Hybrid Models for Semantic Classification of Chinese Unknown Words

Xiaofei Lu
Department of Linguistics and Applied Language Studies
Pennsylvania State University
University Park, PA 16802, USA
xxl13@psu.edu

Abstract

This paper addresses the problem of classifying Chinese unknown words into fine-grained semantic categories defined in a Chinese thesaurus. We describe three novel knowledge-based models that capture the relationship between the semantic categories of an unknown word and those of its component characters in three different ways. We then combine two of the knowledge-based models with a corpus-based model which classifies unknown words using contextual information. Experiments show that the knowledge-based models outperform previous methods on the same task, but the use of contextual information does not further improve performance.

1 Introduction

Research on semantic annotation has focused primarily on word sense disambiguation (WSD), i.e., the task of determining the appropriate sense for each instance of a polysemous word out of a set of senses defined for the word in some lexicon. Much less work has been done on semantic classification of unknown words, i.e., words that are not listed in the lexicon. However, real texts typically contain a large number of unknown words. Successful classification of unknown words is not only useful for lexical acquisition, but also necessary for natural language processing (NLP) tasks that require semantic annotation.

This paper addresses the problem of classifying Chinese unknown words into fine-grained semantic categories defined in a Chinese thesaurus, *Cilin* (Mei et al., 1984). This thesaurus classifies over 70,000 words into 12 major categories, including human (A), concrete object (B), time and space (C), abstract object (D), attributes (E), actions (F), mental activities (G), activities (H), physical states (I), relations (J), auxiliaries (K), and honorifics (L). The 12 major categories are further divided into 94 medium categories, which in turn are subdivided into 1428 small categories. Each small category contains synonyms that are close in meaning. For example, under the major category *D*, the medium category *Dm* groups all words that refer to institutions, and the small category *Dm05* groups all words that refer to educational institutions, e.g., 学校 *xuéxiào* 'school'. Unknown word classification involves a much larger search space than WSD. In classifying words into small categories in *Cilin*, the search space for a polysemous known word consists of all the categories the word belongs to, but that for an unknown word consists of all the 1428 small categories.

Research on WSD has concentrated on using contextual information, which may be limited for infrequent unknown words. On the other hand, Chinese characters carry semantic information that is useful for predicting the semantic properties of the words containing them. We present three novel knowledge-based models that capture the relationship between the semantic categories of an unknown word and those of its component characters in three different ways, and combine two of them with a corpus-based model that uses contextual information to classify unknown words. Experiments show that the combined knowledge-based model achieves an accuracy of 61.6% for classifying unknown words into small categories in *Cilin*, but the use of contextual information does not further improve performance.

The rest of the paper is organized as follows. Section 2 details the three novel knowledge-based models proposed for this task. Section 3 describes a corpus-based model. Section 4 reports the experiment results of the proposed

Proceedings of NAACL HLT 2007, pages 188–195,
Rochester, NY, April 2007.

models. Section 5 compares these results with previous results. Section 6 concludes the paper and points to avenues for further research.

2 Knowledge-based Models

This section describes three novel knowledge-based models for semantic classification of Chinese unknown words, including an overlapping-character model, a character-category association model, and a rule-based model. These models model the relationship between the semantic category of an unknown word and those of its component characters in three different ways.

2.1 The Baseline Model

The baseline model predicts the category of an unknown word by counting the number of overlapping characters between the unknown word and the member words in each category. As words in the same category are similar in meaning and the meaning of a Chinese word is generally the composition of the meanings of its characters, it is common for words in the same category to share one or more character. This model tests the hypothesis that speakers draw upon the repertoire of characters that relate to a concept when creating new words to realize it.

For each semantic category in *Cilin*, the set of unique characters in its member words are extracted, and the number of times each character occurs in word-initial, word-middle, and word-final positions is recorded. With this information, we develop two variants of the baseline model, which differ from each other in terms of whether it takes into consideration the positions in which the characters occur in words.

