Word Sense Disambiguation Based on Semi-automatically Constructed Collocation Dictionary

Minoru Sasaki, Kanako Komiya, Hiroyuki Shinnou
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
Email: {msasaki, kkomiya, shinnou}@mx.ibaraki.ac.jp

***Abstract*—In this paper, we propose a novel Word Sense Disambiguation (WSD) method based on collocation that has a particular meaning. This proposed method is to identify the sense of idiom or common phrase containing a target word before the existing statistical WSD method is applied by capturing the context information. To evaluate the efficiency of the proposed method using collocation dictionary, we make some experiments to compare with the result of the Support Vector Machine (SVM) classification. The results of the experiments show that almost the sense of the extracted collocation has only one particular sense when we obtain the word pair of (the target word, noun word) and (noun word, the target word) with high pointwise mutual information value. Moreover, in the experiment of WSD task, the total average precision of our system is improved in comparison with the baseline system using SVM.**

Keywords-word sense disambiguation; one sense per collocation; sense-tagged collocation dictionary construction

I. INTRODUCTION

Word Sense Disambiguation (WSD) [1] is one of the major tasks in natural language processing. WSD is the process of identifying the most appropriate sense for a polysemous word in a sentence. If we have training data which has already been disambiguated manually, the WSD task reduces to a classification problem based on supervised learning. In this approach, we construct a classifier to assign a word sense to new example by analyzing co-occurrence statistics of a target word. When we assign a sense to a word automatically, we can construct a sense tagged corpus and a case frame dictionary. To construct large-sized training data, language dictionary and thesaurus, it is increasingly important to further improve to select the most appropriate meaning of the ambiguous word.

WSD methods based on supervised learning exploit two powerful constraints: "one sense per collocation" [10] and "one sense per discourse" [3]. In the "one sense per collocation", the nearby words provide clues to the sense of the target word. "One sense per discourse" represents the sense that a target word is consistent with a given document. In the WSD research literature, currently, these two assumptions are widely accepted by natural language processing community and allow a supervised classifier with features based on context information to achieve enhanced classification performance.

Recent work develops above these assumptions into statistical models based on local and topical features surrounding a target word to be disambiguated [4] [7]. However, even when we make use of these assumptions, it is difficult to identify the sense of common expressions or idioms containing a target word. For example, the word "place" means general location. But, the meaning of the idiom "take place" is quite different from the meaning of "take her place". The idiom "take place" means that something occurs or happens at a particular time or place. Thus, an idiom is a group in a fixed order and has a particular meaning that is different from the meanings of the individual words regardless of context of the word to be disambiguated. Although there are many researches to solve WSD problem using phrase in WordNet and idiom dictionary, when we take into consideration the overall occurrence in the target corpus, there still remains some cases where a dictionary may not cover some of the idioms that exist in the target corpus.

In this paper, to solve this problem, we propose a novel word sense disambiguation method that aims to identify the sense of idiom and common phrase. In this method, we first extract idioms containing a target word and assign an appropriate sense to each of the extracted idioms manually to construct a idiom/collocation dictionary. Then, we identify the sense of idiom and common phrase before the existing statistical WSD method is applied by capturing the context information. Thus, this method enables us to identify the sense of a phrase that has a particular meaning regardless of context of the word such as metaphor expressions and idioms. A series of experiments shows our idiom sense identification effectively contributes to WSD precision.

The rest of this paper is organized as follows. Section 2 is devoted to the introduction of the related work in the literature. Section 3 describes a collocation dictionary generation method. Section 4 illustrates the proposed WSD system. In Section 5, we describe an outline of experiments. Experimental results are presented in Section 6. Finally, Section 7 concludes the paper.

II. RELATED WORKS

In this section, some previous research using such information will be compared with our proposed method.

Most WSD research has been focused on automatically assigning an appropriate sense to each occurrence of a target

