2015 Pacific Association for Computational Linguistics

Active Learning to Remove Source Instances for Domain Adaptation for Word Sense Disambiguation

Hiroyuki Shinnou Yoshiyuki Onodera Minoru Sasaki Kanako Komiya
Ibaraki University, Department of Computer and Information Sciences
Hitachi, JAPAN
shinnou@mx.ibaraki.ac.jp, 14nm705n@vc.ibaraki.ac.jp, {msasaki, kkomiya}@mx.ibaraki.ac.jp

***Abstract*—In this paper, an active learning method of domain adaptation issues for word sense disambiguation is presented. In general, active learning is an approach where data with high learning effect is selected from an unlabeled data set, then labeled manually, and added to the training data. However, data in the source domain can deteriorate classification precision (misleading data), which extends errors to the domain adaptation. When data labeled by active learning is added to training data, an attempt is made to detect misleading data in the source domain and delete it from the training data. In this way, compared to standard learning classification precision is improved.**

***Keywords*-active learning; domain adaptation; word sense disambiguation;**

I. INTRODUCTION

When a natural language processing task is performed, the training and test data are usually in the same domain. However, sometimes the data comes from different domains. Recently, studies into domain adaptation have fine-tuned the classifier by using the training data of a learned domain (source domain) to match the test data of another domain (target domain) [1] [2][3].

If the subject of the domain adaptation is problematic due to lack of target domain labels, active learning [4],[5] and semi-supervised learning [6] are effective. In this paper, we use active learning for domain adaptation for Word Sense Disambiguation (WSD).

Generally, active learning is an approach that gradually increases the precision of the classifier by selecting data with a high learning effect from an unlabeled data set, labeling the data, and adding it to the training data, thereby increasing the amount of training data monotonically. However, in domain adaptation, there are data that have a negative influence on the target domain due to classification in the source domain training data. Here we refer to such data as "misleading data"[7]. In this paper, we detect such data in the source domain training data and delete it to construct training data suitable for the target domain using active learning.

In the experiment, we use three domains: Yahoo! Answers (OC), Book (PB) and newspaper (PN) from the Balanced Corpus of Contemporary Written Japanese (BCCWJ [8]). The data set, which is provided by a Japanese WSD SemEval-2 task[9] has word sense tags attached to parts of these corpora. There are 16 multi-sense words with a certain frequency across all domains, and six patterns of domain adaptation (OC → PB, PB → PN, PN → OC, OC → PN, PN → PB, and PB → OC). We investigate domain adaptation for WSD using the proposed active learning method for $16 \times 6 = 96$ patterns and show the effectiveness of the proposed method.

II. ACTIVE LEARNING WITH DELETED MISLEADING DATA

A. Active learning

Active learning is an approach that reduces the amount of manual labeling when building effective training data. Using a classifier trained on the current training data, we selected data with as high a learning effect as possible from an unlabeled data set. Then, we manually assign correct labels to the selected data and add it to the training data. Consequently, the amount of labeled data is increased and the classifier is improved.

The key question of active learning is how to choose data with a high learning effect. There are many active learning methods[4]; however, one particularly effective method is widely used. This method selects data with the lowest classification reliability determined by a powerful classifier such as a support vector machine (SVM) classifier[10].

B. Detecting and deleting misleading data

The initial labeled data in a general active learning is fixed. This is not problematic because all labeled data is useful. However, the initial pool of labeled data for domain adaptation, i.e., labeled data in the source domain can include harmful data. Here we refer to such data 'misleading data.' When general active learning is applied to domain adaptation, misleading data in the source domain prevents active learning from improving the classifier. Therefore, when we add labeled data to the training data, we detect misleading data and delete it from the labeled training data in the source domain.