In variant 1, the score of a category is the sum of the number of occurrences of each character of the target word in the category, as in (1), where t_j denotes a category, w denotes the target word, c_i denotes the ith character in w, n is the length of w, and $f(c_i)$ is the frequency of c_i in t_j.

$$(1) \qquad Score(t_j, w) = \sum_{i=1}^{n} f(c_i)$$

In variant 2, the score of a category is the sum of the number of occurrences of each character of the unknown word in the category in its corresponding position, as in (2), where p_i denotes the position of c_i in w, which could be word-initial, word-middle, or word-final, and $f(c_i,p_i)$ denotes the frequency of c_i in position p_i in t_j.

$$(2) \qquad Score(t_j, w) = \sum_{i=1}^{n} f(c_i, p_i)$$

In each variant, the category with the maximum score for a target word is proposed as the category of the word.

2.2 Character-Category Associations

The relationship between the semantic category of an unknown word and those of its component characters can also be captured in a more sophisticated way using information-theoretical models. We use two statistical measures, mutual information and χ^2, to compute character-category associations and word-category associations. Chen (2004) used the χ^2 measure to compute character-character and word-word associations, but not word-category associations. We use word-category associations to directly predict the semantic categories of unknown words.

The mutual information and χ^2 measures are calculated as in (3) and (4), where $Asso(c,t_j)$ denotes the association between a character c and a semantic category t_j, and $P(X)$ and $f(X)$ denote the probability and frequency of X respectively.

$$(3) \qquad Asso_{MI}(c, t_j) = \log \frac{P(c, t_j)}{P(c)P(t_j)}$$

$$(4) \qquad Asso_{\chi^2}(c, t_j) = \frac{\alpha(c, t_j)}{\max_k \alpha(c, t_k)}$$

$$(5) \qquad \alpha(c, t_j) = \sqrt{\frac{[f(c, t_j)]^2}{f(c) + f(t_j)}}$$

Once the character-category associations are calculated, the association between a word w and a category t_j, $Asso(w,t_j)$, can be calculated as the sum of the weighted associations between each of the word's characters and the category, as in (6), where c_i denotes the ith character of w, $|w|$ denotes the length of w, and λ_i denotes the weight of $Asso(c_i,t_j)$. The λ's add up to 1. The weights are determined empirically based on the positions of the characters in the word.

$$(6) \qquad Asso(w, t_j) = \sum_{i=1}^{|w|} \lambda_i Asso(c_i, t_j)$$

As in variant 2 of the baseline model, the character-category association model can also be made sensitive to the positions in which the characters occur in the words. To this end, we first need to compute the position-sensitive associations between a category and a character in the word-initial, word-middle, and word-final positions separately. The position-sensitive association between an unknown word and a category can then be computed as the sum of the weighted position-sensitive associations between each of its characters and the category.

Once the word-category associations are computed, we can propose the highest ranked category or a ranked list of categories for each unknown word.

2.3 A Rule-Based Model

The third knowledge-based model uses linguistic rules to classify unknown words based on the syntactic and semantic categories of their component characters. Rule-based models have not been used for this task before. However, there are some regularities in the relationship between the semantic categories of unknown words and those of their component characters that can be captured in a more direct and effective way by linguistic rules than by statistical models.

A separate set of rules are developed for words of different lengths. Rules are initially developed based on knowledge about Chinese word formation, and are then refined by examining the development data. In general, the complete rule set takes a few hours to develop.

The rule in (7) is developed for bisyllabic unknown words. This rule proposes the common category of a bisyllabic word's two characters as its category. It is especially useful for words with a parallel structure, i.e., words whose two characters have the same meaning and syntactic category, e.g., 坍塌 *tāntā* 'collapse', where 坍 *tān* and 塌 *tā* both mean 'collapse' and share the category *Id05*. The thresholds for f_A and f_B are determined empirically and are both set to 1 if *AB* is a noun and to 0 and 3 respectively otherwise.

(7) For a bisyllabic word *AB*, if *A* and *B* share a category *c*[1], let f_A and f_B denote the number of times *A* and *B* occur in word-initial and word-final positions in *c* respectively. If f_A and f_B both surpass the predetermined thresholds, propose *c* for AB.

