PKU: Combining Supervised Classifiers with Features Selection

Peng Jin, Danqing Zhu, *Fuxin Li and Yunfang Wu
Institute of Computational Linguistics
Peking University, Beijing, China
*Institute of Automation Chinese Academy of Sciences
Beijing, China
{jandp,zhudanqing,wuyf}@pku.edu.cn *Fuxin.li@ia.ac.cn

Abstract

This paper presents the word sense disambiguation system of Peking University which was designed for the SemEval-2007 competition. The system participated in the Web track of task 11 "English Lexical Sample Task via English-Chinese Parallel Text". The system is a hybrid model by combining two supervised learning algorithms SVM and ME. And the method of entropy-based feature chosen was experimented. We obtained precision (and recall) of 81.5%.

1 Introduction

The PKU system participated in the web track of task 11. In this task, the organizers propose an English lexical sample task for word sense disambiguation (WSD), where the sense-annotated examples are (semi)-automatically gathered from word-aligned English-Chinese parallel texts. After assigning appropriate Chinese translations to each sense of an English word, the English side of the parallel texts can then serve as the training data, as they are considered to have been disambiguated and "sense-annotated" by the appropriate Chinese translations. This proposed task is thus similar to the multilingual lexical sample task in Senseval3, except that the training and test examples are collected without manually annotating each individual ambiguous word occurrence.

The system consists of two supervised learning classifiers, support vector machines (SVM) and maximum entropy (ME). A method of entropy-based feature chosen was experimented to reduce the feature dimensions. The training data was limited to the labeled data provided by the task, and a PoS-tagger (tree-tagger) was used to get more features.

2 Features Selection

We used tree-tagger to PoS-tag the texts before the feature extractor. No other resource is used in the system. The window size of the context is set to 5 around the ambiguous word. Only the following features are used in the system:

Local words
Local PoSs
Bag-of-words
Local collocations

Here local collocation means any two words which fall into the context window to form collocation pair.

Two methods are used to reduce the dimensions of feature space. One comes from the linguistic knowledge, some words whose PoSs are IN, DT, SYM, POS, CC or "``" are not included as the features.

The second method is based on entropy. To each word, the training data was split to two parts for parameter estimation. One (usually consist of 30 – 50 instances) as the simultaneous test and the rest instances form the other part.

First the entropy of each feature was calculated. For example, the target word 'work', it has two senses and the dimensions of its feature space is N. For feature f_i, if it appears in m instances belonging to sense A and n instances in sense B. So the

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 261–263,
Prague, June 2007. ©2007 Association for Computational Linguistics

probability distributions are: $p_1 = m/(m+n)$ and $p_2 = n/(m+n)$. The entropy of f_i is:

$$H(f_i) = \sum_{j=1}^{2} p_j \log \frac{1}{p_j}$$

We rank all the features according to their entropy from small to big. And then first percent lambda features are chosen as the final feature set. Using this smaller feature set, we use the classifier to make a new prediction.

The parameter λ is estimated by comparing the system performance on the simultaneous test. In our system, .68 is chosen. It means that 68% original features used to form the new feature space.

The same classifier was tried on different feature sets to get different outputs and then were combined.

3 Classifiers

The Support Vector Machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. It is developed by Vapnik and has been applied into WSD (Lee et al., 2004). Since most of the target words have more than two senses, we used the implementation of SVM that includes lib-svm (Chang and Lin, 2001) and svm-multiclass (Joachims, 2004). To lib-svm, the parameter of "b" which is used to obtain probability information after training is set 0 or 1 individually to form different classifiers. The default linear kernel is used.

Each vector dimension represents a feature. The numerical value of a vector entry is the numerical value of the corresponding feature. In our system, we use binary features. If the context of an instance has a particular feature, then the feature value is set to 1, otherwise the value is set to 0.

ME modeling provides a framework for integrating information for classification from many heterogeneous information sources. The intuition behind the maximum entropy principle is: given a set of training data, model what is known and assume no further knowledge about the unknown by assigning them equal probability (entropy is maximum). There are also some researchers using ME to WSD (Chao and Dyer, 2002). Dekang Lin's implementation of ME was used. He used Generalized Iterative Scaling (GIS) algorithm.

4 Development

Because of time constraints, we could not experiment all the training data by cross-validation. To each target word, we extract first 50 training instances as the test.

