Word Sense Disambiguation by Learning from Unlabeled Data

Seong-Bae Park[†], Byoung-Tak Zhang[†] and Yung Taek Kim[‡]
Artificial Intelligence Lab (SCAI)
School of Computer Science and Engineering
Seoul National University
Seoul 151-742, Korea
[†]{sbpark,btzhang}@scai.snu.ac.kr [‡]ytkim@cse.snu.ac.kr

Abstract

Most corpus-based approaches to natural language processing suffer from lack of training data. This is because acquiring a large number of labeled data is expensive. This paper describes a learning method that exploits unlabeled data to tackle data sparseness problem. The method uses committee learning to predict the labels of unlabeled data that augment the existing training data. Our experiments on word sense disambiguation show that predictive accuracy is significantly improved by using additional unlabeled data.

1 Introduction

The objective of word sense disambiguation (WSD) is to identify the correct sense of a word in context. It is one of the most critical tasks in most natural language applications, including information retrieval, information extraction, and machine translation. The availability of large-scale corpus and various machine learning algorithms enabled corpus-based approach to WSD (Cho and Kim, 1995; Hwee and Lee, 1996; Wilks and Stevenson, 1998),but a large scale sense-tagged corpus or aligned bilingual corpus is needed for a corpus-based approach.

However, most languages except English do not have a large-scale sense-tagged corpus. Therefore, any corpus-based approach to WSD for such languages should consider the following problems:

- There's no reliable and available sense-tagged corpus.
- Most words are sense ambiguous.
- Annotating the large corpora requires human experts, so that it is too expensive.

Because it is expensive to construct sense-tagged corpus or bilingual corpus, many researchers tried to reduce the number of examples needed to learn WSD (Atsushi et al., 1998; Pedersen and Bruce, 1997). Atsushi et al. (Atsushi et al., 1998) adopted a selective sampling method to use small number of examples in training. They defined a training utility function to select examples with minimum certainty, and at each training iteration the examples with less certainty were saved in the example database. However, at each iteration of training the similarity among word property vectors must be calculated due to their k-NN like implementation of training utility.

While labeled examples obtained from a sense-tagged corpus is expensive and time-consuming, it is significantly easier to obtain the unlabeled examples. Yarowsky (Yarowsky, 1995) presented, for the first time, the possibility that unlabeled examples can be used for WSD. He used a learning algorithm based on the local context under the assumption that all instances of a word have the same intended meaning within any fixed document and achieved good results with only a few labeled examples and many unlabeled ones. Nigam et al. (Nigam et al., 2000) also showed the unlabeled examples can enhance the accuracy of text categorization.

Attribute	Substance
GFUNC	the grammatical function of w
PARENT	the word of the node modified by w
SUBJECT	whether or not *PARENT* of w has a subject
OBJECT	whether or not *PARENT* of w has an object
NMODWORD	the word of the noun modifier of w
ADNWORD	the head word of the adnominal phrase of w
ADNSUBJ	whether or not the adnominal phrase of w has a subject
ADNOBJ	whether or not the adnominal phrase of w has an object

Table 1: The properties used to distinguish the sense of an ambiguous Korean noun w.

In this paper, we present a new approach to word sense disambiguation that is based on selective sampling algorithm with committees. In this approach, the number of training examples is reduced, by determining by weighted majority voting of multiple classifiers, whether a given training example should be learned or not. The classifiers of the committee are first trained on a small set of labeled examples and the training set is augmented by a large number of unlabeled examples. One might think that this has the possibility that the committee is misled by unlabeled examples. But, the experimental results confirm that the accuracy of WSD is increased by using unlabeled examples when the members of the committee are well trained with labeled examples. We also theoretically show that performance improvement is guaranteed by a mild requirement, i.e., the base classifiers need to guess better than random selection. This is because the possibility misled by unlabeled examples is reduced by integrating outputs of multiple classifiers. One advantage of this method is that it effectively performs WSD with only a small number of labeled examples and thus shows possibility of building word sense disambiguators for the languages which have no sense-tagged corpus.

The rest of this paper is organized as follows. Section 2 introduces the general procedure for word sense disambiguation and the necessity of unlabeled examples. Section 3 explains how the proposed method works using both labeled and unlabeled examples. Section 4 presents the experimental results obtained by using the KAIST raw corpus. Section 5 draws conclusions.

