Word Sense Disambiguation by Combining Labeled Data Expansion and Semi-Supervised Learning Method

Sanae Fujita
NTT Communication Science Lab.
fujita.sanae@lab.ntt.co.jp

Akinori Fujino
NTT Communication Science Lab.
fujino.akinori@lab.ntt.co.jp

Abstract

Lack of training data is one of the severest problems facing word sense disambiguation (WSD). To overcome the problem, we propose a method that combines automatic labeled data expansion (Step-1) and semi-supervised learning (Step-2). The Step-1 and 2 methods are both effective, but their combination has a synergistic effect.

In this paper, in Step-1, we automatically extract reliable labeled data from raw corpora using dictionary example sentences, even for infrequent and unseen senses (which do not appear in training data, but appear in a dictionary). Then, in Step-2, we apply a semi-supervised classifier and obtain a improvement using easy-to-get unlabeled data. In this step, we also show that we can guess even unseen senses.

We target a SemEval-2010 Japanese lexical sample WSD task. Both the Step-1 and Step-2 methods performed better than the best published result (76.4 %). Furthermore, the combined method achieved much higher accuracy (84.2 %).

1 Introduction

Many words have multiple meanings that change depending on the context. Recently, it has been confirmed that word sense disambiguation (WSD) improves certain NLP applications such as parse selection (Fujita et al., 2007) or Machine Translation (Chan et al., 2007). In international WSD competitions such as SemEval, many tasks have been proposed, which shows that WSD is a problem that attracts a lot of interest. In this paper, we experiment on the Japanese WSD task from the most recent competition, SemEval-2010 (Okumura et al., 2010).

Various methods have been proposed for WSD (Navigli, 2009). Unsupervised approaches such as clustering based methods (Pedersen, 2006) and extended Lesk (Lesk, 1986) have been shown to do well (Baldwin et al., 2010), although in general, they are beaten by supervised approaches if training data are provided (Tanaka et al., 2007). With the Japanese WSD tasks at SENSEVAL-2 and SemEval-2010, supervised approaches achieved the best results (Murata et al., 2003; Okumura et al., 2010). However, the lack of training data is a severe problem with non-English languages. Also in the Japanese WSD task, there are only 50 given training instances for each target word and this is insufficient.

Two main types of methods have been proposed to compensate for a lack of training data. One type is the semi-supervised learning method (Niu et al., 2005; Pham et al., 2005) or bootstrapping (Mihalcea, 2002, 2004; Yarowsky, 1995). These methods use labeled data and unlabeled data, and this is beneficial because unlabeled training data is easy to obtain. These methods are effective and have high applicability, but unfortunately, there is one problem in that this method cannot obtain training data for senses in the lexicon that do not appear in the training data (we call this an **unseen sense**).

Another type of method designed to make up for a lack of training data, is automatic labeled data expansion (Mihalcea and Moldovan, 1999; Agirre and Martinez, 2000) . They proposed expanding the amount of labeled data through a Web search using monosemous synonyms or unique expressions in definitions from WordNet (Fellbaum, 1998). These methods are also effective, and may be able to obtain labeled data even for unseen senses. But one expected problem will be that the performance is influenced by a sense bias (that is sense frequency) that varies with corpora (Agirre and Martinez, 2004).

Therefore, in this paper, we propose a method

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 676–685,
Chiang Mai, Thailand, November 8 – 13, 2011.

ID 37713	Headword とる【取る・採る・執る・捕る】 *toru "take/pick/collect/do/catch"*
0-0-1-0	置いてあったものなどを手に持つ。*"to get something left into one's hand."*
0-0-1-1	手で握り持つ。*"take and hold by hand."*「手に取って見る」*"pick up and see"*
0-0-1-2	手に持ってそれを使って仕事をする。*"hold something in one's hand, and work with it."* 「筆を取る」*"start writing"*
...	...
0-0-8-0	直接に手がけてする意を添える。*"add emphasis of undertaking some action directly."* 「式を取り行う」*"perform a ceremony"*

Figure 1: Simplified Entry from Iwanami Dictionary: とる *toru* "take"

that combines automatic labeled data expansion and semi-supervised learning aiming a synergistic effect. That is, we propose a two-step approach: in the first step (Step-1), we automatically expand the labeled training data from raw corpora. In this step, we aim to expand the labeled training data even for unseen and infrequent senses.

