing Annual Conference*.

Yang, X., and Li, T. 2002. A study of semantic disambiguation based on hownet. *Computational Linguistics and Chinese Language Processing* 7:47–78.