2 Word Sense Disambiguation

Let $S \in \{s_1, \ldots, s_k\}$ be the set of possible senses of a word to be disambiguated. To determine the sense of the word, we need to consider the contextual properties. Let $\mathbf{x} =< x_1, \ldots, x_n >$ be the vector for representing selected contextual features. If we have a classifier $f(\mathbf{x}, \theta)$ parameterized with θ, then the sense of a word with property vector $\mathbf{x}$ can be determined by choosing the most probable sense s^*:

$$s^* = \arg\max_{s \in S} f(\mathbf{x}, \theta).$$

The parameters θ are determined by training the classifier on a set of labeled examples, $L = \{(\mathbf{x}_1, s_1), \ldots, (\mathbf{x}_N, s_N)\}$.

2.1 Property Sets

In general, the first step of WSD is to extract a set of contextual features. To select particular properties for Korean, the language of our cencern, the following characteristics should be considered:

- Korean is a partially free-order language. The ordering information on the neighbors of the ambiguous word, therefore, does not give significantly meaningful information in Korean.

- In Korean, ellipses appear very often with a nominative case or objective case. Therefore, it is difficult to build a large

scale database of labeled examples with case markers.

Considering both characteristics and results of previous work, we select eight properties for WSD of Korean nouns (Table 1). Three of them (*PARENT, NMODWORD, ADNWORD*) take morphological form as their value, one (*GFUNC*) takes 11 values of grammatical functions[1], and others take only *true* or *false*.

2.2 Unlabeled Data for WSD

Many researchers tried to develop automated methods to reduce training cost in language learning and found out that the cost can be reduced by *active learning* which has control over the training examples (Dagan and Engelson, 1997; Liere and Tadepalli, 1997; Zhang, 1994). Though the number of labeled examples needed is reduced by active learning, the label of the selected examples must be given by the human experts. Thus, active learning is still expensive and a method for automatic labeling unlabeled examples is needed to have the learner automatically gather information (Blum and Mitchell, 1998; Pedersen and Bruce, 1997; Yarowsky, 1995).

As the unlabeled examples can be obtained with ease without human experts it makes WSD robust. Yarowsky (Yarowsky, 1995) presented the possibility of automatic labeling of training examples in WSD and achieved good results with only a few labeled examples and many unlabeled examples. On the other hand, Blum and Mitchell tried to classify Web pages, in which the description of each example can be partitioned into distinct views such as the words occurring on that page and the words occurring in hyperlinks (Blum and Mitchell, 1998). By using both views together, they augmented a small set of labeled examples with a lot of unlabeled examples.

The unlabeled examples in WSD can provide information about the joint probability distribution over properties but they also can mislead the learner. However, the possibility of being misled by the unlabeled examples is reduced by the committee of classifiers since combining or integrating the outputs of several classifiers in general leads to improved performance. This is why we use active learning with committees to select informative unlabeled examples and label them.

3 Active Learning with Committees for WSD

3.1 Active Learning Using Unlabeled Examples

The algorithm for active learning using unlabeled data is given in Figure 1. It takes two sets of examples as inputs. A Set L is the one with labeled examples and $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is the one with unlabeled examples where $\mathbf{x}_i$ is a property vector. First of all, the training set $L_j^{(1)}$ $(1 \leq j \leq M)$ of labeled examples is constructed for each base classifier C_j. This is done by random resampling as in Bagging (Breiman, 1996). Then, each base classifier C_j is trained with the set of labeled examples $L_j^{(1)}$.

After the classifiers are trained on labeled examples, the training set is augmented by the unlabeled examples. For each unlabeled example $\mathbf{x}_t \in D$, each classifier computes the sense $y_j \in S$ which is the label associated with it, where S is the set of possible sense of $\mathbf{x}_t$.

The distribution W over the base classifiers represents the importance weights. As the distribution can be changed each iteration, the distribution in iteration t is denoted by W_t. The importance weight of classifier C_j under distribution W_t is denoted by $W_t(j)$. Initially, the base classifiers have equal weights, so that $W_t(j) = 1/M$.

The sense of the unlabeled example $\mathbf{x}_t$ is determined by majority voting among C_j's with weight distribution W. Formally, the sense y_t of $\mathbf{x}_t$ is predicted by

$$y_t(\mathbf{x}_t) = \arg\max_{y \in S} \sum_{j:C_j(\mathbf{x}_t)=y} W_t(j).$$

If most classifiers believe that y_t is the correct

[1] These 11 grammatical functions are from the parser, KEMTS (Korean-to-English Machine Translation System) developed in Seoul National University, Korea.

