Chinese Verb Sense Discrimination Using an EM Clustering Model with Rich Linguistic Features

Jinying Chen, Martha Palmer
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA, 19104
{jinying,mpalmer}@linc.cis.upenn.edu

Abstract

This paper discusses the application of the Expectation-Maximization (EM) clustering algorithm to the task of Chinese verb sense discrimination. The model utilized rich linguistic features that capture predicate-argument structure information of the target verbs. A semantic taxonomy for Chinese nouns, which was built semi-automatically based on two electronic Chinese semantic dictionaries, was used to provide semantic features for the model. Purity and normalized mutual information were used to evaluate the clustering performance on 12 Chinese verbs. The experimental results show that the EM clustering model can learn sense or sense group distinctions for most of the verbs successfully. We further enhanced the model with certain fine-grained semantic categories called lexical sets. Our results indicate that these lexical sets improve the model's performance for the three most challenging verbs chosen from the first set of experiments.

1 Introduction

Highly ambiguous words may lead to irrelevant document retrieval and inaccurate lexical choice in machine translation (Palmer et al., 2000), which suggests that word sense disambiguation (WSD) is beneficial and sometimes even necessary in such NLP tasks. This paper addresses WSD in Chinese through developing an Expectation-Maximization (EM) clustering model to learn Chinese verb sense distinctions. The major goal is to do sense discrimination rather than sense labeling, similar to (Schütze, 1998). The basic idea is to divide instances of a word into several clusters that have no sense labels. The instances in the same cluster are regarded as having the same meaning. Word sense discrimination can be applied to document retrieval and similar tasks in information access, and to facilitating the building of large annotated corpora. In addition, since the clustering model can be trained on large unannotated corpora and evaluated on a relatively small sense-tagged corpus, it can be used to find indicative features for sense distinctions through exploring huge amount of available unannotated text data.

The EM clustering algorithm (Hofmann and Puzicha, 1998) used here is an unsupervised machine learning algorithm that has been applied in many NLP tasks, such as inducing a semantically labeled lexicon and determining lexical choice in machine translation (Rooth et al., 1998), automatic acquisition of verb semantic classes (Schulte im Walde, 2000) and automatic semantic labeling (Gildea and Jurafsky, 2002). In our task, we equipped the EM clustering model with rich linguistic features that capture the predicate-argument structure information of verbs and restricted the feature set for each verb using knowledge from dictionaries. We also semi-automatically built a semantic taxonomy for Chinese nouns based on two Chinese electronic semantic dictionaries, the Hownet dictionary[1] and the Rocling dictionary.[2] The 7 top-level categories of this taxonomy were used as semantic features for the model. Since external knowledge is used to obtain the semantic features and guide feature selection, the model is not completely unsupervised from this perspective; however, it does not make use of any annotated training data.

Two external quality measures, purity and normalized mutual information (NMI) (Strehl. 2002), were used to evaluate the model's performance on 12 Chinese verbs. The experimental results show that rich linguistic features and the semantic taxonomy are both very useful in sense discrimination. The model generally performs well in learning sense group distinctions for difficult, highly polysemous verbs and sense distinctions for other verbs. Enhanced by certain fine-grained semantic categories called lexical sets (Hanks, 1996), the model's

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

[2] A Chinese electronic dictionary liscenced from The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Nankang, Taipei, Taiwan.

performance improved in a preliminary experiment for the three most difficult verbs chosen from the first set of experiments.

The paper is organized as follows: we briefly introduce the EM clustering model in Section 2 and describe the features used by the model in Section 3. In Section 4, we introduce a semantic taxonomy for Chinese nouns, which is built semi-automatically for our task but can also be used in other NLP tasks such as co-reference resolution and relation detection in information extraction. We report our experimental results in Section 5 and conclude our discussion in Section 6.

