Automatic Domain Adaptation for Word Sense Disambiguation Based on Comparison of Multiple Classifiers

Kanako Komiya
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho, Koganei
Tokyo, 184-8588 Japan
kkomiya@cc.tuat.ac.jp

Manabu Okumura
Tokyo Institute of Technology
4259 Nagatsuta Modori-ku
Yokohama 226-8503 Japan
oku@pi.titech.ac.jp

Abstract

Domain adaptation (DA), which involves adapting a classifier developed from source to target data, has been studied intensively in recent years. However, when DA for word sense disambiguation (WSD) was carried out, the optimal DA method varied according to the properties of the source and target data. This paper proposes automatic DA based on comparing the degrees of confidence of multiple classifiers for each instance. We compared three classifiers for three DA methods, where 1) a classifier was trained with a small amount of target data that was randomly selected and manually labeled but without source data, 2) a classifier was trained with source data and a small amount of target data that was randomly selected and manually labeled, and 3) a classifier was trained with selected source data that were sufficiently similar to the target data and a small amount of target data that was randomly selected and manually labeled. We used the method whose degree of confidence was the highest for each instance when Japanese WSD was carried out. The average accuracy of WSD when the DA methods that were determined automatically were used was significantly higher than when the original methods were used collectively.

1 Introduction

Classifiers in standard supervised machine learning have been trained for data in domain A using manually annotated data in domain A, e.g., to train classifiers for newswires using newswires. However, classifiers for data in domain B have sometimes been necessary when there have been no or few manually annotated data, and there have only been manually annotated data in domain A, which has been related to domain B. Domain adaptation (DA) involves adapting the classifier that has been trained from data in domain A (source domain) to data in domain B (target domain). This has been studied intensively in recent years.

However, the optimal method of DA varied according to the properties of the data in the source domain (the source data) and the data in the target domain (the target data) when DA for word sense disambiguation (WSD) was carried out.

There are many methods of DA for WSD but we assume that the optimal method varies according to each instance. This paper proposes automatic DA based on comparison of the degrees of confidence of multiple classifiers for each instance when Japanese WSD is performed. Our experiments show that the average accuracy of WSD when the DA methods that were determined automatically were used was significantly higher than when the original methods were used collectively.

This paper is organized as follows. Section 2 reviews related work on DA and Section 3 explains how a DA method is automatically determined. Sections 4 and 5 describe the methods and the data we used, respectievly. We present the results in Section 6 and discuss them in Section 7. Finally, we conclude the paper in Section 8.

2 Related Work

The DA problem can be categorized into three types depending on the information for learning, i.e., su-

26th Pacific Asia Conference on Language,Information and Computation pages 80–88

pervised, semi-supervised, and unsupervised approaches. A classifier in a supervised approach is developed from a large amount of labeled source data and a small amount of labeled target data with the aim of classifying target data better than a classifier developed only from the target data. A classifier in a semi-supervised approach is developed from a large amount of labeled source data and unlabeled target data with the aim of classifying target data better than a classifier developed only from the source data. Finally, a classifier is developed from a large amount of labeled source data with the aim of classifying target data accurately in an unsupervised approach. We focused on the supervised DA for WSD in this paper.

Many researchers have investigated DA within or outside the area of natural language processing. Chan and Ng (2006) carried out the DA of WSD by estimating class priors using an EM algorithm. Chan and Ng (2007) also conducted the DA of WSD by estimating class priors using the EM algorithm, but this was supervised DA using active learning.

In addition, Daumé III (2007) worked on the supervised DA. He augmented an input space and made triple length features that were general, source-specific, and target-specific. This was easy to implement, could be used with various DA methods, and could easily be extended to multi-DA problems.

Daumé III et al. (2010) extended the work in (Daumé III, 2007) to semi-supervised DA. It inherited the advantages of the supervised version and outperformed it by using unlabeled target data.

