Automatic Wordnet Mapping: from CoreNet to Princeton WordNet

Jiseong Kim, Younggyun Hahm, Sunggoo Kwon, Key-Sun Choi
Semantic Web Research Center, School of Computing, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon, Korea
{jiseong, hahmyg, fanafa, kschoi}@kaist.ac.kr

Abstract

CoreNet is a lexico-semantic network of 73,100 Korean word senses, which are categorized under 2,937 semantic categories organized in a taxonomy. Recently, to foster the more widespread use of CoreNet, there was an attempt to map the semantic categories of CoreNet into synsets of Princeton WordNet by lexical relations such as synonymy, hyponymy, and hypernymy relations. One of the limitations of the existing mapping is that it is only focused on mapping the semantic categories, but not on mapping the word senses, which are the majority part (96%) of CoreNet. To boost bridging the gap between CoreNet and WordNet, we introduce the automatic mapping approach to link the word senses of CoreNet into WordNet synsets. The evaluation shows that our approach successfully maps previously unmapped 38,028 word senses into WordNet synsets with the precision of 91.2% ($\pm$1.14 with 99% confidence).

Keywords: wordnet, automatic mapping, automatic construction

1. Introduction

CoreNet (Choi et al., 2004) is a semantic hierarchy of Korean word senses, which has been built by KAIST since 1994 based on CoreNet concept hierarchy originated from NTT Goi-Taikei (Ikehara et al., 1997) concept hierarchy.

The CoreNet hierarchy comprises mainly two parts, non-terminal part and terminal part. The non-terminal part of the hierarchy comprises 2,937 semantic categories, called CoreNet concept, as non-terminal nodes, which are organized by a taxonomic relation, while the terminal part of the hierarchy comprises 73,100 Korean word senses as terminal nodes, which are separated from each other (i.e., there is no link between the word senses) and there is only a link between word sense and its semantic category with unknown lexical relations such as *is-a* and *part-of*; the unknown relation means there may be *is-a* or *part-of* relation between word sense and its semantic category, but a label of the relation is not revealed. An example of the CoreNet hierarchy is shown in Figure 1.

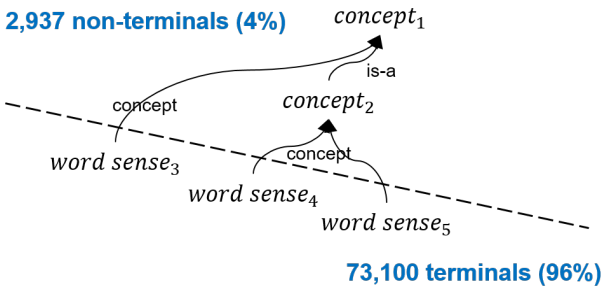

Figure 1: An example of the hierarchy of CoreNet

To extend CoreNet into other languages and to promote its broader utilization for diverse NLP application, Kang et al. (2010) made an attempt to map the CoreNet hierarchy into the Princeton WordNet hierarchy. The scope of the mapping encompassed all of 2,937 semantic categories, which were successfully mapped to WordNet synsets with synonymy, hypernymy, and hyponymy relations.

Although Kang et al. (2010) mapped the almost all of the semantic categories of CoreNet, word senses of CoreNet is not in the scope of the mapping, and still remains out of mapping; this leads to the fact that, from the perspective of NLP application, WordNet operations such as path similarity hard to be applied to the word senses, which are the majority part (96%) of CoreNet, because there is no mapping for the word senses, and, moreover, lexical relations between word senses and mapped part (semantic categories) of CoreNet are totally unknown.

To overcome this limitation, it can be an option that human annotators manually label WordNet synsets to all of the word senses of CoreNet with appropriate lexical relations; however, it requires the excessive cost of human labors.

By the motivation from these facts, in this paper, we introduce an automatic mapping approach that automatically maps the unmapped word senses of CoreNet into WordNet hierarchy to boost bridging the gap between CoreNet and WordNet.

Our contributions are as follows:

(1) We present a wordnet mapping approach that automatically maps word senses in wordnets of different languages, especially Korean and English, using novel semantic features between wordnet hierarchies.