A number of rules are developed for trisyllabic words. While most rules in the model are general, the first rule in this set is rather specific, as it handles words with three specific prefixes, 大 *dà* 'big', 小 *xiǎo* 'little', and 老 *lǎo* 'old', which usually do not change the category of the root word. The other four rules again utilize the categories of the unknown word's component characters. The rules in (8b) and (8c) are similar to the rule in (7). The ones in (8d) and (8e) search for neighbor words with a similar structure as the target word. Eligible neighbors have a common morpheme with the target word in the same position and a second morpheme that shares a category with the second morpheme of the target word. For example, an eligible neighbor for 推销商 *tuīxiāo-shāng* 'sales-man' is 销售商 *xiāoshòu-shāng* 'distribut-or'. These two words share the morpheme 商 *shāng* 'businessman' in the word-final position, and the morphemes 推销 *tuīxiāo* 'to market' and 销售 *xiāoshòu* 'distribute' share the category *He03*. The rule in (8d) therefore applies in this case.

(8) For a trisyllabic word *ABC*:
- a. If *A* equals 大 *dà* 'big', 小 *xiǎo* 'little', or 老 *lǎo* 'old', propose the category of *AB* for *ABC* if *C* is the diminutive suffix 儿 *er* or the category of *BC* for *ABC* otherwise.
- b. If *A* and *BC* share a category *c*, propose *c* for *ABC*.
- c. If *AB* and *C* share a category *c*, propose *c* for *ABC*.
- d. If there is a word *XYC* such that *XY* and *AB* share a category, propose the category of *XYC* for *ABC*.
- e. If there is a word *XBC* such that *X* and *A* share a category, propose the category of *XBC* for *ABC*.

The rules for four-character words are given in (9). Like the rules in (8d) and (8e), these rules also search for neighbors of the target word.

(9) For a four-character word *ABCD*:
- a. If there is a word *XYZD/YZD* such that *XYZ/YZ* and *ABC* share a category, propose the category of *XYZ/YZ* for *ABCD*.
- b. If there is a word *ABCX* such that *X* and *D* share a category, propose the category of *ABCX* for *ABCD*.
- c. If there is a word *XYCD* such that *XY* and *AB* share a category, propose the category of *XYCD* for *ABCD*.
- d. If there is a word *XBCD/BCD*, propose the category of *XBCD/BCD* for *ABCD*.

3 A Corpus-Based Model

The knowledge-based models described above classify unknown words using information about the syntactic and semantic categories of their component characters. Another useful source of information is the context in which unknown words occur. While contextual information is the primary source of information used in WSD research and has been used for acquiring semantic lexicons and classifying unknown words in other languages (e.g., Roark and Charniak 1998; Ci-

[1] *A* and *B* may each belong to more than one category.

aramita 2003; Curran 2005), it has been used in only one previous study on semantic classification of Chinese unknown words (Chen and Lin, 2000). Part of the goal of this study is to investigate whether and how these two different sources of information can be combined to improve performance on semantic classification of Chinese unknown words.

To this end, we first use the knowledge-based models to propose a list of five candidate categories for the target word, then extract a generalized context for each category in *Cilin* from a corpus, and finally compute the similarity between the context of the target word and the generalized context of each of its candidate categories. Comparing the context of the target word with generalized contexts of categories instead of contexts of individual words alleviates the data-sparseness problem, as infrequent words have limited contextual information. Limiting the search space for each target word to the top five candidate categories reduces the computational cost that comes with the full search space.

3.1 Context Extraction and Representation

A generalized context for each semantic category is built from the contexts of its member words. This is done based on the assumption that as the words in the same category have the same or similar meaning, they tend to occur in similar contexts. In terms of context extraction and representation, we need to consider four factors.

Member Words The issue here is whether to include the contexts of polysemous member words in building the generalized context of a category. Including these contexts without discrimination introduces noise. To measure the effect of such noise, we build two versions of generalized context for each category, one using contexts of unambiguous member words only, and the other using contexts of all member words.

Context Words There are two issues in selecting words for context representation. First, words that contribute little information to the discrimination of meaning of other words, including conjunctions, numerals, auxiliaries, and non-Chinese sequences, are excluded. Second, to model the effect of frequency on the context words' contribution to meaning discrimination, we use two sets of context words: one consists of the 1000 most frequent words in the corpus; the other consists of all words in the corpus.

Window Size For WSD, both topical context and microcontext have been used (Ide and Véronis 1998). Topical context includes substantive words that co-occur with the target word within a larger window, whereas microcontext includes words in a small window around the target word. We experiment with topical context and microcontext with window sizes of 100 and 6 respectively (i.e., 50 and 3 words to the left and right of the target word respectively).

