2005－NL－170（11）
2005／11／21

機械学習を用いた中国語意味的依存構造の推定

顔　加軍, デービッド・ブレスウェル, 任　福継, 黒岩　眞吾
知能情報工学科 徳島大学
〒770-8506 徳島市南常三島町 2-1
Email :{ yanjj, davidb, ren, kuroiwa} @is.tokushima-u.ac.jp

構文解析後、文の意味構造を決定するのは重要である。本稿では、Penn Chinese Treebank のために意味的な依存構造を自動的に付与する方法を提案する。まず手動で主辞と意味的依存関係を付与しテストデータを作成する。その後、異なるフィチャーのもとで、二つの教師つき機械学習アルゴリズムをデータに適用し、意味関係を推定する。最後に、中国語の特徴に基づき優先規則を作成し、元コーパスの中に問題がある木構造に対して曖昧性解消を行う。評価実験の結果によると、提案したアルゴリズムが中国語の意味的な依存構造を決定するには有効である。

Automatically Determining Semantic Structure in Chinese Sentences

Jiajun YAN, David B. Bracewell, Fuji REN, Shingo Kuroiwa
Department of Information Science and Intelligent Systems,
Faculty of Engineering, The University of Tokushima
Tokushima 770-8506
Email :{ yanjj, davidb, ren, kuroiwa} @is.tokushima-u.ac.jp

After parsing it is difficult to determine the semantic structure of sentences for Chinese sentences. 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. Two supervised machine learning algorithms with varying features were then adopted to learn the relations. Finally, a set of preferences rules were created based on features of Chinese to solve some problem patterns that were found in the Penn Chinese Treebank dealing with ambiguous structures. The experimental results show that the algorithms and proposed approach are effective for determining semantic dependency structure automatically.

I. 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 addition, headword-modifier relations provide the knowledge which is difficult to acquire manually. In English, much research has been done in semantic parsing using statistical and machine learning methods [1] to semantically annotated corpora such as FrameNet and the proposition Bank 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 and Wong [2] have annotated a subset of the Sinica balanced corpus with semantic dependency relations as defined in HowNet. Li et al. [3] 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 definite 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 and Li [4] 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 and Palmer [5] [6] 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.

Yan et al. [7] reported a method to specify 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.

Research of auto-tagging Chinese corpus 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. Initially a small portion of the Penn Chinese Treebank was manually annotated with headwords and dependency relations. Two supervised machine learning algorithms with varying features were then adopted to learn the relations. Finally, a set of rules were created based on features of Chinese to solve some problem patterns that were found in the Penn Chinese Treebank dealing with ambiguous structures.

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.

II. 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 learning algorithms that were used will be discussed.

A. 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 3639 semantic dependency relations from 116 sentences consisting of 3,510 words. Almost the entire dependency relation tag set reported by Li et al. [3] 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.

When annotating the headword, some non-proper annotations in the original bracketed data of the Penn Chinese Treebank were found in the raw data, which were too shallowly parsed. In some sentences the modifier was parsed at the same level leaf node as the word that should become the headword of the parse tree.

<S>
((IP (PU （) (NP-PN-SBJ (NR 故宫) (NN 博物院)) (VP (VV 提供)) (PU ）)))
PU ← ） SEM_S:
VV ← 提供 SEM_S:
VP ← VV SEM_S:
NN ←博物院 SEM_S:
NR ←故宫 SEM_S:
NP-PN-SBJ ← NR *NN SEM_S: **restrictive,**
PU ← （ SEM_S:
IP ← PU NP-PN-SBJ *VP PU SEM_S: **succeeding, agent, succeeding,**
</S>

Fig.1. Manually annotating the corpus with headword and semantic dependency relation

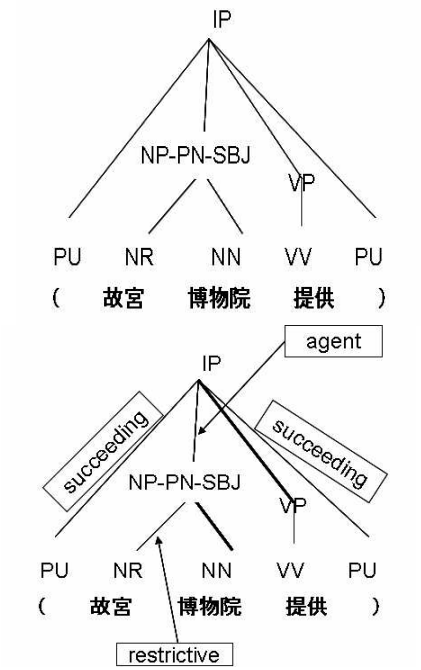


Fig. 2. From parse tree to semantic dependency tree

Figure 3 shows some examples of these difficult sentences. The tree structure of the original sentence for the second example is shown in Figure 4(a).

