Simple Features for Chinese Word Sense Disambiguation

Hoa Trang Dang, Ching-yi Chia, Martha Palmer, and Fu-Dong Chiou
Department of Computer and Information Science
University of Pennsylvania
{htd,chingyc,mpalmer,chioufd}@unagi.cis.upenn.edu

Abstract

In this paper we report on our experiments on automatic Word Sense Disambiguation using a maximum entropy approach for both English and Chinese verbs. We compare the difficulty of the sense-tagging tasks in the two languages and investigate the types of contextual features that are useful for each language. Our experimental results suggest that while richer linguistic features are useful for English WSD, they may not be as beneficial for Chinese.

1 Introduction

Word Sense Disambiguation (WSD) is a central open problem at the lexical level of Natural Language Processing (NLP). Highly ambiguous words pose continuing problems for NLP applications. They can lead to irrelevant document retrieval in Information Retrieval systems, and inaccurate translations in Machine Translation systems (Palmer et al., 2000). For example, the Chinese word 见(jian4) has many different senses, one of which can be translated into English as "see", and another as "show". Correctly sense-tagging the Chinese word in context can prove to be highly beneficial for lexical choice in Chinese-English machine translation.

Several efforts have been made to develop automatic WSD systems that can provide accurate sense tagging (Ide and Veronis, 1998), with a current emphasis on creating manually sense-tagged data for supervised training of statistical WSD systems, as evidenced by SENSEVAL-1 (Kilgarriff and Palmer, 2000) and SENSEVAL-2 (Edmonds and Cotton, 2001). Highly polysemous verbs, which have several distinct but related senses, pose the greatest challenge for these systems (Palmer et al., 2001). Predicate-argument information and selectional restrictions are hypothesized to be particularly useful for disambiguating verb senses.

Maximum entropy models can be used to solve any classification task and have been applied to a wide range of NLP tasks, including sentence boundary detection, part-of-speech tagging, and parsing (Ratnaparkhi, 1998). Assigning sense tags to words in context can be viewed as a classification task similar to part-of-speech tagging, except that a separate set of tags is required for each vocabulary item to be sense-tagged. Under the maximum entropy framework (Berger et al., 1996), evidence from different features can be combined with no assumptions of feature independence. The automatic tagger estimates the conditional probability that a 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. With existing tools for learning maximum entropy models, the bulk of our work is in defining the types of features to look for in the data. Our goal is to see if sense-tagging of verbs can be improved by combining linguistic features that capture information about predicate-arguments and selectional restrictions.

In this paper we report on our experiments on automatic WSD using a maximum entropy approach for both English and Chinese verbs. We compare the difficulty of the sense-tagging tasks in the two languages and investigate the types of contextual features that are useful for each language. We find that while richer linguistic features are useful for English WSD, they do not prove to be as beneficial for Chinese.

The maximum entropy system performed competitively with the best systems on the English verbs in SENSEVAL-1 and SENSEVAL-2 (Dang and Palmer, 2002). However, while SENSEVAL-2 made it possible to compare many different approaches

over many different languages, data for the Chinese lexical sample task was not made available in time for any systems to compete. Instead, we report on two experiments that we ran using our own lexicon and two separate Chinese corpora that are very similar in style (news articles from the People's Republic of China), but have different types and levels of annotation – the Penn Chinese Treebank (CTB)(Xia et al., 2000), and the People's Daily News (PDN) corpus from Beijing University. We discuss the utility of different types of annotation for successful automatic word sense disambiguation.

2 English Experiment

Our maximum entropy WSD system was designed to combine information from many different sources, using as much linguistic knowledge as could be gathered automatically by current NLP tools. In order to extract the linguistic features necessary for the model, all sentences were first automatically part-of-speech-tagged using a maximum entropy tagger (Ratnaparkhi, 1998) and parsed using the Collins parser (Collins, 1997). In addition, an automatic named entity tagger (Bikel et al., 1997) was run on the sentences to map proper nouns to a small set of semantic classes.

Chodorow, Leacock and Miller (Chodorow et al., 2000) found that different combinations of topical and local features were most effective for disambiguating different words. Following their work, we divided the possible model features into topical features and several types of local contextual features. Topical features looked for the presence of keywords occurring *anywhere* in the sentence and any surrounding sentences provided as context (usually one or two sentences). The set of 200-300 keywords is specific to each lemma to be disambiguated, and is determined automatically from training data so as to minimize the entropy of the probability of the senses conditioned on the keyword.

