A Korean Homonym Disambiguation System Based on Statistical Model Using weights

Jun-Su Kim   Wang-Woo Lee   Chang-Hwan Kim   Cheol-young Ock
Dept. of Computer Engineering & Information Technology, University of Ulsan
680-749, San29, Mugeo-dong, Nam-gu
Ulsan, Korea
{jskim, wwlee, edel}@cic.ulsan.ac.kr   okcy@uou.ulsan.ac.kr

Abstract

A homonym could be disambiguated by another words in the context as nouns, predicates used with the homonym. This paper using semantic information (co-occurrence data) obtained from definitions of part of speech (POS) tagged UMRD-S[1]. In this research, we have analyzed the result of an experiment on a homonym disambiguation system based on statistical model, to which Bayes' theorem is applied, and suggested a model established of the weight of sense rate and the weight of distance to the adjacent words to improve the accuracy. The result of applying the homonym disambiguation system using semantic information to disambiguating homonyms appearing on the dictionary definition sentences showed average accuracy of 98.32% with regard to the most frequent 200 homonyms. We selected 49 (31 substantives and 18 predicates) out of the 200 homonyms that were used in the experiment, and performed an experiment on 50,703 sentences extracted from Sejong Project tagged corpus (i.e. a corpus of morphologically analyzed words) of 3.5 million words that includes one of the 49 homonyms. The result of experimenting by assigning the weight of sense rate(prior probability) and the weight of distance concerning the 5 words at the front/behind the homonym to be disambiguated showed better accuracy than disambiguation systems based on existing statistical models by 2.93%.

1   Introduction

The ambiguity, which is the most difficult problem in natural language processing (NLP), occurs inevitably in every analysis process including morphological analysis and syntactic analysis. Ambiguity problems occurring at some parts have been resolved to some degree. As the studies on semantic and discourse analysis are becoming active, more effort is being made to research schemes for word sense disambiguation (WSD). WSD refers to disambiguating the sense of a word contextually suitable for the sentence when it is used with two or more different meanings in sentences. [1, 3, 4, 5, 8]
Studies for solving ambiguity are largely grouped into methods using dictionaries according to the pattern of learning data and ones using a corpus. In terms of methodology, there are largely methods using rules, ones using probability statistics, and ones using semantic hierarchy structure.
A method using dictionaries is disadvantageous in that it is difficult to reflect the dynamic characteristic of a language [10, 11], while advantageous in that it can abstract the detailed information of word senses [2, 3]. To disambiguate using a corpus, a large-sized semantic tagged corpus is required. However, a high-quality corpus is hard to find, and costly and time-consuming to build. But the method reflects the dynamic characteristic of a language.
In the thesis, we abstract semantic information from a dictionary definition corpus based on the method suggested in J. Huh (2000)[3] research. We will study a plan to utilize the semantic information in the homonym disambiguation model based on Bayes' theorem.

---

[1] UMRD-S : Ulsan university Machine Readable Dictionary (Semantic Tagged)

2   Semantic Information from definitions in Dictionary

Semantic information extracted from definitions of part of speech (POS) tagged UMRD. We must classify definitions and titles according to the meaning before extracting semantic information. The structures of definitions are various, and are classified into 11 types by Cho(1999)[2]. The most frequent type is that head-word (hyponym), the title also is contained in semantic information.
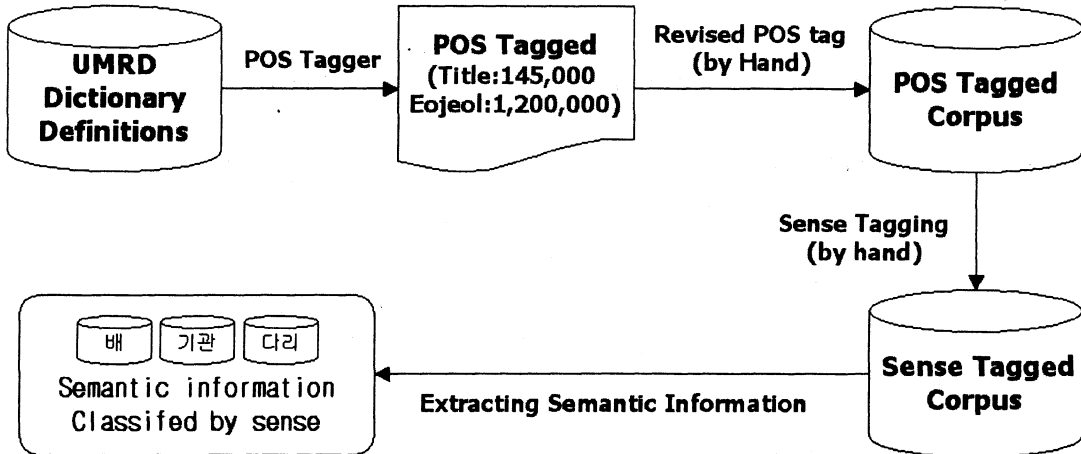

