Master's Thesis

Korean Word Sense Disambiguation using Parallel Corpus as Additional Resource

Chungen Li

Department of Computer Science and Engineering

Pohang University of Science and Technology

2013

병렬 발뭉치를 추가적 자원으로 이용한 한국어 단어 중의성 해소

Korean Word Sense Disambiguation using Parallel Corpus as Additional Resource

Korean Word Sense Disambiguation using Parallel Corpus as Additional Resource

by

Chungen Li

Department of Computer Science and Engineering

Pohang University of Science and Technology

A thesis submitted to the faculty of Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Master of Science in the Computer Science and Engineering

Pohang, Korea
June 4, 2013.
Approved by

Jong-Hyeok Lee
Major Advisor

Korean Word Sense Disambiguation using Parallel Corpus as Additional Resource

Chungen Li

The undersigned have examined this thesis and hereby certify that it is worthy of acceptance for a master's degree from POSTECH.

6/4/2013

Committee Chair	Jong-Hyeok Lee	(Seal)
Member	Gary Geunbae Lee	(Seal)
Member	Hwanjo Yu	(Seal)

MCSE 20111056 Chungen Li. Korean Word Sense Disambiguation using Parallel Corpus as Additional Resource. Department of Computer Science and Engineering, 2013. 22 pages. Advisor: Jong-Hyeok Lee. Text in English.

ABSTRACT

Most previous research on Korean Word Sense Disambiguation (WSD) were focusing on unsupervised corpus-based or knowledge-based approach because they suffered from lack of sense-tagged Korean corpora.

Recently, along with great effort of constructing sense-tagged Korean corpus by government and researchers, finding appropriate features for supervised learning approach and improving its prediction accuracy became an issue. To achieve higher word-sense prediction accuracy, this paper aimed to find most appropriate features for Korean WSD based on Conditional Random Field (CRF) approach.

Also, we utilized Korean-Japanese parallel corpus to enlarge size of sense-tagged corpus, and improved prediction accuracy with it. Experimental result reveals that our method can achieve 95.67% of prediction accuracy.

This thesis is dedicated to my families and friends
who have unlimited potential but could not attend the school.

Contents

List of Figures iii

List of Tables iv

1 Introduction 1

2 Overall Architecture 4

3 Sense-tagged Corpora 6

3.1 21st Century Sejong Corpora 6

3.2 Construct Sense-Tagged Korean Corpus via Korean-Japanese Parallel Corpora 7

3.2.1 Align Korean-Japanese Parallel Corpus in word-level . . 7

3.2.2 Tag ambiguous Korean words by Japanese-translations . 8

3.2.3 Cluster synonymous Japanese-translations & Map to the sense id 9

3.2.4 Combine Sejong and Constructed Corpora 11

4 Proposed Method 13

4.1 Training WSD Classifier 13

4.2 Disambiguate Words in New Context 14

5 **Experimental Results** **15**

5.1 Accuracy of Sense-Tagged Corpora 15

5.2 Finding Appropriate Window Size 16

5.3 WSD Prediction Accuracy 16

5.3.1 Training and Test Data 17

5.3.2 Evaluation Measure 17

5.3.3 WSD Systems with Different Features 17

5.3.4 Other Korean WSD Systems 18

5.3.5 Comparison with SENSEVAL (SEMEVAL) 18

6 **Conclusion** **21**

Bibliography **22**

List of Figures

2.1 Overall Architecture of Proposed WSD System 5

List of Tables

4.1 Comparison table for two feature set 14

5.1 Classifier Accuracy Comparison using 5-fold Cross Validation . . 16

5.2 The Comparison of Different WSD Systems 17

5.3 The Comparison With Previous Work 18

Chapter 1

Introduction

In computational linguistic, lexical ambiguity is one of the first problems that people faced with in Natural Language Processing (NLP) area [6].

Resolving semantic ambiguity - Word Sense Disambiguation (WSD) is the computational process of identifying an ambiguous word's semantic sense according to its usage in a particular context from a set of predefined senses [25].

E.g. For two Korean sentences:

- "**사과**를 먹는 그녀는 참 사랑스러웠다."

  - (The girl who's eating **apple** was so adorable.)

- "**사과**를 하는 그의 진지한 모습에 용서했다."