The local features for a verb w in a particular sentence tend to look only within the smallest clause containing w. They include *collocational* features requiring no linguistic preprocessing beyond part-of-speech tagging (1), *syntactic* features that capture relations between the verb and its complements (2-4), and *semantic* features that incorporate information about noun classes for subjects and objects (5-6):

1. the word w, the part of speech of w, the part of speech of words at positions -1 and +1 relative to w, and words at positions -2, -1, +1, +2, relative to w
2. whether or not the sentence is passive
3. whether there is a subject, direct object, indirect object, or clausal complement (a complement whose node label is S in the parse tree)
4. the words (if any) in the positions of subject, direct object, indirect object, particle, prepositional complement (and its object)
5. a Named Entity tag (PERSON, ORGANIZATION, LOCATION) for proper nouns appearing in (4)
6. WordNet synsets and hypernyms for the nouns appearing in (4)

2.1 English Results

The maximum entropy system's performance on the verbs from the evaluation data for SENSEVAL-1 (Kilgarriff and Rosenzweig, 2000) rivaled that of the best-performing systems. We looked at the effect of adding topical features to local features that either included WordNet class features or used just lexical and named entity features. In addition, we experimented to see if performance could be improved by undoing passivization transformations to recover underlying subjects and objects. This was expected to increase the accuracy with which verb arguments could be identified, helping in cases where selectional restrictions on arguments played an important role in differentiating between senses.

The best overall variant of the system for verbs did not use WordNet class features, but included topical keywords and passivization transformation, giving an average verb accuracy of 72.3%. If only the best combination of feature sets for each verb is used, then the maximum entropy models achieve 73.7% accuracy. These results are not significantly different from the reported results of the best-performing systems (Yarowsky, 2000). Our system was competitive with the top performing systems even though it used only the training data provided and none of the information from the dictionary to identify multi-word constructions. Later experiments show that the ability to correctly identify multi-word constructions improves performance substantially.

We also tested the WSD system on the verbs from the English lexical sample task for SENSEVAL-2.[1]

[1]The verbs were: begin, call, carry, collaborate, develop,

Feature Type (local only)	Accuracy	Feature Type (local and topical)	Accuracy
collocation	48.3	collocation	52.9
+ syntax	53.9	+ syntax	54.2
+ syntax + semantics	59.0	+ syntax + semantics	60.2

Table 1: Accuracy of maximum entropy system using different subsets of features for SENSEVAL-2 verbs.

In contrast to SENSEVAL-1, senses involving multi-word constructions could be identified directly from the sense tags themselves, and the head word and satellites of multi-word constructions were explicitly marked in the training and test data. This additional annotation made it much easier to incorporate information about the satellites, without having to look at the dictionary (whose format may vary from one task to another). All the best-performing systems on the English verb lexical sample task filtered out possible senses based on the marked satellites, and this improved performance.

Table 1 shows the performance of the system using different subsets of features. In general, adding features from richer linguistic sources tended to improve accuracy. Adding syntactic features to collocational features proved most beneficial in the absence of topical keywords that could detect some of the complements and arguments that would normally be picked up by parsing (complementizers, prepositions, etc.). And while topical information did not always improve results significantly, syntactic features along with semantic class features always proved beneficial.

Incorporating topical keywords as well as collocational, syntactic, and semantic local features, our system achieved 60.2% and 70.2% accuracy using fine-grained and coarse-grained scoring, respectively. This is in comparison to the next best-performing system, which had fine- and coarse-grained scores of 57.6% and 67.2% (Palmer et al., 2001). If we had not included a filter that only considered phrasal senses whenever there were satellites of multi-word constructions marked in the test data, our fine- and coarse-grained accuracy would have been reduced to 57.5% and 67.2% (significant at $p = 0.050$).

3 Chinese Experiments

We chose 28 Chinese words to be sense-tagged. Each word had multiple verb senses and possibly other senses for other parts of speech, with an average of 6 dictionary senses per word. The first 20 words were chosen by randomly selecting several files totaling 5000 words from the 100K-word Penn Chinese Treebank, and choosing only those words that had more than one dictionary verb sense and that occurred more than three times in these files. The remaining 8 words were chosen by selecting all words that had more than one dictionary verb sense and that occurred more than 25 times in the CTB. The definitions for the words were based on the CETA (Chinese-English Translation Assistance) dictionary (Group, 1982) and other hard-copy dictionaries. Figure 1 shows an example dictionary entry for the most common sense of jian4. For each word, a sense entry in the lexicon included the definition in Chinese as well as in English, the part of speech for the sense, a typical predicate-argument frame if the sense is for a verb, and an example sentence. With these definitions, each word was independently sense-tagged by two native Chinese-speaking annotators in a double-blind manner. Sense-tagging was done primarily using raw text, without segmentation, part of speech, or bracketing information. After finishing sense tagging, the annotators met to compare and to discuss their results, and to modify the definitions if necessary. The gold standard sense-tagged files were then made after all this discussion.