Figure 1 shows the algorithm of our method. The initial labeled data in the source domain is denoted D_0, and the labeled data added to training data during the active learning

D_0 is set labeled data in source domain
U is set unlabeled data in target domain
$A \leftarrow \{\}$; labeled data added by Active Learning
$D_1 \leftarrow D_0 \cup A$
h_1 is set the classifier learned through D_1
L_1 is set the classification of D_0 by h_1

repeat 10 times do
    b is the labeled data obtained by active learning for U using h_1
    $U \leftarrow U - \{b\}$
    $A \leftarrow A \cup \{b\}$
    $D_2 \leftarrow D_0 \cup A$
    h_2 is set the classifier learned through D_2
    L_2 is set the classification of D_0 by h_2
    z is the misleading data detected through L_1 and L_2
    $D_0 \leftarrow D_0 - \{z\}$
    $D_1 \leftarrow D_0 \cup A$
    $h_1 \leftarrow h_2$
    $L_1 \leftarrow L_2$
done

h_2 is the final classifier

Figure 1. Our proposed active learning

process is denoted A, where initial A is empty. D_1 is the union of D_0 and A, and h_1 is the classifier learned through D_1. By using h_1, we classify D_0; the classification result is denoted L_1. Like general active learning, we classify the unlabeled data set U in the target domain using h_1 and assign a correct label to identify data b with the lowest classification reliability. Data b is added to A. D_2 is the union of D_0 and A, and h_2 is the classifier learned through D_2. We use h_2 to classify D_0 and denote the classification result as L_2. We detect misleading data z using L_1 and L_2 by following procedure. Using to following cases (a),(b) or (c), we can identify misleading data. (a) There are false classifications in L_2. In this case, we identify the data with the highest classification reliability among the false classifications. (b) There are no false classifications. In this case, by comparing L_1 with L_2, we identify the data with the greatest decrease in reliability from L_1 to L_2. (c) There are no false classifications and no data with decreased reliability. In this case, no misleading data is identified. As shown in Figure 1, this procedure is repeated 10 times.

In this study, active learning is complete when 10 data have been added to the labeled training data set. The only difference between general active learning and active learning for domain adaptation is the distribution of the initial labeled data set. Thus when labeled data is increased through active learning, there are very few differences. Therefore, we evaluate the proposed method with 10 repetitions of active learning.

III. Experiment

In the experiment, we use three domains: OC, PB and PN from the Balanced Corpus of Contemporary Written Japanese (BCCWJ [8]). As mentioned previously the data set, which was provided by a Japanese WSD SemEval-2 task [9], has word sense tags attached to part of these corpora. There are 16 multi-sense words with some frequency across all domains. These 16 target words are shown in Table I. [1]. There are six direction patterns of (OC → PB, PB → PN, PN → OC, OC → PN, PN → PB, and PB → OC). Consequently $16 \times 6 = 96$ types of domain adaptation of WSD are used in the experiment.

In each direction of domain adaptation (e.g., OC → PB), we conducted active learning for 16 target words. We evaluated the active learning method for domain adaptation using the average of these 16 precision.

We tried three methods. The first method is active learning to select added data at random (Random), the second is standard active learning (AL), and the third is our proposed active learning (Our AL). For all methods, the classifier is a SVM. We use the SVM tool *'libsvm'* [2] to train the classifier.

[1]The word "入る (Hairu)" has three senses in a dictionary. However, it has four senses in OC and PB domain. The fourth sense is new. In Japanese WSD SemEval-2 task, tagging the new sense was attempted.

[2]http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table I
TARGET WORDS OF EXPERIMENT

Word	# of meanings in dictionary	OC		PB		PN	
		Freq.	Meanings	Freq.	Meanings	Freq.	Meanings
言う (Iu)	3	666	2	1114	2	363	2
入れる (Ireru)	3	73	2	56	3	32	2
書く (Kaku)	2	99	2	62	2	27	2
聞く (Kiku)	3	124	2	123	2	52	2
子供 (Kodomo)	2	77	2	93	2	29	2
時間 (Jikan)	4	53	2	74	2	59	2
自分 (Jibun)	2	128	2	308	2	71	2
出る (Deru)	3	131	3	152	3	89	3
取る (Toru)	8	61	7	81	7	43	7
場合 (Baai)	2	126	2	137	2	73	2
入る (Hairu)	3	68	4	118	4	65	3
前 (Mae)	3	105	3	160	2	106	4
見る (Miru)	6	262	5	273	6	87	3
持つ (Motsu)	4	62	4	153	3	59	3
やる (Yaru)	5	117	3	156	4	27	2
ゆく (Yuku)	2	219	2	133	2	27	2
Average	3.35	193.9	2.94	150.6	2.88	75.56	2.69

Using the `-b` option, we can obtain the reliability of the classification.

