Word Sense Disambiguation Using Active Learning with Pseudo Examples

Minoru Sasaki, Katsumune Terauchi, Kanako Komiya, Hiroyuki Shinnou
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
4-12-1, Nakanarusawa, Hitachi, Ibaraki, Japan
Email: {minoru.sasaki.01, 16nm717f, kanako.komiya.nlp, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

***Abstract*—In recent years, there have been attempts to apply active learning for Word sense disambiguation (WSD). This active learning technique selects the most informative unlabeled examples that were most difficult to disambiguate. The most commonly addressed problem has been the extraction of relevant information, where the system constructs a better classification model to identify the appropriate sense of the target word. Previous research reported that it is effective to create negative examples artificially (i.e., pseudo negative examples). However, this method works only for words that appear in a small number of topics (e.g., technical terms) because the evaluation set is strongly biased. For common noun or verb words, it is hard to apply this system so that problems still remain in the active learning with pseudo negative examples for WSD. In this paper, to solve this problem, we propose a novel WSD system based on active learning with pseudo examples for any words. This proposed method is to learn WSD models constructed from training corpus by adding pseudo examples during the active learning process. To evaluate the effectiveness of the proposed method, we perform some experiments to compare it with the result of the previous methods. The results of the experiments show that the proposed method achieves the highest precision of all systems and can extract more effective pseudo examples for WSD.**

Keywords-word sense disambiguation; active learning; uncertainty sampling; pseudo examples; reliable confidence score.

I. INTRODUCTION

Word sense disambiguation (WSD) is one of the major tasks in natural language processing (NLP). WSD is the process of removing ambiguities and identifying the most appropriate sense for a polysemous word in context. This technique is crucial in many application in other areas of NLP, such as machine translation [16], information retrieval [17], question answering [1], information extraction [2], text summarization [13], and so on.

One of the successful approaches to WSD is based on applying corpus-based learning [6] [11]. In this approach, machine learning (ML) or statistical algorithms have been applied to learn classifiers from corpora in order to perform WSD. WSD approaches based on the ML are classified into two categories, supervised and unsupervised approaches. The supervised learning method is used to learn the rules that correctly classify documents for a given classification algorithm. The unsupervised learning method is used to cluster word contexts into some sets which indicate the same meaning. A variety of techniques for supervised learning algorithms have demonstrated good performance for WSD, when we have enough labeled training data for learning. However, the supervised WSD methods require a large sense-tagged corpus which is expensive to obtain by manual annotation.

In recent years, there have been attempts to apply active learning for WSD [3] [18]. This active learning technique selects the most informative unlabeled examples that were most difficult to disambiguate. Previous research reported that the active learning methods with supervised learning methods could effectively reduce the amount of human labeling effort and can be helpful to improve the WSD models. For example, in the previous research [14], it is realized by automatically extracting pseudo negative examples that have reliable confidence score from unlabeled examples for WSD in Web mining. This method achieves high accuracy compared to the method with manually extracted negative examples for World Wide Web data. However, this method works only for words that appear in a small number of topics (e.g., technical terms) because the evaluation set is strongly biased. For common noun or verb words, it is hard to apply this system so that problems still remain in the active learning with pseudo negative examples for WSD.

In this paper, to solve this problem, we propose a novel WSD system based on active learning with pseudo examples for any words. This proposed method aims at learning WSD models constructed from training corpus by adding pseudo examples during the active learning process. The contribution of our work is three-fold,

- By using active learning with pseudo examples, the proposed WSD system can compute the effective semantic distribution of each sense of words.
- The proposed WSD system can be effective for common noun or verb words by using a new calculation method of confidence score.
- The proposed WSD system adopts support vector machine (SVM) as classifier, which can extract more effective pseudo examples than the previous system.

A series of experiments shows that our method effectively contributes to WSD precision.

The rest of this paper is organized as follows. Section Ⅱ is devoted to the introduction of the related work in the literature. Section Ⅲ describes the proposed WSD system based on active

 ISBN: 978-1-61208-507-4

learning with pseudo examples. In Section IV, we describe an outline of experiments and experimental results. Finally, Section V concludes the paper.