In a manner similar to our English approach, we included topical features as well as collocational, syntactic, and semantic local features in the maximum entropy models. *Collocational* features could be extracted from data that had been segmented into words and tagged for part of speech:

- the target word
- the part of speech tag of the target word
- the words (if any) within 2 positions of the target word
- the part of speech of the words (if any) immediately preceding and following the target word
- whether the target word follows a verb

draw, dress, drift, drive, face, ferret, find, keep, leave, live, match, play, pull, replace, see, serve, strike, train, treat, turn, use, wander, wash, work.

```
<entry id="00007" word="见" pinyin="jian4">
<wordsense id="00007-001">
        <definition id="chinese">看到,观察到,意识到</definition>
        <definition id="english">to see, to perceive</definition>
        <pos>VV</pos>
        <pred-arg>NP0 NP1</pred-arg>
        <pred-arg>NP0 NP1 IP</pred-arg>
        <example>只<word>见</word>一个人转过墙角。</example>
</wordsense>
</entry>
```

Figure 1: Example sense definition for jian4.

When disambiguating verbs, the following *syntactic* local features were extracted from data bracketed according to the Penn Chinese Treebank guidelines:

- whether the verb has a surface subject
- the head noun of the surface subject of the verb
- whether the verb has an object (any phrase labeled with "-OBJ", such as NP-OBJ, IP-OBJ, QP-OBJ)
- the phrase label of the object, if any
- the head noun of the object
- whether the verb has a VP complement
- the VP complement, if any
- whether the verb has an IP complement
- whether the verb has two NP complements
- whether the verb is followed by a predicate (any phrase labeled with "-PRD")

Semantic features were generated by assigning a HowNet[2] noun category to each subject and object, and topical keywords were extracted as for English.

Once all the features were extracted, a maximum entropy model was trained and tested for each target word. We used 5-fold cross validation to evaluate the system on each word. Two methods were used for partitioning a dataset of size N into five subsets: Select $N/5$ consecutive occurrences for each set, or select every 5th occurrence for a set. In the end, the choice of partitioning method made little difference in overall performance, and we report accuracy as the precision using the latter (stratified) sampling method.

[2]http://www.keenage.com/

Feature Type	Acc	Std Dev
collocation (no part of speech)	86.8	1.0
collocation	93.4	0.5
+ syntax	94.4	0.4
+ syntax + semantics	94.3	0.6
collocation + topic	90.3	1.0
+ syntax + topic	92.6	0.9
+ syntax + semantics + topic	92.8	0.9

Table 2: Overall accuracy of maximum entropy system using different subsets of features for Penn Chinese Treebank words (manually segmented, part-of-speech-tagged, parsed).

3.1 Penn Chinese Treebank

All sentences containing any of the 28 target words were extracted from the Penn Chinese Treebank, yielding between 4 and 1143 occurrence (160 average) for each of the target words. The manual segmentation, part-of-speech tags, and bracketing of the CTB were used to extract collocational and syntactic features.

The overall accuracy of the system on the 28 words in the CTB was 94.4% using local collocational and syntactic features. This is significantly better than the baseline of 76.7% obtained by tagging all instances of a word with the most frequent sense of the word in the CTB. Considering only the 23 words for which more than one sense occurred in the CTB, overall system accuracy was 93.9%, compared with a baseline of 74.7%. Figure 2 shows the results broken down by word.

As with the English data, we experimented with different types of features. Table 2 shows the performance of the system using different subsets of features. While the system's accuracy using syntactic features was higher than using only collocational features (significant at $p = 0.050$), the improve-