Context Representation We represent the context of a category as a vector $<w_1, w_2, ..., w_n>$, where n is the total number of context words, and w_i is the weight of the ith context word. To arrive at this representation, we first record the number of times each context word occurs within a specified window of each member word of a category in the corpus as a vector $<f_1, f_2, ..., f_n>$, where f_i is the number of times the ith context word co-occurs with a member word of the category. We then compute the weight of a context word w in context c, $W(w, c)$, using mutual information and t-test, which were reported by Weeds and Weir (2005) to perform the best on a pseudo-disambiguation task. These weight functions are computed as in (10) and (11), where N denotes the size of the corpus.

$$W_{PMI}(w,c) = \log \frac{P(w,c)}{P(w)P(c)} \quad (10)$$

$$W_t(w,c) = \frac{P(w,c) - P(w)P(c)}{\sqrt{P(w,c)/N}} \quad (11)$$

3.2 Contextual Similarity Measurement

We compute the similarity between the context vectors of the unknown word and its candidate categories using cosine. The cosine of two n-dimensional vectors $\vec{x}$ and $\vec{y}$, $cos(\vec{x}, \vec{y})$, is computed as in (12), where x_i and y_i denote the weight of the ith context word in $\vec{x}$ and $\vec{y}$.

$$\cos(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (12)$$

4 Results

4.1 Experiment Setup

The models are developed and tested using the Contemporary Chinese Corpus from Peking University (Yu et al. 2002) and the extended *Cilin* released by the Information Retrieval Lab at Harbin Institute of Technology. The corpus

contains all the articles published in January, 1999 in *People's Daily*, a major newspaper in China. It contains over 1.12 million tokens and is word-segmented and POS-tagged. Table 1 summarizes the distribution of words in *Cilin*. Of the 76,029 words in *Cilin*, 35,151 are found in the Contemporary Chinese Corpus.

Length	Unambiguous	Polysemous	Total
1	2,674	2,068	4,742
2	39,057	5,403	44,460
3	15,112	752	15,864
4	9,397	942	10,338
≥5	590	34	624
Total	66,830	9,199	76,029

Table 1: Word distribution in the extended *Cilin*

We classify words into the third-level categories in the extended *Cilin*, which are equivalent to the small categories in the original *Cilin*. The development and test sets consist of 3,000 words each, which are randomly selected from the subset of words in *Cilin* that are two to four characters long, that have occurred in the Contemporary Chinese Corpus, and that are tagged as nouns, verbs, or adjectives in the corpus. The words in the development and test sets are also controlled for frequency, with 1/3 of them occurring 1-3 times, 3-6 times, and 7 or more times in the corpus respectively.

As Chen (2004) noted, excluding all the words in the development and test data in the testing stage worsens the data-sparseness problem for knowledge-based models, as some categories have few member words, and some characters appear in few words in some categories. To alleviate this problem, the remove-one method is used for testing the knowledge-based models. In other words, the models are re-trained for each test word using information about all the words in *Cilin except* the test word. The corpus-based model is trained once using the training data only, as the data-sparseness problem is alleviated by using generalized contexts of categories. Finally, if a word is polysemous, it is considered to have been correctly classified if the proposed category is one of its categories.

4.2 Results of the Baseline Model

Tables 2 and 3 summarize the results of the baseline model in terms of the accuracy of its best guess and best five guesses respectively.

The columns labeled "Non-filtered" report results where all categories are considered for each unknown word, and the ones labeled "POS-filtered" report results where only the categories that agree with the POS category of the unknown word are considered. In the latter case, if the target word is a noun, only the small categories under major categories A-D are considered; otherwise, only those under major categories E-L are considered. The results show that using POS information about the unknown word to filter categories improves performance. Variant 2 performs better when only the best guess is considered, indicating that it is useful to model the effect of position on a character's contribution to word meaning in this case. However, it is not helpful to be sensitive to character position when the best five guesses are considered.

