LANGUAGE AND LINGUISTICS 9.2:449-461, 2008
2008-0-009-002-000253-1

Semantic Role Labeling in Chinese Using HowNet

Xia Wang
City University of Hong Kong

Semantic Role Labeling (SRL) has significant impact on many application systems, such as Machine Translation, Information Extraction, Question-Answering, Text Summarization and Text Data Mining. Therefore research on SRL is important for natural language understanding, and so far a number of algorithms, mostly statistically oriented, have been proposed in this field.

Statistical algorithms must deal with the problem of data sparseness. In our initial study, we found that most words appear only a small number of times, and other words are absent completely in the training set. Only a small number of frequent words supply sufficient data for training. To solve this problem, we developed a backoff model based on HowNet.

In this study, we demonstrate the benefit of applying the knowledge from HowNet to Semantic Role Labeling by experimenting with four selected Chinese words. Our system employs a statistical approach, which was trained on 208 sentences and tested on 89 sentences. We extracted various lexical and syntactic features, including the phrase type of each constituent, the headword, and the position and distance from the predicate to the constituent in question and voice.

Comparing the result with knowledge support of HowNet to the result without it, we found distinct improvement when using HowNet.

The study also reveals that the system can be improved by applying more information from HowNet, introducing full parsing information, enriching the feature set, and using more appropriate probability estimation model.

Key words: Semantic Role Labeling, HowNet, statistical approach

1. Introduction

Semantic Role Labeling (SRL) has had a great impact on many application systems, such as Machine Translation, Information Extraction, Question-Answering, Text Summarization, Text Data Mining, Speech Recognition, etc. Thus Semantic Role Labeling could be a critical intermediate step for natural language understanding. A number of statistical algorithms have been proposed in this field, most of which are dealing with the English language. In this paper we present an approach for Semantic

Role Labeling, in which shallow syntactic parsing and lexical resources are used. Some definitions adopted in this paper are as following:

(1) Semantic Roles. A semantic role is the semantic relationship that a syntactic constituent has with a predicate. Arguments are grouped into two major types according to their semantic roles: (a) necessary arguments, representing central participants in an event, which include Agent, Patient, Instrument, etc.; (b) optional arguments (adjuncts), optional for an event but supplying more information about the event, which includes Location, Time, Manner, etc.

(2) Semantic Role Labeling. We define the task of Semantic Role Labeling as the assignment of syntactic constituents with semantic roles of predicates in sentences. For example, given a verb in a sentence:

下 星期 學校 舉行 講 故事 比賽

The goal is to locate the constituents which are arguments of the verb, and assign them appropriate semantic roles.

下 星期	學校	舉行	講 故事 比賽
Time	Agent	Target	Patient

2. The necessity of Semantic Role Labeling

Natural language understanding is finally based on meaning. The pure syntactic structure of a sentence does not reflect its meaning. Different meanings with the same syntactic structure mainly arise from what words mean and how these meanings combine into sentences to form sentence meanings. In high-level processing such as Machine Translation, Information Extraction, Question-Answering, Text Summarization, Text Data Mining, and Speech Recognition, semantic analysis is required.

3. Related work

Like all NLP systems, previous work in SRL has two approaches: rule-based and statistical. Rule-based approaches such as Head-Driven Phrase Structure Grammar (HDPSG) have their shortcomings: they are time-consuming and have limited coverage. Most of the SRL approaches are statistical, making use of a variety of models.

Gildea & Jurafsky (2002) presented a system based on statistical classifiers trained on roughly 50,000 sentences that were hand-annotated with semantic roles by the

FrameNet semantic labeling project. They then parsed each training sentence into a syntactic tree and extracted various lexical and syntactic features. These features were combined with knowledge of the predicates. They used various lexical clustering algorithms to generalize across possible fillers of roles. They reported 82% accuracy in identifying the semantic role of presegmented constituents, and 65% precision and 61% recall in segmenting constituents and identifying their semantic role simultaneously.

Sun & Jurafsky (2004) addressed the question of assigning semantic roles to sentences in Chinese. They showed that good semantic parsing results for Chinese can be achieved with a small 1,100-sentence training set.

Kwong & T'sou (2005) described semantic role tagging of Chinese in the absence of a parser. They tackled the task by identifying the relevant headwords in a sentence as a first step to partially locate the corresponding constituents to be labeled. They explored the effect of data homogeneity by experimenting with a textbook corpus and a news corpus, representing simple data and complex data respectively.