Figure 1. The Process to extract Semantic Information

 The semantic information is classified into two types. [2, 3] First has hyponym-hypernym relation between title-word and head-word (homonym) in definition. The other is extracted from definitions in which the homonym is used for defining other words. In other word, the homonym is located middle in the definitions. And the titles of 2nd information are contained in the semantic information. Types merge semantic information.

Table 1. Types of definitions for extracting semantic information

Title	Type
First type	
거선[geo-seon]	아주 큰 배[a-ju keun bae] (large ship)
유조선[yu-jo-seon]	유조 시설을 갖춘 배 [yu-jo si-seor-eul gat-chun bae] (Oiler)
풋배[put-bae]	아직 덜 익은 배[a-jik deol ik-eun bae] (unripe pears)
Second type	
난파 [nan-pa]	배가 항해 중에 폭풍우 등을 만나 깨어짐 [bae-ga hang-hae jung-e pok-pung-u deung-eul man-na kkae-eo-jim] (Shipwreck)
운하 [un-ha]	육지를 파서 강을 내고 배가 다니게 한 수로 [yuk-ji-reul pa-seo gang-eul nae-go bae-ga da-ni-ge han su-ro] (Canal)

3 Word Sense Disambiguation Model Based on Statistics

3.1 Statistic Model Based on Bayes' Theorem

In the WSD model utilizing the Semantic information of homonym sense abstracted from dictionary definitions as the prior probability in Bayes' theorem, a homonym H appearing in an arbitrary sentence C is disambiguated as one of senses Hs_1, Hs_2, ..., Hs_n

$$W(H,C) = \arg\max_{Hs_i} P(Hs_i, C) \qquad (1)$$

$$P(Hs_k, C) = \sum_{j=1}^{m} P(Hs_k \mid w_j) \qquad (2)$$

$$P(Hs_k \mid w_j) = \frac{P(w_j \cap Hs_k)}{\sum_{i=1}^{n} P(w_j \cap Hs_i)} \qquad (3)$$

In formula (3), Hs_k is the k-th sense of homonym H, and w_j appearing in sentence C is a word associated with the semantic information of Hs_k and has its frequency information. In addition, w_j may appear in other frequency with different frequency [2]. Formula (2) represents the sum of probability for each of appearing words extracted from formula (3) to be identified as sense Hs_k. And formula (1) is for disambiguating the sense of homonym H for the sentence C with the maximum of the sums by senses calculated in formula (2)

We experimented the statistical basic model (NB: Naïve Bayes Model) for the selected 31 nouns and 18 predicates among the homonyms frequently appearing in the dictionary definitions against the 3.5 million POS tagged corpus of the Sejong Project. As the result of applying to all the words of 50,703 sentences containing the selected 49 homonyms, the accuracy was 77.67% for the nouns and 61.73% for predicates on the average. When applying only to the 5 words front / behind the homonyms, the accuracy was 72.87% for the nouns and 43.79% for predicates.

3.2 Error Patterns in the Basic Model

In this section, we will examine the cases of erroneous analysis in the statistical basic model (NB) and search for a method to resolve them.

