Korean Semantic Role Labeling Using Korean PropBank Frame Files

Miran Seok¹, Hye-Jeong Song²,³, Chan-Young Park²,³, Jong-Dae Kim²,³, Yu-seop Kim²,³

¹IT Business DIV., Phill-IT Co., Korea
²Department of Convergence Software, Hallym University, Korea
³Bio-IT Research Center, Hallym University, Korea

smr4880@phill-it.com, {hjsong, cypark, kimjd, yskim01}@hallym.ac.kr

Abstract. Semantic role labeling (SRL) is a process that determines the semantic relation of a predicate and its arguments in a sentence and is an important factor in the semantic analysis of natural language processing. In this research, we propose a method for automatic SRL using frame files included in the Korean version of Proposition Bank (PropBank). First, we select the proper sense of the predicate among multiple senses. Senses of the predicate are classified according to the semantic and syntactic properties of its arguments. The semantic similarities between the noun words from the example sentence of each sense in the frame file and the nouns in the given sentence are measured. Finally, the sense with the highest similarity value is selected and the frame information of the sense is applied to the SRL. We acquired about 90% of accuracy.

Keywords: Semantic Role Labeling, Natural Language Processing, Proposition Bank, Frame Files, Semantic Similarity

1 Introduction

Semantic Role Labeling (SRL) is a task that automatically annotates the predicate-argument structure of a sentence with semantic roles [1]. To date, many manual semantic tagging tasks have been constructed; however, these tasks have required a great deal of time and cost. To solve this problem, we propose a method for automatic SRL using frame files included in the Korean version of Proposition Bank (PropBank), which is one of the most widely used corpora.

Frame files provide the guidelines for PropBank annotators and include a list of frame sets, which represent a set of syntactic frames. This study uses semantic similarity to select a semantically suitable frame from multiple frames of the predicate. Numerous studies related to semantic similarity have calculated the similarity between concepts using topological similarity. There are two approaches in topological similarity to measuring the semantic distance between concepts. The first approach evaluates similarity using the information content [2-5] (called the node-
based approach). The second approach evaluates similarity based on conceptual distance (called the edge-based approach) [6].

To select a proper frame, this study measures the semantic similarity between a noun word in the given sentence and a noun word in the example sentence of the frame file. We apply the weighted edge-based method to the semantic similarity measurement. The mapping information of the selected frame was used for SRL of the given predicate and its arguments. Approximately 90% of arguments are consistent with the results of manual SRL.

2 Proposition Bank

Proposition Bank (Propbank) is a corpus where the arguments of each verb predicate are annotated with their thematic roles [7]. PropBank is an annotation of syntactically parsed, or treebanked, structures with “predicate-argument” structures [8].

2.1 Frame Files

PropBank uses frame files for SRL. Frame files provide guidelines for Propbank annotators and include a list of framesets or coarse-grained senses of the verbs. A frameset represents a set of syntactic frames. A description for each frameset includes the list of verb specific roles and examples of different syntactic realizations of the verb [9]. Figure 1 shows an example of a frame file in Korean PropBank.

2.2 Korean PropBank

Korean Propbank Annotations are semantic annotations of the Korean English Treebank Annotations and Korean Treebank Version 2.0. Each verb and adjective occurring in the Treebank has been treated as a semantic predicate and the surrounding text has been annotated for arguments and adjuncts of the predicate. The verbs and adjectives have also been tagged with coarse-grained senses. This work was conducted in the Computer and Information Sciences Department at the University of Pennsylvania.

3 CoreNet

To calculate semantic similarity, CoreNet, which is the Korean concept-based lexical semantic network of BOLA (Bank of Language Resources)2, is used. CoreNet is connected to 2,938 hierarchical concepts and 92,448 lexical items. The CoreNet Korean lexical semantic network systematically and generally provides semantic

2 http://semanticweb.kaist.ac.kr/org/bora/index.php
information with the goal of solving the semantic ambiguity that has occurred in natural language processing. This study uses CBL1 (Korean noun of CoreNet).

4 Weighted Edge based Semantic Similarity Measurement

In this study, the distance between two noun classes is calculated by counting the number of weighted edges. The weight of the edge between the highest class and the second highest class and between the second highest class and the third highest class is 1/2 and 1/4, respectively. Hence, the weight decreases by half as it goes down the hierarchy.

![Diagram](image)

**Fig. 1.** A sample of hierarchical structure which describe the semantic relation among word classes

Weights of edges between noun words in *Nature* and *Animal* class in figure 1 are as follows. The weight of the edge between *Nature* and *Lifeless* is 1/16, the weight between *Lifeless* and *Thing* is 1/8, the weight between *Thing* and *Life* is 1/8, and the weight between *Life* and *Animal* is 1/16. The distance of the *Nature* and *Animal* class is 0.375 by summing all the weights of edges between two classes.

For example, to calculate the similarity between the two words “Machine Civilization” and “Lion,” the result can be obtained by summating the weights of the edges between 11322 and 11311. “Machine Civilization” is a member of *Artifact* and “Lion” is a member of *Animal*. As the sum of the weights of the edges decreases, the similarity increases. In this case, the distance between the two words is short.
5 Semantic Role Labeling Procedure

To assign a proper semantic role to an argument, first, we extract noun words that appear in a given sentence. We also extract noun words from example sentence in each sense in the frame file. In general, predicates have multiple sense frames, so we must deal with noun words in each sense frame. Second, we calculate the semantic similarity between the noun words in the given sentence and the noun words in the example sentence of each frame. The semantic similarity is obtained through the method presented in chapter 4. Third, we select the semantic frame which is most similar to the noun words of the given sentence among several sense frames. Finally, we map the frame information presented in the selected sense frame to the arguments in the given sentence.

6 Experimental Results and Discussion

In this study, we can assign the semantic roles by applying the frame files to a total of 4,468 arguments. We were able to accurately assign the semantic roles of 4,021 arguments. This tells us that our method shows about 90% accuracy.

However, this method has a high accuracy but also has a problem which cannot be ignored. We have found that this method has a lower recall than expected. The recall ratio was only 29.3%. Despite very high accuracy, this method requires a new complement because of its very low recall.

7 Conclusion

In this study, we use the frame file provided in Korean PropBank for Korean semantic role labeling. We estimate the semantic similarity between words and find an appropriate frame among the multiple frames of the predicate. We estimate the semantic similarity between words by weighted edge counting. This method, despite its high accuracy, showed a low recall ratio, which left much room for improvement.
In the future, we will study how to increase the recall ratio while maintaining accuracy.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (2015R1A2A2A01007333), and by Hallym University Research Fund, 2015 (HRF-201512-010).

References