Tsai, Wu, Lin & Hsu (2005) proposed a method that exploits full parsing information by representing it as features of argument classification models and as constraints in integer linear learning programs. They take advantage of SVM-based and Maximum Entropy-based argument classification models. The experimental results show that full parsing information not only increases the F-score of an argument classification model, but also effectively removes all labeling inconsistencies.

The above algorithms can be assigned to one of three classes: approaches that take advantage of complete syntactic analysis of text, pioneered by Gildea & Jurafsky (2002); approaches that use partial syntactic analysis; and approaches in the absence of parsing information, like Kwong & T'sou (2005).

4. The data

4.1 Role set

Different role-sets always vary in the number and type of roles. The definitions of semantic roles range from very general, such as "PROTO-AGENT" and "PROTO-PATIENT", to very specific, such as domain-specific or verb-specific roles. FrameNet roles, which fall in-between, are defined for each semantic frame.

In this study, we worked with a set of 11 predicate-independent abstract semantic roles. As shown in Table 1.

Table 1: The list of semantic roles

Role
主事 (Agent)
受事 (Patient)
客體 (Theme)
經驗者 (Experiencer)
工具 (Instrument)
處所 (Location)
來源 (Source)
目標 (Goal)
時間 (Time)
動量 (Frequency)
數量 (Quantity)

4.2 Training and testing sets

4.2.1 Creation of training and testing sets

We developed our training set by choosing four Chinese verbs, and then selecting all sentences containing these four verbs from a corpus. Several factors should be considered when selecting verbs. (1) Frequency. The verbs should be frequent enough to provide sufficient training data. (2) Syntactic diversity. We should like to select verbs with different numbers of arguments, and with various argument patterns, which are representative of the variety of syntactic behavior in the language. (3) Word sense. Verbs that varied in their number of word senses are preferred.

Based on these three considerations, we selected four verbs for our experiment, as shown in Table 2.

Table 2: List of verbs for experiment

Verb	# of senses	Freq
成立	2	53
出現	1	53
給	3	103
通過	2	50

In total, we have 259 sentences. And then we split the data for each verb into two parts: 80% for training and 20% for test. Each test verb has been seen in the training set. Then we add an unseen predicate “給予” to test our algorithm. Thus there are 208

sentences in the training set and 89 sentences in the testing set.

4.2.2 Format of the annotation

Training sentences were shallow parsed, were annotated with semantic roles and features. Format of the annotation are:

…主持：從兩位對話中，關鍵的地方是<Agent>[主席/n/NP/l2/person]
</Agent> <Time>[四月二十二日/t/TP/l1/time]</Time> <Key>給</Key>
<Goal>[你/r/NP/r1/person]</Goal>
<Patient>[一封信/n/NP/r2/entity]</Patient>，要繼續招標。

Test sentences also were parsed and annotated with these features. Then they would be parsed through the identifier:

[政府]n/NP/l3/organization [還可以要求]/v/VP/l2/request
[行業]n/NP/l1/affairs<Key>成立</Key>[保險賠償基金]n/NP/r1/organization.

5. The experiment

5.1 Feature set

The identification of argument relies on a number of features extracted from the input sentence and its parse. We used features representing various aspects of the syntactic structure of a sentence, which is predictable from semantics, according to linking theory. The features we used include **phrase type**, **position & distance**, **voice**, **headword**, and backoff models for **headword**.

Phrase type indicates the syntactic category of the phrase expressing the semantic roles. Different semantic roles tend to be realized by different syntactic categories. For example, the Location tends to appear in text as a prepositional phrase or noun phrase.

Position and distance. The position indicates that a constituent is located before or after the target verb. And the distance indicates how far the constituent is away from the predicate. We combined them into a single feature, represented as a letter of **L**(left) or **R**(right) followed by a number. For example, **L2** indicates the second chunk to the left of the predicate. We used this combined feature instead of path, because we do not have full parsing information.

Voice. It gives information about surface location of arguments, since direct objects of active verbs often correspond in semantic role to subjects of passive verbs. This feature has two values: active or passive.

Head word. Head words of phrases can be used to express selectional restrictions on the semantic roles. It is a useful but sparse feature. In our corpus, many headwords appear only once. Table 3 shows the top-ten headwords for roles of "成立".

Table 3: Top-ten headwords for roles of "成立"

Word	Freq
組	11
會	8
政府	7
中心	6
罪名	5
公司	5
基金	5
局	5
在	5
特區	4

We used a semantic-based backoff model if no training examples had been seen in training set. For example, 計劃建議<Location>在福島</Location><Key>成立</Key><Goal>三方政府.