Agirre and de Lacalle (2008) worked on the semi-supervised DA for WSD. They applied singular value decomposition (SVD) to a matrix of unlabeled target data and a large amount of unlabeled source data, and trained a classifier with them. Agirre and de Lacalle (2009) worked on the supervised DA using almost the same method, but they used a small amount of labeled source data instead of the large amount of unlabeled source data.

Jiang and Zhai (2007) demonstrated that performance increased as examples were weighted when DA was applied. This method could be used with various other supervised or semi-supervised DA methods. In addition, they tried to identify and remove source data that misled DA, but they concluded that it was only effective if examples were not weighted.

Zhong et al. (2009) proposed an adaptive kernel approach that mapped the marginal distribution of source and target data into a common kernel space. They also conducted sample selection to make the conditional probabilities between the two domains closer.

Raina et al. (2007) proposed self-taught learning that utilized sparse coding to construct higher level features from the unlabeled data collected from the Web. This method was based on unsupervised learning.

Tur (2009) proposed a co-adaptation algorithm where both co-training and DA techniques were used to improve the performance of the model. The research by (Blitzer et al., 2006) involved work on semi-supervised DA, where they calculated the weight of words around the pivot features (words that frequently appeared both in source and target data and behaved similarly in both) to model some words in one domain that behaved similarly in another. They applied SVD to the matrix of the weights, generated a new feature space, and used the new features with the original features.

McClosky et al. (2010) focused on the problem where the best model for each document is not obvious when parsing a document collection of heterogeneous domains. They studied it as a new task of *multiple source parser adaptation*. They proposed a method of parsing a sentence that first predicts accuracies for various parsing models using a regression model, and then uses the parsing model with the highest predicted accuracy. The main difference is that their work was about parsing but ours discussed here is about Japanese WSD. They also assumed that they had labeled corpora in heterogeneous domains but we have not. We determined the best DA method for each instance.

Harimoto et al. (2010) measured the distance between domains to conduct DA using a suitable corpus in parsing. In addition, van Asch and Daelemans (2010) reported that performance in DA could be predicted depending on the similarity between source and target data using automatically annotated corpus in parsing. They focused on how corpora were selected for use as source data according to the distance between domains, but here we have focused on how to select a method of DA depending on the

degrees of confidence of multiple classifiers.

The closest work to this work is our previous work: (Komiya and Okumura, 2011) which determined an optimal DA method using decision tree learning given a triple of the target word type of WSD, source data, and target data. It discussed what features affected how the best method was determined. The main difference was that (Komiya and Okumura, 2011) determined the optimal DA method for each triple of the target word type of WSD, source data, and target data, but this paper determined the method for each instance.

3 Automatic determination of DA method for each instance

We assumed that the optimal method would vary according to each instance. The DA method is automatically determined for each instance as follows:

(1) Train multiple classifiers based on various methods,

(2) Compare the degrees of confidence of multiple classifiers for each instance,

(3) Employ the classifier whose degree of confidence is the highest for the instance.

The degrees of confidence we used here are the predicted values that indicate how confident classification is and are often used to select instances to be labeled in active-learning. We focused on the fact that these degrees of confidence are output from classifiers as the probability, and we can carry out ensemble learning by comparing them.

We would be able to determine the best DA method automatically using ensemble learning based on the degrees of confidence for each instance. Hence, we expected the average accuracy of WSD, when DA methods that were determined automatically were used for each instance, to be higher than when the original methods were used collectively. Navigli (2009) introduced this method as ensemble method for WSD and called it *probability mixture*. We used the *probability mixture* assuming that each classifier is trained for each DA method, rather than for each WSD method.

4 DA methods for WSD

Three methods were used as the DA methods for WSD in this study. All the methods except *Similarity Filtering* were adapted from (Komiya and Okumura, 2011) and *Similarity Filtering* was adapted from (Komiya and Okumura, 2012).

- *Target Only*: Train a classifier with a small amount of target data that is randomly selected and manually labeled but without source data.
- *Random Sampling*: Train a classifier with source data and a small amount of target data that is randomly selected and manually labeled.
- *Similarity Filtering*: Train a classifier with source data and a small amount of target data that is randomly selected and manually labeled. Only the source data that are sufficiently similar to the target data are selected by filtering and used.