(2) We present a new language resource that contains mappings between CoreNet word senses and WordNet synsets with synonymy relation. To the best of our knowledge, it is the first attempt to map CoreNet word senses into WordNet hierarchy.

In the following sections, we describe the problem to be dealt in this paper and our approach much in detail.

2. Problem Statement

Before mapping CoreNet word senses into WordNet synsets, synset candidates for each CoreNet word sense are selected by the following list of actions.

(1) Given a word sense of CoreNet, the word sense is translated into N English words by bilingual dic-

tionaries; the translation is done based on the exact matching of lemma and part-of-speech.

(2) M synsets are selected as synset candidates for the given word sense of CoreNet if lemma and part-of-speech are exactly matched with one of the N translated English words of the given word sense of CoreNet.

An example of the synset candidate selection is shown in Figure 2.

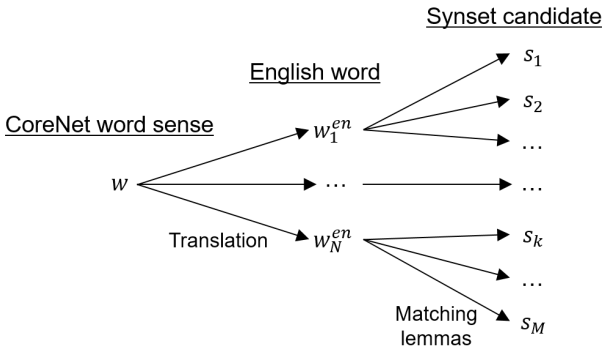

Figure 2: An example of synset candidates for a given word sense of CoreNet

We used 70 domain-specific bilingual dictionaries, Sejong electronic dictionary[1], CoreNet to WordNet mapping (Kang et al., 2010), Naver English dictionary[2], and Google translation[3]. Cumulative coverage of the dictionaries is shown in Table 1.

Table 1: Cumulative coverage of bilingual dictionaries to CoreNet word senses: *Domain* denotes 70 domain-specific dictionaries, *Sejong* denotes Sejong electronic dictionary, *Mapping* denotes CoreNet to WordNet mapping, *Naver* denotes Naver English dictionary, and *Google* denotes Google translation.

Dictionaries	Coverage
Domain	61.82%
Domain, *Sejong*	65.31%
Domain, *Sejong*, *Mapping*	65.35%
Domain, *Sejong*, *Mapping*, *Naver*	72.02%
Domain, *Sejong*, *Mapping*, *Naver*, *Google*	100.0%

After selecting synset candidates, the problem to be dealt in this paper can be translated into word sense disambiguation problem that is to select synonymous synsets from synset candidates for each word sense of CoreNet.
Our approach solves this problem by supervised classification with semantic features of wordnet hierarchies, which is described in the following section in detail.

[1]The Sejong electronic dictionary has been developed by several Korean linguistic researchers, funded by Ministry of Culture and Tourism, Republic of Korea. (http://www.sejong.or.kr)

[2]http://dic.naver.com

[3]https://translate.google.com

3. Mapping Approach

3.1. Semantic Feature Extraction

For a given CoreNet word sense and its synset candidates, three different scores are measured as a feature of semantic similarity between the given CoreNet word sense and its synset candidates.

Vertical similarity is to measure a vertical similarity between hierarchies of CoreNet word sense and its synset candidate; an example is shown in Figure 3.

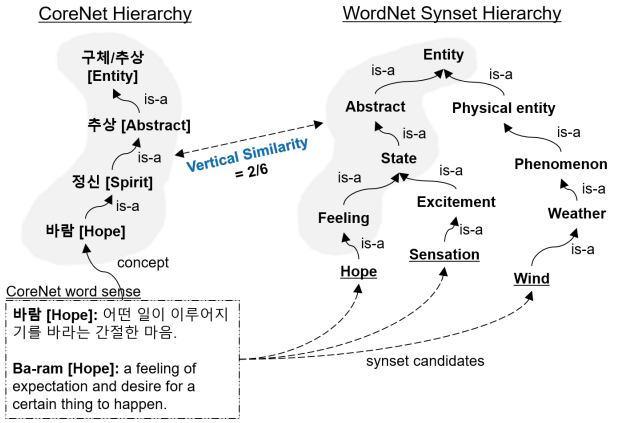

Figure 3: An example of vertical similarity between CoreNet word sense 'Ba-ram' and its synset candidates