Word	pinying (translation)	Events	Senses	Baseline	Acc.	Std Dev
表示	biao3 shi4 (to indicate/express)	100	3	63.0	95.0	5.5
出	chu1 (to go out/to come out)	34	5	50.0	50.0	11.1
达	da2 (to reach a stage/to attain)	181	1	100	100	0.0
到	dao3 (to come/to arrive)	219	10	36.5	82.7	7.1
发展	fa1 zhan3 (to develop/to grow)	437	3	65.2	97.0	1.2
会	hui4 (will/be able to)	86	6	58.1	91.9	6.0
见	jian4 (to see/to perceive)	4	2	75.0	25.0	38.7
解决	jie3 jue2 (to solve/to settle)	44	2	79.5	97.7	5.0
进行	jin4 xing2 (to be in progress)	159	3	89.3	95.6	2.5
可	ke3 (may/can)	57	1	100	100	0.0
来	lai2 (to come/to arrive)	148	6	66.2	96.6	2.1
利用	li4 yong4 (to use/to utilize)	163	2	92.6	98.8	2.4
让	rang4 (to let/to allow)	9	1	100	100	0.0
使	shi3 (to make/to let)	89	1	100	100	0.0
说	shuo1 (to say in spoken words)	306	6	86.9	95.1	2.0
完	wan2 (to complete/to finish)	285	2	98.9	100.0	0.0
为	wei2/wei4 (to be/to mean)	473	7	32.8	86.1	2.4
想	xiang3 (to think/ponder/suppose)	8	3	62.5	50.0	50.0
引进	yin3 jin4 (to import/to introduce)	62	2	85.5	98.4	3.3
在	zai4 (to exist/to be at(in, on))	1143	4	96.9	99.3	0.4
发现	fa1 xian4 (to discover/to realize)	37	3	59.5	100.0	0.0
恢复	hui1 fu4 (to resume/to restore)	27	3	44.4	77.8	19.8
开放	kai1 fang4 (to open to investors)	122	5	74.6	96.7	3.0
可以	ke3 yi3 (may/can)	32	1	100	100	0.0
通过	tong1 guo4 (to pass legislation)	81	5	66.7	95.1	2.5
投入	tou2 ru4 (to input money, etc.)	44	4	40.9	84.1	11.7
要	yao4 (must/should/to intend to)	106	6	65.1	62.3	8.9
用	yong4 (to use)	41	2	58.5	100	0.0
Overall		4497	3.5	76.7	94.4	0.4

Figure 2: Word, number of instances, number of senses in CTB, baseline accuracy, maximum entropy accuracy and standard deviation using local collocational and syntactic features.

ment was not as substantial as for English, and this was despite the fact that the Chinese bracketing was done manually and should be almost error-free.

Semantic class information from HowNet yielded no improvement at all. To see if using a different ontology would help, we subsequently experimented with the ROCLing conceptual structures (Mo, 1992). In this case, we also manually added unknown nouns from the corpus to the ontology and labeled proper nouns with their conceptual structures, in order to more closely parallel the named entity information used in the English experiments. This resulted in a system accuracy of 95.0% (std. dev. 0.6), which again is not significantly better than omitting the noun class information.

3.2 People's Daily News

Five of the CTB words (chu1, jian4, xiang3, hui1 fu4, yao4) had system performance of less than 80%, probably due to their low frequency in the CTB corpus. These words were subsequently sense tagged in the People's Daily News, a much larger corpus (about one million words) that has manual segmentation and part-of-speech, but no bracketing information.[3] Those 5 words included all the words for which the system performed below the baseline

[3]The PDN corpus can be found at http://icl.pku.edu.cn/research/corpus/dwldform1.asp. The annotation guidelines are not exactly the same as for the Penn CTB, and can be found at http://icl.pku.edu.cn/research/corpus/coprus-annotation.htm.

Feature Type	Acc	Std Dev
collocation (no part of speech)	72.3	2.2
collocation	70.3	2.9
+ syntax	71.7	3.0
+ syntax + semantics	72.7	3.1
collocation + topic	73.3	3.2
+ syntax + topic	72.6	3.9
+ syntax + semantics + topic	72.8	3.7

Table 3: Overall accuracy of maximum entropy system using different subsets of features for People's Daily News words (automatically segmented, part-of-speech-tagged, parsed).

Feature Type	Acc	Std Dev
collocation (no part of speech)	71.4	4.3
collocation	74.7	2.3
collocation + topic	72.1	3.1

Table 4: Overall accuracy of maximum entropy system using different subsets of features for People's Daily News words (manually segmented, part-of-speech-tagged).

in the CTB corpus. About 200 sentences for each word were selected randomly from PDN and sense-tagged as with the CTB.

We *automatically* annotated the PDN data to yield the same types of annotation that had been available in the CTB. We used a maximum-matching algorithm and a dictionary compiled from the CTB (Sproat et al., 1996; Xue, 2001) to do segmentation, and trained a maximum entropy part-of-speech tagger (Ratnaparkhi, 1998) and TAG-based parser (Bikel and Chiang, 2000) on the CTB to do tagging and parsing.[4] Then the same feature extraction and model-training was done for the PDN corpus as for the CTB.