We show the result of the experiment in Figures 3, 4, 5, 6, 7 and 8. Each figure shows the result of each domain adaptation. In this experiment, active learning stops after 10 repetitions. After 10 repetitions, the current classifier is presented in Table II and Figure 2. Our proposed active learning method outperforms standard active learning in every domain adaptation type.

Table II
AVERAGE PRECISION OF THE FINAL CLASSIFIER (%)

	AL	Our AL	Random
OC → PB	78.25	**78.98**	75.94
PB → PN	84.06	**84.46**	80.38
PN → OC	75.51	**78.41**	75.31
OC → PN	79.54	**80.24**	77.04
PN → PB	80.81	**81.13**	79.08
PB → OC	78.00	**78.52**	76.33
Average	79.36	**80.29**	77.35

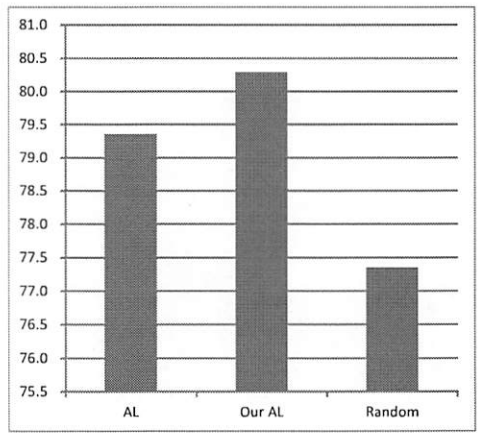

Figure 2. Comparison of Average Precisions

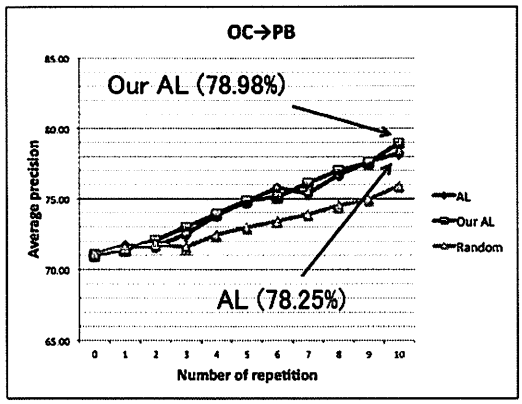

Figure 3. Active learning for "OC → PB"

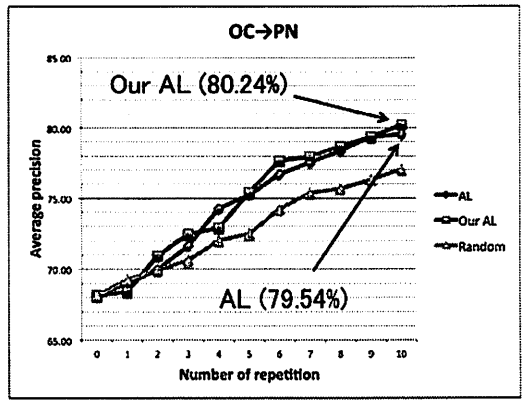

Figure 6. Active learning for "OC → PN"

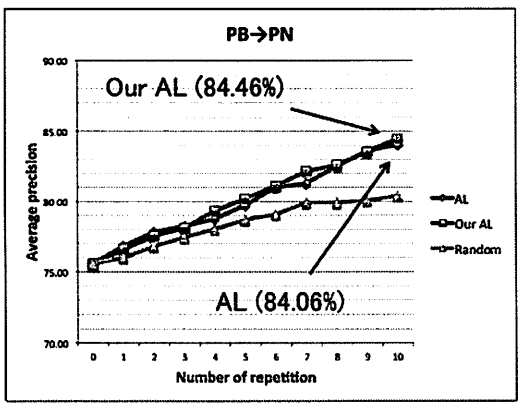

Figure 4. Active learning for "PB → PN"

