Word Sense Disambiguation Based on The Information Theory

Ho Lee, Dae-Ho Baek, Hae-Chang Rim
Natural Language Processing Lab.,
Department of Computer Science and Engineering,
Korea University,
Anam-Dong, Seoul 136, Republic of South Korea
leeho@nlp.korea.ac.kr daeho@nlp.korea.ac.kr rim@nlp.korea.ac.kr

Abstract

The task of word sense disambiguation is to identify the correct sense of a word in context. In this paper, we define a new notion, classification information, based on the Shannon's information theory. The classification information of a word consists of the pair of the most probable class MPC and the discrimination score DS. In the sense decision of the target word, the MPC of a surrounding word represents the sense of the target word most closely related, and the DS represents the degree of correlation between the MPC and the surrounding word. When a new sentence containing the target polysemous word is given, the sense of the target word is determined to the most plausible sense based on the classification information of all surrounding words in the sentence. Experimental results show that the average accuracy of the proposed method is 84.6% for the Korean data set, and 80.0% for the English data set.

1. Introduction

The task of word sense disambiguation is to identify the correct sense of a word in context. The different meanings of a word are listed as its various senses in a dictionary. The improvement in the accuracy of identifying the correct word sense will result in better machine translation systems, information retrieval systems, etc.(Ng 1996).

There have been many approaches to solve word sense disambiguation problem. In the earlier, (Kelly 1975) and (Weiss 1973) made use of hand-coded knowledge. Therefore, it is nearly impossible to apply those approaches to practical systems because it is quite labor intensive to construct rules manually in those approaches(Gale 1992).

Recently, various knowledge sources have been utilized to resolve word sense ambiguity. One group acquired knowledge from machine readable dictionaries, and the other group acquired knowledge from sense tagged corpora. The first group of researchers, (Lesk 1986), (Walker 1987), (Luk 1995), and (Ide 1990), use machine readable dictionaries, such as *Oxford's Advanced Learner's Dictionary of Current English*, to resolve word sense ambiguity. They try to develop a program that can read an arbitrary text and tag each word in the text with a pointer to a particular sense number in a particular dictionary. However, those approaches do not seem to work very well because dictionaries simply do not record enough of the relevant information.

The second group, such as (Miller 1994), (Leacock 1993), (Yarowsky 1992), (Bruce 1994), and (Ng 1996), acquired knowledge from a sense tagged corpus in order to solve word sense disambiguation problem. They extracted unordered set of surrounding words, part of speech of target words, morphological forms, or syntactic relations from corpus. In order to employ those extracted information, they used statistical classifiers, neural networks, IR-based techniques, or exemplar-based learning method. The approaches based on a sense-tagged corpus can reduce human intervention, and report relatively high accuracy.

Recently, there are a few approaches to overcome knowledge acquisition bottleneck problem. Yarowsky(1995) proposed an unsupervised training method, and Gale (1992) used a bilingual corpus in order to solve knowledge acquisition bottleneck problem.

In this paper, we propose a method of resolving word sense ambiguity based on minimal information extracted from a sense tagged corpus. For this research, we define the classification information which can be represented by the most probable class(henceforth, *MPC*) and the discrimination score(henceforth, *DS*).

This paper is organized as follows. In the following section, we define the classification information. In the section 3, we apply the classification information to the word sense disambiguation problem, and then we show the experimental results in the section 4. Finally, we discuss the characteristics and problems of our method, and present the possible way of overcoming the problems in future.

2. Classification Informations

In this section, we define the classification information to determine the sense of the target word. The classification information is formalized form of information involved in each surrounding word. The classification information of a surrounding word consists of two fields, the *MPC* and the *DS*. The *MPC* of a surrounding word represents the sense of the target word most closely related, and the *DS* represents the degree of correlation between the *MPC* and the surrounding word.

Shannon(1951) understood information as a liberty of choice. The liberty of choice is granted on a selected message among various messages which can be produced by information sources. He thinks that the uncertainty grows in proportion to the amount of the increased liberty. Moreover, he measured the uncertainty by the entropy, and the measure becomes the average information value per message. The information value of the i-th message in the entropy equation is $\log_2 p_i$, which is determined by p_i, the occurrence probability of the message. So entropy, H, becomes the average information value of n messages.