> **[*Example sentence* 1]** 그 바람에 배보다 배꼽이 더 커버린 과자 값이 들었지만 그 뒤에도 카메라가 쥐어진 할애비만 보면 외손녀는 울듯 비죽거렸고, 그때마다 아내는 카메라를 동댕이치겠다고 위협했다.
> (Thanks to that, I have to pay for sweets much more than what I tried to save. Since then, for the worse, my granddaughter goes nearly to burst into tears whenever seeing the grandpa holding a camera, and my wife threatens me to slam the camera.)

In [example sentence 1] are two homonyms '배[2]) [bae]' and '들다[3]) [deul-da]'. Of the two, '배 [bae]' is used in the sense of '배_1(body part)'. Words in the semantic information for '배 [bae]' appearing in the [example sentence 1], namely the words in each sense, and their frequencies are as follows.

---

[2]) Senses of homonym '배 [bae]' : 배_1(body part), 배_3(vesse), 배_4(fruit), 배_6 (two times, double)

Table 2. Semantic information abstracted from [example sentence 1]
(5 words front / behind the word)

배_1 (body part)	noun	배꼽(navel)(2)
	predicate	크다(big, large)(26) 들다(stay, lift up , ...)(9)
배_3 (vessel)	noun	바람(as a result of)(9)
	predicate	크다(big, large)(25) 들다(stay, lift up , ...)(1)
배_4 (fruit)	noun	바람(as a result of)(1) 과자(cookie)(2)
	predicate	크다(big, large)(1) 들다(stay, lift up , ...)(1)
배_6 (double)	noun	값(price)(1)
	predicate	크다(big, large)(5)

The result of applying the extracted words and their frequencies to formula (3) and (2) is as [Table 3]. Consequently the basic model selects '배_4(fruit)', so fails to disambiguate.

Table 3. Probability drawn from NB Method

	배_1	배_3	배_4	배_6
배꼽[bae-kkop] (navel)	1.00	0.00	0.00	0.00
바람[ba-ram](as a result of)	0.00	0.48	0.52	0.00
과자[kwa-ja](cookie)	0.00	0.00	1.00	0.00
값[kap](price)	0.00	0.00	0.00	1.00
크다[keu-da](big, large)	0.31	0.26	0.11	0.32
들다[deul-da] (stay, lift up , ...)	0.39	0.05	0.56	0.00
Total	1.70	0.79		1.32
Result of disambiguation: failed 배_4(fruit)				

The major reasons of disambiguation failure are, firstly, that the probability calculation for the frequency of semantic information used in disambiguation does not consider the use frequency of the homonym. For example, concerning the semantic information of '들다[deul-da]' extracted from dictionary definitions in [Table 2], the word appears once for each of '배_3(vessel)' and '배_4(fruit)', but the numbers of words used in the definitions of '배_3(vessel)' and '배_4(fruit)' are 24 and 513 respectively. Therefore, as '들다[deul-da]' is extracted once from 24 words and 513 words, the frequency should be normalized.

Secondly, the present NB model does not analyze syntactic structure, and disambiguate a homonym simply according to what semantic information the sentence containing the homonym has. This is based on the assumption that, if the homonym is not used metaphorically or idiomatically, it is used with the words that are semantically related to the corresponding sense of the homonym, and if the homonym is used in a simple sentence, it is almost possible to disambiguate it without analyzing the syntactic structure. In complex sentences or compound sentences, however, the extracted semantic information

---

[3] Senses of homonym '들다 [deul-da]' : 들다_1(stay, permeate), 들다_4 (lift up, suggest {a fact or an example}), 들다 _5 (receive the action represented by the front noun)

may not be determinant in disambiguating the corresponding homonym. Accordingly, in the present thesis, we extracted semantic information from the 5 words front / behind a homonym. [1, 4] Yet in this case, the semantic relevance may differ according to the distance from the homonym. Therefore, we should consider the location (the distance from the homonym) where the semantic information is found. In the thesis, we suggest a method to resolve the two problems.