The basic idea is that the vertical similarity increases as much as CoreNet word sense and its synset candidate share common ancestors on their hierarchies.
More precisely, the vertical similarity is a translation-based Jaccard similarity between a set of ancestral semantic categories of a given CoreNet word sense and a set of ancestral hypernym and holonym synsets of a synset candidate; the following formula explains the idea:

$$VertSim(w, s) = JaccardSim(Anc_{CN}(w), Anc_{WN}(s))$$

where w denotes a CoreNet word sense, s denotes a synset candidate for w, $Anc_{CN}(w)$ denotes a set of ancestral semantic categories of w, and $Anc_{WN}(s)$ denotes a set of ancestral hypernym and holonym synsets of s.

Horizontal similarity is to measure a horizontal similarity between hierarchies of CoreNet word sense and its synset candidates; an example is shown in Figure 4.
The basic idea is that the horizontal similarity increases as much as CoreNet word sense and its synset candidate share common siblings on their hierarchies.
More precisely, the horizontal similarity is a translation-based Jaccard similarity between a set of sibling word senses of a given CoreNet word sense and a set of sibling synsets of a synset candidate; the following formula explains the idea:

$$HoriSim(w, s) = JaccardSim(Sib_{CN}(w), Sib_{WN}(s))$$

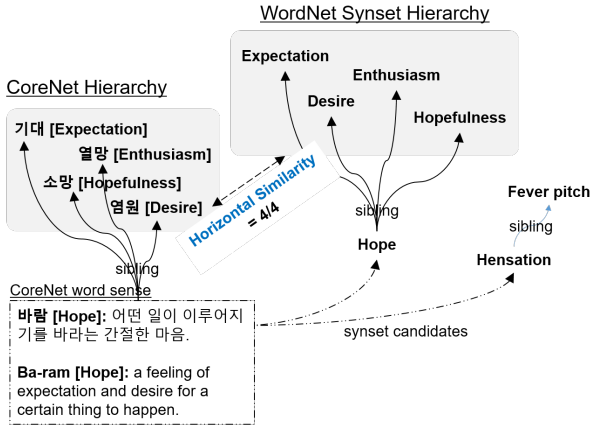

Figure 4: An example of horizontal similarity between CoreNet word sense 'Ba-ram' and its synset candidates

where w denotes a CoreNet word sense, s denotes a synset candidate for w, $Sib_{CN}(w)$ denotes a set of sibling word senses of w, and $Sib_{WN}(s)$ denotes a set of sibling synsets of s.

Conceptual word coverage is to measure a conceptual similarity between CoreNet word sense and its synset candidate based on their conceptual words contained in semantic categories, definition statements, and example sentences.

More precisely, the conceptual word coverage is the measurement of how many words contained in names of semantic categories for a given CoreNet word sense are covered by words contained in definition statements and example sentences of a synset candidate, based on translation; the following formula explains the idea:

$$ConceptCover(w,s) = \frac{|\{w'|w' \in Cwords(w) \cap \{Dwords(s) \cup Ewords(s)\}\}|}{|\{w'|w' \in Cwords(w))\}|}$$

where w denotes a CoreNet word sense, s denotes a synset candidate for w, $Cwords(w)$ denotes a set of words contained in names of w's semantic categories, $Dwords(w)$ denotes a set of words contained in definition statements of s, and $Ewords(w)$ denotes a set of words contained in example sentences of s.

3.2. Basic Feature Extraction

The above-mentioned three semantic features are mainly focused on information about a semantic relationship among word senses, not on word sense itself. To supplement features describing a word sense itself and support the above-mentioned three semantic features for a better performance in a classification task, the basic features of word senses, part-of-speech and semantic categories of word senses, are also used as a feature for training.

3.3. Mapping by Decision Tree Classifier

Given CoreNet word senses and their synset candidates with five different features, our goal is to combine the five features to classify synset candidates as linking or discarding.