The system performance is much lower for the PDN than for the CTB, for several reasons. First, the PDN corpus is more balanced than the CTB, which contains primarily financial articles. A wider range of usages of the words was expressed in PDN than in CTB, making the disambiguation task more difficult; the average number of senses for the PDN words was 8.2 (compared to 3.5 for CTB), and the baseline accuracy was 58.0% (compared to 76.7% for CTB). Also, using automatically preprocessed data for the PDN introduced noise that was not present for the manually preprocessed CTB. Despite these differences between PDN and CTB, the trends in using increasingly richer linguistic preprocessing are similar. Table 3 shows that adding more features from richer levels of linguistic annotation yielded no significant improvement over using only collocational features. In fact, using only *lexical* collocations from automatic segmentation was sufficient to produce close to the best results. Table 4 shows the system performance using the available *manual* segmentation and part-of-speech tagging. While using part-of-speech tags seems to be better than using only lexical collocations, the difference is not significant.

[4]On held-out portions of the CTB, the accuracy of the segmentation and part-of-speech tagging are over 95%, and the accuracy of the parsing is 82%, which are comparable to the performance of the English preprocessors. The performance of these preprocessors is naturally expected to degrade when transferred to a different domain.

4 Conclusion

We have demonstrated the high performance of maximum entropy models for word sense disambiguation in English, and have applied the same approach successfully to Chinese. While SENSEVAL-2 showed that methods that work on English also tend to work on other languages, our experiments have revealed striking differences in the types of features that are important for English and Chinese WSD. While parse information seemed crucial for English WSD, it only played a minor role in Chinese; in fact, the improvement in Chinese performance contributed by manual parse information in the CTB disappeared altogether when automatic parsing was done for the PDN. The fact that bracketing was more important for English than Chinese WSD suggests that predicate-argument information and selectional restrictions may play a more important role in distinguishing English verb senses than Chinese senses. Or, it may be the case that Chinese verbs tend to be adjacent to their arguments, so collocational information is sufficient to capture the same information that would require parsing in English. This is a question for further study.

The simpler level of linguistic processing required to achieve relatively high sense-tagging accuracy in Chinese highlights an important difference between Chinese and English. Chinese is different from English in that much of Chinese linguistic ambiguity occurs at the basic level of word segmentation. Chinese word segmentation is a major task in itself, and it seems that once this is accomplished little more needs to be done for sense dis-

ambiguation. Our experience in English has shown that the ability to identify multi-word constructions significantly improves sense-tagging performance. Multi-character Chinese words, which are identified by word segmentation, may be the analogy to English multi-word constructions.

5 Acknowledgments

This work has been supported by National Science Foundation Grants, NSF-9800658 and NSF-9910603, and DARPA grant N66001-00-1-8915 at the University of Pennsylvania. The authors would also like to thank the anonymous reviewers for their valuable comments.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1).

Daniel M. Bikel and David Chiang. 2000. Two statistical parsing models applied to the chinese treebank. In *Proceedings of the Second Chinese Language Processing Workshop*, Hong Kong.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC.

Martin Chodorow, Claudia Leacock, and George A. Miller. 2000. A topical/local classifier for word sense identification. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, July.

Hoa Trang Dang and Martha Palmer. 2002. Combining contextual features for word sense disambiguation. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, PA.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, July.

Chinese-English Translation Assistance Group. 1982. *Chinese Dictionaries: an Extensive Bibliography of Dictionaries in Chinese and Other Languages.* Greenwood Publishing Group.

Nancy Ide and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1).

Adam Kilgarriff and Martha Palmer. 2000. Introduction to the special issue on SENSEVAL. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.

Ruo-Ping Mo. 1992. A conceptual structure that is suitable for analysing chinese. Technical Report CKIP-92-04, Academia Sinica, Taipei, Taiwan.

M. Palmer, Chunghye Han, Fei Xia, Dania Egedi, and Joseph Rosenzweig. 2000. Constraining lexical selection across languages using tags. In Anne Abeille and Owen Rambow, editors, *Tree Adjoining Grammars: formal, computational and linguistic aspects.* CSLI, Palo Alto, CA.

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

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

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word segmentation algorithm for chinese. *Computational Linguistics*, 22(3).

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing guidelines and ensuring consistency for chinese text annotation. In *Proceedings of the second International Conference on Language Resources and Evaluation*, Athens, Greece.

Nianwen Xue. 2001. *Defining and Automatically Identifying Words in Chinese.* Ph.D. thesis, University of Delaware.

David Yarowsky. 2000. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.