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (1)$$

From the viewpoint of the information theory, each surrounding word can decrease the uncertainty of the given target word. The word, which can decrease much uncertainty, has more discriminating ability. Therefore, assuming that the size of data for each sense is the same, the noise produced by the surrounding word w_k is defined as

$$\begin{aligned} noise_k &= -\sum_{i=1}^{n} p(sense_i|w_k)\log_2 p(sense_i|w_k) \\ &= -\sum_{i=1}^{n} \frac{freq(sense_i, w_k)}{freq(w_k)} \log_2 \frac{freq(sense_i, w_k)}{freq(w_k)} \end{aligned} \quad (2)$$

where n is the number of senses, and p_i, the occurrence frequency of surrounding word w_k, represents $p(sense_i|w_k)$, the conditional probability of $sense_i$ given the surrounding word w_k. In the equation (2), $noise_k$ has the value from 0 to $\log_2 n$ and it has maximum value when all occurrence probabilities of w_k are same. The word whose noise is high has low discriminating ability and provides little assistance for determining the sense of the target polysemous word. Therefore, we can measure the discriminating ability with the reverse function of noise as shown in the equation (3).

$$DS_k = signal_k = \log_2 n - noise_k \quad (3)$$

The MPC can be calculated according to the equation (4).

$$MPC_k = \text{argmax}_i\ p_i = \text{argmax}_i\ p(sense_i|w_k) = \text{argmax}_i \frac{freq(sense_i, w_k)}{freq(w_k)} \quad (4)$$

The equations (3) and (4) are based on the hypothesis that the size of data for each sense is same. However, the difference among the size of data may have an effect on the values of the MPC_k and the DS_k. Therefore, the normalization based on the data size is required. The normalized occurrence probability $\hat{p}_i$ is defined as the equation (5).

$$\hat{p}_i = \frac{p_i \dfrac{\overline{N(senses)}}{N(sense\ i)}}{\sum_{j=1}^{n} p_j \dfrac{\overline{N(senses)}}{N(sense\ j)}} = \frac{p(w_k|sense\ i)}{\sum_{j=1}^{n} p(w_k|sense\ j)} \quad (5)$$

where $N(sense_i)$ represents the data size of i-th sense, and $\overline{N(senses)}$ represents the average of $N(sense_i)$. The equation (6) shows the modified formula of $noise_k$ based on the equation (5).

$$noisy_k = -\sum_{i=1}^{n} \hat{p}_i \log_2 \hat{p}_i = -\sum_{i=1}^{n} \frac{p(w_k|sense\ i)}{\sum_{j=1}^{n} p(w_k|sense\ j)} \log_2 \frac{p(w_k|sense\ i)}{\sum_{j=1}^{n} p(w_k|sense\ j)} \qquad (6)$$

In the equation(6), $noise_k$ also has the value from 0 to $\log_2 n$. The normalized DS_k can be calculated by applying the equation (6) to the equation (3). The normalized MPC_k can be acquired by the equation (7).

$$MPC_k = argmax_i\ \hat{p}_i \qquad (7)$$

3. Sense Decision Using Classification Information

With the following sentence, we will explain the import of the classification information in the word sense disambiguation.

Several financial institutions, both ***banks*** *and insurance companies, have been sounded out.*

In general, human refers surrounding words in order to determine the sense of the polysemous word 'bank'. However, not all of the surrounding words can provide clues for the sense decision. The surrounding words, 'financial', 'institution', 'insurance', and 'company' provide important clues. On the other hand, 'several', 'have', 'be', 'sound', and 'out' provide less information to the sense decision. The words providing important clues occur frequently in the sentence that the word 'bank' is used as one specific sense, but occur rarely in the sentence that the word 'bank' is used as other senses. Consequently, important clues have high DS value in the classification information.

Because the classification information provides the importance of the surrounding word, we can easily determine the sense of the target word with the summation of DS of all surrounding words. The sense of the target word contained in a sentence $S = \{w_1, w_2, \cdots w_n\}$ can be determined by the equation (8).

$$MPC(S) = argmax_i \sum_{k=1}^{n} DS_k(i) \qquad (8)$$

where the discrimination score of w_k over $sense_i$, $DS_k(i)$, is defined as the equation (9).

$$DS_k(i) = \begin{cases} DS_k & \text{if } i = MPC_k \\ 0 & otherwise \end{cases} \qquad (9)$$

