Unsupervised Korean Word Sense Disambiguation using CoreNet

Kijong Han, Sangha Nam, Jiseong Kim, Younggyun Hahm, and Key-Sun Choi
Semantic Web Research Center, School of Computing, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon, South Korea
{han0ah, nam.sangha, jiseong, hahmyg, kschoi}@kaist.ac.kr

Abstract

In this study, we investigated unsupervised learning based Korean word sense disambiguation (WSD) using CoreNet, a Korean lexical semantic network. To facilitate the application of WSD to practical natural language processing problems, a reasonable method is required to distinguish between sense candidates. We therefore performed coarse-grained Korean WSD studies while utilizing the hierarchical semantic categories of CoreNet to distinguish between sense candidates. In our unsupervised approach, we applied a knowledge-based model that incorporated a Markov random field and dependency parsing to the Korean language in addition to utilizing the semantic categories of CoreNet. Our experimental results demonstrate that the developed CoreNet based coarse-grained WSD technique exhibited an 80.9% accuracy on the datasets we constructed, and was proven to be effective for practical applications.

Keywords: Word Sense Disambiguation, CoreNet, Markov Random Field

1. Introduction

Words that have the same form can have different meanings. Word Sense Disambiguation(WSD) is to select the correct meaning of a word in context. It is an important problem that can be utilized in many problems of natural language processing such as machine translation and information extraction (Chaplot et al., 2015). In English, studies are typically conducted on the basis of the senses listed in the Princeton WordNet(PWN) as candidates for the meaning of words (Navigli et al., 2007; Chaplot et al., 2015). In this study, we resolve the ambiguity of words based on the senses listed in CoreNet (Choi et al, 2004), a Korean lexical semantic network.

Both PWN and CoreNet are fine-grained resources, so it is difficult for even human annotators to correctly identify the senses of words. In order for WSD to become an enabling technique for end-to-end applications, it requires the ability to make reasonable sense distinctions (Navigli et al., 2007). In English, coarse-grained WSD is performed by semi-automatically clustering PWN senses (Navigli et al., 2007). However, Korean dictionaries also list homograph and polyseme numbers for each sense. Homographs are words that have the same form but completely different meanings, and polysemes are words that have the same broad meaning but different etymological meanings. For this reason, in Korean language studies, coarse-grained WSD is typically conducted at the homograph level (Shin and Ock, 2016).

WSD methods can generally be divided into supervised and unsupervised learning methods. Supervised learning methods learn sense-tagged corpora and show relatively high levels of performance. However, it is expensive and time-consuming to construct such corpora. The results of one Korean language study showed a 96.5% accuracy based on learning a corpus consisting of 10 million sense tagged words (Shin and Ock, 2016).

In contrast, unsupervised learning methods are relatively low performance but do not require training data. For example, knowledge-based methods that employ lexical semantic networks have been successful (Agirre et al., 2014; Chaplot et al., 2015) because they are able to obtain wide coverage and good performance using structured knowledge (Iacobacci et al., 2016). Along these same lines, state-of-the-art research has been conducted based on Markov Random Fields (MRF) (Chaplot et al., 2015), which convert sentences to an MRF model through part-of-speech (POS) tagging and dependency parsing, and then determines the meanings of all target words in the sentence by way of a maximum a posteriori (MAP) query.

The remainder of this paper is organized as follows. Section 2 describes our investigation of unsupervised Korean WSD using the semantic category "concept" in CoreNet and homographs for coarse-grained WSD. In Section 3, we explain our MRF based method (Chaplot et al., 2015) and its application to CoreNet. Section 4 presents the datasets created during the evaluation of the proposed approach, and outlines the results of the experiments conducted to demonstrate the performance and efficacy of the proposed approach for distinguishing candidates of word senses using the semantic category "concept" in CoreNet. Finally, our conclusions are presented in Section 5, along with our plans for future work.

2. Background

2.1 CoreNet and Concept

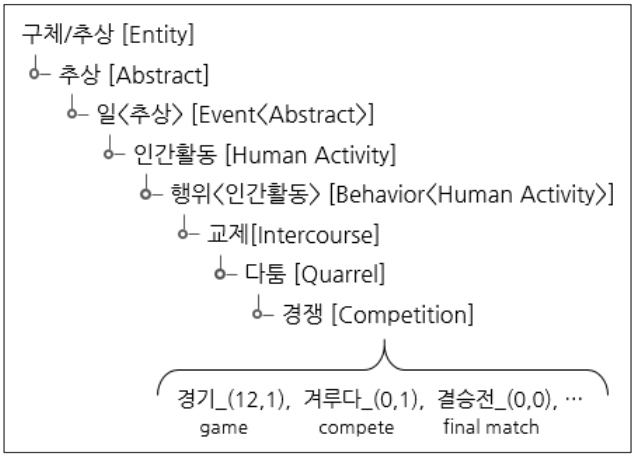

Figure 1: Example concept hierarchy in CoreNet.

