Aligning Features with Sense Distinction Dimensions

[1] Nianwen Xue, [2] Jinying Chen, [3] Martha Palmer
[1]CSLR and [3]Department of Linguistics
University of Colorado
Boulder, CO, 80309
{Nianwen.Xue,Martha.Palmer}@colorado.edu
[2] Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA, 19104
jinying@cis.upenn.edu

Abstract

In this paper we present word sense disambiguation (WSD) experiments on ten highly polysemous verbs in Chinese, where significant performance improvements are achieved using rich linguistic features. Our system performs significantly better, and in some cases substantially better, than the baseline on all ten verbs. Our results also demonstrate that features extracted from the output of an automatic Chinese semantic role labeling system in general benefited the WSD system, even though the amount of improvement was not consistent across the verbs. For a few verbs, semantic role information actually hurt WSD performance. The inconsistency of feature performance is a general characteristic of the WSD task, as has been observed by others. We argue that this result can be explained by the fact that word senses are partitioned along different dimensions for different verbs and the features therefore need to be tailored to particular verbs in order to achieve adequate accuracy on verb sense disambiguation.

1 Introduction

Word sense disambiguation, the determination of the correct sense of a polysemous word from a number of possible senses based on the context in which it occurs, is a continuing obstacle to high performance natural language processing applications. There are several well-documented factors that make accurate WSD particularly challenging. The first has to do with how senses are defined. The English data used for the SENSEVAL exercises, arguably the most widely used data to train and test WSD systems, are annotated based on very fine-grained distinctions defined in WordNet (Fellbaum, 1998), with human inter-annotator agreement at a little over seventy percent and the top-ranked systems' performances falling between 60%~70% (Palmer, *et al*., 2001; Mihalcea *et al.*, 2004). The second source of difficulty for accurate WSD comes from how senses are distributed. It is often the case that a polysemous word has a dominant sense or several dominant senses that occur with high frequency and not enough instances can be found for its low frequency senses in the currently publicly available data. There are on-going efforts to address these issues. For example, the sense annotation component of the OntoNotes project (Hovy, *et al.*, 2006) attempts to create a large-scale coarse-grained sense-annotated corpus with senses defined based on explicit linguistic criteria. These problems will be alleviated when resources like this are available to the general NLP community. There have already been experiments that show such coarse-grained senses lead to substantial improvement in system performance (Palmer *et al*, 2006).

The goal of our experiments is to explore the implications of a related and yet separate problem, specifically the extent to which the linguistic criteria used to define senses are related to what features need to be used in machine-learning systems. There are already published results that show WSD for different syntactic categories may need different types of features. For example, Yarowsky and Florian (2002), in their experiments on SENSEVAL2 English data, showed that sense distinctions of verbs relied more on linguistically motivated features than other parts-of-speech. In this paper,

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 921–928,
Sydney, July 2006.

we will go one step further and show that even for words of the same syntactic category senses are often defined along different dimensions based on different criteria. One direct implication of this observation for supervised machine-learning approaches to WSD is that the features have to be customized for different word categories, or even for different words of the same category. This supports previous arguments for word-specific feature design and parametric modeling for WSD tasks (Chen and Palmer, 2005; Hoste *et al*. 2002). We report experiments on ten highly polysemous Chinese verbs and show that features are not uniformly useful for all words.

The rest of the paper is organized as follows. In Section 2, we describe our WSD system, focusing on the features we used. We also briefly compare the features we use for Chinese with those used in a similar English WSD system. In Section 3, we present our experimental results and show that although rich linguistic features and features derived from a Chinese Semantic Role Labeling improve the WSD accuracy, the improvement is not uniform across all verbs. We show that this lack of consistency is due to the different dimensions along which the features are defined. In Section 4, we discuss related work. Finally Section 5 concludes this paper and describes future directions.