Target Word	Svm-Multi-class	ME	Lib-svm		
			Prob. Output		Non-prob. Output
			Orig. F.S.	Red. FS	
Age	.68	.70	.70	.70	.66
Area	.80	.70	.80	.74	.82
Body	.84	.84	.90	.92	.16
Change	.48	.42	.66	.42	.58
Director	.96	.94	.96	.96	.96
Experience	.90	.88	.88	.90	.88
Future	.94	.94	.94	.98	.94
interest	.84	.82	.82	.88	.84
issue	.88	.88	.84	.90	.88
Life	.92	.94	.98	1.0	.94
Material	.88	.92	.94	.94	.88
Need	.86	.86	.86	.86	.86
performance	.78	.82	.80	.82	.80
Program	.70	.74	.72	.72	.72
Report	.94	.94	.94	.94	.94
System	.76	.70	.76	.76	.70
Time	.70	.64	.68	.60	.76
today	.72	.70	.74	.68	.76
Water	.90	.92	.88	.82	.90
Work	.90	.86	.90	.92	.90

Table 1: The Performance on Nouns

For some adjectives, we just extract first 30 because the training data is small. For ten of adjectives, the training data is too small, we directly use the lib-svm (with probability output) as the final classifier.

Both SVM and ME could output the probability for each instance to each class. So we try to combine them to improve the performance. Several methods of combining classifiers have been investigated (Radu et al., 2002). The enhanced Counted-based Voting (CBV) and Rank-Based Voting, Probability Mixture Model, and best single Classifier are experimented in the training data. Table 1 and Table 2 indicate the results of nouns and adjectives individually, which were achieved with each of the different methods. In these tables, "Orig F.S." and "Red. F.S." mean original feature set and reduced feature set. "Prob. output" and "Non Prob.

output" are two implementation of lib-svm. The former output the probability of each instance belonging to each class, otherwise the latter not. Different from the results of Radu, choosing the best single classifier get the better performance than any kinds of combination. In this paper, we did not list the performances of combining.

According to Table 1 and Table 2, the particular classifier chosen for that word was the one with the highest score in the training data.

Target Word	Svm-Multi-class	ME	Lib-svm		
			Prob. Output		Non-prob. output
			Orig. F.S.	Red. F.S.	
Early	.77	.80	.77	.80	.77
Educational	.87	.87	.87	.83	.87
Free	.74	.80	.84	.90	82
Human	.96	.92	.96	.90	.96
Long	.70	.70	.73	.87	.70
Major	.78	.78	.78	.80	.78
Medical	.76	.86	.78	.84	.78
New	.73	.77	.63	.43	.63
Simple	.73	.77	.77	.77	.80
Third	.98	.94	.98	1.0	.96

Table 2: The performance on Adjectives

Two parameters are different from these two SVMs. One is the "-c", which is the tradeoff between training error and margin. In lib-svm the value of "-c" is set 1; but in svm-multiclass is 0.01. The other is the strategy of how to utility binary-classification to resolve multi-class. In svm-multiclass, no strategy is needed since the algorithm in (Crammer and Singer, 2001) solves the multi-class problem directly. In lib-svm, we use the one-against-all approach which is the default in lib-svm. Down-sampling is used if some result is trivial classification. The reason is that the unbalanced distribution of training data. We compared selecting support vectors and down-sampling. The latter is better.

5 Results

We participated in the subtask of SemEval-2007 English lexical sample task via English-Chinese parallel text. The organizers make use of English-Chinese documents gathered from the URL pairs given by the STRAND Bilingual Databases. They used this corpus for the evaluation of 40 English words (20 nouns and 20 adjectives).

Our system gives exactly one sense for each test example. So the recall is always the same as precision. Micro-average precision is 81.5%. According to the task organizers, the recall of the best participating in this subtask is 81.9%. So the performance of our system compares favorably with the best participating system.

6 Acknowledgements

This research is supported by Humanity and Social Science Research Project of China State Education Ministry (No. 06JC740001) and National Basic Research Program of China (No. 2004CB318102).

We are indebted to Helmut Schmid, IMS, University of Stuttgart, for making Tree-Tagger available free of charge.

Finally, the authors thank the organizers Hwee Tou Ng and Yee Seng Chan, for their hard work to collect the training and test data.

References

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

Gerald Chao and Michael G. Dyer. 2002. Maximum entropy models for word sense disambiguation. *Proceedings of the 19th international conference on Computational linguistics.*Vol (1):1-7

Koby Crammer and Yoram Singer. 2001. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *Journal of Machine Learning Research*, 2, 265-292

Radu Florian, Silviu Cucerzan, Charles Schafer and David Yarowsky. 2002. Combining Classifiers for Word Sense Disambiguation. *Natural Language Engineering*, 8(4): 327 – 341.

Thorsten Joachims. *SVM-Multiclass.* http://svmlight.joachims.org/svm-multiclass.html,2004.

Yoong Keok Lee, Hwee Tou Ng and Tee Kiah Chia, Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources. *Proceedings of SENSEVAL-3.* 137 - 140