II. Related Works

This paper proposes a WSD method using active learning with pseudo examples. In this section, some previous research using active learning for WSD will be compared with our proposed method.

Active learning is the study of machine learning systems that select the data from the data pool and get the labeled data to reduce the amount of labeling efforts. One intuitive approach in pool-based active learning is called uncertainty sampling [10]. This approach selects example for which the classifier is most uncertain. Chan and Ng (2007) propose to combine active learning method with domain adaptation for word sense disambiguation system [3]. This method estimates the reliable confidence score with the prior probability of the target sense to select the most informative data, whereas our method does not consider the prior probability of sense to calculate the reliable confidence score.

Zhu and Hovy (2007) analyzed the effect of resampling techniques, under-sampling and over-sampling with active learning for the WSD imbalanced learning problem [18]. This method uses labeled training data set that includes the positive and negative examples as the input. However, our method of active learning starts with the small positive examples and the pool of unlabeled examples.

Takayama et al. (2009) propose a method of active learning to artificially create negative examples (i.e., pseudo negative examples). This method achieves high accuracy compared to the method with manually extracted negative examples for search results. Our method artificially creates positive and negative examples (i.e., pseudo examples) to improve the WSD performance for common noun or verb words.

III. Active Learning Method with Pseudo Examples for WSD

In this section, we describe the details of the proposed WSD system based on active learning with pseudo examples. The proposed method employs uncertainty sampling active learning strategy.

A. Classifier

In our experiment, to classify the sense label to unlabeled example, we use support vector machine (SVM) as classifier [5]. The SVM is one of the most popular machine learning algorithms. The SVM computes a hyperplane with the largest margin separating the training examples into two classes. A test example is classified depending on the side of the hyperplane. In order to deal with the multi-class problem, we can reduce this problem to a set of binary classification problems by using one-versus-one [7] or one-versus-rest [15] strategy. Therefore, SVM has been successfully applied to many natural language processing problems.

```
1  function Active-Learning-with-Pseudo-Examples (D, s, S, k);
   Input : Data set with positive training examples and
           unlabeled examples D; Sense label of the target
           word s; Set of sense labels S; Total number of
           labeled examples that are required k
   Output: Labeled training data set L
2  P ← training examples with label s;
3  N ← {};
4  PP ← {};
5  PN ← {};
6  repeat
7      foreach d in D-P-N do
8          c(d,s) ← reliable confidence score for d that has
             the sense s;
9          c(d,s̄) ← reliable confidence score for d that
             doesn't have the sense s;
10         diff ← c(d,s̄) − c(d,s);
11         if diff ≥ τ then
12             if s = arg max_{s_i∈S} c(d,s_i) then
13                 PP ← PP ∪ d;
14             else
15                 PN ← PN ∪ d;
16             end
17         end
18     end
19     Construct classifier M using (P+PP, N+PN);
20     c_min ← ∞;
21     foreach d in D-P-N do
22         s′ ← sense label that d is classified into, using
             the WSD model M;
23         c(d,s′) ← reliable confidence score for d that has
             the sense s′;
24         if c(d,s′) < c_min then
25             c_min ← c(d,s′), d_min ← d;
26         end
27     end
28     s_m ← sense label that is manually annotated to d_min;
29     if s_m = s then
30         P ← P ∪ d_min;
31     else
32         N ← N ∪ d_min;
33     end
34 until the number of labeled examples is equal to k;
```

Figure 1. Algorithm of active learning with pseudo examples

To perform our sense label classifier, we convert an example to features. In this paper, we use the following five types of features.

f_1 : Content words (noun, verb, adverb) in the sentence that the target word appears(i.e., current sentence) and also in the previous and the next sentence.

f_2 : The previous content word of target word in the same Japanese phrasal unit (bunsetsu).
f_3 : The next content word of target word in the same Japanese phrasal unit.
f_4 : Unit phrase that depends on the unit the target word appears.
f_5 : Unit phrase the target word depends on.