 ISBN: 978-1-61208-355-1

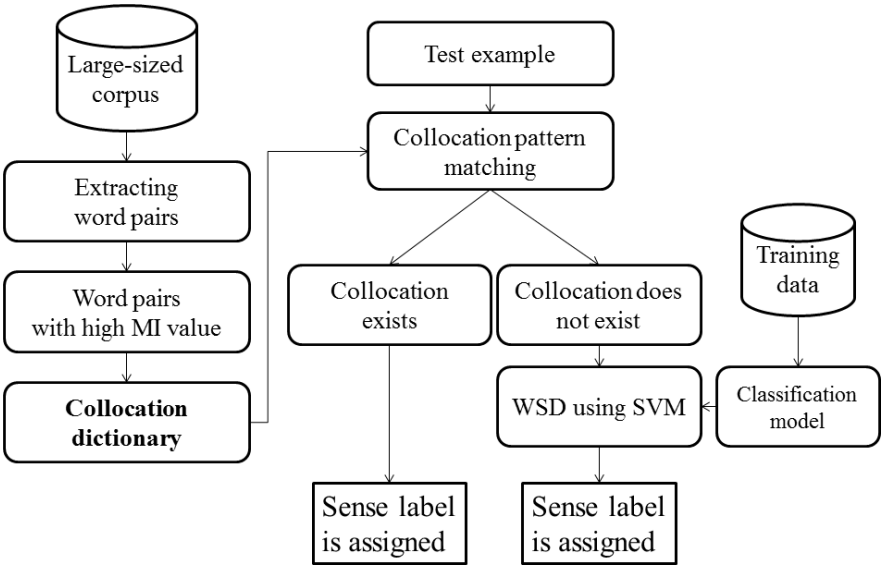

Figure 1. Overview of the proposed system

word in a text. In this research, many systems exploit two powerful constraints: "one sense per collocation" [10] and "one sense per discourse" [3]. Yarowsky's algorithm [10] employs an iterative bootstrapping approach. It starts from a small amount of seed collocation for the target word and assigning a sense using a decision list. The sense assignment process repeats until the whole corpus is consumed. Gale et al. [3] examines that there is a strong tendency for an ambiguous word to share the same sense in a well-written discourse.

In some previous research, collocation dictionary has been applied to the gloss disambiguation task. Yarowsky describes an unsupervised learning algorithm to perform WSD for unannotated English text. This method is to estimate the weighting using log-likelihood from the training set of data [11]. To identify an appropriate sense, it uses only nouns and considers only the two senses of a target word. However, in general, WSD task is a multi-class problem, as there can be more than two senses for a target word. Jimeno-Yepes et al. work on a knowledge-based WSD approach using collocation analysis [5]. This method extracts synonyms and collocations from meta-thesaurus to be added as alternative wordings of the target word. However, this system obtains related terms from the Unified Medical Language System meta-thesaurus [5], so that it does not take into consideration idioms and common expressions. There are some graph based approaches for knowledge based WSD, such as structural pattern recognition framework [8] and HyperLex [2].

III. Generating Collocation dictionary

In this section, we first describe the overview of generating collocation dictionary. From untagged corpora, we extract collocations of a given word in a semi-automatic manner. For more precise collocation data, the massive size of the untagged corpus is required. It is hard to get a large scale tagged corpus so that we use an untagged corpus for extracting collocations.

To extract collocations from large scale corpora, we explore the corpora to obtain the current and previous word pair, the current and next word pair, as well as the Part-Of-Speech tag of the previous and next words. We calculate the frequency of each word pair and use Pointwise Mutual Information (PMI) with each of the word pairs. The PMI is a popular measure of co-occurrence statistics of two words x and y in the data set as follows:

$$\mathrm{PMI}(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}, \quad (1)$$

where $P(x,y)$ is the probability of the word pair occurring together, $P(x)$ is the probability of the word x occurring and $P(y)$ is the probability of the word y occurring.

Then, we take all word pairs that exceed a certain threshold value of mutual information and consider them as collocation. Finally, we assign a sense tag to each of the extracted collocations manually to construct a collocation dictionary. Collocation has a particular meaning that is different from the meanings of the individual words regardless of context of the word to be disambiguated.

IV. Word Sense Disambiguation method Using Sense-tagged Collocation

In this section, we describe the details of the WSD classifier construction using sense-tagged collocation dictionary as mentioned in the previous section. The proposed method is composed of two stages that are WSD using the collocation dictionary and WSD using supervised learning. The overall system of the proposed method is illustrated in the Figure 1.

A. Word Sense Disambiguation using Collocation

Using the constructed collocation dictionary, we first learn the decision list (a set of rules) from the collocation dictionary to disambiguate collocation sense. For all examples of the test data, we explore collocation patterns in the decision list and

apply the decision list classifier. When the collocation patterns are found in the example, the set of rules is used to assign its corresponding sense to the collocation. However, if word pairs are not found in the decision list, no sense label is assigned for the target word in this stage.

B. Supervised Learning Using Support Vector Machine

For the sentences in which sense of the target word is not assigned throughout the test data at the first stage, we next use an implementation of a Support Vector Machine algorithm to train the classifier using context information and assign a particular sense to the target word at the second stage.