2 WSD System for Chinese Verbs

Our WSD system uses a smoothed maximum entropy (MaxEnt) model with a Gaussian prior (McCallum, 2002) for learning Chinese verb senses. The primary reason is that the MaxEnt model provides a natural way for combining different features without the assumption of feature independence. Furthermore, smoothing the MaxEnt model with a Gaussian prior is better than other smoothing methods at alleviating the overfitting problem caused by low frequency features (Chen *et al.*, 1999). This model has been applied successfully for English WSD (Dang, 2004; Chen and Palmer, 2005).

The features used by our Chinese WSD system include:

Collocation Features

- Previous and next word (relative to the target verb), w_{-1} and w_1 and their parts-of-speech p_{-1} and p_1

Syntactic Features

- Whether the target verb takes a direct object (i.e., in a transitive use)
- Whether the verb takes a sentential complement
- Whether the verb, if it consists of a single character, occurs at the last position of a compound verb

Semantic Features

- The semantic role information about the verbs
- The semantic categories for the verb's NP arguments from a general Chinese noun Taxonomy

All of these features require some level of preprocessing of the Chinese raw text, which comes without word boundaries. To extract the collocation features the raw text needs to be segmented and POS-tagged; to extract the syntactic and semantic features, the Chinese text needs to be parsed. We use an integrated parser that does segmentation, POS-tagging and parsing in one step. Since part of the sense-tagged data comes from the Chinese Treebank that the parser is trained on, we divide the Chinese Treebank into nine equal-sized portions and parse each portion with a parsing model trained on the other eight portions so that the parser has not seen any of the data it parses. The data that is not from the Chinese Treebank is parsed with a parsing model trained on the entire Chinese Treebank. The parser produces a segmented, POS-tagged and parsed version of the same text to facilitate the extraction of the different types of features. The extraction of the semantic role labels as features requires the use of a semantic role tagger, which we describe in greater detail in Section 2.2.

In addition to using the semantic role labeling information, we also extract another type of semantic features from the verb's NP arguments. These features are top-level semantic categories from a three-level general taxonomy for Chinese nouns, which was created semi-automatically based on two Chinese semantic dictionaries (Chen and Palmer, 2004).

2.1 A Comparison with Our English WSD System

Similar to our English WSD system, which achieved the best published results on SENSEVAL2 English verbs for both fine-grained and coarse-grained senses (Chen and Palmer, 2005), our Chinese WSD system uses the same smoothed MaxEnt machine learning model and linguistically motivated features for Chinese verb sense disambiguation. However, the features used in the two systems differ

somewhat due to the different properties of the two languages .

For example, our English system uses the inflected form and the part-of-speech tag of the target verb as feature. For Chinese we no longer use such features since Chinese words, unlike English ones, do not contain morphology that marks tense.

The collocation features used by our English system include bi-grams and tri-grams of the words that occur within two positions before or after the target verb and their part-of-speech tags. In contrast, our Chinese system extracts collocation features from a narrower, three-word window, with one word immediately before and after the target verb. This decision was made based on two observations about the Chinese language. First, certain single-character Chinese verbs, such as the verbs "出|chu", "开|kai" and "成|cheng" in our experiments, often form a compound with a verb to its immediate left. That verb is often a good indicator of the sense of this verb. An example is given in (1):

(1) 辽宁 已 呈现 **出**
Liaoning already show **completion**

多元化 发展 趋势 。
multidimensional development trend

"Liaoning Province has shown the trend of multidimensional development."

Being the last word of a verb compound is a strong indicator for Sense 8 of the verb "出|chu1" (used after a verb to indicate direction or aspect), as in "呈现|cheng2xian4 出|chu1".

Second, unlike English common nouns that often require determiners such as *the*, *a* or *an*, Chinese common nouns can stand alone. Therefore, the direct object of a verb often occurs right after the verb in Chinese, as shown in (2).

（2）动员 群众 勒紧 腰带 集
mobilize people tighten waistband collect

资 **开** ***公路*** (direct object)。
funds **build** ***highway***

"Mobilize people to tighten their waistbands (i.e., save money) in order to collect funds to build highways."

Based on these observations, we use words surrounding the target verb and their part-of-speech tags as collocation features. A further investigation on the different sizes of the context window (3,5,7,9,11) showed that increasing the window size decreased our system's accuracy.