2 EM Clustering Model

The basic idea of our EM clustering approach is similar to the probabilistic model of co-occurrence described in detail in (Hofmann and Puzicha 1998). In our model, we treat a set of features $\{f_1, f_2, ..., f_m\}$, which are extracted from the parsed sentences that contain a target verb, as observed variables. These variables are assumed to be independent given a hidden variable *c*, the sense of the target verb. Therefore the joint probability of the observed variables (features) for each verb instance, i.e., each parsed sentence containing the target verb, is defined in equation (1),

$$p(f_1, f_2, ..., f_m) = \sum_c p(c) \prod_{i=1}^{m} p(f_i \mid c) \quad (1)$$

The f_i's are discrete-valued features that can take multiple values. A typical feature used in our model is shown in (2),

$$f_i = \begin{cases} 0 \ \textit{iff} \text{ the target verb has no sentential complement} \\ 1 \ \textit{iff} \text{ the target verb has a nonfinite sentential complement} \\ 2 \ \textit{iff} \text{ the target verb has a finite sentential complement} \end{cases} \quad (2)$$

At the beginning of training (i.e., clustering), the model's parameters $p(c)$ and $p(f_i \mid c)$ are randomly initialized.[3] Then, the probability of *c* conditioned on the observed features is computed in the expectation step (E-step), using equation (3),

$$\tilde{p}(c \mid f_1, f_2, ..., f_m) = \frac{p(c) \prod_{i=1}^{m} p(f_i \mid c)}{\sum_c p(c) \prod_{i=1}^{m} p(f_i \mid c)} \quad (3)$$

In the maximization step (M-step), $p(c)$ and $p(f_i \mid c)$ are re-computed by maximizing the log-likelihood of all the observed data which is calculated by using $\tilde{p}(c \mid f_1, f_2, ..., f_m)$ estimated in the E-step. The E-step and M-step are repeated for a fixed number of rounds, which is set to 20 in our experiments,[4] or till the amount of change of $p(c)$ and $p(f_i \mid c)$ is under the threshold 0.001.

When doing classification, for each verb instance, the model calculates the same conditional probability as in equation (3) and assigns the instance to the cluster with the maximal $p(c \mid f_1, f_2, ..., f_m)$.

3 Features Used in the Model

The EM clustering model uses a set of linguistic features to capture the predicate-argument structure information of the target verbs. These features are usually more indicative of verb sense distinctions than simple features such as words next to the target verb or their POS tags. For example, the Chinese verb "出| chu1" has a sense of *produce*, the distinction between this sense and the verb's other senses, such as *happen* and *go out*, largely depends on the semantic category of the verb's direct object. Typical examples are shown in (1),

(1) a. 他们/their 县/county 出/produce 香蕉/banana
"Their county produces bananas."

b. 他们/their 县/county 出/happen 大/big 事/event 了/ASP
"A big event happened in their county."

c. 他们/their 县/county 出/go out 门/door 就/right away 是/be 山/mountain
"In their county, you can see mountains as soon as you step out of the doors."

The verb has the sense *produce* in (1a) and its object should be something producible, such as "香蕉/banana". While in (1b), with the sense *happen*, the verb typically takes an *event* or *event-like* object, such as "大事/big event", "事故/accident" or "问题/problem" etc. In (1c), the verb's object "门/door" is closely related to *location*, consistent with the sense *go out*. In contrast, simple lexical or POS tag features sometimes fail to capture such information, which can be seen clearly in (2),

[3] In our experiments, for verbs with more than 3 senses, syntactic and semantic restrictions derived from dictionary entries are used to constrain the random initialization.

[4] In our experiments, we set 20 as the maximal number of rounds after trying different numbers of rounds (20, 40, 60, 80, 100) in a preliminary experiment.

(2) a. 去年/last year 出/produce 香蕉/banana 3000 公斤/ kilogram
"3000 kilograms of bananas were produced last year."

b. 要/in order to 出/produce 海南/Hainan 最好/best 的/DE 香蕉/banana
"In order to produce the best bananas in Hainan, ……"

The verb's object "香蕉/banana", which is next to the verb in (2a), is far away from the verb in (2b). For (2b), a classifier only looking at the adjacent positions of the target verb tends to be misled by the NP right after the verb, i.e., "海南/Hainan", which is a Province in China and a typical object of the verb with the sense *go out*.