Then, in the second step (Step-2), we apply a semi-supervised classifier. In this step, we aim to achieve on improvement using easy-to-get unlabeled data. In this step, we also compare the results obtained using given training data only and show the benefits of our combination method. We also show its effectiveness for unseen senses.

This paper is structured as follows. We describe the target task in § 2. We describe our automatic labeled data expansion method (Step-1) in § 3, the evaluate the data quality using an experiment and human evaluation in § 4. Then we apply a semi-supervised learning method (Step-2) in § 5. We conclude the paper in § 6.

2 SemEval-2010: Japanese WSD

In this paper, we experiment on the SemEval-2010 Japanese WSD task. The sense inventory used in this task was the Iwanami Japanese Dictionary (Nishio et al. (1994)). Iwanami was originally paper dictionary. We show an example entry for Iwanami in Figure 1. As shown in Figure 1, each entry in Iwanami has POS information and definition sentences, and most of entries have example sentences. Iwanami has four hierarchical layers in word sense descriptions. In this task, senses at the third layer are used at the evaluation phase. For example, 0-0-1 and 0-0-8 in Figure 1. Iwanami includes 60,321 entries split into 85,870 senses, which are merged into 79,611 senses at the third layer. For this task, 50 words (22 nouns, 23 verbs, and 5 adjectives) are selected as the targets, which are split into 219 senses at the third layer; of these, 144 senses appear in the training data. On the other hand, 9 senses are unseen senses that appear in both Iwanami and the test data, but do not appear in the training data.

Both the training and test data are part of the Balanced Corpus of Contemporary Japanese: BCCWJ[1], which is morphologically analyzed by UniDic[2] and hand-corrected. For each target word, 50 instances are provided in both the training and test data. We show an example of the given training data in (1). The given training data are morphologically analyzed, but have no information about the base forms, therefore we added the base forms (lemma) automatically. The given data are also partly tagged with sense IDs of Iwanami.

(1) <mor pos=' 動詞-一般' rd=' トッ' bfm=' トル' sense='37713-0-0-1-1' lemma=' 取る'> 取っ</mor>

One feature of this task is that the training and test data come from heterogeneous corpora. The training data include books or magazines (PB), newspaper articles (PN), and government white papers (OW). The test data also include documents from a Q&A site on the WWW (OC). However, in this paper, we do not focus on domain adaptation, because 50 instances are insufficient, especially for investigating domain adaptation as reported by Fujita et al. (2010) and Shirai and Nakamura (2010).

3 Method for Automatic Labeled Data Expansion: Step-1

In this section, we introduce our automatic labeled data expansion method (Step-1). The main aim of this step is to obtain reliable labeled data even for unseen and infrequent senses.

[1] http://www.ninjal.ac.jp/kotonoha/
[2] http://www.tokuteicorpus.jp/dist/

As mentioned in § 1, several labeled data expansion methods have been proposed such as Mihalcea and Moldovan (1999) and Agirre and Martinez (2000, 2004). They mainly used WordNet's monosemous synsets (for example,"*recollect*" for *remember*$_1$) for Web search. This kind of method is effective, and offers the possibility of supplying training data for unseen senses. But unfortunately, we cannot obtain monosemous synonyms because our target task is not tagged with WordNet.

Therefore, in this paper, instead of these methods, we propose a method that provides reliable training data using example sentences from a dictionary. Such sentences are informative, but in most case of paper dictionaries such as Iwanami, the examples are fragmentary to save spaces (in Iwanami, an average of 4 words). Therefore, we attempt to extract longer, more natural and high quality labeled data from the raw corpus, under strict conditions using fragmentary examples.

That is, first, we extract example sentences (EX) from Iwanami. Then, we collect sentences that include an exact match for Iwanami's example for sense (s_k) of headword (h). Finally, we morphologically analyze the candidate sentences, and if both the base form and the coarse POS correspond to those of h, we tag the words with s_k and add the sentences to the labeled data.

For example, we can extract an example sentence as in (2), from 37713-0-0-1-2 in Figure 1. In (2), the headword is in **boldface** and is tagged with '37713-0-0-1-2' (at the third layer, '37713-0-0-1').

(2) 筆　を　**取る**
pens ACC pick up
"(I) start writing"

The data used in the Japanese WSD task is part of the BCCWJ corpus. Therefore, we use the remainder of the BCCWJ to extract the training data. Note that its morphological information is not hand-corrected. According to the readme file that comes with the BCCWJ, it includes about 43 million words.

