MSS: Investigating the Effectiveness of Domain Combinations and Topic Features for Word Sense Disambiguation

Sanae Fujita Kevin Duh Akinori Fujino Hirotoshi Taira Hiroyuki Shindo
NTT Communication Science Laboratories
{sanae, kevinduh, taira, a.fujino, shindo}@cslab.kecl.ntt.co.jp

Abstract

We participated in the SemEval-2010 Japanese Word Sense Disambiguation (WSD) task (Task 16) and focused on the following: (1) investigating domain differences, (2) incorporating topic features, and (3) predicting new unknown senses. We experimented with Support Vector Machines (SVM) and Maximum Entropy (MEM) classifiers. We achieved 80.1% accuracy in our experiments.

1 Introduction

We participated in the SemEval-2010 Japanese Word Sense Disambiguation (WSD) task (Task 16 (Okumura et al., 2010)), which has two new characteristics: (1) Both training and test data across 3 or 4 domains. The training data include books or magazines (called **PB**), newspaper articles (**PN**), and white papers (**OW**). The test data also include documents from a Q&A site on the WWW (**OC**); (2) Test data include new senses (called **X**) that are not defined in dictionary.

There is much previous research on WSD. In the case of Japanese, unsupervised approaches such as extended Lesk have performed well (Baldwin et al., 2010), although they are outperformed by supervised approaches (Tanaka et al., 2007; Murata et al., 2003). Therefore, we selected a supervised approach and constructed Support Vector Machines (SVM) and Maximum Entropy (MEM) classifiers using common features and topic features. We performed extensive experiments to investigate the best combinations of domains for training.

We describe the data in Section 2, and our system in Section 3. Then in Section 4, we show the results and provide some discussion.

2 Data Description

2.1 Given Data

We show an example of Iwanami Kokugo Jiten (Nishio et al., 1994), which is a dictionary used as a sense inventory. As shown in Figure 1, each entry has POS information and definition sentences including example sentences.

We show an example of the given training data in (1). The given data are morphologically analyzed and partly tagged with Iwanami's sense IDs, such as “37713-0-0-1-1” in (1).

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

This task includes 50 target words that were split into 219 senses in Iwanami; among them, 143 senses including two **X**s that were not defined in Iwanami, appear in the training data. In the test data, 150 senses including eight **X**s appear. The training and test data share 135 senses including two **X**s; that is, 15 senses including six **X**s in the test data are unseen in the training data.

2.2 Data Pre-processing

We performed two preliminary pre-processing steps. First, we restored the base forms because the given training and test data have no information about the base forms. (1) shows an example of the original morphological data, and then we added the base form (lemma), as shown in (2).

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

Secondly, we extracted example sentences from Iwanami, which is used as a sense inventory. To compensate for the lack of training data, we analyzed examples with a morphological analyzer, Mecab[1] UniDic version, because the training and test data were tagged with POS based on UniDic.

[1] http://mecab.sourceforge.net/

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 383–386,
Uppsala, Sweden, 15-16 July 2010.

HEADWORD	とる【取る・採る・執る・捕る】*take*	(五他 Transitive Verb)
37713-0-0-1-0	<1> 置いてあったものなどを手に持つ。to get something left into one’s hand	
37713-0-0-1-1	<ア> 手で握り持つ。「手を-って導く」 take and hold by hand. “to lead someone by the hand”	

Figure 1: Simplified Entry for Iwanami Kokugo Jiten: とる *take*

For example, from the entry for とる *take*, as shown in Figure 1, we extracted an example sentence and morphologically analyzed it, as shown in (3)[2], for the second sense, 37713-0-0-1-1. In (3), the underlined part is the headword and is tagged with 37713-0-0-1-1.

(3)	手	を	<u>取っ</u>て	導く
	hand	ACC	*<u>take</u>*	*and lead*

“(I) take someone’s hand and lead him/her”

3 System Description

3.1 Features

In this section, we describe the features we generated.

3.1.1 Baseline Features

For each target word w, we used the surface form, the base form, the POS tag, and the top POS categories, such as nouns, verbs, and adjectives of w. Here the target is the ith word, so we also used the same information of $i-2, i-1, i+1$, and $i+2$th words. We used bigrams, trigrams, and skip-bigrams back and forth within three words. We refer to the model that uses these baseline features as bl.

3.1.2 Bag-of-Words

For each target word w, we got all base forms of the content words within the same document or within the same article for newspapers (**PN**). We refer to the model that uses these baseline features as bow.

3.1.3 Topic Features

In the SemEval-2007 English WSD tasks, a system incorporating topic features achieved the highest accuracy (Cai et al., 2007). Inspired by (Cai et al., 2007), we also used topic features.

Their approach uses Bayesian topic models (Latent Dirichlet Allocation: LDA) to infer topics in an unsupervised fashion. Then the inferred topics are added as features to reduce the sparsity problem with word-only features.

