Korean Word-Sense Disambiguation Using Parallel Corpus as Additional Resource

Chungen Li
Pohang University of Science and Technology
jiafei427@gmail.com

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.

1 Introduction

In computational linguistic, lexical ambiguity is one of the first problems that people faced with in Natural Language Processing (NLP) area (Ide and Véronis, 1998).

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. 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) (Carpuat and Wu, 2007), Automatic Speech Recognition (ASR), Information Extraction (IE), and Information Retrieval (IR) (Zhong and Ng, 2012).

According to what kinds of resources are used, WSD can be classified into knowledge-based approach, corpus-based approach, and hybrid approach: Knowledge-based approach relies on knowledge-resources like Machine Readable Dictionary (MRD), WordNet, and Thesaurus; Corpus-based approach 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.

However, most WSD research on Korean were focusing on unsupervised approach and knowledge-based because lack of sense-tagged Korean corpora (Yoon et al., 2006; Min-Ho Kim, 2011; Yong-Min Park, 2012; Jung Heo, 2006). With effort and collaboration of researchers and government, there are several Korean corpora available (Kang and Kim, 2004). 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.

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 113–118,
Hissar, Bulgaria, 9-11 September 2013.

This paper is organized as follows: Section two represented the over-all architecture of our method, corpora that used in our research, and explained the method of constructing sense-tagged Korean corpus, Section three showed evaluation result of our WSD method and compared it with other different systems, Section four made a conclusion for this research and experiments.

2 Construct Sense-Tagged Corpus & Enlarge Training Data

In this research, we used two types of different sense-tagged Korean corpora. First one is from 21st Century Sejong Corpora (Kang and Kim, 2004) 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.

2.1 Overall Architecture

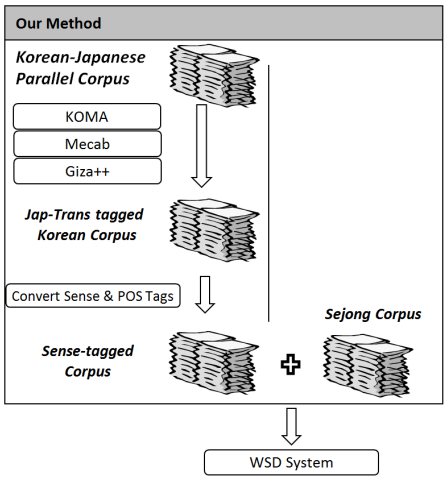

Figure 1: Overall Architecture of Constructing Sense-tagged Corpus

From the overall architecture(Figure 1) we can see mainly it has three important stages: First, we will construct Japanese-translation tagged corpus using Korean-Japanese parallel corpus. Then, we will convert that Japanese-translation tags to sense-id from the original sense-tagged Sejong corpus, and we also need transformation for the Part-Of-Speech tags to match the format of the 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.

2.2 21st Century Sejong Corpora

The 21st Century Sejong Corpora (Kang and Kim, 2004) 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'.

2.3 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.

2.3.1 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 (Melamed, 1998; Och and Ney, 2000) 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

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

Korean, and we used MeCab (Kudo, 2005) to analyze Japanese side.

After morphological analysis of Korean and Japanese sentences, tokenized sentences for both side will be input to the GIZA++ (Och and Ney, 2000) 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.

2.3.2 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.

2.3.3 Cluster Synonymous Japanese Translations & Map to 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 (Isahara et al., 2010) 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 (Wu and Palmer, 1994) 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.

The Wu & Palmer measure 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.

2.3.4 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.

3 Experimental Result

3.1 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 mapping accuracy would be then 93.204% from this observation.

3.2 Finding Appropriate Window Size

As previously mentioned, to use content words as feature, we need to find most appropriate window size for it. We tried to compare several different window sizes with two different features – Y. K. Lee* and our own feature set by training the WSD model using constructed Korean WSD corpus without merging it into the Sejong Corpus. In this experiment, we used 5-fold cross-validation to calculate the prediction accuracy (Table 3.2) .

From the observation for result of the comparison experiment, we found window size 2 had best performance with our feature set (Table 3.2). So we decided to extract content words by window size 2 as the feature for our CRF approach.

3.3 WSD Prediction Accuracy

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

Window Size	Prediction Accuracy (%)		
	Y. K. Lee*	ours	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

Table 1: Classifier Accuracy Comparison using 5-fold Cross Validation

	ours	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

Table 2: The Comparison of Different WSD Systems

3.4 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.

3.5 Comparison of 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 (Lee and Ng, 2002); and The WSD system with our own feature set.

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.

4 Comparison with Related Works

We compared our result to two most recent Korean WSD systems (Table. 4), Kim (Min-Ho Kim, 2011) utilized Korean WordNet and raw corpus to disambiguate word sense, Park (Yong-Min Park, 2012) 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 (Yong-Min Park, 2012) 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).

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

Table 3: The Comparison With Previous Work

Conclusion

In this research, we mainly targeting two things: First, construct sense-tagged corpus using Korean-Japanese parallel corpus. Second, find 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 heuristic rules combining the Japanese WordNet.

Using this constructed sense-tagged corpus, the WSD system outperformed 2.14% than the baseline system which 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 best Korean WSD work using same Sejong sense-tagged corpus.

References

Marine Carpuat and Dekai Wu. 2007. 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.

Nancy Ide and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: the

state of the art. *Computational Linguistics*, 24(1):2–40.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2010. Development of the japanese wordnet.

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

BM Kang and Hunggyu Kim. 2004. Sejong korean corpora in the making. In *Proceedings of LREC*, pages 1747–1750.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*.

Yoong Keok Lee and Hwee Tou Ng. 2002. 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.

Ilya Dan Melamed. 1998. Empirical methods for exploiting parallel texts.

Hyuk-Chul Kwon Min-Ho Kim. 2011. Word sense disambiguation using semantic relations in korean wordnet. *Proceedings of Korea Computer Congress*, 38.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

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

Jae-Sung Lee Yong-Min Park. 2012. Word sense disambiguation using korean word space model. *Journal of Korea Contents Association*.

Yeohoon Yoon, Choong-Nyoung Seon, Songwook Lee, and Jungyun Seo. 2006. Unsupervised word sense disambiguation for korean through the acyclic weighted digraph using corpus and dictionary. *Information processing & management*, 42(3):710–722.

Zhi Zhong and Hwee Tou Ng. 2012. 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.