4   Application of Weight to the Probability Model Based on Statistics

4.1   Consideration of the Weight of Sense Rate

Dictionary definition sense information, which is used as prior probability, varies greatly in the word types and frequencies according to the appearance frequencies of homonym (Hs_1, Hs_2,..., Hs_n) senses.

Table 4. Sampled Number of words and Sums of Frequencies in Semantic information Extracted from Dictionary Definitions

word	sense	number of words		Frequency sum	
		Noun	Pred.	Noun	Pred.
배	배_1	639	307	2,323	1,313
	배_3	668	283	1,593	1,114
	배_4	130	67	164	102
	배_6	363	82	676	178
들다	들다_1	662	466	1,608	1,304
	들다_4	515	226	1,013	516
	들다_5	84	15	146	19

In case word wj ($\in Hs_1 \cap Hs_2 \cap ... \cap Hs_k$) appearing in sentence C appears commonly in several semantic information sets, it is highly likely to be selected as it has high probability through formula (3) when the frequency sum of the word is small. According to [Table 4], if a same word appears with '배_1(a body part)' and '배_4(fruit)', the frequency of with '배_1(a body part)' should be 15 times higher than '배_4(fruit)' to be possibly selected as '배_1(a body part)'. However, frequency 15 is quite high number as a frequency of appearance, and is enough to disambiguate the homonym. Accordingly, we need a method to consider the number of words in the semantic information and the frequency in the Bayes theorem of the basic statistic model, considering the peculiar feature of vocabularies.
In this thesis, we assume that the word in the semantic information provides a solution, and will use words in the semantic information Using the number of words of nouns and predicates belonging to the senses of a homonym (Hs_1, Hs_2, ..., Hs_n), we can obtain the weight of SR(Sense Rate) as formula (4)

$$SR(Hs_k) = \frac{\text{a number of word in } Hs_k}{\sum_{j=1}^{n} \text{a number of word in } Hs_j} \quad (4)$$

Words in sense Hs_k have the prior probability of $P(w_j \cap Hs_k)$. By multiplying the weight of sense rate $SR(Hs_k)$ to the existing probability, we obtain a new probability.

A statistical model (SR : statistical model with Sense Rate) that considers the weight of sense use frequency is completed by applying the probability that reflects the weight to formula (1) and formula (2), which results formula (5).

$$P_{SR}(Hs_k \mid w_j) = \frac{P(w_j \cap Hs_k) \times SR(Hs_k)}{\sum_{i=1}^{n} P(w_j \cap Hs_i) \times SR(Hs_i)} \quad \text{(5)}$$

To [example sentence 1] on which the statistical basic model has failed, we applied the statistical model (SR) that reflects the weight of sense use frequency on the words extracted from [Table 2], and found that the model disambiguated correctly as shown in [Table 5].

Table 5. Probability Considering the Weight of SR

	배_1	배_3	배_4	배_6
배꼽	1.00	0.00	0.00	0.00
과자	0.00	0.00	1.00	0.00
값	0.00	0.00	0.00	1.00
크다	0.47	0.36	0.04	0.13
Total		1.43	1.27	1.13

When comparing [Table 4] and [Table 5], the frequency of '들다[deul-da]' belonging to the semantic information of '배_1[bae]' is 9, while that belonging to '배_4[bae]' is 1. In the existing statistical model, the probability for '배_1[bae]' is 0.39, which is lower than the probability for '배_4[bae]' 0.56. According to [Table 5] reflecting the weight of sense rate, the probability of '배_1[bae]' have risen to 0.70. Accordingly, the statistical model considering the weight of sense rate is advantageous in that it reflects frequency appropriately.