In our proposed approach, we use the inferred topics to find “related’” words and directly add these word counts to the bag-of-words representation.

We applied gibbslda++[3] to the training and test data to obtain multiple topic classification per document or article for newspapers (**PN**). We used the document or article topics for newspapers (**PN**) including the target word. We refer to the model that uses these topic features as tpX, where X is the number of topics and tpdistX with the topics weighted by distributions. In particular, the topic distribution of each document/article is inferred by the LDA topic model using standard Gibbs sampling.

We also add the most typical words in the topic as a bag-of-words. For example, one topic might include 市 *city*, 東京 *Tokyo*, 線 *train line*, 区 *ward* and so on. A second topic might include 解剖 *dissection*, 後 *after*, 医学 *medicine*, 墓 *grave* and so on. If a document is inferred to contain the first topic, then the words (市 *city*, 東京 *Tokyo*, 線 *train line*, ...) are added to the bag-of-words feature. We refer to these features as twdY, including the most typical Y words as bag-of-words.

3.2 Investigation between Domains

In preliminary experiments, we used both SVM[4] and MEM (Nigam et al., 1999), with optimization method L-BFGS (Liu and Nocedal, 1989) to train the WSD model.

First, we investigated the effect between domains (**PN**, **PB**, and **OW**). For training data, we selected words that occur in more than 50 sentences, separated the training data by domain, and tested different domain combinations.

Table 1 shows the SVM results of the domain combinations. For Table 1, we did a 5-fold cross validation for the self domain and for comparison

[2] We use ACC as an abbreviation of accusative postposition.

[3] `http://gibbslda.sourceforge.net/`

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

Table 1: Investigation of Domain Combinations on Training data (features: bl + bow, SVM)

Target Words 77, No. of Instances > 50			
Domain	Acc.(%)	Diff.	Comment
PN	78.7	-	63 words, 1094 instances
PN +**OW**	79.25	0.55	
PN +**PB**	79.43	**0.73**	
PN +**ALL**	79.34	0.64	
PB	79.29	-	75 words, 2463 instances
PB +**PN**	78.85	-0.45	
PB +**OW**	78.56	-0.73	
PB +**ALL**	78.4	**-0.89**	
OW	87.91	-	42 words, 703 instances
OW +**PN**	89.05	**1.14**	
OW +**PB**	88.34	0.43	
OW +**ALL**	89.05	**1.14**	

with the results after adding the other domain data. In Table 1, Diff. shows the differences to the self domain.

As shown in Table 1, for **PN** and **OW**, using other domains improved the results, but for **PB**, other domains degraded the results. So we decided to select the domains for each target word.

In the formal run, for each pair of domain and target words, we selected the combination of domain and dictionary examples that got the best cross-validation result in the training data. Note that in the case of no training data for the test data domain, for example, since no **OC**s have training data, we used all training data and dictionary examples.

We show the number of selected domain combinations for each target domain in Table 2. Because the distribution of target words is very unbalanced in domains, not all types of target words appear in every domain, as shown in Table 2.

3.3 Method for Predicting New Senses

We also tried to predict new senses (**X**) that didn't appear in the training data by calculating the entropy for each target given in the MEM. We assumed that high entropy (when the probabilities of classes are uniformly dispersed) was indicative of **X**; i.e., if [entropy $>$ threshold] $=>$ predict **X**; else $=>$ predict with MEM's output sense tag.

Note that we used the words that were tagged with **X**s in the training data, except for the target words. We compared the entropies of **X** and not **X** of the words and heuristically tuned the threshold based on the differences among entropies. Our three official submissions correspond to different thresholds.

Table 2: Used Domain Combinations

Used	MEM		SVM	
Domain	No.	(%)	No.	(%)
Target: **PB** (48 types of target words)				
ALL +**EX**	26	54.2	23	47.9
ALL	4	8.3	6	12.5
PB	11	22.9	8	16.7
PB +**EX**	1	2.1	1	2.1
PB +**OW**	1	2.1	3	6.3
PB +**PN**	5	10.4	7	14.6
Target: **PN** (46 types of target words)				
ALL +**EX**	30	65.2	30	65.2
ALL	4	8.7	4	8.7
PN	4	8.7	1	2.2
PN +**EX**	0	0	1	2.2
PN +**OW**	2	4.3	2	4.3
PN +**PB**	6	13	8	17.4
Target: **OW** (16 types of target words)				
ALL +**EX**	5	31.3	5	31.3
ALL	2	12.5	1	6.3
OW	6	37.5	3	18.8
OW +**PB**	3	18.8	3	18.8
OW +**PN**	0	0	4	25.0
Target: **OC** (46 types of target words)				
ALL +**EX**	46	100	46	100

4 Results and Discussions

Our cross-validation experiments on the training set showed that selecting data by domain combinations works well, but unfortunately this failed to achieve optimal results on the formal run. In this section, we show the results using all of the training data with no domain selections (also after fixing some bugs).