  - (I accepted the **apology** by his sincerity.)

Then WSD system will disambiguate senses for the Korean word "사과/sakwa" in the first sentence as sense "Apple" and the later as "Apology".

WSD has characteristic of variationoun because it's ubiquitous across all languages. It is also known as one of central challenges in various NLP research because many of them can take WSD's advantage to improve their performances such as Machine Translation (MT) [3], Automatic Speech Recognition (ASR), Information Extraction (IE), and Information Retrieval (IR) [24].

According to what kinds of resources are used, WSD can be classified into knowledge-based approach, corpus-based approach, and hybrid approach [16]: Knowledge-based approach [12, 17] relies on knowledge-resources like Machine Readable Dictionary (MRD), WordNet, and Thesaurus; Corpus-based approach [4, 13, 20, 22] trains a probabilistic or statistical model using sense-tagged or raw corpora; Hybrid approach is combining aspects of both of the knowledge and corpus based methodologies, using the interaction of multiple resources to approach WSD.

The algorithms for WSD can be categorized to supervised approach and unsupervised approach according to whether a training corpus is used [2]. It has been well-known that supervised approach using sense-tagged corpora are the main stream of current state-of-art WSD systems.

However, most WSD research on Korean were focusing on unsupervised approach and knowledge-based because lack of sense-tagged Korean corpora [8, 15, 21, 22]. With effort and collaboration of researchers and government, there are several Korean corpora available [9, 10]. Also it has been proved that supervised learning algorithm can lead a WSD system to the best result.

In this research, we tried to find most appropriate feature set for WSD system based on Conditional Random Field (CRF) approach, and also we constructed sense-tagged Korean corpus via Korean-Japanese parallel corpus to enlarge training examples and achieve better sense prediction accuracy.

This paper is organized as follows: Chapter two represents what kinds of corpora were used in our research, and also explained our method of automatically constructing sense-tagged Korean corpus then convert sense-tag and POS tags to enlarge the sense-tagged training data, chapter three shows details of our WSD classifier which will explain the details of features for supervised approach, chapter four represents evaluation of our method and compared with different systems which is developed by other researchers previously, chapter five we will then make a conclusion for our research and experiments.

Chapter 2

Overall Architecture

From the overall architecture(Figure 2.1), the proposed WSD system is mainly consists of three important stages:

- First, we automatically construct Japanese-translation tagged corpus using Korean-Japanese parallel corpus using morphological analyzers for Korean and Japanese respectively with utilizing the automatic word-alignment tool.

- Then, we convert the Japanese-translation tags which are considered to the "sense tags" for each ambiguous Korean words to sense-id by the sense-tag from Sejong corpus, also in the meanwhile, we need to transform the Part-Of-Speech (POS) tags to match the POS style from the Sejong sense-tagged corpus.

- Finally, we will then merge that constructed Sense-tagged corpus with Sejong sense-tagged corpus, and use that as training data for the WSD system using machine learning method with appropriate features.

Figure 2.1: Overall Architecture of Proposed WSD System

Chapter 3

Sense-tagged Corpora

In this research, we used two types of different sense-tagged Korean corpora. First one is from 21st Century Sejong Corpora [9] which is constructed by Korean researchers and funded by government, and the other is automatically constructed sense-tagged Korean corpus by utilizing Korean-Japanese parallel corpus. In this chapter we will introduce Sejong corpora briefly and present proposed method that construct sense-tagged Korean corpus and convert it to the format in Sejong corpora to enlarge the training examples.

3.1 21st Century Sejong Corpora

The 21st Century Sejong Corpora [9] are one part of the 21st Century Sejong Project that aimed to build Korean national corpora to provide Korean language resources for academia, education and industry. Among the different corpora, we chose semantically tagged Korean corpora which is consists of around 150 million eojeol[1] and tagged word-senses by using 'Standard Korean Dictionary'.

[1] In Korean, an eojeol is a sequence of morphemes, it consists of more than one umjeol, and each eojeol is separated with spaces.

3.2 Construct Sense-Tagged Korean Corpus via Korean-Japanese Parallel Corpora

For constructing sense-tagged Korean corpus using parallel text, we went through with these four steps:

(1) Align Korean-Japanese parallel corpus in word-level.