We show an example of extracted labeled data in (3). The underlined part is exactly the same as the example sentence in (2). Therefore we tag 取る *toru* "pick up/take" with 37713-0-0-1-2 and add this entire sentence to the labeled training data.

(3) 筆　を　**取る**　気　は　起ら
pens ACC pick up feel TOP become
なかっ　た
not did
"(I) did not get into start writing"

Because of our strict condition, the size and variation of the extracted labeled data are limited. However, this method gave us longer and more natural reliable labeled data. Besides, most languages have dictionaries, and most of these dictionaries include examples, we expect our method to be applicable to other languages and dictionaries.

We show the size of the extracted and given training data (Trn) in Table 1. As shown in Table 1, the extracted labeled data give less coverage of sense types to the given training data, but give them many more instances. On average, 130 labeled sentences were extracted per example, for 326 example sentences. And training instances for 9 unseen senses were extracted from Iwanami's EX, and 5 unseen senses were extracted from BCCWJ.

Corpus	Sense Types All	Sense Types Unseen	Instances All	Instances Unseen
Trn	144	-	2,500	-
EX	156	9	1,450	46
BCCWJ	114	5	42,430	94

Table 1: Size of extracted and given training data

4 Evaluation of our Labeled Data Expansion Method: Step-1

In this section, we investigate the reliability and effectiveness of our automatic labeled data expansion. For this purpose, in § 4.1, we investigate the performance over the Japanese WSD task when we apply the supervised learning approach with and without the extracted labeled data. Then in § 4.2, we also provide a quantitative analysis of the extracted training data.

4.1 Performance over Japanese WSD Task

4.1.1 System Description

Machine Learning Methods We constructed supervised and semi-supervised WSD classifiers for each target word, based on machine learning methods. The classifiers for a target word were designed to select a sense from pre-defined senses for an instance of the target word.

In Step-1, we employed a *Maximum Entropy Model (*MEM*)* (Nigam et al., 1999) to design the

supervised WSD classifier. Let x denote the feature vector for an instance of a target word and $s \in \{s_1, \ldots, s_k, \ldots, s_K\}$ denote a sense of the target word. For the supervised WSD classifier, for the target words, the conditional probability of s given x is modeled as

$$P(s_k|x;W) = \frac{\exp(w_k^T x)}{\sum_{k'=1}^{K} \exp(w_{k'}^T x)}, \forall k, \quad (4)$$

where $W = [w_1, \ldots, w_k, \ldots, w_K]$ is a parameter matrix and w_k^T represents the transposed vector of w_k. We estimated the parameter matrix value by using labeled data.

Features For each target word w, we used the surface form, the base form, the POS tag, and the coarse POS categories, such as nouns, verbs, and adjectives of w. Then we also used bag-of-words in the same sentence. Here the target is the ith word, so we also used the same information for the $i-2, i-1, i+1$, and $i+2$nd words. We used bigrams, trigrams, and skipbigrams back and forth within three words. And we also used domain type `PB`, `PN`, `OW`, and `OC` as features.

Analytical Setting One anticipated problem with our expanding method in Step-1 is that the extracted data may have a different sense distribution from test data. Therefore, to investigate trends based on sense distribution, we employed the entropy $E(w)$ of the frequency distribution in given training data, which is given by

$$E(w) = -\sum_{k=1}^{K} p(s_k|w) \log p(s_k|w), \quad (5)$$

where $p(s_k|w)$ is the probability that word w will be sense s_k. In other words, $p(s_k|w)$ and then $E(w)$ reflect sense frequency bias. Note that $p(s_k|w)$ differs from $P(s_k|x)$, which is the probability of sense given *each instance* of the target word.

The entropy $E(w)$ will be lower if one paticular sense appears more frequently. Therefore, following the SENSEVAL-2 Japanese WSD task (Shirai, 2003), we divided the target words into three classes: difficult (D_{diff}: $E(w) \geq 1$), middle (D_{mid}: $0.5 \leq E(w) < 1$), and easy (D_{easy}: $E(w) < 0.5$). There were 9 target words for D_{diff}, 20 for D_{mid}, and 21 for D_{easy}.

4.1.2 Results and Discussion

Learning Curves Because of our strict condition, the variation in the extracted labeled data is limited, therefore, the system may cause overlearning. So first, we investigate the learning curves by limiting the number of added training instances for each `Iwanami` example.

Table 2 shows average accuracies over target words obtained with various number of added labeled instances per example. `L#` in the table shows the upper-bound for adding labeled instances per example. We used all extracted labeled instances when the number was less than `#`.