Five types of features are used in our model:

1. Semantic category of the subject of the target verb
2. Semantic category of the object of the target verb
3. Transitivity of the target verb
4. Whether the target verb takes a sentential complement and which type of sentential complement (finite or nonfinite) it takes
5. Whether the target verb occurs in a verb compound

We obtain the values for the first two types of features (1) and (2) from a semantic taxonomy for Chinese nouns, which we will introduce in detail in the next section.

In our implementation, the model uses different features for different verbs. The criteria for feature selection are from the electronic CETA dictionary file [5] and a hard copy English-Chinese dictionary, The Warmth Modern Chinese-English Dictionary.[6] For example, the verb "出|chu1" never takes sentential complements, thus the fourth type of feature is not used for it. It could be supposed that we can still have a uniform model, i.e., a model using the same set of features for all the target verbs, and just let the EM clustering algorithm find useful features for different verbs automatically. The problem here is that unsupervised learning models (i.e., models trained on unlabeled data) are more likely to be affected by noisy data than supervised ones. Since all the features used in our model are extracted from automatically parsed sentences that inevitably have preprocessing errors such as segmentation, POS tagging and parsing errors, using verb-specific sets of features can alleviate the problem caused by noisy data to some extent. For example, if the model already knows that a verb like "出|chu1" can never take sentential complements (i.e., it does not use the fourth type of feature for that verb), it will not be misled by erroneous parsing information saying that the verb takes sentential complements in certain sentences. Since the corresponding feature is not included, the noisy data is filtered out. In our EM clustering model, all the features selected for a target verb are treated in the same way, as described in Section 2.

4 A Semantic Taxonomy Built Semi-automatically

Examples in (1) have shown that the semantic category of the object of a verb sometimes is crucial in distinguishing certain Chinese verb senses. And our previous work on information extraction in Chinese (Chen et al., 2004) has shown that semantic features, which are more general than lexical features but still contain rich information about words, can be used to improve a model's capability of handling unknown words, thus alleviating potential sparse data problems.

We have two Chinese electronic semantic dictionaries: the Hownet dictionary, which assigns 26,106 nouns to 346 semantic categories, and the Rocling dictionary, which assigns 4,474 nouns to 110 semantic categories.[7] A preliminary experimental result suggests that these semantic categories might be too fine-grained for the EM clustering model (see Section 5.2 for greater details). An analysis of the sense distinctions of several Chinese verbs also suggests that more general categories on top of the Hownet and Rocling categories could still be informative and most importantly, could enable the model to generate meaningful clusters more easily. We therefore built a three-level semantic taxonomy based on the two semantic dictionaries using both automatic methods and manual effort.

The taxonomy was built in three steps. First, a simple mapping algorithm was used to map semantic categories defined in Hownet and Rocling into 27 top-level WordNet categories.[8] The Hownet or Rocling semantic categories have English glosses. For each category gloss, the algorithm looks through the hypernyms of its first sense in WordNet and chooses the first WordNet top-level category it finds.

[5] Licensed from the Department of Defense

[6] The Warmth Modern Chinese-English Dictionary, Wang-Wen Books Ltd, 1997.

[7] Hownet assigns multiple entries (could be different semantic categories) to polysemous words. The Rocling dictionary we used only assigns one entry (i.e., one semantic category) to each noun.

[8] The 27 categories contain 25 unique beginners for noun source files in WordNet, as defined in (Fellbaum, 1998) and two higher level categories *Entity* and *Abstraction*.