The source data were selected as follows in *Similarity Filtering*. Here, the instances in the source and target data are represented as a vector in the WSD feature space. Each instance of WSD represents a word token whose word sense should be disambiguated.

(1) For every instance of target data $\forall t_i \in T$, calculate $sim_{i,j}$, i.e., the cosine similarity to every instance of source data $\forall s_j \in S$.

(2) For every instance of source data $\forall s_j \in S$, find $t_{j,nearest}$, i.e., the nearest instance in all the target data.

(3) For every instance of source data $\forall s_j \in S$, determine if it will be included in the training data set. Only source data s_j whose $sim_{j,nearest}$ is higher than 0.8 are used for the training data in this paper.

Ten instances of the target data were randomly selected and manually labeled in all the experiments.

Libsvm (Chang and Lin, 2001), which supports multi-class classification, was used as the classifier for WSD. We trained three classifiers and employed the classifier whose degree of confidence was the highest. A linear kernel was used according to the

results obtained from preliminary experiments. Seventeen features were introduced to train the classifier.

- Morphological features
  - Bag-of-words
  - Part-of-speech (POS)
  - Finer subcategory of POS
- Syntactic feature
  - If the POS of a target word is a noun, the verb which the target word modifies
  - If the POS of a target word is a verb, the case element of 'ヲ' (wo, objective) for the verb
- Semantic feature
  - Semantic classification code

Morphological features and a semantic feature were extracted from the surrounding words (two words to the right and left) of the target word. POS and finer subcategory of POS can be obtained using a morphological analyzer. We used ChaSen [1] as a morphological analyzer, the Bunruigoihyo thesaurus (National Institute for Japanese Language and Linguistics, 1964) for semantic classification codes (e.g. The code of 'program' is 1.3162.), and CaboCha [2] as a syntactic parser. Five-fold cross validation was used in the experiments.

5 Data

Three data which are the same as (Komiya and Okumura, 2011) were used for the experiments: (1) the sub-corpus of white papers in the Balanced Corpus of Contemporary Japanese (BCCWJ) (Maekawa, 2008), (2) the sub-corpus of documents from a Q&A site on the WWW in BCCWJ, and (3) Real World Computing (RWC) text databases (newspaper articles) (Hashida et al., 1998). DAs were conducted in six directions according to different source and target data. Word senses were annotated in these corpora according to a Japanese dictionary, i.e., the Iwanami Kokugo Jiten (Nishio et al., 1994). It has three levels for sense IDs, and we used the fine-level sense in the experiments. Multi-sense words that appeared equal or more than 50 times in both source and target data were selected as the target words in the experiment. There were 24 word types for white papers ⇔ Q&A site, 22 for white papers ⇔ newspaper articles, and 26 for Q&A site ⇔ newspaper articles. Twenty-eight word types were used in the experiments in total. Table 1 lists the minimum, maximum, and average number of instances of each word type for each corpus and Table 2 summarizes the number of instances of WSD for all combinations of corpora. Table 3 shows the list of target words.

Genre	Min.	Max.	Ave.
BCCWJ white papers	58	7,610	2074.50
BCCWJ Q&A site	82	13,976	2300.43
RWC newspaper	50	374	164.46

Table 1: Minimum, maximum, and average number of instances of each word type for each corpus

Source data	Target data	No. of instances
Q&A site	white paper	49,788
Q&A site	newspaper	4,276
white paper	Q&A site	60,930
white paper	newspaper	4,034
newspaper	Q&A site	63,805
newspaper	white paper	49,283
Total		232,116

Table 2: The number of instances of WSD for all combinations of corpora

(Komiya and Okumura, 2011) found that the optimal method of DA varied depending on each 'case' (i.e., a triple of the target word type of WSD, the source data, and the target data). Here, we have assumed that it varies according to each instance.

