

Optimizing Feature Set for Chinese Word Sense Disambiguation

Zheng-Yu Niu, Dong-Hong Ji

Institute for Infocomm Research
21 Heng Mui Keng Terrace
119613 Singapore
{zniu, dhji}@i2r.a-star.edu.sg

Chew-Lim Tan

Department of Computer Science
National University of Singapore
3 Science Drive 2
117543 Singapore
tancl@comp.nus.edu.sg

Abstract

This article describes the implementation of I^2R word sense disambiguation system ($I^2R - WSD$) that participated in one senseval3 task: Chinese lexical sample task. Our core algorithm is a supervised Naive Bayes classifier. This classifier utilizes an optimal feature set, which is determined by maximizing the cross validated accuracy of NB classifier on training data. The optimal feature set includes part-of-speech with position information in local context, and bag of words in topical context.

1 Introduction

Word sense disambiguation (WSD) is to assign appropriate meaning to a given ambiguous word in a text. Corpus based method is one of the successful lines of research on WSD. Many supervised learning algorithms have been applied for WSD, ex. Bayesian learning (Leacock et al., 1998), exemplar based learning (Ng and Lee, 1996), decision list (Yarowsky, 2000), neural network (Towel and Voorheest, 1998), maximum entropy method (Dang et al., 2002), etc.. In this paper, we employ Naive Bayes classifier to perform WSD.

Resolving the ambiguity of words usually relies on the contexts of their occurrences. The feature set used for context representation consists of local and topical features. Local features include part of speech tags of words within local context, morphological information of target word, local collocations, and syntactic relations between contextual words and target word, etc.. Topical features are bag of words occurred within topical context. Contextual features play an important role in providing discrimination information for classifiers in WSD. In other words, an informative feature set will help classifiers to accurately disambiguate word senses, but an uninformative feature set will deteriorate the performance of classifiers. In this paper, we optimize feature set by maximizing the cross validated accuracy of Naive Bayes classifier on sense tagged training data.

2 Naive Bayes Classifier

Let $C = \{c_1, c_2, \dots, c_L\}$ represent class labels, $F = \{f_1, f_2, \dots, f_M\}$ be a set of features. The value of f_j , $1 \leq j \leq M$, is 1 if f_j is present in the context of target word, otherwise 0. In classification process, the Naive Bayes classifier tries to find the class that maximizes $P(c_i|F)$, the probability of class c_i given feature set F , $1 \leq i \leq L$. Assuming the independence between features, the classification procedure can be formulated as:

$$\hat{i} = \arg \max_{1 \leq i \leq L} \frac{p(c_i) \prod_{j=1}^M p(f_j|c_i)}{\prod_{j=1}^M p(f_j)}, \quad (1)$$

where $p(c_i)$, $p(f_j|c_i)$ and $p(f_j)$ are estimated using maximum likelihood method. To avoid the effects of zero counts when estimating $p(f_j|c_i)$, the zero counts of $p(f_j|c_i)$ are replaced with $p(c_i)/N$, where N is the number of training examples.

3 Feature Set

For Chinese WSD, there are two strategies to extract contextual information. One is based on Chinese characters, the other is to utilize Chinese words and related morphological or syntactic information. In our system, context representation is based on Chinese words, since words are less ambiguous than characters.

We use two types of features for Chinese WSD: local features and topical features. All of these features are acquired from data at senseval3 without utilization of any other knowledge resource.

3.1 Local features

Two sets of local features are investigated, which are represented by LocalA and LocalB. Let n_l denote the local context window size.

LocalA contains only part of speech tags with position information: $POS_{-n_l}, \dots, POS_{-1}, POS_0, POS_{+1}, \dots, POS_{+n_l}$, where POS_{-i} (POS_{+i}) is the part of speech (POS) of the i -th words to the left (right) of target word w , and POS_0 is the POS of w .

LocalB enriches the local context by including the following features: local words with position information ($W_{-n_l}, \dots, W_{-1}, W_{+1}, \dots, W_{+n_l}$), bigram templates ($(W_{-n_l}, W_{-(n_l-1)}), \dots, (W_{-1}, W_{+1}), \dots, (W_{+(n_l-1)}, W_{+n_l})$), local words with POS tags (W_POS) (position information is not considered), and part of speech tags with position information.

All of these POS tags, words, and bigrams are gathered and each of them contributed as one feature. For a training or test example, the value of some feature is 1 if it occurred in local context, otherwise it is 0. In this paper, we investigate two values of n_l for LocalA and LocalB, 1 and 2, which results in four feature sets.

3.2 Topical features

We consider all Chinese words within a context window size n_t as topical features. For each training or test example, senseval3 data provides one sentence as the context of ambiguous word. In senseval3 Chinese training data, all contextual sentences are segmented into words and tagged with part of speech.

Words which contain non-Chinese character are removed, and remaining words occurred within context window size n_t are gathered. Each remaining word is considered as one feature. The value of topical feature is 1 if it occurred within window size n_t , otherwise it is 0.

