SemEval-2007 Task 5: Multilingual Chinese-English Lexical Sample

Peng Jin, Yunfang Wu and Shiwen Yu
Institute of Computational Linguistics
Peking University, Beijing China
{jandp, wuyf, yusw}@pku.edu.cn

Abstract

The Multilingual Chinese-English lexical sample task at SemEval-2007 provides a framework to evaluate Chinese word sense disambiguation and to promote research. This paper reports on the task preparation and the results of six participants.

1 Introduction

The Multilingual Chinese-English lexical sample task is designed following the leading ideas of the Senseval-3 Multilingual English-Hindi lexical sample task (Chklovski et al., 2004). The "sense tags" for the ambiguous Chinese target words are given in the form of their English translations.

The data preparation is introduced in the second section. And then the participating systems are briefly described and their scores are listed.

In the conclusions we bring forward some suggestion for the next campaign.

2 Chinese Word Sense Annotated Corpus

All the training and test data come from the People's Daily in January, February and March of 2000. The People's Daily is the most popular newspaper in China and is open domain. Before manually sense annotating, the texts have been word-segmented and part of speech (PoS) tagged according to the PoS tagging scheme of Institute of Computational Linguistics in Peking University (ICL/PKU). The corpus had been used as one of the gold-standard data set for the second international Chinese word segmentation bakeoff in 2005.[1]

2.1 Manual Annotation

The sense annotated corpus is manually constructed with the help of a word sense annotating interface developed in Java. Three native annotators, two major in Chinese linguistics and one major in computer science took part in the construction of the sense-annotated corpus. A text generally is first annotated by one annotator and then verified by two checkers. Checking is of course a necessary procedure to keep the consistency. Inspired by the observation that checking all the instances of a word in a specific time frame will greatly improve the precision and accelerate the speed, a software tool is designed in Java to gather all the occurrences of a word in the corpus into a checking file with the sense KWIC (Key Word in Context) format in sense tags order. The inter-annotator agreement gets to 84.8% according to Wu. et al. (2006).

The sense entries are specified in the Chinese Semantic Dictionary (CSD) developed by ICL/PKU. The sense distinctions are made mainly according to the Contemporary Chinese Dictionary, the most widely used dictionary in mandarin Chinese, with necessary adjustment and improvement is implemented according to words usage in real texts. Word senses are described using the feature-based formalism. The features, which appear in the form "Attribute =Value", can incorporate extensive distributional information about a word sense. The feature set constitutes the representation of a sense, while the verbal definitions of meaning

[1] http://sighan.cs.uchicago.edu/bakeoff2005/

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 19–23,
Prague, June 2007. ©2007 Association for Computational Linguistics

serve only as references for human use. The English translation is assigned to each sense in the attribute "English translation" in CSD.

Based on the sense-annotated corpus, a sense is replaced by its English translation, which might group different senses together under the same English word.

2.2 Instances selection

In this task together 40 Chinese ambiguous words: 19 nouns and 21 verbs are selected for the evaluation. Each sense of one word is provided at least 15 instances and at most 40 instances, in which around 2/3 is used as the training data and 1/3 as the test data. Table 1 presents the number of words under each part of speech, the average number of senses for each PoS and the number of instances respectively in the training and test set.

	# Average senses	# training instances	# test instances
19 nouns	2.58	1019	364
21 verbs	3.57	1667	571

Table 1: Summary of the sense inventory and number of training data and test set

In order to escape from the sense-skewed distribution that really exists in the corpus of People's Daily, many instances of some senses have been removed from the sense annotated corpus. So the sense distribution of the ambiguous words in this task does not reflect the usages in real texts.

3 Participating Systems

In order to facilitate participators to select the features, we gave a specification for the PoS-tag set. Both word-segmented and un-segmented context are provided.

Two kinds of precisions are evaluated. One is micro-average:

$$P_{mir} = \sum_{i=1}^{N} m_i / \sum_{i=1}^{N} n_i$$

N is the number of all target word-types. m_i is the number of labeled correctly to one specific target word-type and n_i is the number of all test instances for this word-type.

The other is macro-average:

$$P_{mar} = \sum_{i=1}^{N} p_i / N, \quad p_i = m_i / n_i$$

All teams attempted all test instances. So the recall is the same with the precision. The precision baseline is obtained by the most frequent sense. Because the corpus is not reflected the real usage, the precision is very low.

Six teams participated in this word sense disambiguation task. Four of them used supervised learning algorithms and two used un-supervised method. For each team two kinds of precision are given as in table 2.

Team	Micro-average	Macro-average
SRCB-WSD	0.716578	0.749236
I2R	0.712299	0.746824
CITYU-HIF	0.710160	0.748761
SWAT	0.657754	0.692487
TorMd	0.375401	0.431243
HIT	0.336898	0.395993
baseline	0.4053	0.4618

Table 2: The scores of all participating systems

As follow the participating systems are briefly introduced.