Table 3 shows the results for the combination of features on the test data. MEM greatly outperformed SVM. Its effective features are also quite different. In the case of MEM, baseline features (bl) almost gave the best result, and the topic features improved the accuracy, especially when divided into 200 topics. But for SVM, the topic features are not so effective, and the bag-of-words features improved accuracy.

For MEM with bl +tp200, which produced the best result, the following are the best words: 外 *outside* (accuracy is 100%), 経済 *economy* (98%), 考える *think* (98%), 大きい *big* (98%), and 文化 *culture* (98%). On the other hand, the following are the worst words: 取る *take* (36%), 良い *good* (48%), 上げる *raise* (48%), 出す *put out* (50%), and 立つ *stand up* (54%).

In Table 4, we show the results for each POS (bl +tp200, MEM). The results for the verbs are comparably lower than the others. In future work, we will consider adding syntactic features that may improve the results.

Table 3: Comparisons among Features and Test data

TYPE	Precision (%)		
	MEM	SVM	Explain
Base Line	68.96	68.96	Most Frequent Sense
bl	**79.3**	69.6	Base Line Features
bl +bow	77.0	**70.8**	+ Bag-of-Words (B**OW**)
bl +bow +tp100	76.4	70.7	+BOW + Topics (100)
bl +bow +tp200	77.0	70.7	+BOW + Topics (200)
bl +bow +tp300	77.4	70.7	+BOW + Topics (300)
bl +bow +tp400	76.8	70.7	+BOW + Topics (400)
bl +bow +tpdist300	77.0	70.8	+BOW + Topics (300)*distribution
bl +bow +tp300 +twd100	76.2	70.8	+ Topics (300) with 100 topic words
bl +bow +tp300 +twd200	76.0	70.8	+ Topics (300) with 200 topic words
bl +bow +tp300 +twd300	75.9	70.8	+ Topics (300) with 300 topic words
without bow			
bl +tp100	79.3	69.6	+ Topics (100)
bl +tp200	**80.1**	69.6	+ Topics (200)
bl +tp300	**79.6**	69.6	+ Topics (300)
bl +tp400	**79.6**	69.6	+ Topics (400)
bl +tpdist100	79.3	69.6	+ Topics (100)*distribution
bl +tpdist200	79.3	69.6	+ Topics (200)*distribution
bl +tpdist300	79.3	69.6	+ Topics (300)*distribution
bl +tp200 +twd100	74.6	69.6	+ Topics (200) with 100 topic words
bl +tp300 +twd10	74.4	69.4	+ Topics (300) with 10 topic words
bl +tp300 +twd20	75.2	69.3	+ Topics (300) with 20 topic words
bl +tp300 +twd50	74.8	69.2	+ Topics (300) with 50 topic words
bl +tp300 +twd200	74.6	69.6	+ Topics (300) with 200 topic words
bl +tp300 +twd300	75.0	69.6	+ Topics (300) with 300 topic words
bl +tp400 +twd100	74.1	69.6	+ Topics (400) with 100 topic words
bl+tpdist100 +twd20	79.3	69.6	+ Topics (100)*distribution with 20 topic words
bl+tpdist200 +twd20	79.3	69.6	+ Topics (200)*distribution with 20 topic words
bl+tpdist400 +twd20	79.3	69.6	+ Topics (400)*distribution with 20 topic words

Table 4: Results for each POS (bl +tp200, MEM)

POS	No. of Types	Acc. (%)
Nouns	22	85.5
Adjectives	5	79.2
Transitive Verbs	15	76.9
Intransitive Verbs	8	71.8
Total	50	80.1

In the formal run, we selected training data for each pair of domain and target words and used entropy to predict new unknown senses. Although these two methods worked well in our cross-validation experiments, they did not perform well for the test data, probably due to domain mismatch.

Finally, we also experimented with SVM and MEM, and MEM gave better results.

References

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(4):1–21.

Jun Fu Cai, Wee Sun Lee, and YW Teh. 2007. Improving Word Sense Disambiguation using Topic Features. In *Proceedings of EMNLP-CoNLL-2007*, pp. 1015–1023.

Dong C. Liu and Jorge Nocedal. 1989. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Programming*, 45(3, (Ser. B)):503–528.

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).

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).

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. SemEval-2010 Task: Japanese WSD. In *SemEval-2: Evaluation Exercises on Semantic Evaluation*.

Takaaki Tanaka, Francis Bond, Timothy Baldwin, Sanae Fujita, and Chikara Hashimoto. 2007. Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information. In *Proceedings of EMNLP-CoNLL-2007*, pp. 477–485.