We adjusted the parameters based on a 5-fold cross-validation of the given training data. The best result (RALI-2, Brosseau-Villeneuve et al. (2010)) in a formal run of SemEval-2010 is also shown in Table 2 for reference, and the system uses the most frequent sense as a baseline.

	D_{easy}	D_{mid}	D_{diff}	Total
Base Line	**91.5**	67.0	51.3	69.0
RALI-2	-	-	-	76.4
`Trn` (No expansion)	90.6	70.9	61.1	77.4
+ `L1`	90.7	72.0	64.7	78.5†
+ `L5`	**90.8**	72.1	65.8	78.8†
+ `L10`	90.7	72.9	67.3	79.4 †
+ `L30`	90.7	73.8	68.2	**79.9**†
+ `L50`	90.5	**74.0**	68.0	79.8†
+ `L100`	90.3	73.4	68.2	79.6†
+ `L300`	90.4	73.5	68.2	79.6†
+ All extracted instances	89.7	73.6	**68.9**	79.5
+ `EX`	89.8	72.2	64.9	78.3
+ `L30` + `EX`	90.1	73.3	68.4	79.5
+ `L30` + `EXrL`	90.5	**74.8**	**68.4**	**80.2**†

Table 2: Results for given training data (`Trn`) or with extracted labeled data (by MEM) : where † shows that there is significant improvement over `Trn` by t-test at 5 % level of significance.

Using only the given training data (we call this `Trn`, 77.4 %), we achieved an improvement over the best published method (76.4 %). Even only one labeled instance per example gave better results (78.5 %) than the given training data alone, and 30 labeled instances gave the best result (79.9 %) in total[3].

Table 2 shows the accuracy per entropy based difficulty band. We found an interesting trend, namely that expanding the training data tended

[3]All results except "+All extracted instances" significantly improved over `Trn`.

to degrade the accuracy for easy words (D_{easy}), but improved it for the middle (D_{mid}) and difficult words (D_{diff}). With easy words, the best result (91.5 %) was provided by the selection of the most frequent sense. On the other hand, especially for difficult words, expansion was very effective, and using All extracted instances gave a +7.8 % ($= 68.9 - 61.1$) improvement over Trn. Difficult word means that many more senses appear in corpus. In other words, more instances are needed to guess the sense correctly, that is the reason for our method is especially effective for such difficult words.

Adding Original Example Sentences Then, we investigated other conditions. As shown in § 3, in Iwanami, there are 1,450 original example sentences, such as in a sentence (2), for the target words. However we could use only 362 example sentences to extract labeled instances, such as in a sentence (3). Therefore, 1,088 (=1,450-362) example sentences did not used to extract labeled data.

So, we also added original example sentences in 3 patterns: that is adding [1] all the original example sentences (EX), [2] 30 extracted labeled instances ($L30$) and EX, and [3] $L30$ and unused example sentences (EXrL) that could not be used to obtain labeled data. These results are also shown at the bottom of Table 2[4].

As shown in Table 2, the third pattern (+$L30$ +EXrL) gave the best results (80.2 %), and it is superior to the patterns using all original examples. The original examples tend to be short, but because shorter examples are easier to match to the raw corpus, we can filter out examples in EXrL that are too short.

4.2 Human Evaluation of Extracted Training Data

We also provide a quantitative analysis of the extracted training data. The first 5 sentences extracted from BCCWJ were checked manually. Which included 1,038 sentences for 47 words split into 114 senses. Of which 979 sentences (94.3 %) were considered correct.

Because Iwanami was different from WordNet, we could not make a direct comparison with another expansion method such as (Mihalcea and Moldovan, 1999). But the quality of this manual evaluation result is comparable to that (95.7 %) reported in (Mihalcea and Moldovan, 1999).

[4]Only one result using EXrL significantly improved over Trn.

4.3 Conclusion in Step-1

In this step, we extracted labeled data automatically using example sentences from Iwanami. This method gave us longer, more natural and higher quality labeled data from the raw corpus, and we could obtain the labeled data even for unseen and infrequent senses.

Step-1 provided superior performance (80.2 %) to the state-of-the-art result (76.4 %), and the high effectiveness of this method is proved.

However, it may be difficult to achieve any further improvement because the extracted data may have an unnatural sense distribution and limited variations. Therefore, to realize an improvement, we employ a semi-supervised learning method in Step-2.