4.2 Consideration of the Distance between Words

If we can utilize syntactic structure in disambiguating homonyms appearing in sentences, we may reduce unnecessary factors by selecting good-quality semantic information on the disambiguation.
The present disambiguation model is based on a simple statistical model utilizing dictionary semantic information. Thus we attempt to resolve, using efficiently the information about the adjacent words.
In the thesis, the disambiguation accuracy using the 5 words front / behind is not different significantly from that using the whole words. It is because the used semantic information is largely found in the adjacent words. In particular, it is even obvious that a word closer to the homonym among the 5 words front / behind it is more influential in disambiguation. Accordingly, we will apply the weight of distance appropriately.
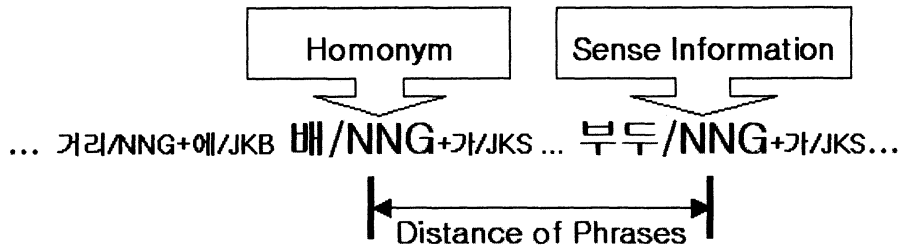
Figure 2. Distance between a homonym and a word of semantic information

Considering the absolute distance $|d(H) - d(w_j)|$ between a homonym and a word used as semantic information, we have derived a weight of formula (6). By applying the weight of distance Dis(H, w_j) to formula (5), which is a new weight considering the weight of sense rate, we reflect the distance from the hononym. The longer the distance is, less influential the word is to the disambiguation. Accordingly, if a word is found near the homonym, it records high probability, and if it is found distantly from the homonym, it records low probability.

$$\text{Dis}(H, w_j) = \frac{1}{\sqrt{|d(H) - d(w_j)|}} \qquad (6)$$

$$P(Hs_k, C) = \sum_{j=1}^{m} P_{SR}(Hs_k \mid w_j) \times \text{Dis}(H, w_j) \qquad (7)$$

Table 6. The result of applying weights of distance to probabilities with a deviation of 20% after applying weights of SR

	배_1	배_3	배_4	배_6
배급	0.00	0.00	0.00	0.00
바람	0.00	0.83	0.17	0.00
타다	0.00	0.00	0.50	0.00
곧	0.00	0.00	0.00	0.50
크다	0.27	0.20	0.02	0.08
들다	0.32	0.04	0.10	0.00
Total		1.07	0.79	0.58
Result of disambiguation: successful 배 (a body part)				

To improve the efficiency of the method, we experimented a method in which the weights of distance are applied only when the difference of disambiguaty is insignificant (the deviation is within 20%) just with the statistical model (SR) considering the weights of sense rate.

According to [Table 5], the highest probability is 36.17%, followed by 23.83% and most are less than 20.%. Thus, when considering the weights of distance [Table 6], we found, the disambiguation is correct.

5   Experiment and Analysis

5.1   Selection if Ambiguous Homonyms and Extraction of Test Sentences

After abstracting 200 homonyms appearing in dictionary definitions, we selects 49 words (31 nouns and 18 predicates). Which are used in balance among senses, and applies them to the disambiguation model [Table 7]. The analysis result by homonyms is as [Appendix 1]

We abstracted 50,700 sentences that include the selected homonyms from the Sejong Project tagged corpus (around 3.5 million words), and performed automatic disambiguation using the statistical basic model (NB). After disambiguating the homonyms correctly through post-processing the automatically disambiguated sentences, we compared the accuracy rate.

Table 7. Homonyms used in disambiguation experiments

Nouns (31)	거리[geo-ri], 결정[gyeol-jeong], 경기[gyeon-ggi], 국[guk], 기구[gi-gu], 기원[gi-won], 날[nal], 눈[nun], 대[dae], 독[dok], 등[deung], 못[mot], 배[bae], 부정[bu-jeong], 비[bi], 상[sang], 성[seong], 의사[ui-sa], 의지[ui-ji], 이상[i-sang], 장기[jang-gi], 장수[jang-su], 절[jeol], 주장[ju-jang], 중[jung], 지도[ji-do], 차[cha], 창[chang], 철[cheol], 판[pan], 표[pyo]
Predicates (18)	갈다[gal-da], 고르다[go-reu-da], 괴다[goe-da], 끼다[kki-da], 달다[dal-da], 들다[deul-da], 말다[mal-da], 맞다[mat-da], 묻다[mut-da], 붓다[but-da], 쉬다[swi-da], 싸다[ssa-da], 타다[ta-da], 쓰다[sseu-da], 이르다[i-reu-da], 차다[cha-da], 켜다[kyeo-da], 지다[ji-da],