These above features were used in previous research [14]. This research reports that the WSD system using these features gives good results. In this paper, to compare with the previous method, we employ the same five types of features in our experiments.

B. Active Learning Method

We describe the proposed active learning method for WSD. This proposed method is based on uncertainty sampling that selects unlabeled examples that were most difficult to disambiguate. By using this method, we can construct a better classifier for active learning because we can obtain pseudo negative examples with high confidence and pseudo positive examples that are near a decision boundary of SVM.

Algorithm 1 shows the proposed active learning method with pseudo examples. This active learning function receives four inputs D, s, S and k. D is a data set with positive training examples and unlabeled examples: s is a sense label of the target word. S is a set of sense labels of the target word and k is the total number of labeled examples that are obtained by active learning.

Firstly, the proposed method generates pseudo examples to construct a classifier M. For each unlabeled example d in $D-P-N$, reliable confidence scores $c(d,s)$ for the sense s and $c(d,\bar{s})$ for the other sense $\bar{s}$ are calculated using the following formula:

$$c(d,s) = \sum_{j=1}^{5} \log p(f_j|s). \quad (1)$$

In the previous research [3] [14], the reliable confidence score is calculated using a different formula, as follows:

$$c(d,s) = \log p(s) \sum_{j=1}^{5} \log p(f_j|s). \quad (2)$$

Here, $p(s)$ represents the prior probability of the sense s. In the experiments from the [14], target words are almost proper nouns such as product name and personal name so that the prior probability $p(s)$ of each word is effective for the reliable confidence score. However, when we use general words such as common noun or verb words as the target word, the prior probability $p(s)$ is not so effective for the reliable confidence score. This reason is that the prior probability $p(s)$ of proper nouns in search result document set is heavily biased. Therefore, we use (2) to calculate the reliable confidence score using the same value of the prior probability value.

For the obtained two reliable confidence scores $c(d,s)$ and $c(d,\bar{s})$, the difference of these scores $diff$ is calculated. When the $diff$ value is not less than the threshold value τ, the example d is added to the pseudo positive example set PP if the sense with the highest reliable confidence score is equal to the sense label s, otherwise the example d is added to the pseudo negative example set PN. If the sense with the highest reliable confidence score is the positive sense label s, the example d is likely to be positive example that is near a decision boundary. It is important to obtain such examples to construct a better decision boundary, If the $diff$ value of the another sense label is the highest, the example d is an almost negative example. In this experiment, the threshold parameter τ which predicted a significant difference between the target sense and the other is set to be 1.0.

Next, we construct the sense label classifier M using SVM from the training set with the pseudo examples $(P+PP, N+PN)$. We use LIBSVM as the implementation of the SVM for our experiments [4]. In the previous research [14], naive bayes classifier is used to develop the classifier. However, we obtain high classification precision using SVM so that we use the SVM as the classifier. For each unlabeled example d in $D-P-N$, d is classified into the sense s' by using the classifier M. Then we calculate the reliable confidence score $c(d,s')$ and extract the example d_{min} that minimize the reliable confidence score $c(d,s')$. For the obtained example d_{min}, sense label s_m is provided manually and the example d_{min} is added to the positive example set P if the sense s_m is equal to the sense label s, otherwise the example d_{min} is added to the negative example set N.

This process is repeated until the number of labeled examples is equal to k. In this experiment, k is set to be 50.

IV. Experiments

To evaluate the effectiveness of the proposed method of active learning with pseudo examples for WSD, we perform some experiments and compare the results of the previous method. In this section, we describe an outline of the experiments.

A. Data

To evaluate our active learning 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 [12]. In this data set, there are 50 training and 50 test examples for each target word. One example in the training and test set is the sentence where the target word appears in.