The combination of the features is performed by a decision tree classifier which shows the best performance among other different classifiers in our experiments described in the following section.

To link CoreNet word senses into WordNet synsets, there are two phases for training a decision tree classifier (training phase) and linking/discarding synset candidates by the trained classifier (mapping phase).

In the training phase shown in Figure 5, a decision tree classifier is trained on the five features extracted from CoreNet word sense w and synset candidate s contained in manually labeled data.

The manually labeled data is built on the samples from all CoreNet word senses and their Top-2 synset candidates where Top-2 means only two synset candidates are selected from the front of the candidate list sorted by linear summation score of vertical similarity, horizontal similarity, and conceptual word coverage in a descending order. The selected Top-2 synset candidates are labeled as linking or discarding.

The reason why we picked only Top-2 synset candidates for each CoreNet word sense is to avoid imbalance of training and test datasets. If negative examples in the datasets overwhelm positive examples, the precision of classification results would be dropped rapidly by enormous false positives; it is showed in Table 2 that precision and coverage are dropped by increasing the ratio of negative examples to positive examples. There are also reports that standard classifiers such as decision trees give sub-optimal classification results when trained on imbalanced datasets (Lane et al., 2012; Haixiang et al., 2017).

Table 2: Dropping of precision and coverage by increasing the ratio of negative examples to positive examples

negative # / positive #	Precision	Coverage
0.5	0.9121	0.9347
1.0	0.908	0.9258
2.0	0.8789	0.9094
3.0	0.8612	0.8992
4.0	0.8489	0.8934
5.0	0.839	0.8846
6.0	0.8279	0.8821
7.0	0.8113	0.8812
8.0	0.8098	0.8783
9.0	0.7911	0.8752

In the mapping phase shown in Figure 6, a trained model of a decision tree classifier is applied to all the pairs of CoreNet word sense w and its Top-2 synset candidates s to classify as linking or discarding. As a result of classification, synset candidates classified as linking are finally mapped to the corresponding CoreNet word sense as synonymy relation.

4. Evaluation

In this section, we evaluate the performance of each of the five features as well as their combination.

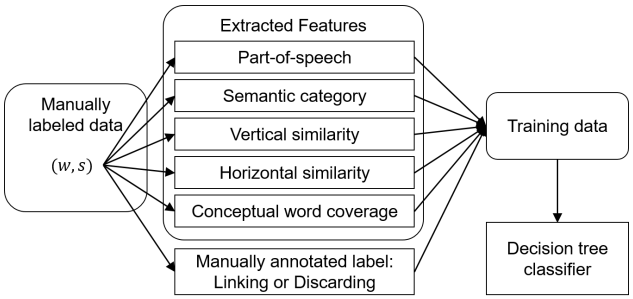

Figure 5: Training phase

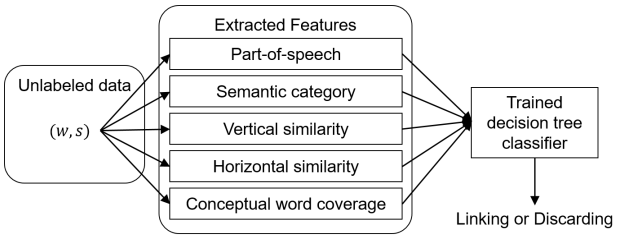

Figure 6: Mapping phase

For evaluation, we use the manually labeled 6,041 CoreNet word senses with 8,655 positive links to synonymous synsets and 2,700 negative links to nonsynonymous synsets.

In the evaluation, we use the two measurements; the one is *precision* defined as the proportion of correctly linked synonymous synsets over all of linking results, and the other is *coverage* defined as the proportion of CoreNet word senses linked to synonymous synsets over all CoreNet word senses to be linked.

All the performance scores are evaluated by 10-fold cross-validation with 90% of labeled data for training and the remaining 10% of labeled data for testing.

The performance scores of decision tree classifiers trained on each feature and combination of all the features are shown in Table 3. 'Random' in the table classifies synset candidates as linking or discarding in a random manner. In summary, the performance of each feature is not good enough, but, when they are combined, the performance is fairly improved up to 91.2% with 99% confidence level.