2.2 Features Based on Automatic Semantic Role Tagging

In a recent paper on the WSD of English verbs, Dang and Palmer (2005) showed that semantic role information significantly improves the WSD accuracy of English verbs for both fine-grained and coarse-grained senses. However, this result assumes the human annotation of the Penn English Propbank (Palmer *et al*, 2005). It seems worthwhile to investigate whether the semantic role information produced by a fully automatic Semantic Role tagger can improve the WSD accuracy on verbs, and test the hypothesis that the senses of a verb have a high correlation to the arguments it takes. To that end, we assigned semantic role labels to the arguments of the target verb with a fully automatic semantic role tagger (Xue and Palmer, 2005) trained on the Chinese Propbank (CPB) (Xue and Palmer, 2003), a corpus annotated with semantic role labels that are similiar in style to the Penn English Propbank. In this annotation, core arguments such as agent or theme are labeled with numbered arguments such as *Arg0* and *Arg1*, up to *Arg5* while adjunct-like elements are assigned functional tags such as *TMP* (for temporal), *MNR*, prefixed by *ArgM*. The Semantic Role tagger takes as input syntactic parses produced by the parser described above as input and produces a list of arguments for each of the sense-tagged target verbs and assigns argument labels to them. Features are extracted from both the core arguments and adjuncts of the target verb. In addition to providing the sematnic role labels (e.g., *Arg0* and *Arg1*) of the extracted core arguments, the Semantic Role tagger also provides Hownet (Dong and Dong, 1991) semantic categories associated with these arguments. (3) shows the arguments for the target verb "打" identified by the Semantic Role tagger:

(3) [$_{\text{ArgM-MNR}}$ 经过 三 年 苦 干] ,
through three year hard work,

[arg0 全 乡] [$_{\text{rel}}$ **打**] 成
whole county **dig** finish

[$_{\text{Arg1}}$ 深 水井] 三 眼 。
deep well three classifier

"The whole county finished digging three deep wells through 3 years of hard work."

Based on the output of the Semantic Role tagger and the Chinese noun taxonomy (as described in Section 2.1), the following features are extracted:

SRL+lex	SRL+HowNet	SRL+Taxonomy
ARG1-水井	ARG1-设施	ARG1_location
ARG0-乡	ARG0-地方	ARG0_location
ARGM|MNR-经过	ARGM|MNR-经受	ARGM|MNR

In this example, semantic role related features include: (1) the head word of the core arguments (ARG1-水井 and ARG0-乡) and the adjunct (ARGM|MNR-经过); (2) the HowNet semantic category for the head word (ARG1-设施, ARG0-地方, ARGM|MNR-经受); (3) the semantic role label of the adjunct (ARGM|MNR); and (4) the top level semantic category from the taxonomy of Chinese nouns for the head word of the NP arguments (ARG1_location and ARG0_location).

3 Experimental Results

The data we used for our experiments are developed as part of the OntoNotes project (Hovy *et al.*, 2006) and they come from a variety of sources. Part of the data is from the Chinese Treebank (Xue *et al*, 2005), which has a combination of Xinhua news and Sinorama News Magazine. Since some verbs have an insufficient number of instances for any meaningful experiments, we also annotated portions of the People's Daily corpus, developed by Peking University. We chose not to use the Chinese WSD dataset used in Senseval 3 [1] because we are mainly interested in investigating how the features used in WSD are related to the criteria used to define the senses of Chinese verbs. The Chinese Senseval dataset includes both nouns and verbs. In addition, the criteria used to define their senses are not made explicit and therefore are not clear to us.