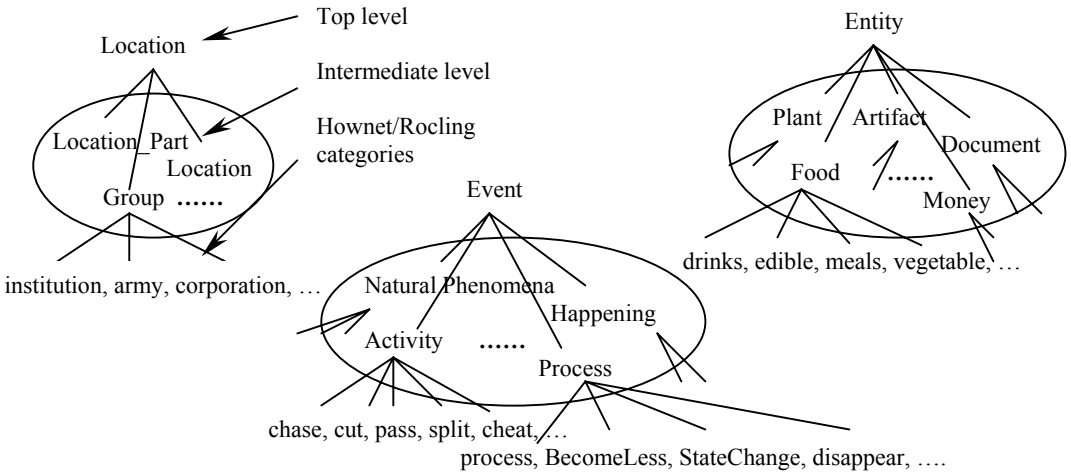

Figure 1. Part of the 3-level Semantic Taxonomy for Chinese Nouns (other top-level nodes are Time, Human, Animal and State)

The mapping obtained from step 1 needs further modification for two reasons. First, the glosses of Hownet or Rocling semantic categories usually have multiple senses in WordNet. Sometimes, the first sense in WordNet for a category gloss is not its intended meaning in Hownet or Rocling. In this case, the simple algorithm cannot get the correct mapping. Second, Hownet and Rocling sometimes use adjectives or non-words as category glosses, such as *animate* and *LandVehicle* etc., which have no WordNet nominal hypernyms at all. However, those adjectives or non-words usually have straightforward meanings and can be easily reassigned to an appropriate WordNet category. Although not accurate, the automatic mapping in step 1 provides a basic framework or skeleton for the semantic taxonomy we want to build and makes subsequent work easier.

In step 2, hand correction, we found that we could make judgments and necessary adjustments on about 80% of the mappings by only looking at the category glosses used by Hownet or Rocling, such as *livestock, money, building* and so on. For the other 20%, we could make quick decisions by looking them up in an electronic table we created. For each Hownet or Rocling category, our table lists all the nouns assigned to it by the two dictionaries. We merged two WordNet categories into others and subdivided three categories that seemed more coarse-grained than others into 2~5 subcategories. Step 2 took three days and 35 intermediate-level categories were generated.

In step 3, we manually clustered the 35 intermediate-level categories into 7 top-level semantic categories. Figure 1 shows part of the taxonomy.

The EM clustering model uses the 7 top-level categories to define the first two types of features that were introduced in Section 3. For example, the value of a feature f_k is 1 if and only if the object NP of the target verb belongs to the semantic category *Event* and is otherwise 0.

5 Clustering Experiments

Since we need labeled data to evaluate the clustering performance but have limited sense-tagged corpora, we applied the clustering model to 12 Chinese verbs in our experiments. The verbs are chosen from 28 annotated verbs in Penn Chinese Treebank so that they have at least two verb meanings in the corpus and for each of them, the number of instances for a single verb sense does not exceed 90% of the total number of instances.

In our task, we generally do not include senses for other parts of speech of the selected words, such as noun, preposition, conjunction and particle etc., since the parser we used has a very high accuracy in distinguishing different parts of speech of these words (>98% for most of them). However, we do include senses for conjunctional and/or prepositional usage of two words, "到|dao4" and "为|wei4", since our parser cannot distinguish the verb usage from the conjunctional or prepositional usage for the two words very well.