(2) Tag ambiguous Korean words by Japanese-translations in the sentence.

(3) For each Korean target words, cluster synonymous Japanese-translations, and map the groups to the sense inventory id in the ’Standard Korean Dictionary’.

(4) Change POS-tags to the Sejong’s POS-tags.

With theses four steps, then we will be able to obtain a sense-tagged Korean corpus with same format as Sejong sense-tagged corpora.

3.2.1 Align Korean-Japanese Parallel Corpus in word-level

In this step, we need to use alignment algorithm to make sentence aligned Korean-Japanese parallel corpus aligned in word-level.

There are many alignment algorithms [14, 18] available and used by much research already.

First of all, to align parallel corpora in word-level, we need to tokenize Korean and Japanese sentences using morphological analyzer respectively.

For Korean, we used in-house Korean morphological analyzer-KOMA to tokenize and obtain the Part-Of-Speech (POS) tags for each sentence in Korean, and we used MeCab [11] to analyze Japanese side.

After morphological analysis of Korean and Japanese sentences, tokenized sentences for both side will be input to the GIZA++ [18] for word alignment procedure.

From the output of GIZA++, then we will be able to acquire the word-level aligned parallel corpus which means each Korean word token are aligned with Japanese word token.

3.2.2 Tag ambiguous Korean words by Japanese-translations

In this step, we filtered and selected Japanese translations which will be served as the “sense-tags” for the corresponding Korean words.

We tagged ambiguous Korean words by Japanese translation from output result of the previous step, so that these Korean words can be regarded to have been disambiguated by different Japanese translations.

From Japanese translation tagged corpus, we observed many ambiguous words are tagged by erroneous and inefficient Japanese translations by error propagation of morphological analyzer and word alignment algorithm.

To reduce this error, we decided filter and eliminate those sentences with incorrect Japanese translation tags by two strategies.

First, we obtained the Japanese translation group for each ambiguous Korean word from the parallel text to apply these two following rules for filtering. (1) From the group of the Japanese translations which have been aligned to ambiguous Korean words, we chose Japanese translations with frequencies above the threshold. Because most of the Japanese translations aligned to the corresponding Korean target word with low occurrence counts are erroneous by morphological analyzer and word alignment of GIZA++.

(2) The one-length Japanese translations which don't belong to Kanji are excluded because Hiragana or other Romaji, Numbers, Punctuations etc. with one length would not be useful for representing senses for ambiguous Korean target words.

3.2.3 Cluster synonymous Japanese-translations & Map to the sense id

In this step, we transformed "sense-tags" represented by Japanese-translations to the sense-id in the Sejong Corpus.

From the previous stage, we could get a set of Japanese translations for the corresponding Korean target word. Mapping each Japanese-translations to sense-id in Sejong may need lots of time which will be very inefficient. So we decided to cluster the Japanese-translations with similar meaning which may create several groups for Japanese-translations then map each group which represents different sense to type of sense-id in Sejong corpus.

With following three processes, we made different Japanese-translation groups for each corresponding Korean target word by utilizing Mecab and Japanese-WordNet [7] as resources.

(1) First of all, we checked pronunciations for each Japanese translation token with Mecab to cluster the same words with different forms because even for the same word, some of them are showed up in full-Kanji, some are full-Hiragana, and some are mixture form of Kanji and Hiragana in the corpus (e.g. 油-しょうゆ-しょう油). Mecab could give pronunciation for each Japanese word, then we used this information to check whether two Japanese words' pronunciations are same or not. If two Japanese words' are having same

pronunciations, they will be recognized as same word and be grouped as one.

(2) Secondly, we used partial matching method to check If two words are representing same meaning by our pattern. Because Japanese Kanji is originally from Chinese characters, so each of words can represent specific meaning, and also there are several different forms in Japanese to show some respect such as adding a Japanese Hiragana character - ‘お’ in front of a noun. So, if two Japanese translations are exactly matched without first or last character of one word, they will be considered as same meaning (e.g. 祈り− お祈り , 船-船舶).

(3) Finally, we used Japanese WordNet and Wu & Palmer’s algorithms [19] to calculate the similarity score between Japanese translations.