For example, the table 1 presents the sense decision in a sentence containing

	Training phase		Testing phase			
surrounding words	MPC_k	DS_k	$DS_k(i)$			
			sense 1	sense 2	sense 3	sense 4
w_1	3	0.7324	0	0	0.7324	0
w_2	2	1.3881	0	1.3881	0	0
w_3	2	0.9077	0	0.9077	0	0
w_4	4	0.3140	0	0	0	0.3140
w_5	3	0.2663	0	0	0.2663	0
w_6	1	0.5817	0.5817	0	0	0
w_7	2	0.8203	0	0.8203	0	0
w_8	3	0.4938	0	0	0.4938	0
$\sum_{i=1}^{8} DS_k(i)$			0.5817	3.1161	1.4925	0.3140
sense of the target word			sense 2			

Table 1. An example of sense decision using the classification information

words $w_1 \sim w_8$. The DS_k is assigned to $DS_k(i)$ if i is the MPC of w_k, and 0 otherwise. Therefore the value of $DS_1(3)$ becomes 0.7324 and other values of $DS_1(i)$ becomes 0, because the MPC of w_1 is $sense_3$ and the DS of w_1 is 0.7324. Finally, we determine the sense which has the maximum $\sum_i DS_k(i)$ as the most plausible sense of the target word.

4. Experimental Results

Our word sense disambiguation method is tested with the data from two languages, one is Korean and the other is English. Probably, our method can be applied to any other language because only the occurrence frequencies of surrounding words are required to determine the word sense.

4.1 Korean Word Sense Disambiguation

Words	Senses
배(Pae):NN	the belly(腹), a pear(梨), a boat(船), an embryo(胚)
전자(Jeon-Ja):NN	an electron(電子), the former(前者)
감다(Kam-Ta):VV	close one's eyes, wash, wind
열리다(yeol-Ri-da):VV	open, hold a meeting, spread a space, make way for a person, start up, enlighten, make out what a person say

Table 2. Four Korean polysemous words and their senses

Word	Inside test			Outside test		
	baseline	accuracy	improvement	baseline	accuracy	improvement
배(Pae)	61.4%	92.8%	31.4%	69.6%	78.3%	8.7%
전자(Jeon-Ja)	87.3%	98.0%	10.7%	69.5%	81.0%	11.5%
감다(Kam-Ta)	60.3%	98.4%	38.1%	80.8%	84.9%	4.1%
열리다(yeol-Ri-da)	68.8%	100.0%	31.2%	70.3%	81.6%	11.3%

Table 3. The results of inside and outside test

For the first experiment, we select four target polysemous words, extract concordances of those words from 10 million size raw corpus, and manually tag the sense of the word. In the outside test, we select the 80% of the concordances as a training set and the remaining concordances as a test set. The table 2 contains the target polysemous words and their senses.

The table 3 contains the result of the inside test and the outside test acquired from 100 trials. The baseline method in the table 3 represents the primitive method that always selects the most frequent sense. In the inside test, the accuracy of our method is much higher than the baseline method. From this result, we can say that the classification informations reflect the implicit informations of the training data set very well. However, the average accuracy in the outside test is about 84.6%. We think that one major reason of the low accuracy is the data sparseness. We also think that morphological ambiguity has bad effect on word sense disambiguation since we use the raw corpus for training and testing.

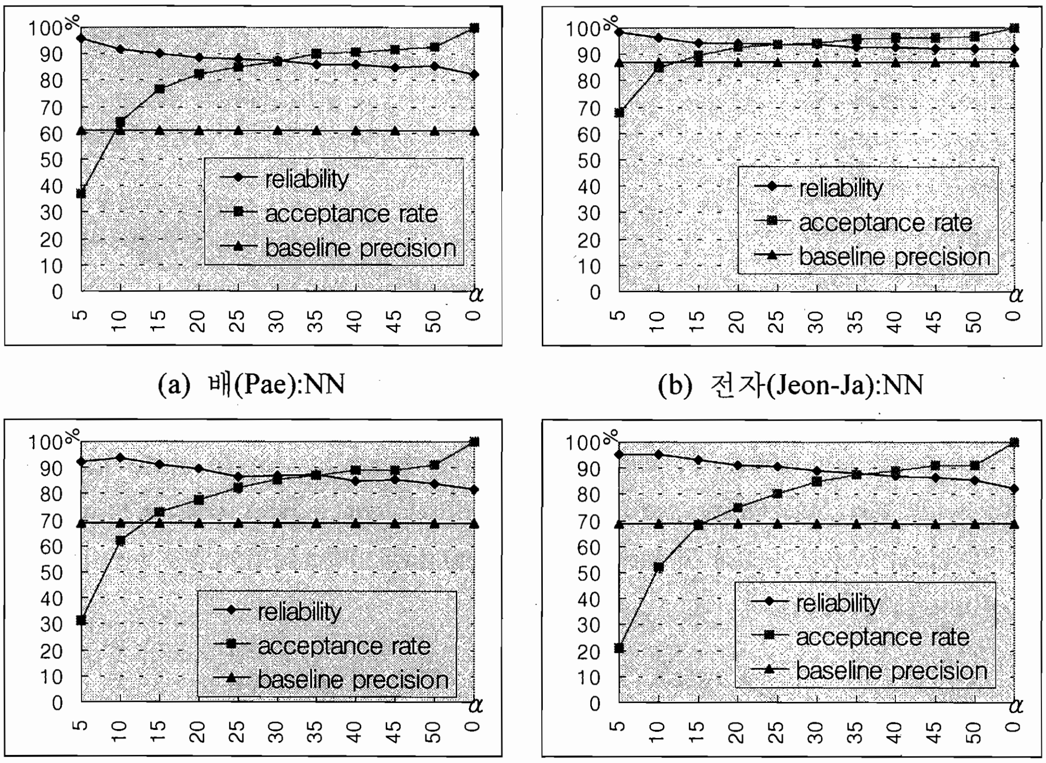
The table 4 shows the average difference between the DS of the correct sense and the maximum DS of incorrect sense per word. The values in the table 4 are calculated by the equation (10)where N is the number of words in the sentence and cs denotes the correct sense.