Five verbs, the first five listed in Table 1, are both highly polysemous and difficult for a supervised word sense classifier (Dang et al., 2002). [9] In our experiments, we manually grouped the verb senses for the five verbs. The criteria for the grouping are similar to Palmer et al.'s (to appear) work on English verbs, which considers both sense coherence and predicate-argument structure distinctions. Figure 2 gives an example of

[9] In the supervised task, their accuracies are lower than 85%, and four of them are even lower than the baselines.

Senses for "到|dao4"

1. to go to, leave for
2. to come
3. to arrive
4. to reach a particular stage, condition, or level
5. marker for completion of activities (after a verb)
6. marker for direction of activities (after a verb)
7. to reach a time point
8. up to, until (prepositional usage)
9. up to, until, (from …) to … (conjunctional usage)

Sense groups for "到|dao4"

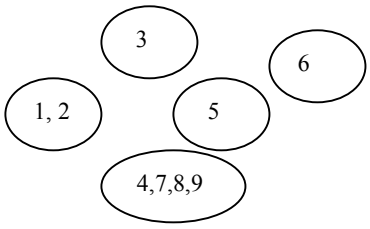

Figure 2. Sense groups for the Chinese verb "到|dao4"

the definition of sense groups. The manually defined sense groups are used to evaluate the model's performance on the five verbs.

The model was trained on an unannotated corpus, People's Daily News (PDN), and tested on the manually sense-tagged Chinese Treebank (with some additional sense-tagged PDN data).[10] We parsed the training and test data using a Maximum Entropy parser and extracted the features from the parsed data automatically. The number of clusters used by the model is set to the number of the defined senses or sense groups of each target verb. For each verb, we ran the EM clustering algorithm ten times. Table 2 shows the average performance and the standard deviation for each verb. Table 1 summarizes the data used in the experiments, where we also give the normalized sense perplexity[11] of each verb in the test data.

5.1 Evaluation Methods

We use two external quality measures, purity and normalized mutual information (NMI) (Strehl. 2002) to evaluate the clustering performance. Assuming a verb has l senses, the clustering model assigns n instances of the verb into k clusters, n_i is the size of the *ith* cluster, n^j is the number of instances hand-tagged with the *jth* sense, and n_i^j is the number of instances with the *jth* sense in the *ith* cluster, purity is defined in equation (4):

$$purity = \frac{1}{n}\sum_{i=1}^{k}\max_{j} n_i^j \qquad (4)$$

It can be interpreted as classification accuracy when for each cluster we treat the majority of instances that have the same sense as correctly classified. The baseline purity is calculated by treating all instances for a target verb in a single cluster. The purity measure is very intuitive. In our case, since the number of clusters is preset to the number of senses, purity for verbs with two senses is equal to classification accuracy defined in supervised WSD. However, for verbs with more than 2 senses, purity is less informative in that a clustering model could achieve high purity by making the instances of 2 or 3 dominant senses the majority instances of all the clusters.

Mutual information (MI) is more theoretically well-founded than purity. Treating the verb sense and the cluster as random variables S and C, the MI between them is defined in equation (5):

$$\begin{aligned} MI(S,C) &= \sum_{s,c} p(s,c)\log\frac{p(s,c)}{p(s)p(c)} \\ &= \sum_{j=1}^{l}\sum_{i=1}^{k}\frac{n_i^j}{n}\log\frac{n_i^j n}{n_i n^j} \end{aligned} \qquad (5)$$

MI(S,C) characterizes the reduction in uncertainty of one random variable *S (or C)* due to knowing the other variable *C (or S)*. A single cluster with all instances for a target verb has a zero MI. Random clustering also has a zero MI in the limit. In our experiments, we used [0,1]-normalized mutual information (NMI) (Strehl. 2002). A shortcoming of this measure, however, is that the best possible clustering (upper bound) evaluates to less than 1, unless classes are balanced. Unfortunately, unbalanced sense distribution is the usual case in WSD tasks, which makes NMI itself hard to interpret. Therefore, in addition to NMI, we also give its upper bound (upper-NMI) and the ratio of NMI and its upper bound (NMI-ratio) for each verb, as shown in columns 6 to 8 in Table 2.