Model variant	Non-filtered		POS-filtered	
	Dev	Test	Dev	Test
1	0.391	0.398	0.450	0.464
2	0.471	0.469	**0.514**	**0.517**

Table 2: Results of the baseline model: best guess

Model variant	Non-filtered		POS-filtered	
	Dev	Test	Dev	Test
1	0.757	0.760	**0.813**	**0.817**
2	0.764	0.762	0.809	0.805

Table 3: Results of the baseline model: best 5 guesses

4.3 Results of the Character-Category Association Model

In this model, only categories that agree with the POS category of the unknown word and that share at least one character with the unknown word are considered. These filtering steps significantly reduce the search space for this model.

We discussed three parameters of the model in Section 2.2, including the statistical measure, the sensitivity to character position in computing character-category associations, and the weights of the associations between categories and characters in different positions. In addition, the computation of the character-category associations can be sensitive or insensitive to the POS categories of the words containing the characters. In the POS-sensitive way, associations are computed among nouns (words in categories A-D) and non-nouns (words in categories E-L) separately, whereas in the POS-insensitive way, they are computed using all the words.

Tables 4 and 5 summarize the results of the character-category association model in terms of the accuracy of its best guess and best five guesses respectively. In all cases, the weights assigned to word-initial, word-middle, and word-final characters are 0.49, 0, and 0.51 respectively.

In terms of the best guess, the model achieves a best accuracy of 58.2%, a 6.5% improvement over the baseline result. The results show that χ^2 consistently performs better than mutual information, and computing position-sensitive character-category associations consistently improves performance. However, computing POS-sensitive associations gives mixed results.

In terms of the best five guesses, the model achieves a best accuracy of 83.8% on the test data, a 2.1% improvement over the best baseline result. Using χ^2 again achieves better results. However, in this case, the best results are achieved when the character-category associations are insensitive to both character position and the POS categories of words.

Sensitivity		Development		Test	
POS	Position	MI	χ^2	MI	χ^2
Yes	Yes	0.482	**0.586**	0.507	**0.582**
Yes	No	0.440	0.578	0.458	0.573
No	Yes	0.487	0.565	0.511	0.567
No	No	0.457	0.555	0.459	0.559

Table 4: Results of the character-category association model: best guess

Sensitivity		Development		Test	
POS	Position	MI	χ^2	MI	χ^2
Yes	Yes	0.735	0.805	0.720	0.810
Yes	No	0.743	0.828	0.754	0.821
No	Yes	0.702	0.813	0.718	0.812
No	No	0.735	**0.830**	0.746	**0.838**

Table 5: Results of the character-category association model: best 5 guesses

Word	Development			Test		
Len	R	P	F	R	P	F
2	0.159	0.796	0.265	0.158	0.772	0.262
3	0.368	0.838	0.511	0.351	0.830	0.493
4	0.582	0.852	0.692	0.540	0.900	0.675
All	0.218	**0.816**	0.344	0.216	**0.803**	0.340

Table 6: Results of the rule-based model: best guess

4.4 Results of the Rule-Based Model

Table 6 summarizes the results of the rule-based model in terms of recall, precision and F-score. The model returns multiple categories for some words, and it is considered to have correctly classified a word only when it returns a single, correct category for the word. Precision of the model is computed over all the cases where the model returns a single guess, and recall is computed over all cases. The model achieves an overall precision of 80.3% on the test data, much higher than the accuracy of the other two knowledge-based models. However, recall of the model is only 21.6%. The comparable results on the development and test sets indicate that the encoded rules are general. The model generally performs better on longer words than on shorter words.

4.5 Combining the Character-Category Association and Rule-Based Models

Given that the rule-based model achieves a higher precision but a lower recall than the character-category association model, the two models can be combined to improve the overall performance. In general, if the rule-based model returns one or more categories, these categories are first ranked among themselves by their associations with the unknown word. They are then followed by the other categories returned by the character-category association model. Tables 7 and 8 summarize the results of combining the two models.

Sensitivity		Development		Test	
POS	Position	MI	χ^2	MI	χ^2
Yes	Yes	0.561	**0.623**	0.572	**0.616**
Yes	No	0.536	0.622	0.542	0.615
No	Yes	0.562	0.610	0.575	0.608
No	No	0.530	0.601	0.532	0.606

Table 7: Results of combining the character-category association and rule-based models: best guess

Sensitivity		Development		Test	
POS	Position	MI	χ^2	MI	χ^2
Yes	Yes	0.834	0.846	0.845	0.843
Yes	No	0.791	**0.860**	0.801	0.851
No	Yes	0.760	0.848	0.742	0.845
No	No	0.773	0.859	0.782	**0.856**

Table 8: Results of combining the character-category association and rule-based models: best 5 guesses

In terms of the best guess, the combined model achieves an accuracy of 61.6%, a 3.4% improvement over the best result of the character-category association model alone. This is achieved using χ^2 with POS-sensitive and position-sensitive computation of character-category associations. In terms of the best five guesses, the model achieves an accuracy of 85.6%, a 1.8% improvement over the best result of the character-category association model alone.