Japanese WordNet is developed by the National Institute of Information and Communications Technology (NICT) since 2006 to support for Natural Language Processing research in Japan. This research was inspired by the Princeton WordNet and the Global WordNet Grid, and aimed to create a large scale, freely available, semantic dictionary of Japanese, just like other languages such as English WordNet or Chinese WordNet.

$$Sim_{wu}(C_1, C_2) = \frac{2 \times depth(C_3)}{depth_{C_3}(C_1) + depth_{C_3}(C_2) + 2 \times depth(C_3)} \tag{3.1}$$

The Wu & Palmer measure (Equation 2.1)calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, so with this calculated similarity score we could know how much two Japanese words are related to the other. Two Japanese words are clustered to same group if the similarity score for that two words is higher than the threshold.

With these three processes above, we will be able to have different groups of

Japanese-translations with different meaning (or sense). We used Sejong's sense definition table from 'Standard Korean Dictionary' to create the matching table from the sense-id in Sejong to the our Japanese-translation groups for each corresponding Korean target word. After that, each ambiguous Korean target word will have different senses represented by Sejong's sense-id which is mapped to the different groups of Japanese-translations.

Then the Japanese-translation tag for each Korean target word in our constructed corpus will be changed to the corresponding Sejong sense-id by the matching table.

3.2.4 Combine Sejong and Constructed Corpora

From the previous stage, we could have a sense-tagged corpus which has exactly same sense-id with Sejong, but here we also have to change the POS tags since our constructed sense-tagged corpus is analyzed and tokenized by our in-house (KOMA) morphological analyzer.

To combine Sejong sense-tagged corpora and automatically constructed corpora, we needed to have not only the same format of sense-id, but also for the same format of POS tagset.

By the careful observation, we found the Sejong have 44 different types of POS tags while our in-house analyzer have 62 different types.

So we mapped the POS tags s from our in-house morphological analyzer which is more fine-grained to Sejong's POS tags, and rewrite the tags in the constructed corpora automatically using that POS tag mapping table.

At the end, we constructed the sense-tagged corpus which have same form of sense-id and POS tags which could be used as enlarging the training data from Sejong sense-tagged corpora.

Chapter 4

Proposed Method

There are already much research have done for the supervised WSD approach [1, 13, 23]. Senseval [5] is one of the contests for evaluating various WSD applications with respect to different words and languages. Korean word sense disambiguation once belonged to one of their task-Senseval2, but it was ended since erroneous task corpora and non-popularity for participants.

In our work, we used supervised method and sources of local collocations, part of speech, and content words (noun, verb, adjective, and adverb) with n window size based on CRF approach.

4.1 Training WSD Classifier

Because there are no state-of-art Korean WSD system using Machine Learning technique, we tried to use the feature set from a state-of-art English WSD system based on SVM approach [12]. We found it is using surrounding context as one of the features from different sentences because it was targeting Senseval task which is given several sentences for disambiguating one target words. Here surrounding context feature was using not only every uni-gram in the target

Table 4.1: Comparison table for two feature set

Y. K. Lee*	Our's
w_i (i = -2, -1, 1, 2)	w_i (i = -2, -1, 0, 1, 2)
POS_i (i = -3, -2, -1, 0, 1, 2, 3)	POS_i (i = -2, -1, 0, 1, 2)
$w_i w_{i+1}$ (i = -2, 1)	$w_i w_{i+1}$ (i = -2, -1, 0, 1)
$w_i w_{i+1} w_{i+2}$ (i = -3, -2, -1, 1)	$POS_i POS_{i+1}$ (i = -2, -1, 0, 1)
$wiw_{i+1} w_{i+2} w_{i+3}$ (i = -1)	$POS_i POS_{i+2}$(i = -1)
Content words	Content words

sentence, but also the uni-gram from the other sentences, and use those unigrams to make a vector to use as feature. Instead of using that feature, we decided to use the content words (Noun, Verb, Adjective, Adverb) surrounded from target word rather to all of the unigrams from surrounding context, which will be more efficient and fit to the WSD.

From the experimental result, we found its prediction accuracy was quite acceptable, but we also tried to find another our own feature set to achieve better result.

In this research, we call the features absorbed from Lee [12] as Y. K. Lee* and our proposed feature set as "Our's" (Table 4.1). Here w refers to tokenized Korean word, and POS refers to part-of-speech tag.

