SRCB-WSD: Supervised Chinese Word Sense Disambiguation with Key Features

Yun Xing
Ricoh Software Research Center Beijing Co., Ltd
Beijing, China
yun.xing@srcb.ricoh.com

Abstract

This article describes the implementation of Word Sense Disambiguation system that participated in the SemEval-2007 multilingual Chinese-English lexical sample task. We adopted a supervised learning approach with Maximum Entropy classifier. The features used were neighboring words and their part-of-speech, as well as single words in the context, and other syntactic features based on shallow parsing. In addition, we used word category information of a Chinese thesaurus as features for verb disambiguation. For the task we participated in, we obtained precision of 0.716 in micro-average, which is the best among all participated systems.

1 Introduction

Word Sense Disambiguation(WSD) is the process of assigning a meaning to a word based on the context in which it occurs. It is very important to many research fields such as Machine Translation, Information Retrieval. The goal of the multilingual Chinese-English lexical sample task in SemEval-2007 is to predict the correct English translation for an ambiguous Chinese word w.

We considered this task as a classification problem, and our system adopted a supervised learning approach with Maximum Entropy classifier, which is widely used in natural language processing(NLP). Within the Maximum Entropy framework, evidence from different features can be combined with no assumptions of feature independence. The used features include neighboring words and their part-of-speech(POS), single words in the context, and other syntactic features based on shallow parsing. In addition, we used word category information of a Chinese thesaurus for verb disambiguation. Note that we did not do any feature selection in this work.

Next, we will describe the Maximum Entropy framework and detail the features used in our WSD system.

2 Maximum Entropy

Maximum entropy modelling is a framework for integrating information from many heterogeneous information sources for classification (Manning and Schütze, 1999). It has been successfully applied to a wide range of NLP tasks, including sentence boundary detection, POS tagging, and parsing (Ratnaparkhi, 1998) . The system estimates the conditional probability that an ambiguous word has sense x given that it occurs in context y, where y is a conjunction of features. The estimated probability is derived from feature weights which are determined automatically from training data so as to produce a probability distribution that has maximum entropy, under the constraint that it is consistent with observed evidence (Dang et al., 2002). We used the implementation of Maximum Entropy framework with OpenNLP MAXENT[1], where each nominal feature was represented as "feature_code=value". Based on this framework, we defined the feature set and implemented the interface of feature extraction. For the convenient of evaluation, the default parameters

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

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 300–303,
Prague, June 2007. ©2007 Association for Computational Linguistics

of training model were used.

3 Used Features

Many research (Stevenson and Wilks, 2001; Lee and Ng, 2002) have indicated that a combination of knowledge sources improves WSD accuracy, but not any kind of knowledge source contributes the improvement of Chinese WSD (Dang et al., 2002). For multilingual Chinese-English lexical sample task, some basic features can be obtained directly. Also, we extracted other syntactic features through shallow parsing. In addition, we used word category information for verb disambiguation.

3.1 Basic Features

Since the data of multilingual Chinese-English lexical sample task are word-segmented and POS-tagged, we can get the following features directly.

- $W_{-1(+1)}$: the words (if any) immediately preceding and following w
- $P_{-1(+1)}$: the POS of the words(if any) immediately preceding and following w
- SW: single words in the context. We did not consider all words in the context as features for WSD, because our experiment shows that it will bring some noise in small scale supervised learning if we add all words in the context to feature set(See Section 4.1 for details). After carefully analyzing the POS set specification which is provided by task organizers, we only picked out words of POS listed in Table 1 as features.

3.2 Syntactic Features based on Shallow Parsing

To get further syntactic features from context, we implemented a simple rule-based parser to do shallow parsing on each instance. The parser only identifies phrases such as noun phrase, verb phrase, adjectival phrase, time phrase, position phrase and quantity phrase. These phrases are considered as constituents of context, as well as words and punctuations which do not belong to any phrase. Table 2 lists the constituent types and relative tags.

POS Tag	Specification
Ng	Nominal morpheme
n	Noun
nr	Personal name
ns	Place name
nt	Institution and Group
nz	Any other proper names
Vg	Verbal morpheme
v	Verb
vd	Verb with the attribute of adverb
vn	Verb with the attribute of noun
r	Pronoun
j	Abbreviation

Table 1: POS of single words in the context to be considered in our WSD system

For example, a word-segmented and POS-tagged instance in Figure 1 would be processed as a constituent list in Figure 2 after shallow parsing.