政府 -> organization

Table 4 shows the top five semantic classes for roles of "成立".

Table 4: Top five semantic classes for roles of "成立"

Semantic class	Freq
organization	50
time	11
accusal	6
person	5
place	5

For words not found in the semantic dictionary, we used the alternative model based on POS. For example, "一九二〇年代" may not be seen in the training and dictionary, but its POS TEMPORAL is a useful indicator that it may be a role of TIME.

5.2 Semantic resource

A semantic resource is needed as support for the semantic-based backoff model. In this study, we used the online knowledge base: HowNet.

HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoted in lexica of Chinese and their English equivalents.

To take one meaning of the Chinese word "打" for example, it is found in the knowledge dictionary as:

NO.=000001
W_C=打
G_C=V
E_C=~醬油，~張票，~飯，去~瓶酒，醋~來了
W_E=buy
G_E=V
E_E=
DEF=buy|買

Where W_C, G_C, E_C denote Chinese word, grammatical category, and example, respectively. And W_ E, G_ E, E_ E are for English. DEF denotes the definition of a concept. The first item in DEF (exercise) is the most important description, which provides the basic semantic features for the concept. So we use it as the semantic class of the word. Table 5 shows more samples we used in the experiment, in which the semantic classes have been underlined.

Table 5: Samples from HowNet

NO.=033206 W_C=國家 G_C=N E_C= W_E=country G_E=N E_E= DEF=place|地方, #human|人, country|國家, politics|政
NO.=071567 W_C=社會

G_C=N E_C= W_E=society G_E=N E_E= DEF=organization|組織

5.3 Probability estimation

To label the semantic role of a constituent automatically, we wish to estimate the probability of the event that the constituent is to fill each possible role. Given the features described above and the predicate, or target word, *t*:

$$P(r \mid h, pt, p\&d, voice, t) = \frac{f(r, h, pt, p\&d, voice, t)}{f(h, pt, p\&d, voice, t)}$$

where *r* indicates semantic role, *h* for head word, pt for phrase type, *p&d* for position and distance. The probability would be estimated from the training set by counting the number of times each role appears with a combination of features and dividing by the total number of times the combination of features appears.

It is difficult to implement in practice, however, because of data sparseness. So we approximate the probability to the product of four independent probabilities:

$$P(r|h,t)*P(r|pt,t)*P(r|pd,t)*P(r|v,t)$$

and smoothed them by giving a very small number.

For the unseen predicates “給予”, we used the data for predicate within same semantic class “給”.

5.4 Experiment result

Table 6 shows the best result of our experiment.

Table 6: Experimental result

Predicate	Without HowNet		With HowNet	
	Precision	Recall (%)	Precision	Recall (%)
成立	87.0	73.1	91.3	88.1
出現	67.2	60.3	78.6	61.9
給	63.1	41.7	71.1	60.7
通過	51.8	42.5	82.3	75.7
給予	60.4	47.9	69.3	63.0
Average	65.9	53.1	78.5	68.9

5.4.1 Result analysis

As with other statistical methods, errors mainly arise because the constituent with higher probability supplanted the real candidate with lower probability.

<Agent>[政府]NP</Agen>t[還可以要求]VP[行業]NP<Key>成立</Key>
<Goal>[保險賠償基金]NP</Goal>

[諮詢文件]NP[亦建議]VP[仿效]VP<Agent>[新加坡及加拿大等國家]NP
</Agent>，[強制]VP[新建成的多層大廈]NP，
<Time>[在出售單位時]PP</Time>自動<Key>成立</Key><Goal>[業主
立案法團]NP</Goal>。

In these two sentences, “行業” and “新建成的多層大廈” should be the agents of “成立”, but “政府” and “新加坡及加拿大等國家” got a higher probability because they may appear in the training set as the agent of “成立”.

6. Conclusions

Several conclusions from our experiment of Semantic Role Labeling in Chinese can be drawn. First, reasonably good performance can be achieved with a very small training set (208 sentences). Second, the system achieved a high precision and recall on the target word “成立” (91.3% of precision and 88.1% of recall) because of its strong patterns in predicate-argument structure, as shown in Table 7:

Table 7: Linking patterns of "成立"

Linking patterns	Freq
Goal + Key	17
Time + Key + Goal	9
Agent + Key + Goal	14
Location + Key + Goal	3

And the system achieved only 71.1% of precision and 60.7% of recall on "給" due to its more complex usage, as shown in Table 8:

Table 8: Linking patterns of "給"