Table 1 summarizes the corpus statistics and the experimental results for the 10 highly polysemous Chinese verbs used in our experiments. The results were obtained by using 5-fold cross validation. The top five verbs are verbs that were identified as difficult verbs in Dang *et al*'s (2002) experiments. The first three columns show the verbs (and their pinyin), the number of instances and the number of senses for each verb in the data. The fourth column shows the sense entropy for each verb in its test data, as calculated in Equation 1.

$$-\sum_{i=1}^{n} P(sense_i)\log P(sense_i) \quad (1)$$

Where n is the number of senses of a verb in our data; $P(sense_i)$ is the probability of the *ith* sense of the verb, which is estimated based on the frequency count of the verb's senses in the data. Sense entropy generally reflects the frequency distribution of senses in the corpus. A verb with an evenly distributed sense distribution tends to have a high entropy value. However, a verb can also have a high sense entropy simply because it is highly polysemous (say, has 20 or more senses) even though the sense distribution may be skewed, with one or two dominant senses. To separate the effects of the number of senses, we also use a normalized sense entropy metric (the sixth column in Table 1), as calculated in Equation 2.

$$\frac{-\sum_{i=1}^{n} P(sense_i)\log P(sense_i)}{-\sum_{i=1}^{n} \frac{1}{n}\log P(\frac{1}{n})} \quad (2)$$

Here a large sense number n corresponds to a high value for the normalization factor $-\sum_{i=1}^{n} \frac{1}{n}\log P(\frac{1}{n})$. Therefore, normalized sense entropy can indicate sense frequency distribution more precisely than sense entropy.

Table 1 (Columns 7 to 10) also shows the experimental results. As we can see, on average, our system achieved about 19% improvement (absolute gain) in accuracy compared to the most frequent sense baseline. Its performance is consistently better than the baseline for all 10 verbs.

3.1 Corpus Statistics and Disambiguation Accuracy

The data in Table 1 shows that verbs with a high normalized sense entropy have the low frequency baselines. Furthermore, this relation is stronger than that between un-normalized sense entropy and the baseline. However, sense entropy is a better predictor for system performance than normalized sense entropy. The reason is intuitive: unlike the baseline, the automatic WSD system, trained on the training data, does not only rely on sense frequency information to predict senses.

[1] http://www.senseval.org/senseval3

	# of instance	# of sense	sense entropy	norm. sense entropy	baseline	all feat	all-SRL
出|chu	271	11	1.12	0.47	74.54	79.70	78.59
恢复|huifu	113	3	0.93	0.84	50.44	69.91	72.57
见|jian	167	7	1.01	0.52	72.46	82.63	82.03
想|xiang	231	6	1.00	0.56	65.80	76.19	77.49
要|yao	254	9	1.56	0.71	33.46	46.46	49.21
成|cheng	161	8	1.38	0.67	43.48	73.29	72.67
打|da	313	21	2.29	0.75	20.77	45.05	32.59
开|kai	382	18	2.31	0.80	19.37	50.00	39.27
通过|tongguo	384	4	0.97	0.70	55.73	81.51	79.17
发展|fazhan	1141	7	0.88	0.45	74.76	79.58	77.56
average		9.4			51.08	70.18	67.13
total	3417						

Table 1 Corpus Statistics and Experimental Results for the 10 Chinese Verbs

The number of senses has a direct impact on how many training instances exist for each verb sense. As a consequence, it is more difficult for the system to make good generalizations from the limited training data that is available for highly polysemous verbs. Therefore, sense entropy, which is based on both sense frequency distribution and polysemy is more appropriate for predicting system accuracy. A related observation is that the system gain (compared with the baseline) is bigger for verbs with a high normalized sense entropy, such as "恢复|huifu", "打|da", "开|kai", and "通过|tongguo", than for other verbs; and the system gain is very small for verbs with low normalized sense entropy and a relatively large number of senses, such as "出|chu" and "发展|fazhan", since they already have high baselines.