CoreNet (Choi et al., 2004) is a lexical semantic network that represents the senses and relationships of Korean nouns, adjectives, and verbs. There are about 73,000 senses in CoreNet, and each sense contains additional resources, such as definitions and usage.

The key feature of CoreNet is its concept hierarchy. In CoreNet, the term "concept" refers to a semantic category, and every sense in CoreNet is mapped to one or more concepts. The concept hierarchy is based on Japanese NTT Goidaikei, which is a Japanese lexicon (Ikehara et al.,

1997), and includes a total 2,954 concepts, 277 of which are Korean in origin. These concepts constitute a hierarchy with a maximum depth of 12, each of which is mapped to Chinese, Japanese, and PWN senses.
An example of the concept hierarchy is illustrated in Figure 1 for the concept "Competition," which has a depth of eight. The verb sense "compete" and noun sense "game" and "final match" are mapped to this concept. In each sense, the first number after the word represents the homograph number and the second represents the polysemy number.

2.2 WSD with CoreNet

There is a need for coarse-grained WSD to facilitate practical applications of WSD (Navigli et al., 2007). To this end, we used homographs as in other Korean studies, as well as the concepts in CoreNet. In CoreNet, senses that have different homograph numbers are typically mapped to different concepts; however, there are a few exceptions.

- 사과[sa-gwa]-(1,0) : abbreviation of Korean cantaloupe
- 사과[sa-gwa]-(3,0) : apple

The two senses of the word "sa-gwa" have different homograph numbers, but both are mapped to the same concept "Edible Fruit." In this case, these senses are clustered into the same candidate. Senses mapped to the same concept are considered to be the same candidate; thus, the concept can be utilized in the WSD method. That is, the sense candidates of the word "sa-gwa" can be classified as shown in Table 1. Concepts such as "Apology" and "Edible Fruit" can also be utilized when evaluating the suitability of each candidate for the word "sa-gwa." In addition, as shown in Section 5, these types of candidate distinctions based on CoreNet concepts are meaningful in other natural language processing (NLP) tasks.

	Concept	(Homograph, Polyseme) Number	Definition
O	**Apology**	(8,0)	- apology.
1	**Edible Fruit**	(1,0)	- abbreviation of Korean cantaloupe.
		(3,0)	- apple.
2	**Study general/ Subject of Study**	(6,1)	- 4 courses of cultivating one's moral sense.
		(6,2)	- 4 courses of Confucianism.

Table 1: Examples of sense candidates for word '사과[sa-gwa]' based on the CoreNet Concept.

3. Approach

The primary approach outlined in this paper is the MRF based method, which will be described in Section 3.2. However, an alternate approach is the term frequency–inverse document frequency (TF-IDF) vector similarity method, which will be described in Section 3.1, and can be used to obtain the frequency values of concepts necessary to implement the MRF based method, as described in Section 3.2.

3.1 TF-IDF Vector Similarity

The TF-IDF Vector Similarity method uses the cosine similarity between the TF-IDF vector of all definition and usage sentences mapped to the concept associated with each candidate and the TF-IDF vector of the sentence containing the words to be disambiguated. The candidate that has the largest cosine similarity is selected. If the candidates are mapped to multiple concepts, then it is considered correct to disambiguate the sense when the system select one of those concepts. This method is based on the principle that the more the same words appear in the two sentences, the more the TF-IDF vector cosine similarity increases.

3.2 MRF Method

A MRF is an undirected graphical model that consists of set of random variables. Each node in the graph represent a random variable, and each random variable is only dependent on another random variable that represents another node that is directly connected by an edge. This model has been used to solve many NLP problems (Jung et al., 1996; Chaplot et al., 2015)
In this study, we applied an MRF based WSD method (Chaplot et al., 2015) to the Korean language using the concept hierarchy in CoreNet. In this method, target words for the WSD in a sentence are selected as nodes in the MRF, and edges are only generated for two directly connected words in the dependency tree. Finally, the senses of all the words are jointly disambiguated by way of a MAP query on this MRF model. We adopted the detailed methods described by Chaplot et al. (2015), and outline how they were applied to the Korean language in the following text.