Given an unlabeled example set $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$
and a labeled example set L
and a word sense set $S \in \{s_1, \ldots, s_k\}$ for $\mathbf{x}_i$,

Initialize $W_1(j) = \frac{1}{M}$,
where M is the number of classifiers in the committee.
Resample $L_j^{(1)}$ from L for each classifier C_j,
where $|L_j^{(1)}| = |L|$ as done in Bagging.
Train base classifier C_j $(1 \leq j \leq M)$ from $L_j^{(1)}$.
For $t = 1, \ldots, T$:

1. Each C_j predicts the sense $y_j \in S$ for $\mathbf{x}_t \in D$.

$$Y =< y_1, \ldots, y_M >$$

2. Find the most likely sense y_t from Y using distribution W:

$$y_t = \arg\max_{y \in S} \sum_{j: C_j(\mathbf{x}_t) = y} W_t(j).$$

3. Set $\alpha_t = \frac{1-\epsilon_t}{\epsilon_t}$, where

$$\epsilon_t = \frac{\text{No. of } C_j\text{'s whose predictions are not } y_t}{M}.$$

4. **If** α_t is larger than a *certainty* threshold θ, **then** update W_t:

$$W_{t+1}(j) = \frac{W_t(j)}{Z_t} \times \begin{cases} \alpha_t & \text{if } y_j = y_t \\ 1 & \text{otherwise,} \end{cases}$$

where Z_t is a normalization constant.

5. **Otherwise,** every classifier C_j is restructured from new training set $L_j^{(t+1)}$:

$$L_j^{(t+1)} = L_j^{(t)} + \{(\mathbf{x}_t, y_t)\}.$$

Output the final classifier:

$$f(\mathbf{x}) = \arg\max_{y \in S} \sum_{j: C_j(\mathbf{x}) = y} W_T(j).$$

Figure 1: The active learning algorithm with committees using unlabeled examples for WSD.

sense of $\mathbf{x}_t$, they need not learn $\mathbf{x}_t$ because this example makes no contribution to reduce the variance over the distribution of examples. In this case, instead of learning the example, the weight of each classifier is updated in such a way that the classifiers whose predictions were correct get a higher importance weight and the classifiers whose predictions were wrong get a lower importance weight under the assumption that the correct sense of $\mathbf{x}_t$ is y_t. This is done by multiplying the weight of the classifier whose prediction is y_t by *certainty* α_t. To ensure the updated W_{t+1} form a distribution, W_{t+1} is normalized by constant Z_t. Formally, the importance weight is updated as follows:

$$W_{t+1} = \frac{W_t(j)}{Z_t} \times \begin{cases} \alpha_t & \text{if } y_j = y_t, \\ 1 & \text{otherwise.} \end{cases}$$

The certainty α_t is computed from error ϵ_t. Because we trust that the correct sense of $\mathbf{x}_t$ is y_t, the error ϵ_t is the ratio of the number of classifiers whose predictions are not y_t. That is, α_t is computed as

$$\alpha_t = \frac{1 - \epsilon_t}{\epsilon_t}$$

where ϵ_t is given as

$$\epsilon_t = \frac{\text{No. of } C_j\text{'s whose predictions are not } y_t}{M}.$$

Note that the smaller ϵ_t, the larger the value of α_t. This implies that, if the sense of $\mathbf{x}_t$ is certainly y_t and a classifier predicts it, a higher weight is assigned to the classifier. We assume that most classifiers believe that y_t is the sense of $\mathbf{x}_t$ if the value of y_t is larger than a certainty threshold θ which is set by trial-and-error.

However, if the certainty is below the threshold, the classifiers need to learn the example $\mathbf{x}_t$ yet with belief that the sense of it is y_t. Therefore, the set of training examples, $L_j^{(t)}$, for the classifier C_j is expanded by

$$L_j^{(t+1)} = L_j^{(t)} + \{(\mathbf{x}_t, y_t)\}.$$

Then, each classifier C_j is restructured with $L_j^{(t+1)}$.

This process is repeated until the unlabeled examples are exhausted. The sense of a new example $\mathbf{x}$ is then determined by weighted majority voting among the trained classifiers:

$$f(\mathbf{x}) = \arg\max_{y \in S} \sum_{j:C_j(\mathbf{x})=y} W_T(j),$$

where $W_T(j)$ is the importance weight of classifier C_j after the learning process.