5.2   Comparison of the Basic Model and the Model based on Weights of Sense Rate

The average accuracy rate of the basic statistical model is as [Table 8]. In the experiment, we attempted to disambiguate using the whole words and the 5 words front / behind the homonyms.

Table 8. Basic Statistical Model (NB, Accuracy Rate)

	Number of Sentences	All words	5 words	Accuracy rate (all words)	Accuracy rate (5 words)
Noun	30,451	23,652	22,189	77.67%	72.87%
Predicate	20,252	12,506	8,868	61.75%	43.79%

In case of nouns at the existing statistical model (NB), the difference in accuracy rates between when disambiguating through the whole words and when doing through the 5 words is 4.8%. This indicates that lots of significant information is in the adjacent words and that, in a long sentence, words adjacent to homonym give semantic information enough to disambiguate. In addition, using adjacent words may exclude unnecessary information, which might occur when abstracting information from the whole words,

173

For Predicate, the semantic informationabstracted from dictionary definitions is often insufficient for disambiguating homonyms. For example, '붓다[but_da]' is closely associated with '쏟다[ssot_da](pour)', but '쏟다[ssot_da]' is not included in the sense information. As the result of adding it to the semantic informationand analyzing, the accuracy rate increased by 6%. The reason is that dictionary definition techniques limit the words to be used. Accordingly, a more efficient method to add necessāry semantic informaton should be researched further in the future.

Table 9. The model considering the weights of SR

	Number of Sentences	All words	5 words	Accuracy rate (all words)	Accuracy rate (5 words)
Noun	30,451	24,412	23,034	80.17%	75.64%
Predicate	20,252	12,557	9,238	62.00%	45.62%

[Table 9] is the analysis result using the model that obtains new probabilities considering the weights of sense rate (SR). As the result of considering the weights, the accurate rate increased by 1.7% for analysis on the whole words, and by 2.4% for analysis on the five words. When analyzing the whole words, accuracy rates increased in 29 homonyms, and when analyzing the five words, the rates increased in 31 homonyms. The result shows that applying the weights of sense rate to the basic statistical model makes disambiguation more efficient. In addtion, the weights for the 5 words front / behind homonyms are applied more effectively.

Table 10. The model applying the weights of distance to probabilities with a deviation of 20% after applying weights of SR

	Number of Sentences	All words	5 words	Accuracy rate (all words)	Accuracy rate (5 words)
Noun	30,451	24,653	23,322	80.96%	76.59%
predicate	20,252	11,814	9,219	58.33%	45.52%

The result in [Table 10] came from applying the weights of distance to the 5 words front / behind homonyms when the deviation of probability is 20% after the first disambiguation considering the weights of sense rate. The accuracy rate for the whole words has increased only by 0.61% from that of the basic statistical model (NB), and the increase is less than that when applying the weights of sense rate. For the five words front / behind homonyms, the analysis result shows the highest accuracy rate. Accordingly, a model that combines the two weights suggested in the thesis is most efficent.

5.3 Error Patterns

According to the result of analyzing the cases that the accuracy rates fall in the model reflecting the weights of sense use frequency and the distance between words, the most significant cause appears to be lack of semantic information It is assumed that dictionaries restrain the use of extensive vocabularies and define the meanings of a word using a limited number of words.

6   Conclusions and Further Researches

First, according to the result of experimenting the statistical model reflecting the weights of sense rate and those of distances between words, which are suggested in the thesis, it is concluded that appropriate weights support disambiguation and that a further determinant weights should be explored for.