$$\frac{\left| \sum_{k=1}^{N} DS_k(cs) - \{ argmax_{\,i,i \neq cs} \sum_{k=1}^{N} DS_k(i) \} \right|}{N} \qquad (10)$$

As shown in the table 4, the average differences of DSs are much smaller in the case that incorrect senses are selected. We have made an experiment that admit

Words	Successful case	Failed case
배(Pae)	0.2905	0.1337
전자(Jeon-Ja)	0.1595	0.0936
감다(Kam-Ta)	0.2683	0.1062
열리다(yeol-Ri-da)	0.3017	0.1360

Table 4. The differences between $DS_k(i)$



(a) 배(Pae):NN

(b) 전자(Jeon-Ja):NN

(c) 감다(Kam-Ta):VV

(d) 열리다(yeol-Ri-da):VV

Figure 1. The reliability and the acceptance rate of reservation strategy

the *empty decision*. The empty decision represents the case when the word sense decision is deferred if the average difference of *DS* per word is less than the threshold calculated by the equation (11).

$$Threshold = \frac{\log_2 n}{\alpha} \tag{11}$$

where α is a arbitrary constant value and n is the number of senses. In the equation (11), we do not use the single constant value as the threshold. The more sense the polysemous word has, the greater value the average difference of *DS* per word has. Therefore, the empty decision rate increases in proportion to the number of senses, if the threshold has the single constant value. In order to acquire the consistent result for all polysemous words, we make the variable threshold in proportion to the maximum of the average difference of *DS* per word. For example, if the value of α is 5, then the empty decision breaks out when the average difference of *DS* per word is less than $\frac{1}{5}$(=20%) of the maximum value.

The figure 1 show the experimental results of the reservation strategy. The reliability means the proportion of the correct decision to the total number of decision. The acceptance rate means the proportion of the decided sentences to whole input sentences.

WSD research	accuracy
baseline	53%
Black(1988)	72%
Zernik(1990)	70%
Yarowsky(1992)	72%
Bruce & Wiebe(1994)	79%
Ng & Lee(1996)	89%
proposed method	80%

Table 5. Comparison with previous works

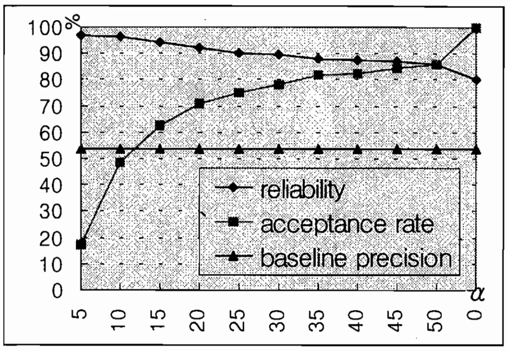


Figure 2. The result of reservation strategy - interest:NN

As shown in the figure 1, we can improve reliability by a little loss of acceptance rate with the reservation strategy. Therefore, we expect that we will get high accuracy if other word sense disambiguation method is additionally employed to our method as a post-process.