3.2 Theoretical Analysis

Previous studies show that using multiple classifiers rather than a single classifier leads to improved generalization (Breiman, 1996; Freund et al., 1992) and learning algorithms which use *weak* classifiers can be boosted into *strong* algorithms (Freund and Schapire, 1996). In addition, Littlestone and Warmuth (Littlestone and Warmuth, 1994) showed that the error of the weighted majority algorithm is linearly bounded on that of the best member when the weight of each classifier is determined by held-out examples.

The performance of the proposed method depends on that of initial base classifiers. This is because it is highly possible for unlabeled examples to mislead the learning algorithm if they are poorly trained in their initial state. However, if the accuracy of the initial majority voting is larger than $\frac{1}{2}$, the proposed method performs well as the following theorem shows.

Theorem 1 *Assume that every unlabeled data $\mathbf{x}_t$ is added to the set of training examples for all classifiers and the importance weights are not updated. Suppose that p_0 be the probability that the initial classifiers do not make errors and β_t $(0 \leq \beta_t \leq 1)$ be the probability by which the accuracy is increased in adding one more correct example or decreased in adding one more incorrect example at iteration t. If $p_0 \geq \frac{1}{2}$, the accuracy does not decrease as a new unlabeled data is added to the training data set.*

Proof. The probability for the classifiers to predict the correct sense at iteration $t = 1$, p_1, is

$$\begin{aligned} p_1 &= p_0(p_0 + \beta_0) + (1 - p_0)(p_0 - \beta_0) \\ &= p_0(2\beta_0 + 1) - \beta_0 \end{aligned}$$

because the accuracy can be increased or decreased by β_0 with the probability p_0 and $1 - p_0$, respectively. Therefore, without loss of generality, at iteration $t = i + 1$, we have

$$p_{i+1} = p_i(2\beta_i + 1) - \beta_i.$$

To ensure the accuracy does not decrease, the condition $p_{i+1} \geq p_i$ should be satisfied.

$$\begin{aligned} p_{i+1} - p_i &= p_i(2\beta_i + 1) - \beta_i - p_i \\ &= p_i(2\beta_i) - \beta_i \geq 0 \\ \therefore p_i &\geq \frac{1}{2} \end{aligned}$$

The theorem follows immediately from this result. ■

3.3 Decision Trees as Base Classifiers

Although any kind of learning algorithms which meet the conditions for Theorem 1 can be used as base classifiers, Quinlan's C4.5 release 8 (Quinlan, 1993) is used in this paper. The main reason why decision trees are used as base classifiers is that there is a fast restructuring algorithm for decision trees. Adding an unlabeled example with a predicted label to the existing set of training examples makes the classifiers restructured. Because the restructuring of classifiers is time-consuming, the proposed method is of little practical use without an efficient way to restructure. Utgoff et al. (Utgoff et al., 1997) presented two kinds of efficient algorithms for restructuring decision trees and showed experimentally that their methods perform well with only small restructuring cost.

We modified C4.5 so that word matching is accomplished not by comparing morphological forms but by calculating similarity between words to tackle data-sparseness problem. The similarity between two Korean words is measured by averaged distance in *WordNet* of their English-translated words (Kim and Kim, 1996).

Word	No. of Senses	No. of Examples	Sense	Percentage
bae	4	876	pear	6.2%
			ship	**55.2%**
			times	13.7%
			stomach	24.9%
bun	3	796	person	46.2%
			minute	**50.8%**
			indignation	3.0%
jonja	2	350	the former	28.6%
			electron	**71.4%**
dari	2	498	bridge	30.9%
			leg	**69.1%**

Table 2: Various senses of Korean nouns used for the experiments and their distributions in the corpus.

4 Experiments

4.1 Data Set

We used the KAIST Korean raw corpus[2] for the experiments. The entire corpus consists of 10 million words but we used in this paper the corpus containing one million words excluding the duplicated news articles. Table 2 shows various senses of ambiguous Korean nouns considered and their sense distributions. The *percentage* column in the table denotes the ratio that the word is used with the sense in the corpus. Therefore, we can regard the maximum percentage as a lower bound on the correct sense for each word.

4.2 Experimental Results

For the experiments, 15 base classifiers are used. If there is a tie in predicting senses, the sense with the lowest order is chosen as in (Breiman, 1996). For each noun, 90% of the examples are used for training and the remaining 10% are used for testing.