To facilitate comparison with previous studies, the results of the combined model in terms of its best guess in classifying unknown words into major and medium categories are summarized in Table 9. As χ^2 consistently outperforms mutual information, results are reported for χ^2 only. With POS-sensitive and position-sensitive com-

putation of character-category associations, the combined model achieves an accuracy of 83.0% and 69.9% for classifying unknown words into major and medium categories respectively.

Sensitivity		Development		Test	
POS	Position	Major	Med	Major	Med
Yes	Yes	**0.840**	**0.705**	**0.830**	**0.699**
Yes	No	0.831	0.698	0.828	0.698
No	Yes	0.832	0.692	0.825	0.692
No	No	0.821	0.687	0.821	0.689

Table 9: Results of the combined model for classifying unknown words into major and medium categories: best guess

4.6 Results of the Corpus-Based Model

The corpus-based model re-ranks the five highest ranked categories proposed by the combined knowledge-based model. Table 10 enumerates the parameters of the model and lists the labels used to denote the various settings in Table 11.

Parameter	Label	Setting	Label
Member words	MW	All members words	all
		Unambiguous members	un
Context words	CW	All words	all
		1000 most frequent	1000
Window size	WS	100	100
		6	6
Weight function	WF	Mutual information	mi
		t-test	t

Table 10: Parameter settings of the corpus-based model

Table 11 summarizes the results of 16 runs of the model with different parameter settings. The best accuracy on the test data is 37.1%, achieved in run 5 with the following parameter settings: using unambiguous member words for building contexts of categories, using all words in the corpus for context representation, using a window size of 100, and using mutual information as the weight function. As the combined knowledge-based model gives an accuracy of 85.6% for its best five guesses, the expected accuracy of a naive model that randomly picks a candidate for each word as its best guess is 17.1%. Compared with this baseline, the corpus-based model achieves a 13.0% improvement, but it performs much worse than the knowledge-based models.

Table 12 summarizes the accuracy of the top three runs of the model on words with different frequency in the corpus. Each of the three groups consists of 1,000 words that have occurred 1-2, 3-6, and 7 or more times in the corpus respectively. The model consistently performs better on words with higher frequency, suggesting that it may benefit from a larger corpus.

Run ID	Parameter Setting				Accuracy	
	MW	CW	WS	WF	Dev	Test
1	un	1000	100	mi	0.326	0.303
2	un	1000	100	t	0.317	0.288
3	un	1000	6	mi	0.304	0.301
4	un	1000	6	t	0.299	0.301
5	un	all	100	mi	0.359	**0.371**
6	un	all	100	t	0.292	0.296
7	un	all	6	mi	**0.370**	0.365
8	un	all	6	t	0.322	0.297
9	all	1000	100	mi	0.302	0.294
10	all	1000	100	t	0.314	0.304
11	all	1000	6	mi	0.313	0.314
12	all	1000	6	t	0.308	0.308
13	all	all	100	mi	0.336	0.333
14	all	all	100	t	0.287	0.300
15	all	all	6	mi	0.356	0.356
16	all	all	6	t	0.308	0.308

Table 11: Results of the corpus-based model

Run ID	Development			Test		
	1-2	3-6	≥7	1-2	3-6	≥7
5	0.331	0.360	0.385	0.323	0.389	0.402
7	0.323	0.363	0.423	0.335	0.357	0.402
15	0.328	0.346	0.395	0.334	0.355	0.379

Table 12: Results of the corpus-based model on words with different frequency

5 Related Work

The few previous studies on semantic classification of Chinese unknown word have primarily adopted knowledge-based models. Chen (2004) proposed a model that retrieves the word with the greatest association with the target word. This model is computationally more expensive than our character-category association model, as it entails computing associations between every character-category, category-character, character-character, and word-word pair. He reported an accuracy of 61.6% on bisyllabic V-V compounds. However, he included all the test words in training the model. If we also include the test words in computing character-category associations, the computationally cheaper model achieves an overall accuracy of 75.6%, with an accuracy of 75.1% on verbs.