3.2 The Effect of Semantic Role Features

When Semantic Role information is used in features, the system's performance on average improves 3.05%, from 67.13% to 70.18% compared with when the features derived from the Semantic Role information is not used. If we look at the system's performance on individual verbs, the results show that adding Semantic Role information as features improves the accuracy of 7 of the 10 verbs. For the remaining 3 verbs, adding semantic role information actually hurts the system's performance. We believe this apparent inconsistency can be explained by looking at how senses are defined for the different verbs. The two verbs that present the most challenge to the system, are "打|da" and "要|yao" While Semantic Role features substantially improve the accuracy of "打|da", they actually hurt the accuracy of "要|yao". For "要|yao", its three most frequent senses account for 86% of its total instances (232 out of 270) and they are the "intend to", "must, should" and "need" senses:

(4) Three most frequent senses of "要|yao"

(a) 双方 表示 要 进一步 合作 。
two sides indicate intend further cooperation

"The two sides indicated that they intended to step up their cooperation."

(b) 路 很 滑， 大家 要 小心 。
road very slippery, everybody should careful

"The road is slippery. Everybody should be careful."

(c) 苏钢 每 年 要 靠
Suzhou Steel Works every year need depend

大运河 运输 原料 。
the Great Canal transport raw material

"Suzhou Steel Works needs to depend on the Great Canal to transport raw material."

Two of the senses, "must" and "need", are used as auxiliary verbs. As such, they do not take arguments in the same way non-auxiliary verbs do. For example, they do not take noun phrases as arguments. As a result, the Semantic Role tagger, which assigns argument labels to head words of noun phrases or clauses, cannot produce a meaningful argument for an auxiliary verb. For the "intend to" sense, even if it is not

an auxiliary verb, it still does not take a noun phrase as an object. Instead, its object is a verb phrase or a clause, depending on the analysis. The correct head word of its argument should be the lower verb, which apparently is not a useful discriminative feature either.

In contrast, the senses of "打|da" are generally defined based on their arguments. The three most frequent senses of "打|da" are "call by telephone", "play" and "fight" and they account for 40% of the "打|da" instances. Some examples are provided in (5)

(5) Top three senses of "打|da"

(a) 你 有过 排长 打
you have queue in long line call

公共 电话 的 经验 吗 ?
public phone DE experience ma

"Do you have the experience of queuing in a line and waiting to make a call with a public phone?"

(b) 几个 值班 人员 围坐
a few on duty personnel sit

一 圈 打 扑克 。
one circle play poker

"A few of the personnel on duty were sitting in a circle and playing poker."

(c) 动员 全 社会 力量 打好
mobilize whole society power fight

扶贫 攻坚 战 。
helping the poor crucial battle

"…mobilize the power of the whole society and fight the crucial battle of helping the poor."

The senses of "打|da" are to a large extent determined by its PATIENT (or *Arg1*) argument, which is generally realized in the object position. The *Arg1* argument usually forms highly coherent lexical classes. For example, the *Arg1* of the "call" sense can be "电话|dianghua/phone, "手机|shouji/cellphone", etc. its *Arg1* argument can be "蓝球|langqiu/basketball", "桥牌|qiaopai/bridge", "游戏|youxi/game", etc for the "play" sense, Finally , for its sense "fight", the *Arg1* argument can be "攻坚|gongjian/crucial 战|zhan/battle", "巷战|xiangzhang/street warfare", "游击战|youjizhan/guerilla warfare", etc.. It's not surprising that recognizing the arguments of "打|da" is crucial in determining its sense.

The accuracy for both verbs is still very low, but for very different reasons. In the case of "要|yao4", the challenge is identifying discriminative features that may not be found in the narrow local context. These could for instance include discourse features. In the case of "打|da', one important reason why the accuracy is still low is because "打|da" is highly polysemous and has over forty senses. Given its large number of senses, the majority of its senses do not have enough instances to train a reasonable model. We believe that more data will improve its WSD accuracy.

There are other dimensions along which verb senses are defined in addition to whether or not a verb is an auxiliary verb and what type of auxiliary verb it is, and what types of arguments it takes. One sense of "出|chu" is a verb particle that indicates the direction or aspect of the main verb that generally immediately precedes it. In this case the most important feature for identifying this sense is the collocation feature.