[10] The sense-tagged PDN data we used here are the same as in (Dang et al., 2002).

[11] It is calculated as the entropy of the sense distribution of a verb in the test data divided by the largest possible entropy, i.e., $\log_2$ (the number of senses of the verb in the test data).

Verb| Pinyin	Sample senses of the verb	# Senses in test data	# Sense groups in test data	Sense perplexity	# Clusters	# Training instances	# Test instances
出 |chu1	go out /produce	16	7	0.68	8	399	157
到 |dao4	come /reach	9	5	0.72	6	1838	186
见 |jian4	see /show	8	5	0.68	6	117	82
想 |xiang3	think/suppose	6	4	0.64	6	94	228
要 |yao4	Should/intend to	8	4	0.65	7	2781	185
表示|biao3shi4	Indicate /express	2		0.93	2	666	97
发现|fa1xian4	discover /realize	2		0.76	2	319	27
发展|fa1zhan3	develop /grow	3		0.69	3	458	130
恢复|hui1fu4	resume /restore	4		0.83	4	107	125
说 |shuo1	say /express by written words	7		0.40	7	2692	307
投入|tou2ru4	to input /plunge into	2		1.00	2	136	23
为 |wei2_4	to be /in order to	6		0.82	6	547	463

Table 1. A summary of the training and test data used in the experiments

Verb	Sense perplexity	Baseline Purity (%)	Purity (%)	Std. Dev. of purity (%)	NMI	Upper-NMI	NMI-ratio (%)	Std. Dev. of NMI ratio (%)
出	0.68	52.87	63.31	1.59	0.2954	0.6831	43.24	1.76
到	0.72	40.32	90.48	1.08	0.4802	0.7200	75.65	0.00
见	0.68	58.54	72.20	1.61	0.1526	0.6806	22.41	0.66
想	0.64	68.42	79.39	3.74	0.2366	0.6354	37.24	8.22
要	0.65	69.19	69.62	0.34	0.0108	0.6550	1.65	0.78
表示	0.93	64.95	98.04	1.49	0.8670	0.9345	92.77	0.00
发现	0.76	77.78	97.04	3.87	0.7161	0.7642	93.71	13.26
发展	0.69	53.13	90.77	0.24	0.4482	0.6918	64.79	2.26
恢复	0.83	45.97	65.32	0.00	0.1288	0.8234	15.64	0.00
说	0.40	80.13	93.00	0.58	0.3013	0.3958	76.13	4.07
投入	1.00	52.17	95.65	0.00	0.7827	0.9986	78.38	0.00
为	0.82	32.61	75.12	0.43	0.4213	0.8213	51.30	2.07
Average	0.73	58.01	82.50	1.12	0.4088	0.7336	54.41	3.31

Table 2. The performance of the EM clustering model on 12 Chinese verbs measured by purity and normalized mutual information (NMI)

5.2 Experimental Results

Table 2 summarizes the experimental results for the 12 Chinese verbs. As we see, the EM clustering model performs well on most of them, except the verb "要|yao4".[12] The NMI measure NMI-ratio turns out to be more stringent than purity. A high purity does not necessarily mean a high NMI-ratio. Although intuitively, NMI-ratio should be related to sense perplexity and purity, it is hard to formalize the relationships between them from the results. In fact, the NMI-ratio for a particular verb is eventually determined by its concrete sense distribution in the test data and the model's clustering behavior for that verb. For example, the verbs "出|chu1" and "见|jian4" have the same sense perplexity and "见|jian4" has a higher purity than "出|chu1" (72.20% vs. 63.31%), but the NMI-ratio for "见|jian4" is much lower than "出|chu1" (22.41% vs. 43.24%). An analysis of the classification results for "见|jian4" shows that the clustering model made the instances of the verb's most dominant sense the majority instances of three clusters (of total 5 clusters), which is penalized heavily by the NMI measure.