4.2 Disambiguate Words in New Context

After the word sense classifier is trained with training corpus, it will be able to determine the most appropriate sense for the target word when system get entry of a Korean sentence which contains the target word.

Chapter 5

Experimental Results

5.1 Accuracy of Sense-Tagged Corpora

We checked the accuracy for grouping for synonymous Japanese translations manually to evaluate the automatically constructed sense-tagged corpora.

To construct sense-tagged Korean corpora, we used Korean-Japanese parallel text that consists of 608,692 sentences, and extracted 40,622 sentences of sense-tagged corpora targeting 200 of ambiguous Korean nouns.

Evaluation result shows that we clustered 606 Japanese words correctly into same groups among 686 words, which give us 88.34% (606/686) of accuracy. However, when we check the frequencies of those incorrectly grouped Japanese translations that appeared in the parallel corpora for the corresponding Korean WSD target word, it showed only 2.65% (1,410/53,264) error rate which is quite low.

Also when we tried to evaluate those groups of Japanese-translations by how many of them can be actually map to the sense-id in the Sejong's "Standard Korean Dictionary". Result showed that among 515 different Japanese-translation groups, 480 of them can be mapped to Sejong's sense-id, so the

Table 5.1: Classifier Accuracy Comparison using 5-fold Cross Validation

Window Size	Prediction Accuracy (%)		
	Y. K. Lee*	Our's	Comparison
2	**88.87**	**90.88**	**+2.01**
4	88.65	90.47	+1.82
6	88.02	90.14	+2.12
8	87.73	89.90	+2.17
10	87.50	89.79	+2.29

mapping accuracy would be then 93.204% from this observation.

5.2 Finding Appropriate Window Size

As previously mentioned, to use content words, we need to find most appropriate window size for WSD system. So we tried to compare several different window sizes using two different features – Y. K. Lee* and our own.

From the observation for result of the comparison experiment, we found window size 2 had best performance with our feature set (Table 5.1). So we decided to extract content words by window size 2 as the feature for our CRF approach.

5.3 WSD Prediction Accuracy

For the evaluation of WSD system, we made three different types of training data to compare three different systems.

Table 5.2: The Comparison of Different WSD Systems

Training Data / Method	Our's	Y. K. Lee*	Base-Line
Sejong	95.57	94.88	76.19
Sejong+	95.67	94.96	76.19
CK	78.33	72.32	76.19

5.3.1 Training and Test Data

First of all, we randomly chose 90% (256,304 sentences) of corpora for the training data , and 10% (28,627 sentences) for test data from Sejong corpora.

Second, we used constructed sense-tagged corpus by our method as training corpus to check its credibility.

Also, we combined training data from Sejong and our constructed sense-tagged corpus to see how does it affect the WSD system.

5.3.2 Evaluation Measure

In this paper, the assessment of WSD systems is performed in terms of evaluation measure based on prediction accuracy. Here, we define the *prediction accuracy* ***P*** as the percentage of correctly tagged senses among total number of the instances:

$$P = \frac{\#\ Correct\ Instance}{\#\ Total\ Instance} \tag{5.1}$$

5.3.3 WSD Systems with Different Features

We compared three different WSD systems: The base-line system which is choosing the Most Frequent Sense (MFS) only; The WSD system using features from Lee [12]; and The WSD system with our own feature set.

Table 5.3: The Comparison With Previous Work

Author	Target Size	Test Size (Sentence)	Accuracy (%)
Kim et al. 2011	10	574	86.2
Park et al. 2012	583	200	94.02
Our Method	200	28,627	95.67

From the result we observed that our WSD system outperformed the baseline system (MFS) around 13.6% of prediction accuracy, and it also proved that system with our feature was able to reach higher prediction accuracy by 0.57% of improvement compare to system used features from Y. K. Lee*. Meanwhile, adding the sense-tagged corpora to Sejong resulted 0.1% improvement of prediction accuracy.

5.3.4 Other Korean WSD Systems

We compared our result to two most recent Korean WSD systems (Table. 5.3), Kim [15] utilized Korean WordNet and raw corpus to disambiguate word sense, Park [21] built word vectors from Sejong sense-tagged corpus to resolve word senses. Among three different types of WSD approaches, our method showed best performance. Although Park [21] was targeting 583 words which is triple size of our target word, they used only 200 sentences for evaluation which is quite small compare to our test size (28,627 Sentences).