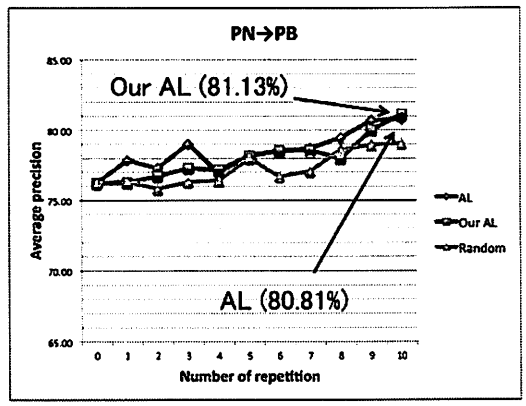

Figure 7. Active learning for "PN → PB"

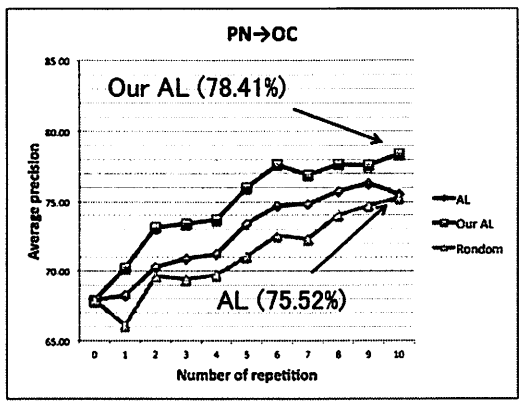

Figure 5. Active learning for "PN → OC"

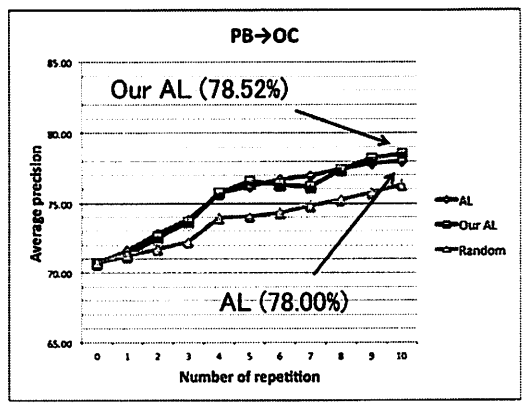

Figure 8. Active learning for "PB → OC"

IV. Discussion

A. Existence and detection of misleading data

We do not know whether the data as misleading data in the experience are actually misleading data. Here, we use the data labels to determine if the detected data are in fact misleading data, and we examine whether the method for detecting misleading data is effective.

At first, we identify the misleading data individually following a previously proposed method [11]. The labeled data D in S of target word w exists in domain adaptation for fine-tuning the domain S to T. Next we measure the correct answer rate p_0 of the classifier T learned by D, delete data x from D, and measure the correct answer rate p_1 of the classifier T learned by $D-\{x\}$. When $p_1 > p_0$, we consider data x to be misleading data. We perform this procedure for all data across D and find the misleading data of target word w. Table III shows the amount of misleading data found by this process. The numerical values in the parentheses are the amount of all data.

From the data presented in Table III, we investigate whether misleading data detected by the experimental procedure are true or not. The result are shown in Table IV. The numerical values in the parenthesis are the amount of detected data, and the numerical values next to the parenthesis are the amount of the true misleading data. From Table IV, it is evident that the amount of detected data is 959, the amount of true misleading data is 121, and the precision is 0.1262. It is thought that this value is low. However, precision is not always reduced deleting false detected data. Therefore, we believe that the detected data were not related to classification.

B. Instance weight

In domain adaptation tasks, labeled data in the target domain are more important than labeled data in the source domain. Therefore, instance weight learning is effective in domain adaptation [7]. Generally, the weight of the instance is defined by the probability density ratio [12]. Here, we investigate active learning weighting of the detected target domain data. We simply weight detected data by doubling the frequency of such data. Table VI shows the average precision of the final classifier obtained by active learning.