5 Employing Semi-supervised Learning Method: Step-2

In Step-2, we constructed a semi-supervised WSD classfier for each target word by using a successful semi-supervised learning method called *Maximum Hybrid Log-likelihood Expectation (*MHLE*)* (Fujino et al., 2010). It was reported that the MHLE method was useful for obtaining better classification performance especially when there was a large difference between the distributions of the labeled and test data. As mentioned in § 2, the data of the Japanese WSD task is across very different types of corpus; ranging from formal government paper to rough Web data. That was the reason that we employed the MHLE method.

In this section, we first describe the outline of the MHLE-based semi-supervised WSD classifier (in § 5.1), and then we present our method for extracting unlabeled data (in § 5.2). Finally, we undertake an experiment and investigate the effectiveness of our combination of semi-supervised learning and labeled data expansion (in § 5.3).

5.1 MHLE-based semi-supervised WSD classifier

In the MHLE-based semi-supervised WSD classifier for a taget word, the conditional probability, $P(s|x)$, of sense $s \in \{s_1, \ldots, s_k, \ldots, s_K\}$ given the feature vector x of a word instance is modeled by a combination of discriminative and generative models, $P_d(s|x;W)$ and $p_g(x,s;\Theta)$, where W and Θ

are the parameters of these models. By applying the classifier form and training method presented in Fujino et al. (2010), we defined $P(s|x)$ as

$$P(s_k|x;W,\Theta,\beta) = \frac{P_d(s_k|x;W)p_g(x,s_k;\Theta)^\beta}{\sum_{k'=1}^{K} P_d(s_{k'}|x;W)p_g(x,s_{k'};\Theta)^\beta}, \ \forall k. \quad (6)$$

We also provided the objective function, J for the parameter estimation of $P(s_k|x;W,\Theta,\beta)$ by using labeled and unlabeled datasets, $L = \{(x_n,s_n)\}_{n=1}^N$, and $U = \{x_m\}_{m=1}^M$ as

$$\begin{aligned} J &= \sum_{n=1}^{N} \log P_d(s_n|x_n;W)p_g(x_n,s_n;\Theta)^\beta \\ &+ \sum_{m=1}^{M} \log \sum_{k=1}^{K} P_d(s_k|x_m;W)p_g(x_m,s_k;\Theta)^\beta \\ &+ \log p(W) + \beta \log p(\Theta). \end{aligned} \quad (7)$$

Here, $\beta(>0)$ is a combination weight. W and Θ can be estimated as the values that maximize J for a fixed β value. The local optimal solution of W and Θ around an initial value can be obtained by an iterative computation such as the EM algorithm (Dempster et al., 1977). Namely, the MHLE-based semi-supervised WSD classifier is constructed by combining the discriminative and generative models trained on both labeled and unlabeled samples (See Fujino et al. (2010) for the details of the combination and training methods).

We employed a maximum entropy model (multinomial logistic regression model) and a naive Bayes model as $P_d(s|x;W)$ and $p_g(x,s;\Theta)$, as well as the text classifier presented in Fujino et al. (2010). In the naive Bayes model, the probability distribution of $x = (x_1,\ldots,x_i,\ldots,x_V)$ given s is regarded as a multinomial distribution: $p_g(x|s;\Theta) \propto \prod_{i=1}^V (\theta_{si})^{x_i}$, and the joint probability distribution of x and s is modeled as $p_g(x,s;\Theta) = p_g(x|s;\Theta)P(s)$. Here, $\theta_{si} > 0$ and $\sum_{i=1}^V \theta_{si} = 1$. V represents the dimension of feature vector x, and θ_{si} is the probability that the ith feature appears in an instance whose sense is s. $\Theta = \{\theta_{si}\}_{s,i}$ is the paremeter set of the naive Bayes model. In our experiments, we set $P(s) = 1/K$. We used Gaussian and Dirichlet priors as $p(W)$ and $p(\Theta)$, respectively. We tuned the β ($\in 0.5,1,2,5,10$) value with a 5-fold cross-validation of the labeled data.

5.2 Extracting Unlabeled Data

As unlabeled data, we extract sentences that include the target words from BCCWJ corpus. We show an example in (8), the **boldface** part indicates the target word.

(8) 年貢 に **取る** べし
annual tax as assess should
"(You) should assess annual tax"

Because of the looser restriction, we can extract many more sentences than the labeled data in § 3. For example, from BCCWJ alone, we can extract more than 10,000 instances for 22 words and more than 1,000 instances for the remaining words except for one (一つ *hitotsu* "one").