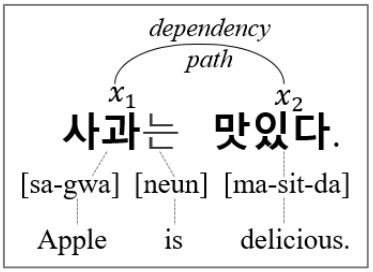

Figure 2 : An example of converting sentence to MRF

First, all common nouns, verbs, and adjectives in a sentence were designated WSD target words. Then, we set the random variables representing the concepts of these words to $X = \{x_1, x_2, \ldots, x_n\}$, where x_i can take m_i possible concept values. The concepts that x_i could take on were $s_i^1, s_i^2, \ldots, s_i^{m_i}$. In the case of the sentence shown in Figure 2, the noun "sa-gwa" and adjective "ma-sit-da" were selected as the target words, and the postposition "neun" was not selected. The random variable x_1 represented the word "sa-gwa," and the random variable x_2 was used to represent "ma-sit-da." The candidates of x_1 were s_1^1 = "Apology," s_1^2 = "Edible Fruit," and so on, as shown in Table 1.
There are node potential function $\psi(x_i)$, and edge potential function $\psi(x_i, x_j)$ for this MRF are as follows.

$$\psi(x_i = s_i^a) \propto \log(frequency(s_i^a) + e)$$

where $frequency(s_i^a)$ refers to the frequency of occurrence of concept s_i^a. These values were measured for 1.7 million words from 10% of the full text of Wikipedia using the TF-IDF method described in Section 3.1. In this process, we set the cosine similarity threshold value to 0.14,

which resulted in a precision of 0.951 and a recall value of 0.287 for our datasets..

$$\psi\left(x_i = s_i^a, x_j = s_j^b\right) \propto Relatedness(s_i^a, s_j^b)$$

The edge potential function, which indicates when two related words simultaneously have certain concepts, is proportional to the relatedness between the two concepts. Edges are only generated when two words are directly connected on the dependency tree. If we let this set of edges be E, then $\{x_i, x_j\} \in E$. The relatedness can be measured by the following two methods, and experiments were conducted for both.

(1) $Relatedness\left(s_i^a, s_j^b\right) = 1/(shortestpath\left(s_i^a, s_j^b\right) + 1)$
(2) $Relatedness\left(s_i^a, s_j^b\right) = \log(frequency\left(s_i^a, s_j^b\right) + e)$

where $shortestpath\left(s_i^a, s_j^b\right)$ refers to the inverse of the shortest path between concept s_i^a and s_j^b in CoreNet, and $frequency\left(s_i^a, s_j^b\right)$ refers to the frequency of co-occurrence of the two concepts s_i^a, s_j^b in same sentence. This measurement method is same as that employed by Chaplot et al. (2015), and can be used to obtain the node potential value..

$$\psi(\mathrm{X}) = \prod_{x_i \in X} \psi(x_i) \prod_{\{x_i, x_j\} \in E} \psi(x_i, x_j)$$

The final potential function of this model is as shown above. Let S be the set of disambiguated concepts for each word. Then, we can find S jointly through the MAP query shown below. Note that we used the library(Ankan and Panda, 2015) when implementing this MRF model.

$$\arg\max_S \psi(\mathrm{X} = \mathrm{S})$$

4. Experiments

4.1 Datasets

Each of our datasets consisted of a sentence and a word that was disambiguated in the sentence. To create our datasets, we randomly selected sentences from articles featured on Wikipedia, and then randomly selected either a noun, verb, or adjective. Three annotators were used to tag the proper senses to the words in these datasets. A total of 470 datasets were constructed, and the number of datasets with ambiguity, i.e., the number of datasets that had more than two candidates, was 215. The statistics for the datasets are shown in Table 2. All of the sentences were different, but the WSD target words may have been the same, so the number of words in Table 2 refers to the number of different words in our datasets.