4.2 English Word Sense Disambiguation

In the second experiment, we used an English data set which has been commonly used in several previous researches. So far, very few existing works on word sense disambiguation have been tested and evaluated on a common data set. We could acquire only one sense-tagged data set used in (Bruce 1994), which has been made available in the public domain by Bruce and Wiebe. The data set consists of 2369 sentences each containing an occurrence of the noun "interest" (or its plural form "interests") with its correct sense manually tagged(Bruce 1994)(Ng 1996). In order to compare our method with other researches, we applied classification informations to the common data set. The results of previous researches and our approach are shown in table 5.

As shown in the table 5, our proposed method is relatively better than previous works except the Ng's method. Ng's method is better than any other method in terms of the accuracy because he used complex informations such as parts of speech and surface forms of target words, surrounding words, collocations and structural relations. In our approach, however, only surrounding words are used to determine word senses. Therefore, our approach can be easily applied to other languages.

We also apply the reservation strategy to the English data set, and the result is shown in the figure 2. We can also achieve high reliability by a little loss of acceptance rate in the English data set by admitting the empty decision.

5. Conclusions and Future Works

In this paper, we have presented a method of word sense disambiguation by using classification informations. We have achieved about 96.7% accuracy in the inside test and about 84.6% accuracy in the outside test. Moreover, we could achieve higher accuracy at the cost of few recall rate under the reservation strategy.

We can say that our method has three characteristics. The first characteristic is the ease of modeling. As we use classification informations, it is possible to decompose whole word sense disambiguation model easily into word unit models. The second characteristic is the ease of information acquisition. For classification informations of word sense disambiguation, the minimal information, the occurrence frequencies of surrounding words, is only required. The third characteristic is language independency. However, our method can be applied to any other language because the information used in our method is so simple that it can be extracted by the same procedure regardless of the language. Our method have two problems, the knowledge acquisition bottleneck and the data sparseness problem.

For the future work, we will try to use a word class as a unit of the classification information in order to solve the data sparseness problem and combine our method to the unsupervised training technique. Moreover, we will also study the technique of combining classification informations with other useful informations.

References

Black, Ezra, "An Experiment in Computational Discrimination of English Word Sense," *IBM Journal of Research and Development*, 32(2), pp. 185-194, 1988.

Bruce, R. and Wiebe, J., "Word Sense Disambiguation using Decomposable Models," In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 139-146, 1994.

Gale, W., Church, K. and Yarowsky, D., "A Method for Disambiguating word senses in a large corpus," *Computers and the Humanities*, 26, pp. 415-439, 1992.

Ide, N. M., and Veronis J., "Very Large Neural Networks for Word Sense Disambiguation,", In *Proceedings of the 9th European Conference on Artificial Intelligence, ECAI90*, pp. 366-368, 1990.

Kelly, E. and Stone, P., *Computer Recognition of English Word Senses*, 1975.

Leacock, C., Towell, G., and Voorhees, E., "Corpus-based statistical sense resolution," In *Proceedings of the ARPA Human Language Technology Workshop*, 1993.

Lesk, M., "Automatic Sense Disambiguation: How to tell a Pine Cone from an Ice Cream Cone," In *Proceeding of the 1986 SIGDOC Conference*, 1986.

Luk, K. A., "Statistical sense disambiguation with relatively small corpora using dictionary definitions," In *Proceedings of the 33rd Annual Meetings of the Association for Computational Linguistics*, pp. 181-188, 1995.

Miller, A. G., Chodorow, M., Landes, S., Leacock, C. and Robert G. T., "Using a semantic concordance for sense identification," In *Proceedings of the ARPA Human Language Technology Workshop*, 1994.

Ng, H. T. and Lee, H. B., "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach," In *Proceedings of the 34th Annual Meetings of the Association for Computational Linguistics*, pp. 40-47, 1996.

Shannon, C. E., Prediction and Entropy in Printed English, In *Bell System Technical Journal*, pp. 50-65, 1951.

Walker, D., "Knowledge Resource Tools for Accessing Large Text Files," In *Machine Translation: Theoretical and Methodological Issues*, 1987.

Weiss, S., "Learning to Disambiguate.", In *Information Storage and Retrieval*, pp. 33-41, 1973.

Yarowsky, D., "Word-sense Disambiguation using statistical models of Roget's categories trained on Large Corpora," In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, pp. 456-460, 1992.

Yarowsky, D., "Unsupervised Word Sense Disambiguation rivaling supervised methods," In *Proceedings of the 33rd Annual Meetings of the Association for Computational Linguistics*, pp. 189-196, 1995.

Zernik, Uri, "Tagging word senses in corpus:the needle in the haystack revisitied," *Technical Report 90CRD198*, GE R&D Center, 1990.