6 Results

Table 4 lists the micro and macro averaged accuracies of WSD for the whole data set and Tables 5 and 6 summarize the micro and macro averaged accuracies of WSD according to the corpora and DA methods, respectively. [3] The DA methods in bold are

[1] http://sourceforge.net/projects/masayu-a/

[2] http://sourceforge.net/projects/cabocha/

[3] The macro-averaged accuracies were always lower than micro-averaged accuracies in the three tables. We think this

Number of senses	Target words (in Japanese)	Sense example in English
2	場合	case
	自分	self
3	事業	project
	情報	information
	地方	area
	社会	society
	思う	suppose
	子供	child
4	分かる	understand
	考える	think
5	含む	contain
	使う	use
	技術	technique
6	関係	connection
	時間	time
	一般	general
	現在	present
	作る	make
7	今	now
8	前	before
10	持つ	have
11	進む	advance
12	見る	see
14	入る	enter
16	言う	say
21	出す	serve
22	手	hand
	出る	leave

Table 3: The list of target words

our proposed methods. *RS and TO* selected the DA method for each instance from *Random Sampling* and *Target Only*, *RS and SF* selected it from *Random Sampling* and *Similarity Filtering*, *SF and TO* selected it from *Similarity Filtering* and *Target Only*, and *All* selected it from *Random Sampling*, *Target Only*, and *Similarity Filtering* in Tables 4, 5, and 6. We used the -b option of libsvm when the method was *Random sampling*, *Target Only*, and *Similarity Filtering* to train a model for probability estimation. *MFS*, which is most frequent sense of fully annotated target data, *Source Only*, which is standard supervised learning only with the source data, *Self*, which is standard supervised learning with the whole target data, assuming that fully annotated data were obtained and could be used for learning, *oracle(i)*, which is oracle(instance) assuming that the system knows the optimal DA method for each instance, and *oracle(c)*, which is oracle(case) assuming that the system knows the optimal DA method given a 'case', were tested as references.

DA method	Micro	Macro
Random Sampling	79.85%	73.39%
Target Only	79.66%	72.09%
Similarity Filtering	78.47%	71.24%
RS and TO	****83.50 %***	****75.60%***
RS and SF	****81.22 %***	***74.09%***
SF and TO	****80.97 %***	***72.87%***
All	****82.96 %***	****74.77%***
MFS	77.05%	72.23%
Source Only	76.61%	69.82%
Self	92.82%	84.10%
oracle(i)_RS and TO	89.15%	83.31%
oracle(i)_RS and SF	89.15%	81.81%
oracle(i)_SF and TO	86.71%	79.82%
oracle(i)_All	91.74%	85.81%
oracle(c)_RS and TO	84.57%	77.73%
oracle(c)_RS and SF	84.03%	76.41%
oracle(c)_SF and TO	81.67%	75.17%
oracle(c)_All	85.14%	78.25%

Table 4: Average accuracies of WSD for the whole data set

The underline in these three tables means the highest accuracy for each combination of the source and target corpus and the bold means the proposed method outperformed the original methods. For example, the accuracy of *RS and TO* is in bold when it outperformed *Random Sampling* and *Target Only*. The asterisk means the difference between accuracies of the proposed and original methods is statistically significant according to a chi-square test. The level of significance in the test was 0.05.

7 Discussion

Table 4 indicates that our proposed method of automatic DA based on comparison of multiple classifiers always outperformed the original methods

is because the tasks with many data tend to give high accuracy.