Our experimental results seem to lend support to a WSD approach where features are tailored to each target word, or at least each class of words, based on a careful analysis of the dimensions along which senses are defined. Automatic feature selection (Blum and Langley, 1997) could also prove useful in providing this type of tailoring. An issue that immediately arises is the feasibility of this approach. At least for Chinese, the task is not too daunting, as the number of highly polysemous verbs is small. Our estimation based on a 250K-word chunk of the Chinese Treebank and a large electronic dictionary in our possession shows only 6% or 384 verb types having four or more definitions in the dictionary. Even for these verbs, the majority of them are not difficult to disambiguate, based on work by Dang *et al*. (2002). Only a small number of these verbs truly need customized features.

4 Related work

There is a large body of literature on WSD and here we only discuss a few that are most relevant to our work. Dang and Palmer (2005) also use predicate-argument information as features in their work on English verbs, but their argument labels are not produced by an automatic SRL system. Rather, their semantic role labels are directly extracted from a human annotated

corpus, the English Proposition Bank (Palmer *et al*, 2005), citing the inadequate accuracy of automatic semantic role labeling systems. In contrast, we used a fully antomated SRL system trained on the Chinese Propbank. Nevertheless, their results show, as ours do, that the use of semantic role labels as features improves the WSD accuracy of verbs.

There are relatively few attempts to use linguistically motivated features for Chinese word sense disambiguation. Niu *et al* (2004) applied a Naive Bayesian model to Chinese WSD and experimented with different window sizes for extracting local and topical features and different types of local features (e.g., bigram templates, local words with position or parts-of-speech information). One basic finding of their experiments is that simply increasing the window size for extracting local features or enriching the set of local features does not improve disambiguation performance. This is consistent with our usage of a small size window for extracting bigram collocation features. Li *et al.* (2005) used sense-tagged *true* bigram collocations [2] as features. These features were obtained from a collocation extraction system that used lexical co-occurrence statistics to extract candidate collocations and then selected *true* collocations by using syntactic dependencies (Xu *et al*., 2003). In their experiments on Chinese nouns and verbs extracted from the People's Daily News and the SENSEVAL3 data set, the Naive Bayesian classifier using *true* collocation features generally performed better than that using simple bigram collocation features (i.e., bigram co-occurence features). It is worth noting that the *true* collocations overlap to a large degree with rich syntactic information used here such as the subject and direct object of a target verb. Therefore, their experiments show evidence that rich linguistic information benefits WSD on Chinese, consistent with our results.

Our work is more closely related to the work of Dang *et al* (2002), who conducted experiments on 28 verbs and achieved an accuracy of 94.2%. However the high accuarcy is largely due to the fact that their verbs are randomly chosen from the Chinese Treebank and some of them are not even polysemous (having a single sense). Extracting features from the gold standard parses also contributed to the high accuracy, although not by much. For 5 of their 28 verbs, their initial experimental results did not break the most frequent sense baseline. They annotated additional data on those five verbs and their system trained on this new data did outperfom the baseline. However, they concluded that the contribution of linguistic motivated features, such as features extracted from a syntactic parse, is insignificant, a finding they attributed to unique properties of Chinese given that the same syntactic features significantly improves the WSD accuracy. Our experimental results show that this conclusion is premature, without a detailed analysis of the senses for the individual verbs.

[2] In their definition, a collocation is a recurrent and conventional fixed expression of words that holds syntactic and semantic relations.

5 Conclusion and future work

We presented experiments with ten highly polysemous Chinese verbs and showed that a previous conclusion that rich linguistic features are not useful for the WSD of Chinese verbs is premature. We demonstrated that rich linguistic features, specifically features based on syntactic and semantic role information, are useful for the WSD of Chinese verbs. We believe that the WSD systems can benefit even more from rich linguistic features as the performance of other NLP tools such as parsers and Semantic Role Taggers improves. Our experimental results also lend support to the position that feature design for WSD should be linked tightly to the study of the criteria that sense distinctions are based on. This position calls for the customization of features for individual verbs based on understanding of the dimensions along which sense distinctions are made and a closer marriage between machine learning and linguistics. We believe this represents a rich area of exploration and we intend to experiment with more verbs with further customization of features, including experimenting with automatic feature selection.