Table 3 shows the 10-fold cross validation result of WSD experiments for nouns listed in Table 2. The accuracy of the proposed method shown in Table 3 is measured when the accuracy is in its best for various ratios of the number of labeled examples for base classifiers to total examples. The results show that WSD by selective sampling with committees using both labeled and unlabeled examples is comparable to a single learner using all the labeled examples. In addition, the method proposed in this paper achieves 26.3% improvement over the lower bound for '*bae*', 41.5% for '*bun*', 22.1% for '*jonja*', and 4.2% for '*dari*', which is 23.6% improvement on the average. Especially, for '*jonja*' the proposed method shows higher accuracy than the single C4.5 trained on the whole labeled examples.

Figure 2 shows the performance improved by using unlabeled examples. This figure demonstrates that the proposed method outperforms the one without using unlabeled examples. The *initial learning* in the figure means that the committee is trained on labeled examples, but is not augmented by unlabeled examples. The difference between two lines is the improved accuracy obtained by using unlabeled examples. When the accuracy of the proposed method gets stabilized for the first time, the improved accuracy by using unlabeled examples is 20.2% for '*bae*', 9.9% for '*bun*, 13.5% '*jonja*', and 13.4% for '*dari*'. It should be mentioned that the results also show that the accuracy of the proposed method may be dropped when the classifiers are trained on too small a set of labeled data, as is the case in the early stages of Figure 2. However, in typical situations where the classifiers are trained on minimum training set

[2]This corpus is distributed by the Korea Terminology Research Center for Language and Knowledge Engineering.

Word	Using Partially Labeled Data	Using All Labeled Data	Lower Bound
bae	81.5 ± 7.7%	**82.3% ± 5.9%**	55.2%
bun	92.3 ± 7.7%	**94.3% ± 5.7%**	50.8%
jonja	**93.5 ± 6.5%**	90.6% ± 9.4%	71.4%
dari	73.3 ± 14.2%	**80.8 ± 10.9%**	69.1%
Average	85.2%	**87.0%**	61.6%

Table 3: The accuracy of WSD for Korean nouns by the proposed method.

size, this does not happen as the results of other nouns show. In addition, we can find in this particular experiment that the accuracy is always improved by using unlabeled examples if only about 22% of training examples, on the average, are labeled in advance.

In Figure 2(a), it is interesting to observe jumps in the accuracy curve. The jump appears because the unlabeled examples mislead the classifiers only when the classifiers are poorly trained, but they play an important role as information to select senses when the classifiers are well trained on labeled examples. Other nouns show similar phenomena though the percentage of labeled examples is different when the accuracy gets flat.

5 Conclusions

In this paper, we proposed a new method for word sense disambiguation that is based on unlabeled data. Using unlabeled data is especially important in corpus-based natural language processing because raw corpora are ubiquitous while labeled data are expensive to obtain. In a series of experiments on word sense disambiguation of Korean nouns we observed that the accuracy is improved up to 20.2% using only 32% of labeled data. This implies, the learning model trained on a small number of labeled data can be enhanced by using additional unlabeled data. We also theoretically showed that the predictive accuracy is always improved if the individual classifiers do better than random selection after being trained on labeled data.

As the labels of unlabeled data are estimated by committees of multiple decision trees, the burden of manual labeling is minimized by using unlabeled data. Thus, the proposed method seems especially effective and useful for the languages for which a large-scale sense-tagged corpus is not available yet.

Another advantage of the proposed method is that it can be applied to other kinds of language learning problems such as POS-tagging, PP attachment, and text classification. These problems are similar to word sense disambiguation in the sense that unlabeled raw data are abundant but labeled data are limited and expensive to obtain.

Acknowledgements

This research was supported in part by the Korean Ministry of Education under the BK21 Program and by the Korean Ministry of Information and Communication through IITA under grant 98-199.