Source data	Q&A site	Q&A site	white paper	white paper	newspaper	newspaper
Target data	white paper	newspaper	Q&A site	newspaper	Q&A site	white paper
DA method	Accuracy					
Random Sampling	87.21%	73.95%	83.97%	72.09%	76.61%	72.66%
Target Only	88.35%	66.46%	75.74%	67.75%	74.46%	84.57%
Similarity Filtering	88.20%	71.14%	70.04%	70.45%	75.04%	84.77%
RS and TO	**88.54%**	72.80%	*83.03%	**72.48%**	***78.10%**	***87.81%**
RS and SF	***88.65%**	73.20%	*80.14%	**72.46%**	***77.83%**	*80.86%
SF and TO	***90.17%**	70.39%	*74.53%	**70.72%**	***75.78%**	***88.09%**
All	***89.96%**	72.54%	*80.66%	***72.63%**	***77.22%**	***87.90%**
MFS	78.81%	67.35%	76.70%	68.59%	75.88%	78.74%
Source Only	80.64%	73.46%	83.37%	71.02%	75.50%	66.36%
Self	95.98%	78.09%	91.75%	79.57%	90.69%	96.07%
oracle(i)_RS and TO	91.09%	83.33%	90.59%	85.32%	83.18%	93.96%
oracle(i)_RS and SF	92.21%	81.48%	87.65%	79.35%	85.20%	94.47%
oracle(i)_SF and TO	93.85%	76.75%	83.70%	81.46%	81.42%	91.34%
oracle(i)_All	94.41%	84.87%	92.38%	87.63%	87.22%	95.06%
oracle(c)_RS and TO	88.58%	75.80%	85.41%	76.67%	76.66%	88.62%
oracle(c)_RS and SF	89.40%	75.28%	84.02%	73.53%	79.42%	86.21%
oracle(c)_SF and TO	89.83%	71.75%	76.35%	74.07%	76.66%	87.98%
oracle(c)_All	89.87%	75.84%	85.43%	77.09%	79.46%	88.80%

Table 5: Micro-averaged accuracies of WSD according to the corpora and the DA methods

when the average accuracies for all the directions of DA were compared. All the differences between micro-averaged accuracies of the proposed and original methods were statistically significant according to a chi-square test. When macro-averaged accuracies were compared, some differences were no longer significant due to the decrease of the samples of the test. Tables 5 and 6 denoted the same tendencies.

Table 4 also shows the micro and macro averaged accuracies of all the proposed method outperformed baseline methods, *Source Only* and *MFS*, as well as the three original methods. Particularlly, our proposed methods have beaten *MFS*, the baseline which needs fully annotated target data although our methods do not need them.

In addition, Tables 5 and 6 indicate that the automatic DA method based on comparison of multiple classifiers outperformed the original methods in four directions except when the source data were a Q&A site and the target data were newspapers and when the source data were white papers and the target data were a Q&A site. [4] These results mean that our proposed method is not always effective for every combination of all corpora but it is generally effective.

However, the results of *oracle(i)* are much better than those of the proposed methods. This indicates that the degree of confidence does not always predict the correct answer.

In addition, Table 4 shows the accuracy of *All*, i.e., the proposed method where the DA method was selected from three methods, is not the highest; the accuracy of *RS and TO*, the proposed method where the DA method was selected from two methods, is higher than this. According to Tables 5 and 6, the accuracies of *All* are not always the highest as seen in Table 4. In fact, the highest accuracy varies according to the combination of the source and target corpora and even depending on how they were averaged (micro vs. macro). Tables 5 and 6 show that

[4] However, *RS and TO* gives the highest accuracy when the source data were white papers and the target data were a Q&A site in Table 6.