他/r 没有/d 说明/v 事情/n 的/u 真相/n 。/w

Figure 1: A word-segmented and POS-tagged instance. Note that the instance is not illustrated in XML format as data of multilingual Chinese-English lexical sample task, instead, it is illustrated in the form of "word/pos" for convenient.

他/entity 没有说明/action 事情的真相/entity 。/w

Figure 2: After shallow parsing, instance is organized in the form of "constituent/tag", that is, the word "他" is identified as an entity, and words "没有" and "说明" are merged together as an action.

Suppose C_0 is the constituent which the target word w belongs to , then we add following information to feature set:

- CT_0: the constituent tag of C_0
- $CT_{-i(+i)}, 0 < i \leq 3$: the tag of ith constituent to the left(right) of C_0
- $KCT_{-i(+i)}, 0 < i \leq 3$: the tag of ith constituent to the left(right) of C_0, and the type must be entity or action

Constituent type	Tag
noun phrase	entity
verb phrase	action
adjective phrase	adjective
time phrase	time
place phrase	place
quantity phrase	quantity
non-phrase	same as POS tag

Table 2: Constituent type and relative tag

- $LPOS_{-i(+i)}$: the POS of ith word in the same constituent of w.

3.3 Word Category Information

We considered word category information as an important knowledge source for verb disambiguation. The word category information comes from a Chinese thesaurus (Mei et al., 1983). If w is a verb, then the word category information of nouns in the right side of w is added into feature set. Figure 3 shows an example of how to use word category information for verb disambiguation.

他/r 坐/v 飞机/n 回/v 北京/ns

Figure 3: A word-segmented and POS-tagged instance of ambiguous verb "坐". The word category information of noun "飞机" has to be added into feature set.

Note that some nouns can belong to more than two categories, in this case, we do not use the word category information of this kind of noun for disambiguation.

Our experiment showed that this extra knowledge source did improve the accuracy of WSD (See 4.1 for detail).

4 Evaluation

Since the multilingual Chinese-English lexical sample task of SemEval-2007 is quite similar to the Chinese lexical sample task of SENSEVAL-3, we firstly evaluated feature set on the data of SENSEVAL-3 Chinese lexical sample task, and then gave the official SemEval-2007 scores of our system based on the best feature set.

Feature Set	Micro-average precision
FS1	0.630
FS2	0.635
FS3	0.654

Table 3: Result of feature set evaluation on SENSEVAL-3 test data

System	Micro-average precision	Macro-average precision
SRCB-WSD	0.716	0.749

Table 4: Official result on SemEval-2007 test data

4.1 Evaluation on SENSEVAL-3 Data

We did three experiments on the data of SENSEVAL-3 Chinese lexical sample task to evaluate if all the single words in the context should be included in feature set, and if the word category information of Chinese thesaurus is helpful for WSD. The first experiment used feature set (FS1) included almost the same features listed in Section 3.1 and 3.2, the only difference is that all single words in the context were considered. The second experiment used feature set (FS2) included all the features listed in Section 3.1 and 3.2. The third experiment used feature set (FS3) included all the features listed in Section 3.1, 3.2 and 3.3. The experimental result is given in Table 3. It shows that considering all single words in the context as features did not improve the performance of WSD, while word category information of Chinese thesaurus improved the accuracy obviously.

4.2 Official SemEval-2007 Scores

In multilingual Chinese-English lexical sample task, there are 2686 instances in training data for 40 Chinese ambiguous words. All these ambiguous words are noun or verb. Test data consist of 935 untagged instances of the same target words. The official result of our system in multilingual Chinese-English lexical sample task is reported in Table 4.

According to the task organizers, our system achieved the best performance out of all the participated systems.

5 Conclusion

In this paper, we described our participating system in the SemEval-2007 multilingual Chinese-English lexical sample task. We adopted Maximum Entropy method, and collected features not only from context provided by task organizers, but also from extra knowledge source. Evaluation results show that this feature set is much effective for supervised Chinese WSD.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments and suggestions.

References

Manning, C. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing.*. The MIT Press, Cambridge, Massachusetts.

Ratnaparkhi, A. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis University of Pennsylvania.

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

Mei, J.J., Li, Y.M., Gao, Y.Q. and et al. 1983. *Chinese thesaurus(Tongyici Cilin)*. Shanghai thesaurus Press.

Stevenson, M. and Wilks, Y. 2001. *The interaction of knowledge sources in word sense disambiguation*. Computational Linguistics, 27(3):321-349.

Lee, Y.K. and Ng, H.T. 2002. *An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), pages 41-48.