SRCB-WSD system exploited maximum entropy model as the classifier from OpenNLP[2] The following features are used in this WSD system:

- All the verbs and nouns in the context, that is, the words with tags "n, nr, ns, nt, nz, v, vd, vn"
- PoS of the left word and the right word
- noun phrase, verb phrase, adjective phrase, time phrase, place 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.

- the type of these phrases which are around the target phrases

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

• word category information comes from Chinese thesaurus

I2R system used a semi-supervised classification algorithm (label propagation algorithm) (Niu, et al., 2005). They used three types of features: PoS of neighboring words with position information, unordered single words in topical context, and local collocations.

In the label propagation algorithm (LP) (Zhu and Ghahramani, 2002), label information of any vertex in a graph is propagated to nearby vertices through weighted edges until a global stable stage is achieved. Larger edge weights allow labels to travel through easier. Thus the closer the examples, the more likely they have similar labels (the global consistency assumption). In label propagation process, the soft label of each initial labeled example is clamped in each iteration to replenish label sources from these labeled data. Thus the labeled data act like sources to push out labels through unlabeled data. With this push from labeled examples, the class boundaries will be pushed through edges with large weights and settle in gaps along edges with small weights. If the data structure fits the classification goal, then LP algorithm can use these unlabeled data to help learning classification plane.

CITYU-HIF system was a fully supervised one based on a Naïve Bayes classifier with simple feature selection for each target word. The features used are as follows:

- Local features at specified positions:
  PoS of word at w_{-2}, w_{-1}, w_1, w_2
  Word at w_{-2}, w_{-1}, w_1, w_2
- Topical features within a given window:
  Content words appearing within w_{-10} to w_{10}
- Syntactic features:
  PoS bi-gram at $w_{-2}w_0$, $w_{-1}w_0$, w_0w_1, w_0w_2
  PoS tri-gram at w_{-2} $w_{-1}w_0$ and $w_0w_1w_2$

One characteristic of this system is the incorporation of the intrinsic nature of each target word in disambiguation. It is assumed that WSD is highly lexically sensitive and each word is best characterized by different lexical information. Human judged to consider for each target word the type of disambiguation information if they found useful. During disambiguation, they run two Naïve Bayes classifiers, one on all features above, and the other only on the type of information deemed useful by the human judges. When the probability of the best guess from the former is under a certain threshold, the best guess from the latter was used instead.

SWAT system uses a weighted vote from three different classifiers to make the prediction. The three systems are: a Naïve Bayes classifier that compares similarities based on Bayes' Rule, a classifier that creates a decision list of context features, and a classifier that compares the angles between vectors of the features found most commonly with each sense. The features include bigrams, and trigrams, and unigrams are weighted by distance from the ambiguous word.

TorMd used an unsupervised naive Bayes classifier. They combine Chinese text and an English thesaurus to create a `Chinese word'--`English category' co-occurrence matrix. This system generated the prior-probabilities and likelihoods of a Naïve Bayes word sense classifier not from sense-annotated (in this case English translation annotated) data, but from this word--category co-occurrence matrix. They used the Macquarie Thesaurus as very coarse sense inventory.

They asked a native speaker of Chinese to map the English translations of the target words to appropriate thesaurus categories. Once the Naïve Bayes classifier identifies a particular category as the intended sense, the mapping file is used to label the target word with the corresponding English translation. They rely simply on the bag of words that co-occur with the target word (window size of 5 words on either side).

HIT is a fully unsupervised WSD system, which puts bag of words of Chinese sentences and the English translations of target ambiguous word to search engine (Google and Baidu). Then they could get all kinds of statistic data. The correct translation was found through comparing their cross entropy.

4 Conclusion

The goal of this task is to create a framework to evaluate Chinese word sense disambiguation and to promote research.