Table V
ACTIVE LEARNING WITH INSTANCE WEIGHT (%)

	Our AL	Our AL with instance weight
OC → PB	**78.98**	77.70
PB → PN	84.46	**84.75**
PN → OC	**78.41**	78.05
OC → PN	**80.24**	80.15
PN → PB	81.13	**82.25**
PB → OC	78.52	**79.81**
Average	80.29	**80.45**

From Table VI, we can confirm the effect of weighting on target domain labeled data. This experiment is simply weighting double heaviness. We intended to investigate the potential for improvement in future work.

C. Feature weight

Because target domain labeled data are added by active learning, we can use the supervised domain adaptation method.

Here, we combine Daumé's method[13] with active learning. We convert vector $\boldsymbol{x}_s$ of the source domain into a triple length vector $(\boldsymbol{x}_s, \boldsymbol{x}_s, \mathbf{0})$, and vector $\boldsymbol{x}_t$ of the target domain into a triple length vector $(\mathbf{0}, \boldsymbol{x}_t, \boldsymbol{x}_t)$ using Daumé's method. We classify the target domain data with the standard classification using the tripled vector. This method weights the common (overlapped) features of the source domain and the target domain.

When the Daumé's method is combined with active learning, we only have to convert source domain data $\boldsymbol{x}_s$ into $(\boldsymbol{x}_s, \boldsymbol{x}_s, \mathbf{0})$, and target domain data $\boldsymbol{x}_t$ into $(\mathbf{0}, \boldsymbol{x}_t, \boldsymbol{x}_t)$. The result for ten repetitions are shown in Table VI.

From Table VI, it is evident that using the proposed method with Daumé's method is not effective; however standard active learning combined with Daumé's method is effective. It is thought that the influence of misleading data becomes small with Daumé's method; consequently, the proposed method with Daumé's method was not effective. In future, we intend to investigate this possibility.

Table III
MISLEADING DATA

単語	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC
言う (Iu)	159 (666)	75 (1114)	82 (363)	158 (666)	35 (363)	127 (1114)
入れる (Ireru)	6 (73)	15 (56)	3 (32)	28 (73)	1 (32)	19 (56)
書く (Kaku)	21 (99)	2 (62)	12 (27)	39 (99)	15 (27)	0 (62)
聞く (Kiku)	26 (124)	0 (123)	4 (52)	21 (124)	27 (52)	26 (123)
子供 (Kodomo)	5 (77)	1 (93)	12 (29)	0 (77)	13 (29)	12 (93)
時間 (Jikan)	1 (53)	0 (74)	0 (59)	8 (53)	5 (59)	0 (74)
自分 (Jibun)	13 (128)	0 (308)	0 (71)	25 (128)	1 (71)	0 (308)
出る (Deru)	14 (131)	32 (152)	22 (89)	10 (131)	10 (89)	39 (152)
取る (Toru)	6 (61)	18 (81)	12 (43)	5 (61)	22 (43)	10 (81)
場合 (Baai)	0 (126)	13 (137)	14 (73)	0 (126)	9 (73)	7 (137)
入る (Hairu)	36 (68)	27 (118)	27 (65)	11 (68)	42 (65)	38 (118)
前 (Mae)	8 (105)	1 (160)	15 (106)	5 (105)	2 (106)	10 (160)
見る (Miru)	10 (262)	12 (273)	8 (87)	3 (262)	28 (87)	3 (273)
持つ (Motsu)	8 (62)	11 (153)	1 (59)	0 (62)	1 (59)	2 (153)
やる (Yaru)	0 (117)	0 (156)	0 (27)	0 (117)	0 (27)	0 (156)
ゆく (Yuku)	17 (219)	1 (133)	3 (27)	0 (219)	3 (27)	15 (133)