Rich linguistic features turn out to be very effective in learning Chinese verb sense distinctions. Except for the two verbs, "发现|fa1xian4" and "表示|biao3shi4", the sense distinctions of which can usually be made only by syntactic alternations,[13] features such as semantic features or combinations of semantic features and syntactic alternations are very beneficial and sometimes even necessary for learning sense distinctions of other verbs. For example, the verb "见|jian4" has one sense *see*, in which the verb typically takes a *Human* subject and a sentential complement, while in another sense *show*, the verb typically takes an *Entity* subject and a *State* object. An inspection of the classification results shows

[12] For all the verbs except "要|yao4", the model's purities outperformed the baseline purities significantly ($p<0.05$, and $p<0.001$ for 8 of them).

[13] For example, the verb "发现|fa1xian4" takes an object in one sense *discover* and a sentential complement in the other sense *realize*.

that the EM clustering model has indeed learned such combinatory patterns from the training data.

The experimental results also indicate that the semantic taxonomy we built is beneficial for the task. For example, the verb "投入|tou1ru4" has two senses, *input* and *plunge into*. It typically takes an *Event* object for the second sense but not for the first one. A single feature obtained from our semantic taxonomy, which tests whether the verb takes an *Event* object, captures this property neatly (achieves purity 95.65% and NMI-ratio 78.38% when using 2 clusters). Without the taxonomy, the top-level category *Event* is split into many fine-grained Hownet or Rocling categories, which makes it very difficult for the EM clustering model to learn sense distinctions for this verb. In fact, in a preliminary experiment only using the Hownet and Rocling categories, the model had the same purity as the baseline (52.17%) and a low NMI-ratio (4.22%) when using 2 clusters. The purity improved when using more clusters (70.43% with 4 clusters and 76.09% with 6), but it was still much lower than the purity achieved by using the semantic taxonomy and the NMI-ratio dropped further (1.19% and 1.20% for the two cases).

By looking at the classification results, we identified three major types of errors. First, preprocessing errors create noisy data for the model. Second, certain sense distinctions depend heavily on global contextual information (cross-sentence information) that is not captured by our model. This problem is especially serious for the verb "要|yao4". For example, without global contextual information, the verb can have at least three meanings *want*, *need* or *should* in the same clause, as shown in (3).

(3) 他/he 要/want/need/should 马上/at once
读完/finish reading 这本/this 书/book.
"He wants to/needs to/should finish reading this book at once."

Third, a target verb sometimes has specific types of NP arguments or co-occurs with specific types of verbs in verb compounds in certain senses. Such information is crucial for distinguishing these senses from others, but is not captured by the general semantic taxonomy used here. We did further experiments to investigate how much improvement the model could gain by capturing such information, as discussed in Section 5.3.

5.3 Experiments with Lexical Sets

As discussed by Patrick Hanks (1996), certain senses of a verb are often distinguished by very narrowly defined semantic classes (called lexical sets) that are specific to the meaning of that verb sense. For example, in our case, the verb "恢复|hui1fu4" has a sense *recover* in which its direct object should be something that can be recovered naturally. A typical set of object NPs of the verb for this particular sense is partially listed in (4),

(4) Lexical set for naturally recoverable things
{体力/physical strength, 身体/body, 健康/health, 精力/mental energy, 听力/hearing, 知觉/feeling, 记忆力/memory, ……}

Most words in this lexical set belong to the Hownet category *attribute* and the top-level category *State* in our taxonomy. However, even the lower-level category *attribute* still contains many other words irrelevant to the lexical set, some of which are even typical objects of the verb for two other senses, *resume* and *regain*, such as "邦交/diplomatic relations" in "恢复/resume 邦交/diplomatic relations" and "名誉/reputation" in "恢复/regain名誉/reputation". Therefore, a lexical set like (4) is necessary for distinguishing the *recover* sense from other senses of the verb.