At the first step, we extract a set of features (nouns and verbs) that have co-occurred with the target word from each sentence in the training and test data. Then, each feature set is represented as a vector by calculating co-occurrence frequencies of the words. For each target word, we can obtain a matrix derived from the set of word co-occurrence vectors.

For the obtained matrix, classification model is constructed by using Support Vector Machine (SVM). When the classification model is obtained by training data, we predict one sense for each test example using this model. When a new sentence including the target word is given, the sense of the target word is classified to the most plausible sense based on the obtained classification model. To employ the SVM for distinguishing more than two senses, we use one-versus-rest binary classification approach for each sense.

V. Experiments

To evaluate the efficiency of the proposed method using collocation dictionary, we make some experiments to compare with the result of the SVM classification. In this section, we describe an outline of the experiments.

A. Data

To construct a collocation dictionary, we used the white papers and best-selling books in the BCCWJ corpus which is a balanced corpus of one hundred million words of contemporary written Japanese [6]. The document sets of white papers and best-selling books consist of 1,500 documents (16.4MB) and 1,408 documents (13.4MB) respectively.

To evaluate our WSD method, we used the Semeval-2010 Japanese WSD task data set, which includes 50 target words comprising 22 nouns, 23 verbs, and 5 adjectives from the BCCWJ corpus [9]. In this data set, there are 50 training and 50 test instances for each target word. When we apply the SVM to identify the sense of a target word, this training data of the target word is used to construct a classification model. The test data is used for evaluating the performance of the proposed WSD system.

B. Experiment on collocation extraction

In order to investigate the quality of the constructed collocation dictionary, we make some experiments using our collocation extraction method. To extract collocations, some conditions are to be fulfilled in each of the experiments. These conditions are summarized as follows:

TABLE I. PRECISION RATIO OF THE EXTRACTED COLLOCATION

PMI	1	2	3	4	5
Noun Only	0.975	0.979	0.988	0.980	0.923
All POS	0.787	0.770	0.765	0.789	0.842

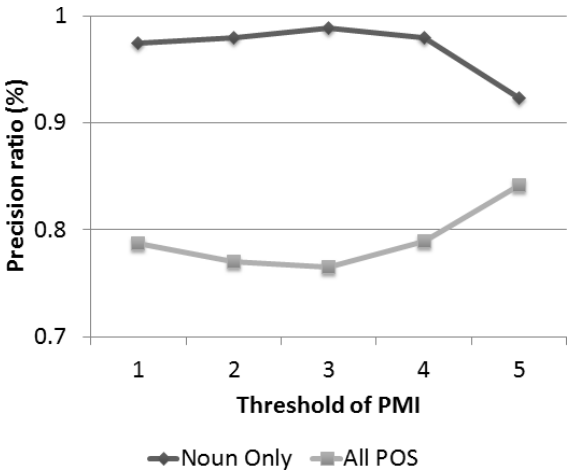

Figure 2. Precision ratio of the extracted collocation

- Part-Of-Speech(POS) of the previous and the next word (noun only or all POS)
- word pairs whose pointwise mutual information value is not less than the threshold k are considered as collocations ($k = 1,2,3,4,5$).

Under each of the above conditions, we construct a collocation dictionary and compare the quality of the extracted collocations. To evaluate the quality, we examine whether the sense of the extracted collocation has only one particular sense regardless of context of the target word. Then, we calculate the total number of correct collocations and the precision ratio of the number of collocations that have one particular sense to all extracted collocations for each target word. If the higher precision ratio is obtained, it turns out that the high quality collocation dictionary is constructed.

C. Experiment on WSD

To evaluate the results of the proposed method for the test data, we compare their performances with the results of simple SVM training. We obtain the precision value of each condition over all the examples to analyze the average performance of systems.

VI. Experimental Results

A. Quality of Collocation Extraction

Figure 2 and Table I show the result of the experiment of our collocation extraction method. In case that part-of-speech of both the previous and the next word is restricted to noun only, we obtain the high precision ratio. Therefore, almost the sense of the extracted collocation has only one particular sense, when we obtain the word pair of (the target word, noun

 ISBN: 978-1-61208-355-1

TABLE II. EXPERIMENTAL RESULTS OF WSD USING ADJACENCY NOUN

Accuracy	SVM	PMI=1	PMI=2	PMI=3	PMI=4	PMI=5
Ave.Prec.	0.690	0.704	0.696	0.694	0.693	0.691
Increase		17	13	9	6	2
Equal		31	35	41	44	48
Decrease		2	2	0	0	0

TABLE III. EXPERIMENTAL RESULTS OF WSD USING ALL ADJACENCY WORDS

Accuracy	SVM	PMI=1	PMI=2	PMI=3	PMI=4	PMI=5
Ave.Prec.	0.690	0.695	0.688	0.689	0.690	0.691
Increase		22	15	9	5	4
Equal		10	19	29	37	42
Decrease		18	16	12	8	4

word) and (noun word, the target word) with high PMI value. However, when the threshold value is 5, the precision value is decreased to 92.3%. The small number of the extracted collocation is obtained (197 collocations for $k = 1$ and 13 for $k = 5$) so that the precision ratio varies greatly.