Second, further researches are required for refining and expanding the semantic information abstracted from dictionary definitions. For refining, we should examine how nouns (사람[sa-ram], 일[il], 말[mal], 때[ttae], …), verbs (하다[ha-da], 되다[dae-da], 있다[it-sa], 이르다[I-reu-da], …) and adverbs (없다[eop-da], 있다[it-da], 크다[keu-da], 작다[jak-da], 같다[gat-da], 다르다[da-reu-da], …) which are highly frequency because of the peculiar characteristic of dictionary definitions, affect disambiguation, and prepare a method to exclude unnecessary semantic information appropriately. It is also required to study on methods to abstract semantic information and to expand information using semantic networks, along with establishing a large-sized semantic tagged corpus by creating semantic tagged program for expanding semantic information.

References

[1] D. Yarowsky(1992), "Word-Sense Disambiguation Using Statical Model of Roget's Corpora",COLING-92

[2] P.O. Cho and C.Y. Ock(1999), "A Korean Noun Semantic Hirarchy based on Semantic Features", Proceeding of the 18th ICCPOL Vol.1.

[3] J. Hur(2001), "A Homonym Disambiguation System based on Semantic Information extracted from Definitions in dictionary", ICCPOL-2001

[4] G. Rigau(2000), "Naïve Bayes and Examplar-based Approaches to Word Sense Disambiguation Revisited", ECAL

[5] L. Marquez(2000), "Machine Learning and Natural Language Processing"

[6] Alpha k, Luk(1995), "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions", 33rd Annual Meeting of the ACL

[7] R. Bruce(1994),"Word Sense Disambiguation Using Decomposable Models"32rd Annual Meeting of the ACL, pp. 139-145

[8] P. Brown, V. Della Pietra, S. Della Pietra and R. Mercer(1991) Word sense disambiguation using statistical methods. In Proceedings of the 29th Annual Meeting of the ACL, pp.264-270

[9] G. Cottrell(1989) A Connectionist Approach to Word sense Disambiguation. Pitman, London

[10] E. Brill(1993) A Corpus-Based Approach to Language Learning. Ph.D. thesis Computer and Information Science, University of Pennsylvania

[11] S.B Park, B.T Zhang, Y.H Kim(2000) "Word Sense Disambiguation From Unlabeled Data", KISS '2000 Spring B', pp330 – 332

[12] Song, Young-bin, Choi, Gi-sun (2000) "The use of thesaurus for disambiguating verbs and its limitation", Treatise collection of the 12th Korean Alphabet and Korean Language Information Processing Conference, pp.255 - 261

[13] Lee, Chang-gi, Lee, Kun-bae (2000) "WordNet automatic mapping using disambiguation, Treatise collection of the 12th Korean Alphabet and Korean Language Information Processing Conference, pp.262 - 268

[14] Cho, Jung-mi(1998) "Disambiguating verbs using corpus and dictionaries", Ph.D. these, Korea Advanced Institute of Science and Technology

[Appendix]

