Investigation on Semantic Role Labeling of Chinese

Zhi Teng[1], Fuji Ren[1,2] and Shingo Kuroiwa[1]

[1] Faculty of Engineering, The University of Tokushima
2-1 Minamijosanjima, Tokushima, 770-8506, Japan

[2] School of Information Engineering, Beijing University of
Posts and Telecommunications, Beijing, 100876, China

{teng, ren, kuroiwa}@is.tokushima-u.ac.jp

Abstract. Semantic roles (AGENT, THEME, LOCATION, etc) provide a natural level of shallow semantic representation for a sentence. In the natural language processing field, researchers have experienced a growth of interest in semantic role labeling by applying statistical and machine-learning methods. So far much of the research has been focused on English due to the lack of semantically annotated resources in other languages. In this paper we will report the method and state on semantic role labeling using a pre-release version of the Chinese Proposition Bank.

1 Introduction

Recent efforts on semantic annotation have made it possible to train domain-independent semantic systems [3] [4] [5] [9] [10] [13]. Most of the semantic annotation projects focus on the predicate-argument structure, which represents a predicate and a number of arguments that are expected of this predicate. Generally each expected argument is assigned a label that marks the role this argument plays in relation to its predicate. It is in the level of generalization these role labels represent that the various annotation efforts differ. The most general are a limited set of roles such as agent and theme that are globally meaningful [2]. The role labels used in FrameNet [1] are less general in that they are meaningful only with respect to a specific situation, more formally known as a semantic *frame.* For example, the label *Byr* is only meaningful in the "Commercial_transaction" frame. One reflection of this reduced generality is that it is realized with a small class of predicates that indicate transaction, e.g. *purchase, rent.* The least general are the labels used in the Propbank annotation. The Propbanks [8] [12] use predicate-specific labels *ARG0, ARG1,... ARGn* for arguments and *ARGM* combined with a secondary tag to mark adjunct-like elements. The secondary tags indicate types of adjuncts and represent generalizations across all verbs.

The predicate-specific approach of the Propbank annotation builds a solid foundation for making high-level generalizations in a bottom-up manner, if broader generalizations are needed. There is generally a straightforward mapping from the numbered role labels to more general roles such as agent and theme. It is much harder to derive these semantic concepts from syntactic representation because an argument may not always be realized, and when it is, it may not always be realized in the same syntactic position as a result of syntactic alternations [7], etc. In addition, different senses of a verb take different sets of arguments that demonstrate different syntactic patterns. Thus, predicate-argument structure recognition at this level represents a crucial leap towards proper representation of semantic structure from the syntactic structure.

2 Semantic Role Labeling and the Corpus

2.1 Semantic Role Labeling

Semantic roles in the English [6] and Chinese [11] PropBanks are grouped into two major types:

(1) Arguments: which represent central participants in an event. A verb may require one, two or more arguments and they are represented with a contiguous sequence of numbers prefixed by arg, as arg0, arg1.

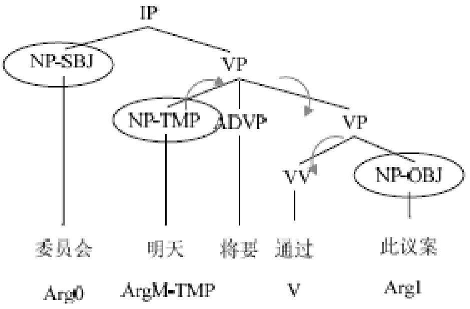
(2) Adjuncts: which are optional for an event but supply more information about an event, such as time, location, reason, condition, etc. An adjunct role is represented with argM plus a tag (Table 1.). For example, argM-TMP stands for temporal, argM-LOC for location.

argM-ADV	Adverbials	argM-MNR	Manner
argM-BNE	Beneficiary	argM-PRP	Purpose or Reason
argM-CND	Condition	argM-TMP	Temporal
argM-DIR	Direction	argM-TPC	Topic
argM-DGR	Degree	argM-CAU	Cause
argM-EXT	Extent	argM-NEG	Negation
argM-FRQ	Frequency	argM-MOD	Modal
argM-LOC	Locative		

Table 1. the Adjuncts Semantic roles

2.2 The Corpus

Here we discuss the linguistic annotation of the Chinese Proposition Bank [12]. The Chinese Propbank is based on the Chinese Treebank, which is a 500K-word corpus annotated with syntactic structures. The semantic annotation in the Propbank is added to the appropriate constituents in a syntactic tree. This is illustrated in Fig.1.



The council will pass this bill tomorrow.

Fig.1 A treebank tree annotated with semantic role labels and frameset ID

The frameset represents a major sense defined by the set of arguments a predicate takes. The task of semantic role labeling is to use the role labels as categories and classify each argument as belonging to one of these categories.

3 Semantic role tagging with hand-crafted parses

To be used in real-world natural language applications, a semantic role tagger has to use automatically produced constituent boundaries either from a parser or some other means.