	# data	# word
All data	470	322
Data with ambiguity	215	144

Table 2 : Statistics of our datasets

4.2 Performance

Method	RANDOM	TF-IDF	MRF SP	MRF co-occur
Accuracy for all data	73.6	88.5	**91.3**	91.1
Accuracy for data with ambiguity	42.2	74.9	**80.9**	80.5

Table 3 Accuracy by method

The results after measuring the accuracy of the data in the datasets is shown in Table 3. The accuracy was measured as follows. In the RANDOM baseline, a candidate was randomly selected, and the results of five trials were averaged. TF-IDF refers to the TF-IDF vector similarity based method described in Section 3.1. "MRF SP" and "MRF co-occur" refer to the MRF based method described in Section 3.2. The differences between the columns was used as the method of relatedness measurement. "MRP SP" used shortest path, as described in (1), and "MRF co-occur" used the method of co-occurrence, as described in (2). The MRF based method exhibited an 80.9% accuracy on our datasets with ambiguity, and the performance of this method was higher than that of the RANDOM and TF-IDF based methods. Our accuracy was lower than that of the recent Korean WSD study, which employed a supervised approach (Shin and Ock, 2016); however, our method is unsupervised, and therefore has the advantage that it can be applied to any document without learning.

4.3 Results of Applying Our WSD Method

The performance of the task of Relation Extraction from sentences improved when our WSD method was applied. This confirms that coarse-grained WSD based on the CoreNet concept is meaningful in real applications. The details of our application method are as follows. We applied our WSD method to the convolutional neural network based relation extraction model, which was implemented by our research team for the Korean language based on the description in (Zeng et al., 2014), and word embedding vectors of each token from the sentences were used as inputs to this relation extraction model. Word embedding has an advantage in that tokens with high semantic relevance are generated with similar real vector values. However, if the sentence is only tokenized in morpheme units, the algorithm cannot distinguish between ambiguous words. Thus, we applied a sense number to each word token using our WSD method so that words that had different senses but were in the same form could be distinguished at the time of embedding. This sense-tagged embedding exhibited a 7% higher F1-score performance for relation extraction, as shown in Table 4.

Unit of word embedding token	Morpheme	Morpheme + Sense
Relation Extraction F1-score	0.474	**0.544**

Table 4 : F1-score of Relation Extraction

5. Conclusion and Future Work

In this study, we implemented an unsupervised coarse-grained WSD algorithm using the semantic category "concept" in CoreNet for coarse-grained sense distinction. We then confirmed that it was meaningful by computing how our WSD algorithm improved the relation extraction F1-score. For our unsupervised approach, we also utilized the CoreNet concept as applied to knowledge-based MRF WSD model, and computed an accuracy of 80.9% using the datasets we constructed.

In the Korean language, the Sejong corpus consist of 10 million POS and sense tagged words. However, the sense numbers of the Sejong corpus and those of CoreNet originated from different Korean dictionaries. In the future, we plan to revise our method to utilize the Sejong corpus in CoreNet because we believe that this will improve the performance of Korean language WSD.

6. Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

7. Bibliographical References

Agirre, E., de Lacalle, O. L., and Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1), 57-84.

Ankan, A., and Panda, A. (2015). pgmpy: Probabilistic Graphical Models using Python. In *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*.

Chaplot, D. S., Bhattacharyya, P., and Paranjape, A. (2015). Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser. In *AAAI* (pp. 2217-2223).

Choi, K. S., Bae, H. S, Kang, W., Lee, J., Kim, E., Kim, H., Kim, D., Song, Y., and Shin, H. (2004). Korean-Chinese-Japanese Multilingual Wordnet with Shared Semantic Hierarchy. In *LREC*.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2016). Embeddings for Word Sense Disambiguation: An Evaluation Study. In *ACL (1)*.

Ikehara, S. et al. (1997). The Semantic System, volume 1 of Goidaikei – A Japanese Lexicon. Tokyo: Iwanami Shoten.

Jung, S. Y., Park, Y. C., Choi, K. S., and Kim, Y. (1996). Markov random field based English part-of-speech tagging system. In *Proceedings of the 16th conference on Computational linguistics*-Volume 1 (pp. 236-242). Association for Computational Linguistics.

Navigli, R., Litkowski, K. C., and Hargraves, O. (2007). Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 30-35). Association for Computational Linguistics.

Shin, J. C., and Ock, C. Y. (2016). Improvement of Korean Homograph Disambiguation using Korean Lexical Semantic Network (UWordMap). *Journal of KIISE*, 43(1), 71-79.

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). Relation Classification via Convolutional Deep Neural Network. In *COLING* (pp. 2335-2344).