Target Word	Sense #	Training #	Test #	Base-line	Scores					
					SRCB-WSD	I2R	CITYU-HIF	SWAT-MP	TORMD	HIT
补	3	63	20	.50	.70	.80	.75	.75	.55	.55
成立	3	73	27	.370	.778	.815	.741	.778	.481	.407
吃	4	69	23	.435	.696	.609	.696	.696	.174	.174
出	9	222	77	.130	.506	.506	.481	.532	.169	.091
带	8	197	67	.150	.567	.552	.537	.433	.119	.104
动	4	58	20	.50	.60	.50	.55	.60	.30	.30
动摇	2	47	16	.625	.875	.875	.875	.563	.50	.438
发	5	105	36	.278	.694	.667	.611	.889	.25	.139
赶	3	56	18	.50	.667	.722	.667	.667	.389	.333
叫	4	106	39	.256	.718	.615	.641	.538	.256	.256
进	5	132	44	.227	.659	.75	.727	.568	.25	.114
开通	2	56	20	.50	.90	.95	.95	.60	.50	.50
看	4	103	34	.294	.765	.706	.765	.559	.294	.294
平息	2	20	8	.50	.75	.75	.75	.625	.375	.50
使	2	46	16	.625	.938	.813	.813	.875	.563	.438
说明	2	60	18	.556	.667	.722	.778	.722	.444	.556
挑	2	40	14	.429	.571	.643	.571	.571	.143	.286
推翻	2	29	10	.60	.80	.70	.90	.80	.30	.30
望	2	37	13	.769	.769	.769	.769	.769	.462	.462
想	4	110	37	.270	.730	.676	.676	.541	.216	.216
震惊	2	38	14	.714	.930	1.0	.929	.786	.714	.571
Ave.	3.57	1667	571	.342/.44	.685/.728	.676/.721	.671/.723	.618/.66	.30/.355	.263/.335

Table 3: Performance on verbs. Micro / macro average precisions are spitted by "/" at the last row.

Together six teams participate in this WSD task, four of them adopt supervised learning methods and two of them used unsupervised algorithms. All of the four supervised learning systems exceed obviously the baseline obtained by the most frequent sense. It is noted that the performances of the first three systems are very close. Two unsupervised methods' scores are below the baseline. More unlabeled data maybe improve their performance.

Although the SRCB-WSD system got the highest scores among the six participants, it does not perform always better than other system from table 2 and table 3. But to each word, the four supervised systems always predict correctly more instances than the two un-supervised systems.

Besides the corpus, we provide a specification of the PoS tag set. Only SRCB-WSD system utilized this knowledge in feature selection. We will provide more instances in the next campaign.

Target Word	Sense #	Training #	Test #	Base-line	Scores					
					SRCB-WSD	I2R	CITYU-HIF	SWAT-MP	TORMD	HIT
本	3	68	25	.40	.88	.84	.88	.76	.72	.32
表面	2	53	18	.611	.611	.722	.722	.833	.556	.333
菜	2	56	19	.526	.842	.842	.684	.789	.474	.632
长城	3	48	21	.476	.571	.591	.619	.619	.429	.619
单位	2	50	17	.588	.824	.824	.824	.647	.706	.529
道	3	53	18	.50	.778	.722	.778	.611	.50	.222
队伍	3	64	22	.455	.591	.591	.636	.545	.318	.364
儿女	2	60	20	.50	1.0	.95	1.0	1.0	.50	.50
机组	2	38	14	.714	1.0	1.0	1.0	1.0	.643	.571
镜头	2	45	15	.533	.733	.733	.60	.467	.467	.467
面	3	67	23	.435	.783	.783	.739	.696	.348	.696
牌子	2	44	17	.353	.529	.589	.588	.588	.353	.529
旗帜	3	50	18	.556	.611	.611	.722	.722	.50	.111
气息	2	39	14	.714	.929	.786	.714	.786	.857	.571
气象	2	47	16	.625	.813	.813	.938	1.0	.438	.563
日子	3	88	32	.313	.656	.563	.625	.656	.281	.344
天地	3	65	25	.40	.88	1.0	.92	.60	.56	.44
眼光	2	41	14	.714	.786	.714	.786	.643	.714	.50
中医	2	43	16	.625	.875	.938	1.0	.875	.438	.50
Ave.	2.45	1019	364	.506/.528	.766/.773	.761/.769	.772/.778	.72/.728	.50/.516	.456/.464

Table 4: Performance on nouns. Micro / macro average precisions are spitted by "/" at the last row.

5 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 would like to thank Tao Guo and Yulai Pei for their hard work to guarantee the quality of the corpus. Huiming Duan provides us the corpus which has been word-segmented and PoS-tagged and gives some suggestions during the manual annotation.

References

Rada Mihalcea, Timothy Chklovski and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. *Proceedings of SENSEVAL-3*. 25-28.

Timothy Chklovski, Rada Mihalcea, Ted Pedersen and Amruta Purandare. 2004. The Senseval-3 Multilingual English-Hindi lexical sample task. *Proceedings of SENSEVAL-3*. 5-8.

Xiaojin Zhu, Zoubin Ghahramani. 2002. Learning from Labeled and Unlabeled Data with Label Propagation. *CMU CALD tech report CMU-CALD-02-107.*

Yunfang Wu, Peng Jin, Yangsen Zhang, and Shiwen Yu. 2006. A Chinese Corpus with Word Sense Annotation. *Proceedings of ICCPOL*, Singapore, 414-421.

Zhen-Yu Niu, Dong-Hong Ji and Chew-Lim Tan. 2005. Word Sense Disambiguation Using Label Propagation Based Semi Supervised Learning. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.395-402