Linking patterns	Freq
Key + Goal + Patient	14
Key + Patient	2
Agent + Key + Goal + Patient	36
Agent + Time + Patient + Key + Goal	2
Agent + Time + Key + Goal + Patient	6
Theme + Key + Goal + Theme	12
Theme + Key + Patient	2
Time + Key + Goal	2
Theme + Key + Theme	1
Key + Theme	2
Agent + Key + Patient	2
Time + Agent + Patient + Key + Goal	2
Agent + Patient + Key + Goal	2
Time + Agent + Key + Goal + Patient	2
Patient + Key	1
Patient + Agent + Key + Goal	5
Agent + Time + Key + Patient + Goal	1

Third, the performance for the unseen predicate "給予" (69.3 of precision and 58.2 of recall) is not much worse than "給". It shows that predicates within the same semantic class are semantically similar in patterns of argument structure. Thus to use semantic class as a backoff model is an effective way to handle unseen words.

7. Future work

Our system is preliminary at this stage. We plan to improve it in several respects, including:

First of all, enrich the feature set. Only five features were used in this study. We believe that more features suitable to Chinese can be found from the string of words, sentences, and the parsing.

Introduce the full parsing technique into the system. We used shallow parsing information in this study, and the experiment shows improvement over that based on words. We believe that additional information provided by complete syntax should be useful to improvement.

Improve the probability estimation model. Owing to time constraints, experiments with different models were not conducted at this stage, which should be done in the future.

Expand the experiment to a larger corpus. We used only four words for training and five for test in this study, due to a tight schedule. A more complete semantic analysis system should have the ability to process texts on a large scale.

References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics* (*COLING-ACL '98*), 86-90. Montreal: Université de Montréal.

Gildea, Daniel, and Martha Palmer. 2002. The necessity of ldea, parsing for predicate argument recognition. *Proceedings of the 40th Meeting of the Association for Computational Linguistics* (*ACL-02*). Philadelphia: University of Pennsylvania.

Gildea, Daniel, and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28.3:245-288.

Kingsbury, Paul, and Martha Palmer. 2002. From TreeBank to PropBank. *Proceedings of the 3rd Conference on Language Resources and Evaluation* (*LREC-02*). Las Palmas, Canary Islands, Spain.

Kingsbury, Paul, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the Penn TreeBank. *Proceedings of the Human Language Technology Conference* (*HLT-02*). San Diego, California, USA.

Kipper, Karin, Martha Palmer, and Owen Rambow. 2002. Extending PropBank with VerbNet semantic predicates. *Proceedings of the AMTA-2002 Workshop on*

Applied Interlinguas. Tiburon, California, USA.

Kwong, Oi Yee, and Benjamin K. T'sou. 2005. Data homogeneity and semantic role tagging in Chinese. *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*.

Sun, Honglin, and Daniel Jurafsky. 2004. Shallow semantic parsing of Chinese. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL 2004*), 249-256. Boston, Massachusetts, USA.

Tsai, Tzong-Han, Chia-Wei Wu, Yu-Chun Lin, and Wen-Lian Hsu. 2005. Exploiting full parsing information to label semantic roles using an ensemble of ME and SVM via integer linear programming. Paper presented at the CoNLL-2005. Ann Arbor, Michigan, USA.

Xue, Nianwen, and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese Treebank. *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*. Sapporo, Japan.

[Received 21 December 2006; revised 21 September 2007; accepted 1 November 2007]

Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon
Hong Kong
cnwangxia@gmail.com

基於知網的中文語義論元標注

王　霞
香港城市大學

語義論元標注對於許多應用系統，如：機器翻譯、信息提取、問答系統、數據挖掘等都有重大的影響。語義論元標注的研究對自然語言理解的重要性可想而知。迄今爲止，這個領域的研究者提出了各種算法，其中多數是基於統計的。

基於統計的算法必須處理數據稀疏的問題。在我們的初步研究中，發現大多數詞都是低頻詞，有的甚至沒有在訓練語料中出現，只有極少數的高頻詞有充足的語料進行訓練。爲了解決這個問題，我們採用了基於知網的回退模型。

我們選擇了四個中文動詞進行了實驗。本實驗採用了 208 句的訓練語料和 89 句的測試語料。我們從訓練文本中抽取了各種詞彙和句法特徵，包括論元的短語類型、中心詞、論元相對於謂詞的位置和距離等。實驗結果證明，把知網的知識用於語義論元標注，能很好的改善標注的準確率。

關鍵詞：語義論元標注，知網，統計方法