Computers and the Humanities **34:** 141–146, 2000.

Word Sense Disambiguation Using the Classification Information Model

Experimental Results on The SENSEVAL Workshop

HO LEE[1], HAE-CHANG RIM[1] and JUNGYUN SEO[2]
[1]*Korea University, Seoul, 136, Korea (E-mail: {leeho,rim}@nlp.korea.ac.kr);* [2]*Sogang University, Seoul, 121 Korea (E-mail: seojy@ccs.sogang.ac.kr)*

Abstract. A Classification Information Model is a pattern classification model. The model decides the proper class of an input instance by integrating individual decisions, each of which is made with each feature in the pattern. Each individual decision is weighted according to the distributional property of the feature deriving the decision. An individual decision and its weight are represented as classification information which is extracted from the training instances. In the word sense disambiguation based on the model, the proper sense of an input instance is determined by the weighted sum of whole individual decisions derived from the features contained in the instance.

Key words: Classification Information Model, classification information, word sense disambiguation

1. Introduction

Word sense disambiguation can be treated as a kind of classification process. Classification is the task of classifying an input instance into a proper class among pre-defined classes, using features extracted from the instance. When the classification technique is applied to word sense disambiguation, an instance corresponds to a context containing a polysemous word. At the same time, a class corresponds to a sense of the word, and a feature to a clue for disambiguation. In this paper, we propose a novel classification model, the Classification Information Model (Lee et al., 1997), and describe the task of applying the model to the case of word sense disambiguation.

2. Classification Information Model

Classification Information Model is a model of classifying the input instance by use of the binary features representing the instance (Lee et al., 1997). We assume that each feature is independent from any other features. In the model, the proper class of an input instance, X, is determined by equation 1.

$$proper\ class\ of\ X \overset{\text{def}}{=} \arg\max_{c_j} \text{Rel}(c_j, X) \tag{1}$$

where c_j is the j-th class and $\text{Rel}(c_j, X)$ is the relevance between the j-th class and X. Since it is assumed that there is no dependency between features, the relevance can be defined as in equation 2.[1]

$$\text{Rel}(c_j, X) \stackrel{\text{def}}{=} \sum_{i=1}^{m} x_i w_{ij} \tag{2}$$

where m is the size of the feature set, x_i is the value of the i-th feature in the input instance, and w_{ij} is the weight between the i-th feature and the j-th class. In equation 2, x_i has binary value (1 if the feature occurs within context, 0 otherwise) and w_{ij} is defined by using classification information.

Classification information of a feature (f_i) is composed of two components. One is the MPC_i,[2] which corresponds to the most probable class of the instance determined by the feature. The other is the DS_i,[3] which represents the discriminating ability of the feature. Assuming we consider only a feature f_i, we can determine the proper class to be MPC_i and assign DS_i to the weight of the decision which is made with the feature f_i. Accordingly, w_{ij} in equation 2 is defined as in equation 3 with classification information of features.

$$w_{ij} \stackrel{\text{def}}{=} \begin{cases} \text{DS}_i & \text{if } c_j = \text{MPC}_i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In order to define classification information, the model uses the normalized conditional probability, $\hat{p}_{ji}$, defined in equation 4, instead of the conditional probability of classes given features, $p(c_j|f_i)$.[4]

$$\begin{aligned} \hat{p}_{ji} &\stackrel{\text{def}}{=} \frac{p(c_j|f_i)\frac{\overline{N(c)}}{N(c_j)}}{\sum_{k=1}^{n} p(c_k|f_i)\frac{\overline{N(c)}}{N(c_k)}} \\ &= \frac{p(f_i|c_j)}{\sum_{k=1}^{n} p(f_i|c_k)} \end{aligned} \tag{4}$$

In equation 4, $N(c_j)$ is the number of instances belonging to the class c_j and $\overline{N(c)}$ is the average number of instances per class. With the normalized conditional probability, both components of classification information are defined as in equations 5 and 6.

$$\begin{aligned} \text{MPC}_i &\stackrel{\text{def}}{=} \arg\max_{c_j} \hat{p}_{ji} \\ &= \arg\max_{c_j} p(f_i|c_j) \end{aligned} \tag{5}$$

Table I. Example of features and their classification information

Feature	MPC	DS	Feature	MPC	DS
(–1 very)	512274	0.8173	(+1 and)	512274	0.5202
(±5 very)	512274	0.8756	(±5 and)	512274	0.0275
(±5 been)	512274	0.8651	(±5 we)	512309	1.591
(±5 have)	512309	1.017	(±5 raised)	512309	2.585
(±5 about)	512309	1.619	(–B been very)	512274	2.585
(±B very and)	512274	2.585			

$$\begin{aligned} \mathrm{DS}_i &\overset{\text{def}}{=} \log_2 n - H(\hat{p}_i) \\ &= \log_2 n + \sum_{j=1}^{n} \hat{p}_{ji} \log_2 \hat{p}_{ji} \end{aligned} \tag{6}$$

3. Word Sense Disambiguation Based on the Classification Information Model

When the classification technique is applied to word sense disambiguation, input instances correspond to contexts containing polysemous words. At the same time, classes correspond to senses of the word, and features to clues for disambiguation. There are, however, various types of clues for sense disambiguation within context. Therefore, disambiguation models should be revised in order to utilize them. In addition to word bigram, a set of positional relationships, part-of-speech sequences, co-occurrences in a window, trigrams and verb-object pairs can be useful clues for word sense disambiguation (Yarowsky, 1996). Therefore, we adopt feature templates used in Yarowsky (1994) in order to represent all types of clues together. The templates of the condition field in our model are as follows:

1. *word* immediately to the right (+1 W)
2. *word* immediately to the left (–1 W)
3. *word* found in $\pm k$ word window ($\pm k$ W)
4. Pair of *words* at offsets –2 and –1 (–B W W)
5. Pair of *words* at offsets –1 and +1 (±B W W)
6. Pair of *words* at offsets +1 and +2 (+B W W)

The features extracted from the sentence 700005 among testing data set of *generous* and their classification information are shown in Table I.[5]

There are two advantages of separating the feature extractor from the disambiguation model. One is the language independent characteristic of the model. In order to apply this approach to other languages, only the substitution of feature templates,

Table II. Experimental results on the SENSEVAL data set

Sense degree	Systems	All words	Nouns	Verbs	Adjectives
Fine-grained	best baseline	0.691	0.746	0.676	0.688
	best system	0.781	0.845	0.720	0.751
	our system	0.701	0.773	0.646	0.673
Mixed-grained	best baseline	0.720	0.804	0.699	0.703
	best system	0.804	0.865	0.748	0.764
	our system	0.740	0.817	0.682	0.712
Coarse-grained	best baseline	0.741	0.852	0.717	0.705
	best system	0.818	0.885	0.761	0.766
	our system	0.752	0.835	0.692	0.715

not the modification of the model itself, is required. The other is flexibility for utilizing linguistic knowledge. If new useful linguistic knowledge is provided, the model can easily utilize it by extending feature templates.

4. Experimental Results

Some experimental results on the data set of the SENSEVAL workshop are shown in Table II.[6] Since our system uses a supervised learning method, the precision for only trainable words are contained in the table. Among the supervised learning systems, our system was ranked middle in performance, and can generally determine senses better than the best baseline method. However, our system was especially weak in determining the sense of verbs. One possible reason for this weakness is that the system exploited only words and parts-of-speech, though other higher level information, such as syntactic relations, is important for determining senses of verbs.

Figure 1 shows the correlation between the size of training data and precision: as the size of the data set is decreased, so too is the level of performance. This tendency is fairly regular and is independent of the part-of-speech of target polysemous words. Therefore, additional techniques for relaxing the data sparseness problem are required for our system.

5. Summary

Our model is a supervised learning model, based on classification information. It has several good characteristics. The model can exploit various types of clues because it adopted the feature templates. Moreover, the model is language independent since the feature extractor instead of the disambiguation model handles all

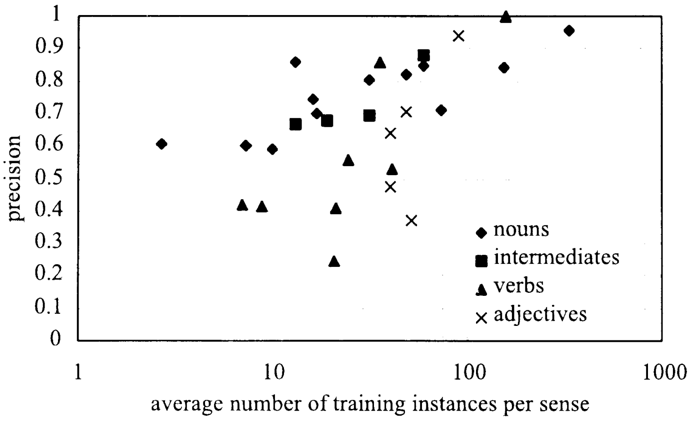

Figure 1. Correlation between the size of training data and system performance.

of the language dependent aspects. The time complexity of the algorithm for learning and applying the model is low[7] because the disambiguation process requires only a few string matching operations and lookups to the sets of classification information.

However, it is essential for our model that we overcome the data sparseness problem. For Korean polysemous words, we have already tried to relax the data sparseness problem by exploiting automatically constructed word class information. The precision was somewhat improved, but it was not remarkable because it has some difficulty in clustering words with low frequency. For future work, we will combine statistical and rule-based word clustering methods and also adopt similarity-based approaches to our model.

Notes

[1] Classification Information Model can be regarded as a kind of linear classifier because the right side of equation 2 is completely matched with that of a linear classifier. w_{ij} of linear classifer is generally learned by the least-mean-square algorithm. However, the Classification Information Model directly assigns w_{ij} with equation 3. According to Lee (1999), Classification Information Model makes decisions much faster on learning and somewhat more precisely than a linear classifier based on the least-mean-square algorithm for the data set used in Leacock et al. (1998).

[2] The MPC represents the Most Probable Class.

[3] The DS represents the Discrimination Score.

[4] According to Lee et al. (1997), the normalized conditional probability is useful for preventing the model from overemphasizing the imbalance of the size of training data set among classes.

[5] The features that did not occur in the training data were removed from the table.

[6] There was a mistake on the mapping from internal sense number to the official sense number in our system. The content of Table II was based on the result of revision on 16 October 1998.

[7] The time complexity for the learning algorithm is $O(mn)$, where m is the size of feature set and n is the number of senses. And, the time complexity for applying the algorithm is $O(n + log_2m)$ (Lee, 1999).

References

Leacock, C., M Chodorow and G. A. Miller. “Using Corpus Statistics and WordNet Relations for Sense Identification”. *Computational Linguistics*, 24(1) (1998), 147–165.

Lee, H., D.-H. Baek and H.-C. Rim. “Word Sense Disambiguation Based on The Information Theory”. In *Proceedings of Research on Computational Linguistics Conference*, 1997, pp. 49–58.

Lee, H. *A Classification Information Model for Word Sense Disambiguation*. Ph.D. thesis. The Department of Computer Science and Engineering (in Korean), Korea University, 1999.

Yarowsky, D. E. “Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French”. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 1994, pp. 88–95.

Yarowsky, D. E. *Three Machine Learning Algorithms for Lexical Ambiguity Resolution*. Ph.D. in Computer and Information Science, University of Pennsylvania, 1996.