Table 3: Performance of each feature and the combination

	Precision	Coverage
Random	0.7627	0.7235
Part-of-speech	0.7613	1.0
Semantic category	0.7903	0.8425
VertSim	0.8308	0.9388
HoriSim	0.8107	0.9408
ConceptCover	0.7585	1.0
Combination	0.9121	0.9347

In Table 4, the performance scores of five different classifiers are shown. The classifiers are trained on the combination of all the features. Although the decision tree classifier shows the relatively low coverage, it achieves the best performance of 91.2% precision with 99% confidence level.

Table 4: Performance of different classification models

	Precision	Coverage
Logistic regression	0.8099	0.9815
Naive Bayes	0.8085	0.9321
Decision tree	0.9121	0.9347
SVM	0.7669	0.9983
Multilayer perceptron	0.8315	0.9668

By using the decision tree classifier trained on the combination of all the features, we classified all CoreNet word senses and obtained the mappings between 38,028 CoreNet word senses and their synonymous WordNet synsets. In other words, we constructed a Korean wordnet composed of 38,028 Korean word senses (33,956 nouns, 3,617 verbs, 355 adjectives) with the precision of 91.2% (± 1.14 with 99% confidence level).

5. Related Work

(Lee et al., 2000) introduced the automatic mapping between Korean word senses in bilingual dictionaries and synsets in Princeton WordNet by word sense disambiguation. They reported that 21,654 Korean word senses are mapped to WordNet synset with the precision of 93.59% by decision tree learning on six heuristic features.

In other languages, especially Persian, many works tried to map word senses in bilingual dictionaries to synsets in Princeton WordNet in a similar way (Dehkharghani and Shamsfard, 2009; Mousavi and Faili, 2017).

The above-mentioned works have a common point that they target to map word senses in bilingual dictionaries that are not organized in a semantic network. Inevitably, the features used in their approaches lack the use of hierarchical features in their own languages.

The difference of our work from them is that semantic features introduced in this paper fully utilize hierarchical features of both source language (Korean) and target language (English).

6. Conclusion

This paper has explored an automatic mapping of wordnets, especially CoreNet and Princeton WordNet, by supervised classification with novel semantic features between wordnet hierarchies.

The experiments showed that the combination of all the features introduced in this paper achieves the better performance than each of individual features, and a decision tree classifier is the best choice for performing the combination of all the features.

Our approach is not restricted to CoreNet and Princeton WordNet, but it can be applied on any wordnets with traditional wordnet structures whose word senses are organized in the same lexical relations and have definition statements and example sentences.

After applying our mapping approach on all the CoreNet word senses, we obtained the new synonym mapping between 38,028 word senses of CoreNet and corresponding WordNet synsets. A series of experiments showed that the accuracy of mapping is over 90%.

7. Acknowledgement

This research was financially supported by the Ministry of Trade, Industry and Energy(MOTIE) and Korea Institute for Advancement of Technology(KIAT) through the International Cooperative R&D program.

8. Bibliographical References

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*.

Dehkharghani, R. and Shamsfard, M. (2009). Mapping persian words to wordnet synsets. *International Journal of Interactive Multimedia and Artificial Intelligence*, 1(2).

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: review of methods and applications. *Expert Systems with Applications*, 73:220–239.

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., and Hayashi, Y. (1997). Goi-taikei-a japanese lexicon.

Kang, I.-S., Kang, S.-J., Nam, S.-J., and Choi, K.-S. (2010). Linking corenet to wordnet-some aspect and interim consideration. In *Proceedings of the 5th Global WordNet Conference*, pages 239–242.

Lane, P. C., Clarke, D., and Hender, P. (2012). On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems*, 53(4):712–718.

Lee, C., Lee, G., and Yun, S. J. (2000). Automatic wordnet mapping using word sense disambiguation. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 142–147. Association for Computational Linguistics.

Mousavi, Z. and Faili, H. (2017). Persian wordnet construction using supervised learning. *arXiv preprint arXiv:1704.03223*.