거리[geo-ri]	3	64.20%	62.96%	63.58%	64.30%	65.53%	64.71%
결정[gyeol-jeong]	2	94.63%	88.35%	94.83%	88.68%	94.76%	88.48%
경기[gyeon-ggi]	3	79.33%	78.25%	77.95%	77.95%	80.09%	78.41%
국[guk]	2	88.03%	83.76%	88.89%	84.19%	91.45%	85.04%
기구[gi-gu]	3	81.58%	78.22%	78.42%	73.86%	80.00%	77.82%
기원[gi-won]	3	58.16%	59.22%	58.51%	57.80%	59.93%	60.64%
날[nal]	2	93.52%	83.11%	95.73%	85.74%	95.44%	85.60%
눈[nun]	2	84.58%	82.28%	87.68%	85.20%	88.62%	85.75%
대[dae]	4	64.89%	57.45%	56.38%	57.45%	59.57%	58.51%
독[dok]	2	89.93%	82.01%	91.37%	84.89%	91.37%	84.89%
등[deung]	3	83.66%	82.35%	87.15%	84.53%	86.71%	83.66%
못[mot]	3	88.59%	86.47%	91.50%	89.57%	92.44%	91.46%
배[bae]	4	63.38%	62.54%	64.17%	64.27%	66.88%	66.09%
부정[bu-jeong]	3	39.66%	42.37%	37.05%	40.26%	38.86%	39.76%
비[bi]	4	87.09%	87.09%	90.08%	90.55%	91.97%	92.13%
상[sang]	4	62.37%	66.13%	61.83%	66.13%	71.51%	71.51%
성[seong]	4	70.71%	65.67%	71.97%	66.77%	74.49%	71.50%
의사[ui-sa]	2	69.44%	66.03%	65.48%	63.71%	67.80%	66.17%
의지[ui-ji]	2	75.80%	71.28%	76.20%	71.81%	76.86%	73.54%
이상[I-sang]	3	60.39%	52.90%	74.36%	66.09%	74.22%	67.36%
장기[jang-gi]	3	54.64%	49.34%	49.34%	44.37%	46.36%	43.38%
장수[jang-su]	3	49.66%	46.21%	48.97%	46.21%	48.28%	46.21%
절[jeol]	3	68.81%	66.10%	60.34%	61.69%	63.73%	64.41%
주장[ju-jang]	3	94.65%	81.99%	98.09%	87.93%	97.60%	87.59%
중[jung]	3	55.49%	46.71%	64.16%	51.93%	61.86%	53.50%
지도[ji-do]	2	82.90%	74.04%	83.10%	74.45%	82.09%	77.67%
차[cha]	3	70.41%	68.65%	70.53%	68.77%	72.51%	69.71%
창[chang]	2	78.25%	81.75%	78.60%	82.11%	85.61%	85.61%
철[Cheol]	3	65.64%	71.78%	62.58%	66.87%	65.03%	65.64%
판[pan]	3	67.24%	63.22%	66.09%	62.64%	63.22%	60.92%
표[pyo]	3	63.09%	59.78%	60.88%	57.58%	62.26%	58.68%
갈다[gal-da]	3	64.25%	41.90%	60.89%	45.81%	62.01%	45.81%
고르다[go-reu-da]	3	57.86%	35.22%	52.83%	34.28%	48.11%	33.65%
괴다[goe-da]	2	77.97%	55.93%	79.66%	54.24%	72.88%	49.15%
끼다[kki-da]	2	73.20%	53.10%	74.19%	53.60%	72.21%	53.85%
달다[dal-da]	4	64.75%	42.04%	67.36%	46.48%	68.67%	46.48%
들다[deul-da]	3	61.92%	46.36%	63.00%	48.54%	59.77%	49.05%
말다[mal-da]	3	48.51%	34.70%	49.25%	33.58%	45.90%	33.58%
맞다[mat-da]	3	55.45%	38.63%	55.38%	40.44%	52.49%	39.70%
묻다[mut-da]	3	64.86%	39.71%	66.85%	40.43%	60.71%	41.28%
붓다[but-da]	2	76.92%	57.69%	76.92%	56.04%	75.27%	57.69%
쉬다[swi-da]	3	79.19%	54.53%	80.92%	57.03%	76.69%	56.65%
싸다[ssa-da]	3	67.77%	40.40%	63.13%	44.15%	60.93%	45.03%
쓰다[sseu-da]	3	53.78%	41.93%	56.68%	44.45%	52.18%	43.86%
이르다[I-reu-da]	3	60.50%	41.31%	49.24%	37.97%	44.07%	35.21%
지다[ji-da]	4	70.22%	46.22%	64.00%	45.78%	63.11%	46.96%
차다[cha-da]	4	60.74%	36.60%	62.20%	39.39%	58.36%	39.79%
켜다[kyeo-da]	2	83.82%	63.97%	87.50%	70.59%	83.82%	68.38%
타다[ta-da]	5	72.52%	50.23%	77.93%	56.83%	75.08%	57.81%