Source data	Q&A site	Q&A site	white paper	white paper	newspaper	newspaper
Target data	white paper	newspaper	Q&A site	newspaper	Q&A site	white paper
DA method	Accuracy					
Random Sampling	84.45%	71.06%	72.56%	69.54%	69.25%	73.74%
Target Only	83.74%	63.76%	68.99%	67.31%	68.04%	82.18%
Similarity Filtering	83.85%	68.17%	58.75%	69.20%	67.16%	81.62%
RS and TO	**84.48%**	69.40%	**73.21%**	**71.04%**	***72.04%**	***84.64%**
RS and SF	**84.64%**	69.99%	*68.53%	**70.12%**	***72.18%**	79.73%
SF and TO	**85.69%**	67.08%	*63.59%	**69.44%**	**68.84%**	**84.07%**
All	**85.70%**	68.84%	*69.25%	**70.73%**	**71.41%**	**83.91%**
MFS	78.21%	66.28%	71.46%	70.27%	69.81%	77.58%
Source Only	75.27%	70.71%	70.66%	68.07%	67.86%	65.96%
Self	91.13%	74.79%	85.24%	78.59%	84.23%	91.53%
oracle(i)_RS and TO	88.32%	79.80%	82.55%	83.09%	77.66%	89.75%
oracle(i)_RS and SF	89.27%	78.61%	74.69%	76.89%	79.94%	92.36%
oracle(i)_SF and TO	89.83%	72.85%	77.19%	79.33%	74.81%	86.39%
oracle(i)_All	91.92%	81.39%	84.22%	85.56%	82.25%	90.53%
oracle(c)_RS and TO	85.71%	72.78%	76.44%	75.10%	73.07%	84.41%
oracle(c)_RS and SF	86.61%	72.36%	72.71%	71.66%	73.14%	82.72%
oracle(c)_SF and TO	85.89%	68.74%	70.39%	73.39%	69.71%	84.52%
oracle(c)_All	87.08%	72.88%	76.53%	75.82%	73.34%	85.06%

Table 6: Macro-averaged accuracies of WSD according to the corpora and the DA methods

only one combination for each table had the highest accuracy with *All* (white paper ⇒ newspaper in Table 5 and Q&A site ⇒ white paper in Table 6). They indicate that the accuracy does not always increase with the augmentation of the methods to be compared.

We think the reasons why *RS and TO* outperformed *All* are as follows. First, it is because the accuracy of *Similarity Filtering* was not as high as that of the other two methods according to Table 4. The accuracies of *RS and SF*, and *SF and TO* were also lower than that of *RS and TO*. Therefore, it seems that the accuracy of *All* decreased because the accuracy of the third method, *Similarity Filtering*, was lower than that of the others.

Moreover, we think that *RS and TO* achieved the highest accuracy because the two DA methods, *Random Sampling* and *Target Only*, were sufficiently different. In contrast, *Similarity Filtering* is similar to *Target Only* when the source and target data are not similar to each other and it is similar to *Random Sampling* when the source and target data are similar to each other. In other words, the DA method *Similarity Filtering* is intermediate between *Random Sampling* and *Target Only* and is similar to either of them in some way. We think that the experiments revealed that the accuracy of WSD increases when the DA methods are selected from those that are sufficiently different to one another.

Furthermore, we think that the property of *Target Only* affected the high accuracy of *RS and TO*. The accuracy of *Target Only* is very high especially when the percentage of occurrences of the most frequent sense is high as Khapra et al. (2010) stated that "Sense distributions of words are highly skewed and depend heavily on the domain at hand. This fact makes it very difficult for WSD approaches to beat the corpus baseline." On the other hand, the method *Target Only* will never be able to output the correct word sense for the instances whose word senses do not appear in the training data. Thus, the method with more training data, i.e., *Random Sampling*, should be used for these instances. We think the accuracy of *RS and TO* is high because the

degree of confidence of *Target Only* is low for the instances whose word senses do not appear in the training data (because their features are not similar to those of instances in the training data) .

We compare these results with those of Komiya and Okumura (2011). Even though we cannot have a direct comparison because the svm-predict -b 0 and -b1 (with/without probability estimation) give different accuracy values, the best result of the proposed method (83.50) is comparable to that of Komiya and Okumura (2011) (83.50). In addition, oracle(i) always outperformed oracle(c) in all the experiments, which indicates that our assumption where the optimal method of DA varies according to each instance seems to be better than that of Komiya and Okumura (2011) where it varies according to each 'case'. Even though the degree of confidence does not always predict the correct answer, we think the proposed method is sufficiently useful because it is much simpler than the previous method.