3.1 Classifier

We can use a Maximum Entropy classifier with a tunable Gaussian prior in the Mallet Toolkit. The Maximum Entropy classifier does multi-category classification and thus can be straightforwardly applied to the problem here. The classifier can be tuned to minimize overfitting by adjusting the Gaussian prior.

3.2 Architecture

The Propbank annotation is predicate-centered in that sense that only constituents that are semantic arguments and adjuncts (in a loose sense) to a predicate are annotated. Since the treebank sentences are very long and generally contain several verbs, the majority of the constituents are not related to the predicate in question. One obvious strategy is to assign a *NULL label* to the unannotated constituents, but it is a known fact that when negative samples *(NULL* constituents in this case) overwhelm positive samples (constituents that are actually annotated), the classifier will be heavily biased towards *NULL* constituents. Most systems find a way to filter out some of the negative samples to make the classification task more balanced. For example, [5] uses a two-stage architecture where a binary classifier is first used to label all the constituents as either *NULL* or *NON-NULL,* and then a multi-category classifier is run on the *NON-NULL* constituents to assign the semantic role labels. Xue and Palmer (2004) uses a three-stage architecture in which some negative samples are first filtered out with heuristics that exploit the syntactic structures represented in a parse tree. A binary classifier is then applied to further separate the positive samples from the negative samples and finally a multiply-category classified is applied to assign the semantic role labels to the positive samples.

3.3 Features

One characteristic of feature-based semantic role modeling is that the feature space is generally large. This is in contrast with the low-level NLP tasks such as POS tagging, which generally have a small feature space. A wide range of features have been shown to be useful in previous work on semantic role labeling [3] [9] [13] and we suspect that many more will be tested before they will settle down with a core set of features. In their preliminary work on Chinese semantic role labeling, Sun and Jurafsky (2004) has successfully adapted a number of the features to Chinese.

The features that we use are listed below:

- *Position:* The position is defined in relation to the predicate verb and the values are *before* and *after.*
- *Path:* The path between the constituent in focus and the predicate.
- *Head word and its part ofspeech:* The head word and its part-of-speech is often a good indicator of the semantic role label of a constituent.
- *Predicate:* The verb itself.
- *Subcat frame:* The rule that expands the parent of the verb.
- *Phrase type:* The syntactic category of the constituent in focus.
- *First and last word of the constituent in focus*
- *Phrase type of the sibling to the left*
- *Syntactic frame:* The syntactic frame consists of the NPs that surround the predicate verb. This feature is defined by the position of the constituent in focus in relation to this syntactic frame [13].
- *Combination features:* predicate-head word combination, predicate-phrase type combination.

4 Conclusions

In the experiments of Xue and Palmer (2004), the labeled precision and recall are 81.83% and 82.91% respectively for all sentences. The result showed that the semantic role labeling has potential in the natural language processing field. In the next it is very important to find better features for semantic role labeling of Chinese. We think the semantic role labeling can be successfully used in the Japanese.

Acknowledgment

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 14380166, 17300065, Exploratory Research, 17656128.

References

[1] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet proj ect. In *Proceedings of COLING/ACL,* pages 86–90, Montreal, Canada, 1998.

[2] Keh-Jiann Chen, Chu-Ren Huang, Feng-Yi Chen, Chi-Ching Luo, Ming-Chung Chang, and Chao-Jan Chen. Sinica Treebank: Design Criteria, Representational Issues and Implementation. In Anne Abeill´e, editor, *Building and Using Parsed Corpora.* Kluwer, 2004.

[3] D. Gildea and D. Jurafsky. Automatic labeling for semantic roles. *Computational Linguistics,* 28(3):245–288, 2002.

[4] Dan Gildea and Martha Palmer. The Necessity of Parsing for Predicate Argument Recognition. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics,* Philadelphia, PA, 2002.

[5] Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. Technical Report CSLR-2003-1, Center for Spoken Language Research at the University of Colorado, 2003.

[6] Kingsbury, Paul, Martha Palmer, and Mitch Marcus.2002. Adding semantic annotation to the Penn Treebank. In *Proceedings of HLT-02*

[7] Beth Levin. *English Verbs and Alternations: A Preliminary Investigation.* Chicago: The Unversity of Chicago Press, 1993.

[8] Martha Palmer, Dan Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *ComputationalLinguistics,* 31(1), 2005.

[9] Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of NAACL-HLT 2004,* Boston, Mass., 2004.

[10] Honglin Sun and Daniel Jurafsky. Shallow semantic parsing of chinese. In *Proceedings of NAACL 2004,* Boston, USA, 2004.

[11] Xue, Nianwen. 2002. Guidelines for the Penn Chinese Proposition Bank (1st Draft), UPenn.

[12] Nianwen Xue and Martha Palmer. Annotating the Propositions in the Penn Chinese Treebank. In *The Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing,* Sapporo, Japan, 2003.

[13] Nianwen Xue and Martha Palmer. Calibrating features for semantic role labeling. In *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing,* Barcelona, Spain, 2004.