In the experiments of this paper, we use randomly selected 10 words (2 nouns and 8 verbs) in the Semeval-2010 Japanese WSD task data set. For each sense of the target word, as the input data of the system, we use some labeled data that were randomly selected from the training examples and the other examples in the data set as unlabeled data. For the input data, the system extracts the previous and next sentences of each example and extracts noun, verb and adverb words from these sentences by using Japanese morphological analysis tool MeCab [9] to obtain the features f_1, f_2 and f_3. Moreover, the system uses the dependency analysis tool Cabocha [8] to obtain the features f_4 and f_5. Table I shows the number of the

 ISBN: 978-1-61208-507-4

initial training examples for each sense of the target word and Table II shows the number of test examples for each sense of the target word, where the $s_i(i = 1, \cdots, 7)$ indicates the *i*-th sense of the word in the Iwanami Japanese Dictionary.

TABLE I: The number of the initial training examples for each sense of the target word

Words	s1	s2	s3	s4	s5	s6	s7
ageru	5	5	2	5	5	1	1
ataeru	5	5	5	-	-	-	-
imi	5	5	5	-	-	-	-
kodomo	5	5	-	-	-	-	-
suru	5	5	2	2	3	-	-
dasu	5	5	3	1	-	-	-
deru	5	5	2	-	-	-	-
toru	3	5	4	5	3	1	1
noru	5	2	4	5	-	-	-
motsu	5	5	2	-	-	-	-

TABLE II: The number of test examples for each sense of the target word

Words	s1	s2	s3	s4	s5	s6	s7
ageru	10	10	4	10	10	2	2
ataeru	10	10	10	-	-	-	-
imi	10	10	10	-	-	-	-
kodomo	10	10	-	-	-	-	-
suru	10	10	4	3	5	-	-
dasu	10	10	5	2	-	-	-
deru	10	10	5	-	-	-	-
toru	7	10	8	10	7	2	2
noru	9	5	9	10	-	-	-
motsu	10	10	2	-	-	-	-

B. Experiment on Active Learning for WSD

To evaluate the results of the proposed method for the test examples, we compare the four systems as follows:

System 1:
Active learning with pseudo negative examples using naive bayes classifier and the original reliable confidence score in the equation (2) (baseline).

System 2:
Active learning with pseudo negative examples using SVM classifier and the proposed reliable confidence score in the equation (1).

System 3:
Active learning with pseudo examples using SVM classifier and the original reliable confidence score in the equation (2).

System 4:
Active learning with pseudo examples using SVM classifier and the proposed reliable confidence score in the equation (1) (proposed method).

We obtain the precision value of each system and analyze the average performance of systems.

C. Experimental Results

In this section, we present the experimental results on the WSD system using active learning with pseudo examples. Table III shows the precision for each of the target words by using each WSD system.

TABLE III: Precision of the each WSD system for target words

Words	System1	System2	System3	System4
ageru	14.6%	**44.8%**	14.6%	33.3%
ataeru	40.0%	56.7%	40.0%	**60.0%**
imi	26.7%	53.3%	26.7%	**60.0%**
kodomo	5.0%	60.0%	5.0%	**95.0%**
suru	15.6%	18,8%	15.6%	**59.4%**
dasu	7.4%	14.8%	7.4%	**77.8%**
deru	40.0%	20.0%	40.0%	**60.0%**
toru	6.5%	15.2%	6.5%	**58.7%**
noru	17.6%	**64.7%**	20.6%	47.1%
motsu	36.4%	27.3%	36.4%	**50.0%**

As shown in the Table III, the proposed method achieves the highest precision of all systems for the eight target words. For the target word ”kodomo (子供; child)”, although the precision of the system 1 and 3 is very low, the precision is 95% by using the proposed system. For the word ”suru (する; do, play ...)”, ”dasu (出す; put out, appear, ...)” and ”toru (取る; take, catch, ...)”, despite these words have many senses, the proposed system obtains the highest precision. Therefore, the proposed active learning method can extract more effective pseudo examples for WSD.

The baseline system (system 1) and system 3 give almost the same results in this WSD experiment. Using these systems, precision of WSD is low in comparison with the proposed systems. Hence, these systems are not effective to estimate an appropriate word sense for common words. Moreover, these results show that it is not so effective to append pseudo examples to the training data by using the reliable confidence score with the prior probability.