Finally, this paper compared only three methods, *Target Only*, *Random Sampling*, and *Similarity Filtering*, and we used the method whose degree of confidence was the highest for each instance. It remains unanswered and should be investigated in the future how effective this method is when the DA methods used changes or when the number of DA methods increases.

8 Conclusion

This paper proposed automatic DA based on comparing the degrees of confidence of multiple classifiers for each instance. We compared three classifiers for three DA methods, *Target Only*, *Random Sampling*, and *Similarity Filtering* and used the method whose degree of confidence was the highest for each instance. *Target Only* was a method where a classifier was trained with a small amount of target data that was randomly selected and manually labeled but without source data, *Random Sampling* was a method where a classifier was trained with source data and a small amount of target data that was randomly selected and manually labeled, and *Similarity Filtering* was a method where a classifier was trained with selected source data that were sufficiently similar to the target data and a small amount of target data that was randomly selected and manually labeled. The average accuracy of WSD when the DA methods that were determined automatically were used was significantly higher than when the original methods were used collectively. However, the experiment revealed that the accuracy of *All*, the proposed method where the DA method was selected from the three methods, was not the highest. The accuracy of *RS and TO*, i.e., the proposed method where the DA method was selected from the two methods, was higher than this. We think that the accuracy of WSD increases when the DA methods are selected from the methods that are sufficiently different. Even though the degree of confidence does not always predict the correct answer, we think the proposed method is sufficiently useful. It remains unanswered and should be investigated in the future how effective this method is when DA methods used changes or when the number of DA methods increases.

Acknowledgments

We would like to thank the reviewers for very constructive and detailed comments. This research is supported be Grants-in-Aid for Scientific Research, Priority Area "Japanese Corpus".

References

Eneko Agirre and Oier Lopez de Lacalle. 2008. On robustness and domain adaptation using svd for word sense disambiguation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 17–24.

Eneko Agirre and Oier Lopez de Lacalle. 2009. Supervised domain adaption for wsd. In *Proceedings of the 12th Conference of the European Chapter of the Association of Computational Linguistics*, pages 42–50.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural coppespondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128.

Yee Seng Chan and Hwee Tou Ng. 2006. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 89–96.

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010*, pages 23–59.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Keiko Harimoto, Yusuke Miyao, and Jun'ichi Tsujii. 2010. Kobunkaiseki no bunyatekiou ni okeru seido teika youin no bunseki oyobi bunyakan kyori no sokutei syuhou, in japanese. In *Proceedings of NLP2010*, pages 27–30.

Koichi Hashida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, and Wakako Kashino. 1998. The rwc text databases. In *Proceedings of the First International Conference on Language Resource and Evaluation*, pages 457–461.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271.

Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. Domain-specific word sense disambiguation combining corpus based and wordnet based parameters. In *Proceedings of the 5th International Conference on Global Wordnet (GWC2010)*.

Kanako Komiya and Manabu Okumura. 2011. Automatic determination of a domain adaptation method for word sense disambiguation using decision tree learning. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, IJCNLP 2011*, pages 1107–1115.

Kanako Komiya and Manabu Okumura. 2012. Automatic selection of domain adaptation method for wsd using decision tree learning. In *Journal of NLP (In Japanese), In press*.

Kikuo Maekawa. 2008. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pages 101–102.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36.

National Institute for Japanese Language and Linguistics. 1964. *Bunruigoihyo*. Shuuei Shuppan, In Japanese.

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

Minoru Nishio, Etsutaro Iwabuchi, and Shizuo Mizutani. 1994. *Iwanami Kokugo Jiten Dai Go Han*. Iwanami Publisher, In Japanese.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: Transfer learning from unlabeled data. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 759–766.

Gokhan Tur. 2009. Co-adaptation: Adaptive co-training for semi-supervised learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.*, pages 3721–3724.

Vincent van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010*, pages 31–36.

Erheng Zhong, Wei Fan, Jing Peng, Kun Zhang, Jiangtao Ren, Deepak Turaga, and Olivier Verscheure. 2009. Cross domain distribution adaptation via kernel mapping. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1027–1036.