5.3.5 Comparison with SENSEVAL (SEMEVAL)

Senseval (or Semeval) is an international competition for the Word Sense Disambiguation system. The first workshop – Senseval-1 took place in the summer of 1998 for English, French, and Italian, culminating in a workshop held at

Herstmonceux Castle, Sussex, England on September 2-4.

The purpose of Senseval is to evaluate the strengths and weaknesses of computer programs which is made for automatically determining the sense of a word in context (Word Sense Disambiguation or WSD) with respect to different words, different varieties of language, and different languages. Semeval is having broader scope which is like later version of the Senseval, and it include semantic analysis tasks outside of WSD.

Senseval(or Semeval) workshops are held in every three years, and it's still focusing on the evaluation of semantic analysis systems. With great efforts of many researchers, it's able to provide a more detailed historial overview nowadays.

Especially for Korean, once it was belonged to the Senseval-2 which is held in 2001 for the competition of WSD systems, but it couldn't show any big effect because of non-popularity of participation and erroneous corpus problem. Based on these reasons, Korean task was eliminated from Senseval-3, and there are no any official data for the Korean WSD evaluation.

Nowadays, the most research in Senseval are related to English task, previously the WSD sense inventory is decide by the synset id in WordNet, and construct the training and test data for the evaluation. Compare to tasks in Senseval, in this research, we used the actual sense id for each ambiguous Korean words from the 'Standard Korean Dictionary' which is distinct from the synset id from the WordNet.

In the Senseval WSD research, there are mainly three types of approaches: supervised approach, unsupervised approach, and knowledge-based approach.

Generally, supervised approaches to WSD have obtained better results than unsupervised methods and knowledge-based approach.

Supervised WSD approach are using machine learning techniques for inducing a classifier from manually sense-tagged data. The classifier will see that as classification task in order to tag the suitable sense to each instance of that ambiguous word. The most popular machine learning methods are these: Decision Lists, Decision Trees, Naive Bayes, Neural Networks, Exemplar-Based or Instance-Based Learning, Support Vector Machines (SVM), Conditional Random Field (CRF), Semi-Supervised Learning and Etc.

Among them, the Semi-Supervised approach is to build a sense classifier with little training data to solve the main problems of supervision: the lack of annotated data and and the data sparsity problem.

By contrast to the Semi-Supervised approach, this research used large sense-tagged Korean corpus from Sejong corpora. Also the aim of this research is using parallel corpus to construct the sense-tagged corpus instead of raw corpus, and Semi-Supervised approach is used when there are little training corpus available, so compare to that we decided to use the CRF approach the expect the best performance.

Comparing and evaluating the results and performances of various different WSD systems using different method is extremely difficult, it would have to evaluate all of them on a same input to ensure that evaluate them on the same words for the same part of speech sense. Also it need to have same sense inventories and knowledge resources. Because WSD systems on the Senseval (or Semeval) are built based on different language and resources, we will not compare to the systems from Senseval in this paper.

Chapter 6

Conclusion

In this research, we mainly targeting two things: First, construct sense-tagged corpus using Korean-Japanese parallel corpus. Second, find most appropriate feature set for the Korean WSD system.

To construct sense-tagged corpus using parallel text, we represented a way to cluster synonymous Japanese words using several matching patterns combining the Japanese WordNet.

Using this constructed sense-tagged corpus, the WSD system outperformed 2.14% than the base-line system which is choosing most frequent sense only, and also the WSD system using enlarged training data with this corpus have achieved best performance with 95.67% of prediction accuracy.

This research also had focused on finding most appropriate feature template by comparing several different features. Feature set created our own with enlarged training corpus, we achieved better prediction accuracy compared to the previous Korean WSD work using same sense-tagged corpus.

Bibliography

[1] E. Agirre, G. Rigau, L. Padro, and J. Atserias. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and the Humanities*, 34(1-2):103–108, 2000.

[2] R. V. Bhala and S. Abirami. Trends in word sense disambiguation. *Artificial Intelligence Review*, pages 1–13, 2012.

[3] M. Carpuat and D. Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, 2007.

[4] M. Diab and P. Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262. Association for Computational Linguistics.