It has been argued that the extensional definition of lexical sets can only be done using corpus evidence and it cannot be done fully automatically (Hanks, 1997). In our experiments, we use a bootstrapping approach to obtain five lexical sets semi-automatically for three verbs "出|chu1", "见|jian4" and "恢复|hui1fu4" that have both low purity and low NMI-ratio in the first set of experiments. [14] We first extracted candidates for the lexical sets from the training data. For example, we extracted all the direct objects of the verb "恢复|hui1fu4" and all the verbs that combined with the verb "出|chu1" to form verb compounds from the automatically parsed training data. From the candidates, we manually selected words to form five initial seed sets, each of which contains no more than ten words. A simple algorithm was used to search for all the words that have the same detailed Hownet semantic definitions (semantic category plus certain supplementary information) as the seed words. We did not use Rocling because its semantic definitions are so general that a seed word tends to extend to a huge set of irrelevant words. Highly relevant words were manually selected from all the words found by the searching algorithm and added to the initial seed sets. The enlarged sets were used as lexical sets.

The enhanced model first uses the lexical sets to obtain the semantic category of the NP arguments

[14] We did not include "要|yao4", since its meaning rarely depends on local predicate-argument structure information.

of the three verbs. Only when the search fails does the model resort to the general semantic taxonomy. The model also uses the lexical sets to determine the types of the compound verbs that contain the target verb "出|chu1" and uses them as new features.

Table 3 shows the model's performance on the three verbs with or without using lexical sets. As we see, lexical sets improves the model's performance on all of them, especially on the verb "出|chu1". Although the results are still preliminary, they nevertheless provide us hints of how much a WSD model for Chinese verbs could gain from lexical sets.

Verb	w/o lexical sets (%)		with lexical sets (%)	
	Purity	NMI-ratio	Purity	NMI-ratio
出	63.61	43.24	76.50	52.81
见	72.20	22.41	77.56	34.63
恢复	65.32	15.64	69.03	19.71

Table 3. Clustering performance with and without lexical sets for three Chinese verbs

6 Conclusion

We have shown that an EM clustering model that uses rich linguistic features and a general semantic taxonomy for Chinese nouns generally performs well in learning sense distinctions for 12 Chinese verbs. In addition, using lexical sets improves the model's performance on three of the most challenging verbs.

Future work is to extend our coverage and to apply the semantic taxonomy and the same types of features to supervised WSD in Chinese. Since the experimental results suggest that a general semantic taxonomy and more constrained lexical sets are both beneficial for WSD tasks, we will develop automatic methods to build large-scale semantic taxonomies and lexical sets for Chinese, which reduce human effort as much as possible but still ensure high quality of the obtained taxonomies or lexical sets.

7 Acknowledgements

This work has been supported by an ITIC supplement to a National Science Foundation Grant, NSF-ITR-EIA-0205448. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Jinying Chen, Nianwen Xue and Martha Palmer. 2004. Using a Smoothing Maximum Entropy Model for Chinese Nominal Entity Tagging. In *Proceedings of the 1st Int. Joint Conference on Natural Language Processing*. Hainan Island, China.

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

Christiane Fellbaum. 1998. *WordNet – an Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, London.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3): 245-288, 2002.

Patrick Hanks. 1996. Contextual dependencies and lexical sets. *The Int. Journal of Corpus Linguistics*, 1:1.

Patrick Hanks. 1997. Lexical sets: relevance and probability. in B. Lewandowska-Tomaszczyk and M. Thelen (eds.) *Translation and Meaning, Part 4*, School of Translation and Interpreting, Maastricht, The Netherlands.

Thomas Hofmann and Puzicha Jan. 1998. *Statistical models for co-occurrence data*, MIT Artificial Intelligence Lab., Technical Report AIM-1625.

Adam Kilgarriff and Martha Palmer. 2000. Introduction to the sepcial issue on SENSEVAL. *Computers and the Humanities*, 34(1-2): 15-48.

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

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1998. *EM-based clustering for NLP applications.* AIMS Report 4(3).Institut für Maschinelle Sprachverarbeitung.

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th Int. Conference on Computational Linguistics*, 747-753.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24 (1): 97-124.

Alexander Strehl. 2002. Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining. *Dissertation*. The University of Texas at Austin. http://www.lans.ece.utexas.edu/~strehl/diss/.