Acknowledgement

This work was supported by National Science Foundation Grant NSF-0415923, Word Sense Disambiguation, the DTO-AQUAINT NBCHC-040036 grant under the University of Illinois subcontract to University of Pennsylvania 2003-07911-01 and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do

not necessarily reflect the views of the National Science Foundation, the DTO, or DARPA.

References

Avrim L. Blum and Pat Langley. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245-271, 1997.

Jinying Chen and Martha Palmer. 2004. Chinese Verb Sense Discrimination Using an EM Clustering Model with Rich Linguistic Features, In Proc. of the 42nd Annual meeting of the Assoication for Computational Linguistics, ACL-04. July 21-24, Barcelona, Spain

Jinying Chen and Martha Palmer. 2005. Towards Robust High Performance Word Sense Disambiguation of English Verbs Using Rich Linguistic Features. In Proc. of the 2nd International Joint Conference on Natural Language Processing. Jeju Island, Korea, in press.

Stanley. F. Chen and Ronald Rosenfeld. 1999. *A Gaussian Prior for Smoothing Maximum Entropy Modals*. Technical Report CMU-CS-99-108, CMU.

Hoa T. Dang, Ching-yi Chia, Martha Palmer and Fu-Dong Chiou. 2002. Simple Features for Chinese Word Sense Disambiguation. In *Proceedings of COLING-2002, the Nineteenth Int. Conference on Computational Linguistics*, Taipei, Aug.24–Sept.1.

Hoa T. Dang. 2004. *Investigations into the role of lexical semantics in word sense disambiguation*. PhD Thesis. University of Pennsylvania.

Hoa Dang and Martha Palmer. 2005. The role of semantic roles in disambiguating verb senses. In *Proceedings of ACL-05*, Ann Arbor, Michigan.

Zhendong Dong and Qiang Dong, HowNet. 1991. http://www.keenage.com.

Christiane Fellbaum, ed. 1998. *WordNet: An* Electronic *Lexical Database.* Cambridge, MA: MIT Press.

Veronique Hoste, Iris Hendrickx, Walter Daelemans, and Antal van den Bosch. 2002. Parameter optimization for machine-learning of word sense disambiguation. *NLE, Special Issue on Word Sense Disambiguation Systems*, 8(4):311–325.

Eduard Hovy, Mtchchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In Proceedings of the HLT-NAACL 2006, New York City.

Wanyin Li, Qin Lu and Wenjie Li. 2005. Integrating Collocation Features in Chinese Word Sense Disambiguation. In Proceedings of the Fourth Sighan Workshop on Chinese Language Processing. pp: 87-94. Jeju, Korea.

Andrew K. McCallum: MALLET: A Machine Learning for Language Toolkit. http://www.cs.umass.edu/~mccallum/mallet (2002).

Rada Mihalcea, Timothy Chklovski and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain. July.

Zheng-Yu Niu, Dong-Hong Ji and Chew Lim Tan, Optimizing Feature Set for Chinese Word Sense Disambiguation. 2004. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*. Barcelona, Spain.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, 21-24.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 31(1): 71–106.

Martha Palmer, Christiane Fellbaum and Hoa Trang Dang. (to appear, 2006). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*.

Ruifeng Xu , Qin Lu, and Yin Li. 2003. An automatic Chinese Collocation Extraction Algorithm Based On Lexical Statistics. In *Proceedings of the NLPKE Workshop*. Beijing, China.

Nianwen Xue, Fei Xia, Fu-Dong Chiou and Martha Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207-238.

Nianwen Xue and Martha Palmer. 2003. Annotating Propositions in the Penn Chinese Treebank, In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, in conjunction with ACL'03. Sapporo, Japan.

Nianwen Xue and Martha Palmer. 2005. Automatic Semantic Role Labeling for Chinese Verbs. In Proceedings of the 19th International Joint Conference on Artificial Intelligence. Edinburgh, Scotland.

David Yarowsky and Radu Florian. 2002. Evaluating sense disambiguation across diverse parameter spaces. *Journal of Natural Language Engineering,* 8(4): 293–310.