[5] P. Edmonds and S. Cotton. Senseval-2: overview, 2001.

[6] N. Ide and J. Véronis. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):2–40, 1998.

[7] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki. Development of the japanese wordnet. 2010.

[8] H. S. Jung Heo, Myengkil Cang. Homonym disambiguation based on mutual information and sense-tagged compound noun dictionary. *Proceedings of Korea Computer Congress*, 33:1073–1089, 2006.

[9] B. Kang and H. Kim. Sejong korean corpora in the making. In *Proceedings of LREC*, pages 1747–1750, 2004.

[10] H.-G. Kim and B.-M. Kang. 21st century sejong project-compiling korean corpora. *development*, 1999:2000, 1998.

[11] T. Kudo. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*, 2005.

[12] Y. K. Lee and H. T. Ng. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 41–48. Association for Computational Linguistics, 2002.

[13] Y. K. Lee, H. T. Ng, and T. K. Chia. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 137–140, 2004.

[14] I. D. Melamed. Empirical methods for exploiting parallel texts. 1998.

[15] H.-C. K. Min-Ho Kim. Word sense disambiguation using semantic relations in korean wordnet. *Proceedings of Korea Computer Congress*, 38, 2011.

[16] R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

[17] R. Navigli and P. Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(7):1075–1086, 2005.

[18] F. J. Och and H. Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics, 2000.

[19] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[20] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

[21] J.-S. L. Yong-Min Park. Word sense disambiguation using korean word space model. *Journal of Korea Contents Association*, 2012.

[22] Y. Yoon, C.-N. Seon, S. Lee, and J. Seo. Unsupervised word sense disambiguation for korean through the acyclic weighted digraph using corpus

and dictionary. *Information processing & management*, 42(3):710–722, 2006.

[23] Z. Zhong and H. T. Ng. Word sense disambiguation for all words without hard labor. In *Proceeding of the Twenty-first International Joint Conference on Artificial Intelligence.*

[24] Z. Zhong and H. T. Ng. Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 273–282. Association for Computational Linguistics, 2012.

[25] X. Zhou and H. Han. Survey of word sense disambiguation approaches. In *Proceedings of the 18 th International Florida AI Research Society Conference*, pages 307–313, 2005.

Acknowledgment
감사의 글

I am deeply appreciative of the many individuals who have supported my work and continually encouraged me through the writing of this dissertation. Without their time, attention, encouragement, thoughtful feedback, and patience, I would not have been able to see it through.

Above all, I would like to thank my advisor, Jong-Hyeok Lee, for his inspirational and timely advice and constant encouragement over the last two years. I have learned a great deal from his unique perspective on research, his sharp insight on almost any issue, and his personal integrity and expectations of excellence. He has been a great advisor for me outside of the academic world as well. He has always been patient when pointing out and explaining the problems which I had during my research, and has shared with me many witty jokes, metaphors, and lessons on how to strive for work-life balance. I really appreciate his support of my research.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Gary Geunbae Lee and Prof. Hwanjo Yu, for their encouragement, insightful comments, and hard questions.

My fellow labmates' support has been continuous fuel during my long journey in finishing this master program. Particularly, I'd like to thank my research group members who helped me a lot during my research work . I will never forget the time we had spent together: sharing our sad and happy story, having barbecue party, and playing the Starcraft.

Life would not have been as colorful without the many good friends I met in POSTECH. I would like to extend my thanks to the student I met here. Our friendship is built not on the many social gatherings we attended together, but on the many values we share.

Many people helped me and offered their friendship before I joined graduate school. I want to thank my mentors and friends at Yanbian University of Science and Technology; childhood friends from middle school who grew up with me. Even though the ocean separates us, thinking about them always brings a warm smile. Whenever tough time comes, I feel motivated and encouraged by reading their words and seeing their faces online.

Last but not least, I want to thank my loving and caring family. Thanks to my father Guangzhu Li and my mother Lanying Piao for teaching me to be curious and sincere to life, giving me constant love and encouragement, and for always being there for me. Words cannot express my appreciation for my parents. Thank you for always supporting me in my academic pursuits and for the wonderful life that we share together.

모든 영광을 저의 부모님에게 돌립니다.

사랑합니다. 어머님, 아버님.