References

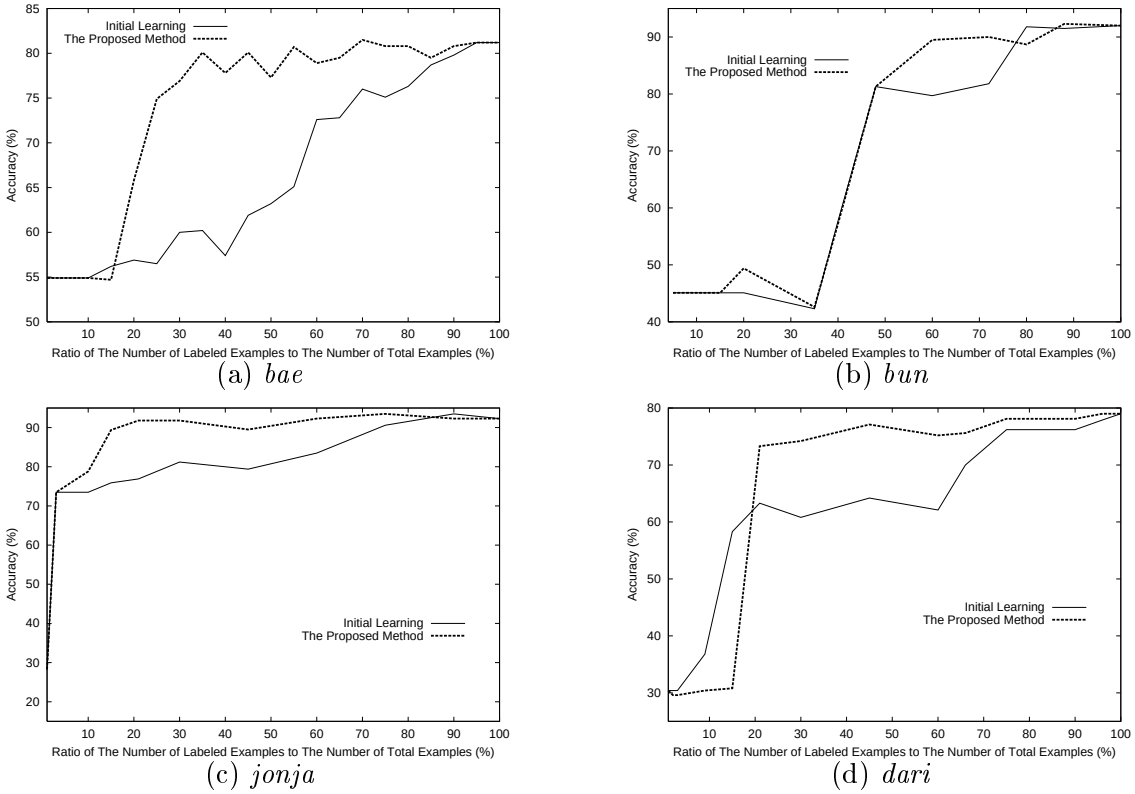
F. Atsushi, I. Kentaro, T. Takenobu, and T. Hozumi. 1998. Selective sampling of effective example sentence sets for word sense disambiguation. *Computational Linguistics*, 24(4):573–597.

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT-98*, pages 92–100.

L. Breiman. 1996. Bagging predictors. *Machine Learning*, 24:123–140.

J.-M. Cho and G.-C. Kim. 1995. Korean verb sense disambiguation using distributional information from corpora. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pages 691–696.

I. Dagan and S. Engelson. 1997. Committee-based sampling for training probabilistic classi-



(a) *bae* (b) *bun*

(c) *jonja* (d) *dari*

Figure 2: Improvement in accuracy by using unlabeled examples.

fiers. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 150–157.

Y. Freund and R. Schapire. 1996. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156.

Y. Freund, H. Seung, E. Shamir, and N. Tishby. 1992. Selective sampling with query by committee algorithm. In *Proceedings of NIPS-92*, pages 483–490.

T. Hwee and H. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the ACL*, pages 40–47.

Nari Kim and Y.-T. Kim. 1996. Ambiguity resolution of korean sentence analysis and korean-english transfer based on korean verb patterns. *Journal of KISS*, 23(7):766–775. in Korean.

R. Liere and P. Tadepalli. 1997. Active learning with committees for text categorization. In *Proceedings of AAAI-97*, pages 591–596.

N. Littlestone and M. Warmuth. 1994. The weighted majority algorithm. *Information and Computation*, 108(2):212–261.

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. 2000. Learning to classify text from labeled and unlabeled documents. *Machine Learning*, 39:1–32.

T. Pedersen and R. Bruce. 1997. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 399–401.

R. Quinlan. 1993. *C4.5: Programs For Machine Learning*. Morgan Kaufmann Publishers.

P. Utgoff, N. Berkman, and J. Clouse. 1997. Decision tree induction based on efficient tree restructuring. *Machine Learning*, 29:5–44.

Y. Wilks and M. Stevenson. 1998. Word sense disambiguation using optimised combinations of knowledge sources. In *Proceedings of COLING-ACL '98*, pages 1398–1402.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 189–196.

B.-T. Zhang. 1994. Accerlated learning by active example selection. *International Journal of Neural Systems*, 5(1):67–75.