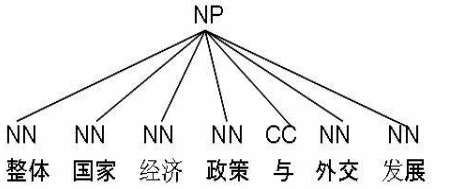
The sentence was left ambiguous. If there had been a deeper parse than the resulting parse tree would most likely look that in Figure 4 (b) and selecting the headword and relations would be more straightforward. However, as it is in Figure 4(a) it is difficult to decide which word is the headword and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

Part of the problem is that it is a fragment and not a sentence. 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 and relations to. A later section will show how to propose a method for disambiguation.

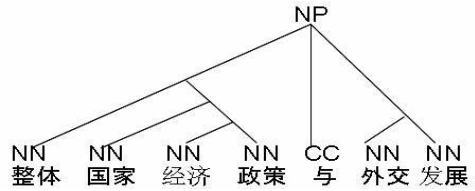
((FRAG (NN 文) (PU ．) (NR 张静茹) (NN 图) (PU ．) (NR 宜新) (NN 文化) (NN 事业) (JJ 有限) (NN 公司) (VV 提供)))
English: Articles were offered by Jingru Zhang, pictures, by YiXin Culture Business Limited Company.
FRAG ← NN PU NR NN PU NR NN NN JJ NN VV
SEM_S: ?

(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 SEM_S: ?

Fig. 3. Some examples of shallow parsing



(a) The original tree



(b) Tree structure parsed more deeply

Fig. 4. Tree structure of the original data and improved one

B. 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 Naïve Bayesian Classifier and a simple probabilistic model. 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. Since more complicated algorithms, like support vector machines, require a great deal of time for

training, it made sense to start with simpler algorithms that are quick to train. This allows new training data to be added and the system retrained in a timely manner. The features that were looked at as well as more information about the two algorithms will be explained in the following subsections.

1) Feature Selection: The features Xue and Palmer [6] 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 the most informative features are used. The intention is to find the most useful information from the manually annotated corpus and transfer it into formatted knowledge that the models can use. 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 syntax
- Modifier syntax
- Context

The context feature is the modifiers 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.

2) Naïve Bayesian classifier (NBC): The Naïve Bayesian Classifier is widely used in machine learning due to its efficiency and its ability to combine evidence from a large number of features [8]. The combinations of features that were used are listed below.

- Headword syntax and modifier syntax
- Headword and its syntax and modifier and its syntax
- Headword syntax, modifier syntax, and context
- Headword and its syntax, modifier and its syntax, and context

For example, in Figure 2 from the phrase "NP-PN-SBJ ← NR *NN 故宫 博物院 SEM_S: restrictive", the syntax features are "NR NN," the word features are "故宫 博物院," and the context feature is "[]" meaning empty. For an example of context, In Figure 4, if "NN 发展" is taken as the headword and a relation is being assigned between "NN 政策" and "NN 发展," the context would be "[CC NN]."

3) Simple probabilistic model (SPM): In addition to the Naïve Bayesian classifier, a simpler probabilistic model was also examined. The simple probabilistic model uses the same combination of features. However, for the syntax feature the headword syntax and its modifier's syntax are made into a bigram. The same is done for the word feature.

Another difference between it and the Naïve Bayesian classifier is the omission of using P(relation) in calculating the probability of a relation given a set of features. The calculation can be seen in equation 1. The probabilities that are calculated are relative probabilities. In equation 2, how to determine the relation when there are multiple features is shown. For brevity bigram is shortened to *bg* and relation to *rel*. If there is no possible answer from the computed probabilities, i.e. the bigram/relation pair has not been seen before, and then the most probable relation is assigned.

$$P(rel \mid bg) = \arg\max_{P} \left(P(bg \mid rel)\right) \quad (1)$$

$$P(rel \mid feature) = Max\left\langle \arg\max_{P}(P(bg_1 \mid rel)), \ldots, \arg\max_{P}(P(bg_N \mid rel)) \right\rangle \quad (2)$$

C. *Rule based correction*

To resolve the problem patterns in the Penn Chinese Treebank, some preference rules were created and added to the system.

Input sentence (SYN, Words, SEM_S)
If (there is a CC)
Then {the last word of the phrase must be headword, the relation between CC and headword must be "coordinate", the relation between the word before CC and headword must be "conjuncture", the relation between the other words and headword could not be "conjuncture" }

Fig. 5. A rule for the problem phrase

The rules were according to the features of such problem sentences. For the phrase in Figure 4, the rule in figure 5 was created.