In later experiment, we set different values for n_t , ex. 1, 2, 3, 4, 5, 10, 20, 30, 40, 50. Our experimental result indicated that the accuracy of sense disambiguation is related to the value of n_t . For different ambiguous words, the value of n_t which yields best disambiguation accuracy is different. It is desirable to determine an optimal value, \hat{n}_t , for each ambiguous word by maximizing the cross validated accuracy.

4 Data Set

In Chinese lexical sample task, training data consists of 793 sense-tagged examples for 20 ambiguous Chinese words. Test data consists of 380 untagged examples for the same 20 target words. Table 1 shows the details of training data and test data.

5 Criterion for Evaluation of Feature Sets

In this paper, five fold cross validation method was employed to estimate the accuracy of our classifier, which was the criterion for evaluation of feature sets. All of the sense tagged examples of some target word in senseval3 training data were shuffled and divided into five equal folds. We used four folds as training set and the remaining fold as test

set. This procedure was repeated five times under different division between training set and test set. The average accuracy over five runs is defined as the accuracy of our classifier.

6 Evaluation of Feature Sets

Four feature sets were investigated:

FEATUREA1: LocalA with $n_l = 1$, and topical feature within optimal context window size \hat{n}_t ;

FEATUREA2: LocalA with $n_l = 2$, and topical feature within optimal context window size \hat{n}_t ;

FEATUREB1: LocalB with $n_l = 1$, and topical feature within optimal context window size \hat{n}_t ;

FEATUREB2: LocalB with $n_l = 2$, and topical feature within optimal context window size \hat{n}_t .

We performed training and test procedure using exactly same training and test set for each feature set. For each word, the optimal value of topical context window size \hat{n}_t was determined by selecting a minimal value of n_t which maximized the cross validated accuracy.

Table 2 summarizes the results of Naive Bayes classifier using four feature sets evaluated on senseval3 Chinese training data. Figure 1 shows the accuracy of Naive Bayes classifier as a function of topical context window size on four nouns and three verbs. Several results should be noted specifically:

If overall accuracy over 20 Chinese characters is used as evaluation criterion for feature set, the four feature sets can be sorted as follows: $FEATUREA1 > FEATUREA2 \approx FEATUREB1 > FEATUREB2$. This indicated that simply increasing local window size or enriching feature set by incorporating bigram templates, local word with position information, and local words with POS tags did not improve the performance of sense disambiguation.

In table 2, it showed that with FEATUREA1, the optimal topical context window size was less than 10 words for 13 out of 20 target words. Figure 1 showed that for most of nouns and verbs, Naive Bayes classifier achieved best disambiguation accuracy with small topical context window size (<10 words). This gives the evidence that for most of Chinese words, including nouns and verbs, the near distance context is more important than the long distance context for sense disambiguation.

7 Experimental Result

The empirical study in section 6 showed that FEATUREA1 performed best among all the feature sets. A Naive Bayes classifier with FEATUREA1 as feature set was learned from all the senseval3 Chinese training data for each target word. Then we used

Table 1: Details of training data and test data in Chinese lexical sample task.

Ambiguous word	POS occurred in training data	# training examples	# senses occurred in training data	# test examples
ba3wo4	n v vn	31	4	15
bao1	n nr q v	76	8	36
cai2liao4	n	20	2	10
chong1ji1	v vn	28	3	13
chuan1	v	28	3	14
di4fang1	b n	36	4	17
fen1zi3	n	36	2	16
huo2dong4	a v vn	36	5	16
lao3	Ng a an d j	57	6	26
lu4	n nr q	57	6	28
mei2you3	d v	30	3	15
qi3lai2	v	40	4	20
qian2	n nr	40	4	20
ri4zi5	n	48	3	21
shao3	Ng a ad j v	42	5	20
tu1chu1	a ad v	30	3	15
yan2jiu1	n v vn	30	3	15
yun4dong4	n nz v vn	54	3	27
zou3	v vn	49	5	24
zuo4	v	25	3	12

this classifier to determine the senses of occurrences of target words in test data. The official result of $I^2R - WSD$ system in Chinese lexical sample task is listed below:

Precision: 60.40% (229.00 correct of 379.00 attempted).

Recall: 60.40% (229.00 correct of 379.00 in total).

Attempted: 100.00% (379.00 attempted of 379.00 in total).

8 Conclusion

In this paper, we described the implementation of $I^2R - WSD$ system that participated in one senseval3 task: Chinese lexical sample task. An optimal feature set was selected by maximizing the cross validated accuracy of supervised Naive Bayes classifier on sense-tagged data. The senses of occurrences of target words in test data were determined using Naive Bayes classifier with optimal feature set learned from training data. Our system achieved 60.40% precision and recall in Chinese lexical sample task.

References

Dang, H. T., Chia, C. Y., Palmer M., & Chiou, F.D. (2002) Simple Features for Chinese Word Sense Disambiguation. *In Proc. of COLING*.

Leacock, C., Chodorow, M., & Miller G. A. (1998) Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24:1, 147–165.

Mooney, R. J. (1996) Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. *In Proc. of EMNLP*, pp. 82-91, Philadelphia, PA.

Ng, H. T., & Lee H. B. (1996) Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. *In Proc. of ACL*, pp. 40-47.