In case that any part-of-speech is considered to the previous and the next word, the precision ratio is lower than the result using noun only. However, we obtain over 75% precision ration so that many word pairs have the potential to become collocation that has the particular sense.

B. Performance of WSD

Tables II and III show that the result of the experiment of WSD. In case that part-of-speech of both the previous and the next word is restricted to noun only, the total average precision of our system is improved in comparison with the baseline system using SVM. In the 50 target words, the precision of the only two words, "与える (ataeru; give, assign, ...)" and "経済 (keizai; economics, economy)", is decreased in comparison with the baseline system. These results are due to the failure to extract collocations that have a particular sense. However, if the threshold value k is larger than 3, the precision of our method has equal to the baseline system. In the data set used in these experiments, the number of training data is small so that many context words contained in the test data are not appeared in the training data. To improve the performance of the WSD system, we need to consider some additional information such as the glosses in WordNet and thesaurus.

In case that any part-of-speech is considered to the previous and the next word, the precision of our system is lower than that of the baseline. Using the threshold value $k = 1$, the precision of our system is higher. But, the precision of the 18 target words is decreased. Thus, the obtained collocation dictionary does not have good quality for disambiguating words, even though many collocations are extracted.

VII. Conclusion and Future Work

In this paper, we propose a novel word sense disambiguation method based on collocation that has a particular meaning. This proposed method is to identify the sense of idiom or common phrase containing a target word before the existing statistical WSD method is applied by capturing the context information. To evaluate the efficiency of the proposed method using collocation dictionary, we make some experiments to compare with the result of the SVM classification. The results of the experiments show that almost the sense of the extracted collocation has only one particular sense when we obtain the word pair of (the target word, noun word) and (noun word, the target word) with high PMI value. Moreover, in the experiment of WSD task, the total average precision of our system is improved in comparison with the baseline system using SVM. However, in case that any part-of-speech is considered to the previous and the next word, the precision of our system is lower than that of the baseline because the obtained collocation dictionary does not have good quality for disambiguating words.

Further work would be required to consider some additional information such as the glosses in WordNet, Wikipedia and other thesaurus to improve the performance of word sense disambiguation. Moreover, we need to consider a more syntactic information such as subject-verb-object relations and dependency structure to obtain more precise collocations.

References

[1] E. Agirre and P. Edmonds, *Word Sense Disambiguation: Algorithms and Applications*, 1st ed. Springer Publishing Company, Incorporated, 2007.

[2] E. Agirre, D. Martínez, O. L. de Lacalle, and A. Soroa, "Evaluating and optimizing the parameters of an unsupervised graph-based wsd algorithm," in *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, ser. TextGraphs-1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 89–96.

[3] W. A. Gale, K. W. Church, and D. Yarowsky, "One sense per discourse," in *Proceedings of the Workshop on Speech and Natural Language*, ser. HLT '91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 233–237.

[4] N. Ide and J. Véronis, "Word sense disambiguation: The state of the art," *Computational Linguistics*, vol. 24, pp. 1–40, 1998.

[5] A. Jimeno-Yepes, B. T. McInnes, and A. R. Aronson, "Collocation analysis for umls knowledge-based word sense disambiguation," *BMC Bioinformatics*, vol. 12, no. S-3, S4, 2011.

[6] K. Maekawa *et al.*, "Design, compilation, and preliminary analyses of balanced corpus of contemporary written japanese," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), May 2010, pp. 1483–1486.

[7] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, no. 2, pp. 10:1–10:69, Feb. 2009.

[8] R. Navigli and P. Velardi, "Structural semantic interconnections: A knowledge-based approach to word sense disambiguation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 7, pp. 1075–1086, Jul. 2005.

[9] M. Okumura, K. Shirai, K. Komiya, and H. Yokono, "Semeval-2010 task: Japanese wsd," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, ser. SemEval '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 69–74.

[10] D. Yarowsky, "One sense per collocation," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT '93. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, pp. 266–271.

[11] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '95. Stroudsburg, PA, USA: Association for Computational Linguistics, 1995, pp. 189–196.

 ISBN: 978-1-61208-355-1