Chen and Chen (2000) adopted similar exemplar-based models. Chen and Chen used a morphological analyzer to identify the head of the target word and the semantic categories of its modifier. They then retrieved examples with the same head as the target word. Finally, they computed the similarity between two words as the

similarity between their modifiers, using the concept of information load (IC) of the least common ancestor (LCA) of the modifiers' semantic categories. They reported an accuracy of 81% for classifying 200 unknown nouns. Given the small test set of their study, it is hard to directly compare their results with ours.

Tseng used a morphological analyzer in the same way, but she also derived the morpho-syntactic relationship between the morphemes. She retrieved examples that share a morpheme with the target word in the same position and filtered those with a different morpho-syntactic relationship. Finally, she computed the similarity between two words as the similarity between their non-shared morphemes, using a similar concept of IC of the LCA of two categories. She classified unknown words into the 12 major categories only, and reported accuracies 65.8% on adjectives, 71.4% on nouns, and 52.8% on verbs. These results are not as good as the 83.0% overall accuracy our combined knowledge-based model achieved for classifying unknown words into major categories.

Chen and Lin (2000) is the only study that used contextual information for the same task. To generate candidate categories for a word, they looked up its translations in a Chinese-English dictionary and the synsets of the translations in WordNet, and mapped the synsets to the categories in *Cilin*. They used a corpus-based model similar to ours to rank the candidates. They reported an accuracy of 34.4%, which is close to the 37.1% accuracy of our corpus-based model, but lower than the 61.6% accuracy of our combined knowledge-based model. In addition, they could only classify the unknown words listed in the Chinese-English dictionary.

6 Conclusions

We presented three knowledge-based models and a corpus-based model for classifying Chinese unknown words into fine-grained categories in the Chinese thesaurus *Cilin*, a task important for lexical acquisition and NLP applications that require semantic annotation. The knowledge-based models use information about the categories of the unknown words' component characters, while the corpus-based model uses contextual information. By combining the character-category association and rule-based models, we achieved an accuracy of 61.6%. The corpus-based model did not improve performance.

Several avenues can be taken for further research. First, additional resources, such as bilingual dictionaries, morphological analyzers, parallel corpora, and larger corpora with richer linguistic annotation may prove useful for improving both the knowledge-based and corpus-based models. Second, we only explored one way to combine the knowledge-based and corpus-based models. Future work may explore alternative ways to combine these models to make better use of contextual information.

References

C.-J. Chen. 2004. Character-sense association and compounding template similarity: Automatic semantic classification of Chinese compounds. In *Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing*, pages 33–40.

M. Ciaramita and M. Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In Proceedings of EMNLP-2003, pages 594-602.

K.-J. Chen and C.-J. Chen. 2000. Automatic semantic classification for Chinese unknown compound nouns. In *Proceedings of COLING-2000,* pages 173-179.

H.-H. Chen and C.-C. Lin. 2000. Sense-tagging Chinese corpus. In *Proceedings of the 2nd Chinese Language Processing Workshop*, pages 7-14.

J. Curran. 2005. Supersense tagging of unknown nouns using semantic similarity. In Proceedings of ACL-2006, pages 26-33.

N. Ide and J. Véronis. 1998. Introduction on the special issue on word sense disambiguation: The state of the art. *Computational Linguistics 24*(1):2–40.

J. Mei, Y. Zhu, Y. Gao, and H. Yin. (eds.) 1984. *Tongyici Cilin [A Thesaurus of Chinese Words]*. Commercial Press, Hong Kong.

B. Roark and E. Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of COLING/ACL-1998,* pages 1110-1116.

H. Tseng. 2003. Semantic classification of Chinese unknown words. In *Proceedings of ACL-2003 Student Research Workshop,* pages 72-79.

J. Weeds and D. Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics 31*(4):439–475.

S. Yu, H. Duan, X. Zhu, and B. Sun. 2002. The basic processing of Contemporary Chinese Corpus at Peking University. *Journal of Chinese Information Processing 16*(5):49–64.