Pedersen, T. (2001) A Decision Tree of Bigrams is an Accurate Predictor of Word Sense. *In Proc. of NAACL*.

Towel, G., & Voorheest, E. M. (1998) Disambiguating Highly Ambiguous Words. *Computational Linguistics*, 24:1, 125–146.

Yarowsky, D. (2000) Hierarchical Decision Lists for Word Sense Disambiguation. *Computers and the Humanities*, 34(1-2), 179–186.

Table 2: Accuracy of Naive Bayes classifier with different feature sets on Senseval3 Chinese training data.

Ambiguous word	FEATUREA1		FEATUREA2		FEATUREB1		FEATUREB2	
	\hat{n}_t	Accuracy	\hat{n}_t	Accuracy	\hat{n}_t	Accuracy	\hat{n}_t	Accuracy
ba3wo4	5	30.0	4	23.3	4	30.0	3	30.0
bao1	2	30.7	20	34.0	2	33.3	20	32.0
cai2liao4	2	85.0	2	80.0	2	75.0	2	60.0
chong1ji1	20	40.0	3	40.0	30	36.0	1	28.0
chuan1	3	72.0	5	68.0	3	56.0	5	64.0
di4fang1	2	74.3	1	62.9	1	71.4	1	65.7
fen1zi3	20	91.4	50	91.4	20	88.6	20	85.7
huo2dong4	5	40.0	20	51.4	10	42.9	4	40.0
lao3	3	49.1	4	47.3	3	52.7	20	52.7
lu4	1	83.6	2	78.2	2	81.8	1	76.4
mei2you3	20	50.0	20	47.9	4	43.3	3	50.0
qi3lai2	4	75.0	1	75.0	1	80.0	1	77.5
qian2	3	57.5	4	57.5	3	60.0	5	57.5
ri4zi5	4	62.2	4	57.8	10	55.6	4	55.6
shao3	4	45.0	3	50.0	10	42.5	20	50.0
tu1chu1	10	83.3	10	80.0	10	80.0	10	76.7
yan2jiu1	20	43.3	20	46.7	10	50.0	20	36.7
yun4dong4	10	64.0	10	66.0	10	62.0	10	58.0
zou3	5	44.4	5	44.4	4	51.1	4	51.1
zuo4	20	64.0	30	60.0	20	64.0	20	64.0
Overall		57.7		56.9		57.0		55.1

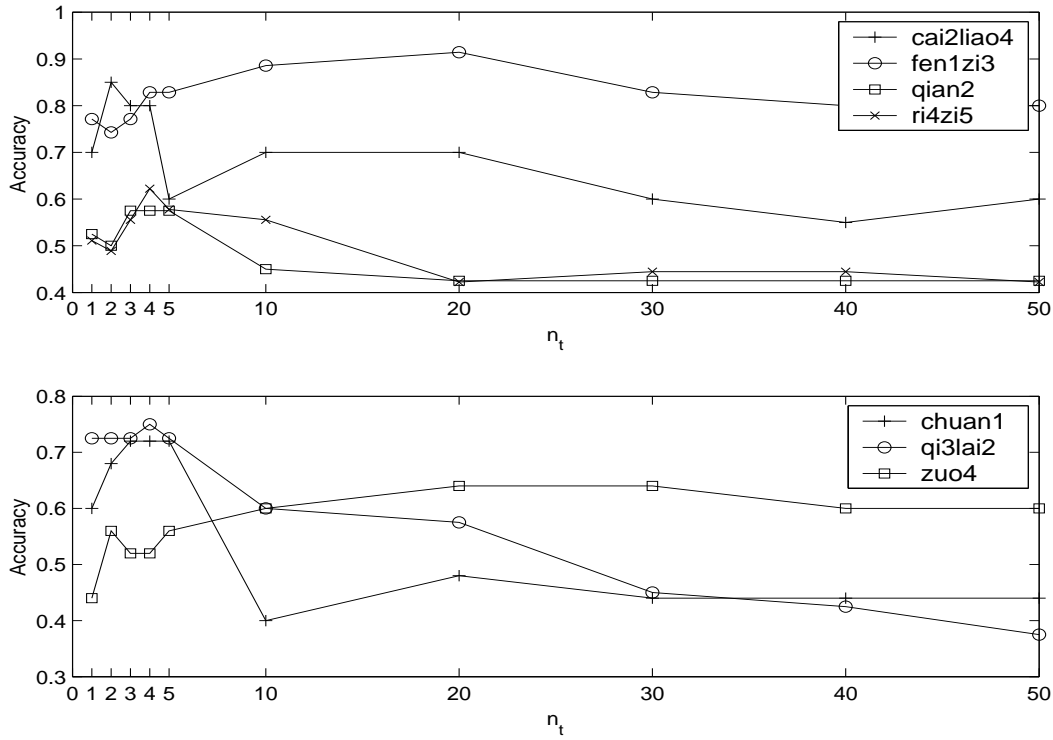


Figure 1: Accuracy of Naive Bayes classifier with the optimal feature set FEATUREA1 on four nouns (top figure) and three verbs (bottom figure). The horizontal axis represents the topical context window size.