System 2 obtains higher precision than the system using the reliable confidence score with the prior probability for the eight target words (except for the words ”deru” and ”motsu”). For the target words ”ageru (あげる; give, get up, ...)” and ”noru (のる; ride, go into gear, ...)”, this system also obtains higher precision than the proposed system. However, for the other target words, the precision is less than 50% so that system 2 did not obtain high precision. This result shows that it is effective for WSD system to append pseudo examples to the training data by using the reliable confidence score without the prior probability.

V. CONCLUSION

In this paper, we propose a novel WSD system based on active learning with pseudo examples for any words. This proposed method is to learn WSD models constructed from training corpus by adding pseudo examples during the active learning process. To evaluate the effectiveness of the proposed active learning method, we perform some experiments and compare the results with the results of the previous method. The results of the experiments show that the proposed WSD system can be effective for common noun or verb words by using a new calculation method of confidence score. Moreover, the proposed method achieves the highest precision of all

 ISBN: 978-1-61208-507-4

systems and can extract more effective pseudo examples for WSD. However, by using the reliable confidence score with the prior probability, it is not so effective to append pseudo examples to the training data. Therefore, the proposed WSD system can compute the effective semantic distribution of each sense of words.

Further work would be required to consider some additional features such as thesaurus information and add more unlabeled data to obtain more meaningful examples by active learning to improve the performance of word sense disambiguation.

REFERENCES

[1] S. Beale, B. Lavoie, M. McShane, S. Nirenburg, and T. Korelsky, "Question answering using ontological semantics," in *Proceedings of the 2Nd Workshop on Text Meaning and Interpretation*, ser. TextMean '04. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 41–48, 2004.

[2] J. Y. Chai and A. W. Biermann, "The use of word sense disambiguation in an information extraction system," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, ser. AAAI '99/IAAI '99. Menlo Park, CA, USA: American Association for Artificial Intelligence, pp. 850–855, 1999.

[3] Y. S. Chan and H. T. Ng, "Domain adaptation with active learning for word sense disambiguation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: In Proceedings of Association for Computational Linguistics, pp. 49–56, June 2007.

[4] C.-C. Chang and C.-J. Lin, "Libsvm – a library for support vector machines," https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

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

[7] U. Kreßel, "Pairwise classification and support vector machines," *in Advances in Kernel Methods - Support Vector Learning*, 1999.

[8] T. Kudo, "Cabocha: Yet another japanese dependency structure analyzer," http://taku910.github.io/cabocha/.

[9] T. Kudo, "Mecab: Yet another part-of-speech and morphological analyzer," http://taku910.github.io/mecab/.

[10] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '94. New York, NY, USA: Springer-Verlag New York, Inc., pp. 3–12, 1994.

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

[12] 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, pp. 69–74, 2010.

[13] M. Pourvali and M. S. Abadeh, "Automated text summarization base on lexicales chain and graph using of wordnet and wikipedia knowledge base," *International Journal of Computer Science Issues*, vol. abs/1203.3586, 2012.

[14] Y. Takayama, M. Imamura, N. Kaji, M. Toyoda, and M. Kitsuregawa, "Active learning with pseudo negative examples for word sense disambiguation in web mining," *IPSJ TOD*, vol. 2, no. 2, pp. 1–9, jun 2009.

[15] V. N. Vapnik, *Statistical learning theory*, 1st ed. Wiley, Sep. 1998.

[16] D. Vickrey, L. Biewald, M. Teyssier, and D. Koller, "Word-sense disambiguation for machine translation," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 771–778, 2005.

[17] Z. Zhong and H. T. Ng, "Word sense disambiguation improves information retrieval," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ser. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 273–282, 2012.

[18] J. Zhu, "Active learning for word sense disambiguation with methods for addressing the class imbalance problem," in *In Proceedings of Association for Computational Linguistics*, pp. 783–790, 2007.

 ISBN: 978-1-61208-507-4