5.3 Experiment and Evaluation: Step-2

In this section, we describe an experiment in which we employed a semi-supervised classifier based on MHLE.

This experiment has two purposes; that is to investigate the effectiveness of (a) MHLE based semi-supervised WSD, and (b) automatically expanded data as labeled data.

5.3.1 Results and Discussion

Table 3 shows the results we obtained using unlabeled data extracted from BCCWJ. *U#* in the table shows the number of used unlabeled instances. In this experiment, we limited the unlabeled data to 10, 100, 200, 300, 500 and 1000. The smaller unlabeled data are subsets of the larger unlabeled data. We also use the given training data and several types of expanded data as labeled data.

Effectiveness of MHLE When we used only the given training data as labeled data (Trn), 300 unlabeled data gave the best performance (82.8 %), and it achieved a +5.4 % (= 82.8 − 77.4) improvement.

In addition, in contrast with the results in § 4, some improvements were achieved even for easy words. As described in § 5.1, it has been reported that the MHLE was robust even when the labeled and test data were very different, as with the Japanese WSD task, which came from heterogeneous corpus. Therefore, we can say that this semi-supervised WSD using a hybrid generative/discriminative approach (MHLE) is effective (with respect to purpose (a) above).

Effectiveness of Automatically Expanded Labeled Data We used several types of expanded data as labeled data; that is, Trn +EX, +*L*1, +*L*30, +*L*1 +EXrL, and +*L*30 +EXrL. Note that Trn +*L*30 +EXrL gave the best result in Step-1.

As shown in Table 3, all of the automatically expanded labeled data provided better results than the given training data alone. Therefore, we can say that these automatically expanded data are better than the given training data as labeled data (as regards purpose (b) above).

The labeled data Trn +$L30$ +EXrL gave the best results in Step-1, but the data that achieved the best result overall was Trn +$L1$. This shows that ultimately the original (fragmentary) example sentences are no match for the real world sentences extracted from raw corpora. By comparing +$L1$ with $L30$, for easy words in particular, +$L1$ produced better results than +$L30$ probably because of the sense bias. But for difficult words, +$L30$ produced better results than +$L1$.

In conclusion, also in our combination method, larger labeled data expansion was effective for more difficult words.

Learning Curves for Sample Words As shown above, MHLE works very effectively, however, that's not to say that the larger unlabeled data give better results. We show some learning curves for sample target words in D_{mid}, using Trn +$L1$ as the labeled data in Figure 2. As this figure shows, the behavior is very different from that of the target words.

For some words, the accuracy still is not saturated (For example, 良い *yoi* "good"), but for some words, the accuracy is decreasing (For example, 持つ *motu* "have"). In future work, we will investigate the causes of improvement or degradation.

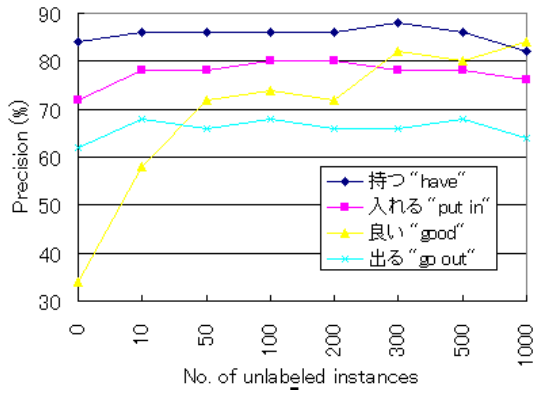

Figure 2: Learning curves for sample target words using expanded training data (Trn +$L1$) as the labeled data (MHLE) (%)