Table IV
CORRECT ANSWER RATES OF DETECTION OF MISLEADING DATA

単語	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC
言う (Iu)	2 (10)	2 (10)	2 (10)	3 (10)	1 (10)	2 (10)
入れる (Ireru)	2 (10)	3 (10)	2 (10)	4 (10)	0 (10)	2 (10)
書く (Kaku)	1 (10)	1 (10)	5 (10)	3 (10)	5 (10)	0 (10)
聞く (Kiku)	1 (10)	0 (10)	2 (10)	1 (10)	4 (10)	1 (10)
子供 (Kodomo)	1 (10)	1 (10)	3 (10)	0 (10)	5 (10)	0 (10)
時間 (Jikan)	0 (10)	0 (10)	0 (10)	0 (10)	0 (10)	0 (10)
自分 (Jibun)	0 (10)	0 (10)	0 (10)	2 (10)	0 (10)	0 (10)
出る (Deru)	1 (10)	1 (10)	2 (10)	2 (10)	0 (10)	1 (10)
取る (Toru)	2 (10)	2 (10)	4 (10)	1 (10)	4 (10)	2 (10)
場合 (Baai)	0 (10)	2 (10)	3 (10)	0 (10)	1 (10)	0 (10)
入る (Hairu)	5 (10)	2 (10)	4 (10)	1 (10)	7 (10)	3 (10)
前 (Mae)	0 (10)	0 (10)	1 (10)	1 (10)	0 (10)	1 (10)
見る (Miru)	0 (10)	1 (10)	0 (10)	0 (10)	1 (10)	0 (10)
持つ (Motsu)	1 (10)	0 (10)	1 (10)	0 (10)	0 (10)	0 (10)
やる (Yaru)	0 (10)	0 (10)	0 (9)	0 (10)	0 (10)	0 (10)
ゆく (Yuku)	1 (10)	0 (10)	0 (10)	0 (10)	1 (10)	1 (10)

V. CONCLUSION

In this paper, we proposed a new active learning method of domain adaptation for WSD. In standard active learning, labeled training data increases monotonically. However, data in the source domain can deteriorate classification precision (misleading data), which extends errors to the domain adaptation. Our proposed method detects and deletes misleading data in the source domain during the standard active learning process. Through an experiment using three domains (OC, PB and PN) in BCCWJ and 16 common target words, the proposed method outperformed standard active learning. In future, we intend to investigate methods to detect misleading data more accurately and to assign proper weight to instances and features during the active learning process.

Table VI
USE OF DAUMÉ'S METHOD IN ACTIVE LEARNING(%)

	AL	Our AL	AL with Daumé	Our AL with Daumé
OC → PB	78.25	**78.98**	77.09	76.24
PB → PN	84.06	**84.46**	82.08	79.00
PN → OC	75.51	78.41	**78.98**	75.50
OC → PN	79.54	**80.24**	79.37	78.75
PN → PB	80.81	**81.13**	81.01	74.57
PB → OC	78.00	78.52	**80.83**	80.75
Average	79.36	**80.29**	79.89	77.47

REFERENCES

[1] A. Søgaard, *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Morgan & Claypool, 2013.

[2] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[3] S. Mori, "Domain adaptation in natural language processing (in japanese)," *The Japanese Society for Artificial Intelligence*, vol. 27, no. 4, pp. 365–372, 2012.

[4] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, 2010.

[5] P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian, "Domain adaptation meets active learning," in *NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, 2010, pp. 27–32.

[6] O. Chapelle, B. Schölkopf, A. Zien *et al.*, *Semi-supervised learning*. MIT press Cambridge, 2006, vol. 2.

[7] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in nlp," in *ACL-2007*, 2007, pp. 264–271.

[8] K. Maekawa, "Design of a Balanced Corpus of Contemporary Written Japanese," in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, 2007, pp. 55–58.

[9] M. Okumura, K. Shirai, K. Komiya, and H. Yokono, "SemEval-2010 Task: Japanese WSD," in *The 5th International Workshop on Semantic Evaluation*, 2010, pp. 69–74.

[10] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *ICML*, 2000, pp. 839–846.

[11] H. Yoshida and H. Shinnou, "Detection of misleading data by outlier detection methods (in japanese)," in *The 5th Japanese Corpus Linguistics Workshop*, 2014, pp. 49–56.

[12] M. Sugiyama and M. Kawanabe, *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation* . MIT Press, 2011.

[13] Daumé III, Hal, "Frustratingly Easy Domain Adaptation," in *ACL-2007*, 2007, pp. 256–263.