III. EXPERIMENTS

In this paper a 10-Fold-Cross-Validated test was adopted. The manually annotated corpus was divided up using the standard 80-20 rule 10 times to create 10 different training and testing data sets. First to test if the Naïve Bayesian classifier and the simple probabilistic model 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 Naïve Bayesian Classifier is listed as NBC and the simple probabilistic model is listed as SPM. The Accuracy is simply the number of correctly guess relations divided by the total number of relations in the testing data set.

From Table 1 it can be seen that using syntax and words the closed test results are very high. 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.

TABLE I
CLOSED TEST RESULTS

Algorithm	Avg. Accuracy
NBC (syntax only)	70.45% (±3.02%)
NBC (syntax + words)	96.82% (±0.47%)
SPM (syntax only)	71.73% (±2.91%)
SPM (syntax + words)	98.32% (±0.28%)

TABLE II
OPEN TEST RESULTS

Algorithm	Avg. Accuracy
NBC (syntax only)	67.19% (±3.09%)
NBC (syntax + words)	69.63% (±1.83%)
NBC (syntax + context)	71.22% (±5.03%)
NBC (syntax + words + context)	73.11% (±3.45%)
SPM (syntax only)	68.87% (±2.77%)
SPM (syntax + words)	68.73% (±2.65%)
SPM (syntax + words + context)	70.88% (±3.76%)

As can be seen from Table 2 the best results came from the Naïve Bayesian Classifier using syntax, 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. For the problem patterns in the original Treebank, at this time only 4 rules were created solely for the purposes of seeing if they would help. The order of the rules can affect the outcome and as such manually crafting the rules is troublesome. However, as can be seen in Table 3 the results do improve slightly even with just 4 rules. 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 [9].

TABLE III:
OPEN TEST RESULTS WITH RULES

Algorithm	Avg. Accuracy
NBC (syntax + context + rules)	71.56% (±5.02%)
NBC (syntax + words + context + rules)	74.05% (±3.55%)
SPM (syntax + words + context + rules)	71.98% (±3.61%)

IV. CONCLUSION AND FUTURE WORK

We see the principal results of our work to be the following:

- This paper has firstly presented the method of automatically annotating semantic dependency relations for the Penn Chinese Treebank.
- The experiments of automatically annotating semantic dependency relations were carried out. The results indicate that Naive Bayesian Classifier is more effective for annotating semantic dependency structure automatically.

- We proved that the headword provides the knowledge, which is most useful to decide the semantic dependency relations.
- In this study, we also designed preference rules for the problem patterns of the Penn Chinese Treebank. The results show that our strategy of creating rules for problem pattern is useful.

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. In addition the larger annotated corpus may improve the Naïve Bayesian classifier's results due to a larger training data set.

In addition we will look at using genetic algorithms or transformation-based learning to automatically acquire 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.

ACKNOWLEDGEMENT

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).

REFERENCES

[1] Daniel Gildea, Daniel Jurafsky, "Automatic Labeling of Semantic Roles", Computational Linguistic 28(3):pp.496–530, 2002.

[2] Kok Wee Gan, Ping Wai Wong, "Annotating Information Structures in Chinese Texts Using HowNet", In Proceedings of Second Chinese Language Processing Workshop, ACL-00, HongKong, 2000.

[3] Mingqin Li, Juanzi Li, Zhendong Dong, Zuoying Wang and Dajin Lu, "Building A Large Chinese Corpus Annotated With Semantic Dependency", Proceedings of the Second SIGHAN Workshop, Sapporo, Japan, 2003.

[4] Yang Xiaofeng, Li Tangqiu, "A Study of Semantic Disambiguation Based on HowNet", Computational Linguistics and Chinese Language Processing, Vol. 7, No. 1, pp. 47-78, February 2002.

[5] Nianwen Xue and Martha Palmer, "Annotating Propositions in the Penn Chinese Treebank", In Proceedings of the Second Sighan Workshop, Sapporo, Japan, 2003.

[6] Nianwen Xue and Martha Palmer. "Automatic Semantic Role Labeling for Chinese Verbs", in Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 2005.

[7] Jiajun Yan, Peilin Jiang, Shingo Kuroiwa, Fuji Ren, "Semantic Analysis Using Compound Rules" (in Japanese), the 11th Language Processing Annual Conference, Kagawa, Japan,2005(3)

[8] Christopher D.Manning and Hinrich Schutze, Foundations of Statistical Natural Language Processing, Published by The MIT Press Cambridge, Massachusetts, May 1999.

[9] Eric Brill, "A Simple Rule-Based Part-of-Speech Tagger," in Proceedings of 3rd Applied Natural Language Processing, 152-155, 1992.