	D_{easy}	D_{mid}	D_{diff}	Total
Trn	90.6	70.9	61.1	77.4
+$U10$	91.2	76.7	68.2	81.3
+$U100$	91.3	78.2	68.4	82.0
+$U200$	91.9	79.1	69.1	82.7
+$U300$	91.9	79.6	68.9	**82.8**
+$U500$	91.3	78.9	68.0	82.2
+$U1000$	90.9	79.0	68.0	82.0
Trn + EX	89.8	72.2	64.9	78.3
+$U10$	91.8	77.5	70.4	82.2
+$U100$	91.9	79.7	70.9	83.2
+$U200$	92.3	80.4	72.2	**83.9**†
+$U300$	92.4	80.6	71.1	83.8†
+$U500$	92.0	80.4	72.2	83.8†
+$U1000$	91.6	80.6	70.0	83.3
Trn +$L30$ +EXrL	90.5	74.8	68.4	80.2†
+$U10$	91.7	78.1	71.1	82.6
+$U100$	91.8	79.6	72.9	83.5
+$U200$	91.8	80.3	72.7	**83.8**
+$U300$	91.9	80.0	71.8	83.5
+$U500$	91.4	80.5	72.2	83.6†
+$U1000$	91.4	80.5	72.2	83.4
Trn +$L1$ +EXrL	90.6	72.9	62.9	78.5†
+$U10$	91.7	76.2	67.8	81.2
+$U100$	**92.5**	79.8	70.0	83.4†
+$U200$	92.3	80.6	70.9	**83.8**†
+$U300$	92.3	80.4	70.0	83.5
+$U500$	91.9	81.3	70.4	**83.8**†
+$U1000$	91.7	80.6	70.0	83.4
Trn + $L30$	90.7	73.8	68.2	79.9†
+$U10$	91.6	78.5	71.8	82.8
+$U100$	91.8	79.2	72.2	83.2
+$U200$	92.1	79.9	73.1	83.8
+$U300$	92.2	80.0	**73.6**	**84.0**
+$U500$	91.5	79.8	72.4	83.4
+$U1000$	91.0	80.0	72.7	83.3
Trn + $L1$	90.7	72.0	64.7	78.5†
+$U10$	92.1	77.1	70.7	82.2
+$U100$	92.2	79.7	71.1	83.4†
+$U300$	92.4	**80.9**	72.2	**84.2**†
+$U500$	92.1	80.8	73.1	**84.2**†
+$U1000$	91.4	79.7	70.4	83.0

Table 3: Results of semi-supervised learning (MHLE), using given/expanded training data as the labeled data (%): where † shows that there is significant improvement over results using Trn as labeled data by t-test at 5 % level of significance; that is we compared Trn +$L1$ to Trn, Trn +$L1$ +$U10$ to Trn + $U10$, and so on.

5.3.2 Effectiveness for Unseen Senses

One of the advantages of our method is that it can provide training data even for unseen senses (which did not appear in the given training data but were in the dictionary). Therefore, we investigated the accuracy for unseen senses. In the Japanese WSD task, there are 18 instances for 9 unseen senses (See § 3). Table 4 shows the result for the 18 instances.

When we use expanded labeled data, the system can sometimes guess the unseen senses. The best performance (9 correct (50.0 %)) was achieved by +*L*30, because more labeled data were provided even for unseen senses. But also with +*L*1, 6 instances (33.3 %) were correct.

Of course, these unseen senses have no given training data (Trn), so the accuracy is 0 % on Trn. Therefore, no method based on given training data alone (Trn) can guess these senses correctly. So this constitutes a significant improvement.

	No.	%
Trn	0	0.0
Trn + EX	0	0.0
+*U*10 - *U*1000	5	27.8
Trn + *L*30 + EXrL	0	0.0
+*U*10	3	16.7
+*U*100	7	38.9
+*U*200 - *U*1000	**9**	50.0
Trn +*L*1 +EXrL	0	0.0
+*U*10 - *U*1000	2	11.1
Trn + *L*30	0	0.0
+*U*10	2	11.1
+*U*100	5	27.8
+*U*200 - *U*500	**9**	50.0
+*U*1000	4	22.2
Trn + *L*1	0	0.0
+*U*10	2	11.1
+*U*100	4	22.2
+*U*200, *U*300,*U*1000	5	27.8
+*U*500	6	33.3

Table 4: Effectiveness for unseen senses (9 senses, 18 instances)

6 Conclusion

In this paper, we proposed a combination WSD method consisting of automatic labeled data expansion (Step-1) and semi-supervised learning (Step-2). We targeted the SemEval-2010 Japanese WSD task, and showed the effectiveness of our proposed method.

In Step-1, we automatically extracted labeled data from raw corpora. We could extract longer, more natural and higher quality labeled data even for unseen senses, with a strict condition using the fragment examples.

In this step, we had already achieved a better performance (80.2 %) than the best result (76.4 %) in the formal run of the Japanese WSD task.

In Step-2, we employed semi-supervised learning. As the semi-supervised learning method, we employed the hybrid generative/discriminative approach (MHLE), because this method has been reported to be robust even when the labeled and test data were very different as in the Japanese WSD task, which came from heterogeneous corpus.

As a result, in this step MHLE achieved a good improvement (82.8 %), even when using only given training data as labeled data. Moreover, we showed the effectiveness of our expanded data as labeled data. We investigated which type of expanded data was the best as labeled data, then showed that adding only one sentence per original example was the best as labeled data (84.2 %), and could make it possible for the system to guess even unseen senses (33.3 %). In other words, when using labeled data for semi-supervised learning, minimum expansion provides the best performance, and protects the system against sense bias.

In feature work, we intend to investigate the reasons for improvement or degradation. We also intend to perform experiments in which we change the amount of expanded labeled data based on entropy based difficulties.

References

Eneko Agirre and David Martinez. 2000. Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web. *CoRR*, pp. 11–19.

Eneko Agirre and David Martinez. 2004. Unsupervised WSD based on Automatically Retrieved Examples: The Importance of Bias . In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing: EMNLP-2004*, pp. 25–32.

Timothy Baldwin, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez, and Takaaki Tanaka. 2010. A Reexamination of MRD-based Word Sense Disambiguation. *Transactions on Asian Language Information Process, Association for Computing Machinery (ACM)*, 9(1):1–21.

Bernard Brosseau-Villeneuve, Noriko Kando, and Jian-Yun Nie. 2010. RALI: Automatic Weighting of Text Window Distances. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 375–378. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: ACL-2007*, pp. 33–40.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Christine Fellbaum. 1998. A semantic network of English verbs. In Christine Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 3, pp. 70–104. MIT Press.

Akinori Fujino, Naonori Ueda, and Masaaki Nagata. 2010. A robust semi-supervised classification method for transfer learning. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, pp. 379–388.

Sanae Fujita, Francis Bond, Stephan Oepen, and Takaaki Tanaka. 2007. Exploiting Semantic Information for HPSG Parse Selection. In *Proceedings of ACL-2007 Workshop on Deep Linguistic Processing*, pp. 25–32.

Sanae Fujita, Kevin Duh, Akinori Fujino, Hirotoshi Taira, and Hiroyuki Shindo. 2010. MSS: Investigating the Effectiveness of Domain Combinations and Topic Features for Word Sense Disambiguation. In *the 5th International Workshop on Semantic Evaluation*, pp. 383–386. Association for Computational Linguistics.

Michael Lesk. 1986. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems documentation*, pp. 24–26.

Rada Mihalcea. 2002. Bootstrapping Large Sense Tagged Corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation: LREC-2002*, pp. 1407–1411.

Rada Mihalcea. 2004. Co-training and Self-training for Word Sense Disambiguation. In *Proceedings of the Conference on Natural Language Learning (CoNLL-2004)*, pp. 33–40.

Rada Mihalcea and Dan Moldovan. 1999. An Automatic Method for Generating Sense Tagged Corpora. In *Proceedings of the American Association for Artificial Intelligence (AAAI-1999)*, pp. 461–466.

Masaaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. 2003. CRL at Japanese dictionary-based task of SENSEVAL-2. *Journal of Natural Language Processing*, 10(3):115–143. (in Japanese).

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1–69.

Kamal Nigam, John Lafferty, and Andrew McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61–67.

Minoru Nishio, Etsutaro Iwabuchi, and Shizuo Mizutani. 1994. *Iwanami Kokugo Jiten Dai Go Han [Iwanami Japanese Dictionary Edition 5]*. Iwanami Shoten, Tokyo. (in Japanese).

Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. 2005. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics: ACL-2005*, pp. 395–402.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. SemEval-2010 Task: Japanese WSD. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 69–74. Association for Computational Linguistics.

Ted Pedersen. 2006. *Word Sense Disambiguation: Algorithms and Applications*, volume 33, chapter 6, pp. 133–166. Springer.

Thanh Phong Pham, Hwee Tou Ng, and Wee Sun Lee. 2005. Word Sense Disambiguation with Semi-Supervised Learning. In *Proceedings of the Twentieth National Conference on Artificial Intelligence: AAAI-2005*, pp. 1093–1098.

Kiyoaki Shirai. 2003. SENSEVAL-2 Japanese dictionary task. *Journal of Natural Language Processing*, 10(3):3–24. (in Japanese).

Kiyoaki Shirai and Makoto Nakamura. 2010. JAIST: Clustering and Classification Based Approaches for Japanese WSD. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 379–382. Association for Computational Linguistics.

Takaaki Tanaka, Francis Bond, Timothy Baldwin, Sanae Fujita, and Chikara Hashimoto. 2007. Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning: EMNLP-CoNLL-2007*, pp. 477–485.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics: ACL-93*, pp. 189–196.