International Journal of Knowledge and Language Processing
Volume 5, Number 4, **2014**

www.ijklp.org
KLP International ⓒ2014 ISSN 2191-2734
pp.66—84

Survey on Sense Guessing of Chinese Unknown Words

Fenfen Shang[1], Weiguang Qu[1] , Bin Li[2] and Yanhui Gu[1]

[1] School of Computer Science and Technology
Nanjing Normal University
Nanjing, Jiangsu, 210023, China
shangfenfen1009@163.com; wgqu@njnu.edu.cn; gu@njnu.edu.cn

[2] School of Chinese Language and Literature
Nanjing Normal University
Nanjing, Jiangsu, 210097, China
gothere@126.com

Received December 2014; revised December 2014

ABSTRACT. *With the widespread use of the network, the number of Chinese unknown words is rapidly growing. However, the senses of these words are unknown, which brought a lot of difficulties in a grammatical analysis, information retrieval, machine translation and other Chinese natural language processing tasks, etc. Therefore, Chinese word semantic prediction is one of the important research issues in NLP. However, there is no unknown word vocabulary and no meaning set pre-defined, making the unknown word sense prediction is much more difficult than semantic disambiguation. Related research focused on the unknown words in the context of morpheme structure or composition between the unknown words, but the effect of semantic prediction is not very satisfactory. In recent years, more researches integrate these models into a hybrid model to improve the semantic prediction performance. In this paper, we introduce and classify a variety of mainstream semantic prediction methods and cutting-edge methods. We summarily introduce the database and the methods of evaluation for semantic prediction of Chinese unknown words. Finally, semantic logged forecast for future research directions and trends is discussed.*

Keywords: Chinese unknown words; Sense guessing

1. **Introduction.** Unknown words are the words which exist in the corpus but not in the dictionary. For the approach which is not based on dictionary, unknown words are the

words that exist in the test set but not in the training set. Unknown words appear extensively in Chinese corpus. In Chinese text, there is no blank to mark word boundaries and no inflectional markers nor capitalization markers to denote the syntactic or semantic types of new words. Hence the unknown words for Chinese became one of the most difficulties in Chinese natural language processing tasks, for example, a grammatical analysis, information retrieval, machine translation. Because of the dynamic characteristic of language, new words are created daily. However, there is on dictionary can conclude all of the everyday new words, so automatic sense guessing is needed for occur unknown words to determine the sense of unknown word.

Besides part-of-speech tagging, another important level of lexical annotation is semantic annotation. Semantic annotation is crucial and sometimes required for a number of natural language applications. Based on the type of information provided in the annotation and the proportion of words in the text that are annotated, the following four levels of semantic annotation can be differentiated (Stevenson 2003).

(1) Semantic Disambiguation. This is the most general level and refers to any task in which some form of semantic annotation is applied to text. No restrictions are imposed on the type of semantic annotation or the portion of text being annotated.

(2) Semantic Tagging. This is the same as semantic disambiguation except that all words in the text are annotated.

(3) Sense Disambiguation. Like semantic disambiguation, this does not require all words to be annotated, but the annotations applied must now be senses from some lexicon, instead of just any semantic tag.

(4) Sense Tagging. This is the most specific level. It requires all words in the text to be tagged with senses from some lexicon.

The previous research of Chinese unknown words focuses on proper names, but in recent years the research tends to compound nouns and compound verbs. In fact, compound nouns are the unknown words which occurred in Chinese text most frequently, but not proper names. According to an inspection on the Sinica corpus (Chen etc. 1996), 3.51% of the word tokens in the corpus are unknown. Among them, about 51% of the word types are compound nouns, 34% are compound verbs and 15% are proper names. So resent researches are facilitated to identify, to disambiguate and to evaluate the structure of a compound noun.

In [18], the limited coverage of lexical-semantic resources is a significant problem for NLP systems which can be alleviated by automatically classifying the unknown words. Supersense tagging assigns unknown nouns one of 26 broad semantic categories used by lexicographers to organize their manual insertion into WordNet. They describe an unsupervised approach, based on vector-space similarity, which does not require annotated examples but significantly outperforms their tagger. We also demonstrate the use of an extremely large shallow-parsed corpus for calculating vector-space semantic similarity. Based on the works of [19, 20], to classify a previously unseen common noun our approach extracts synonyms. They have experimented with the synonyms, filtered by frequency and number of contexts to increase their reliability. Using this approach we have significantly outperformed the supervised multi-class perceptron in [21].

The methods of Chinese unknown words research can be simply classified by the method based on rules and the method based on statistics. In the research of compound words, previous researchers mainly classify the unknown words based on the contextual characteristic. Modern researches mainly based on morpheme or based on contextual information or the hybrid models which combined them. According to the multiple models and algorithms which are proposed by researchers, this paper concludes and try to analysis which method can predict the sense of unknown words more effective. This paper concludes several methods as follows:

(1) Exemplary-based model in [4] focuses on non-contextual features, which may play a key role for unknown words that occur only once and hence have limited context. The feature he focuses on is morphological similarity to words whose semantic category is known. His nearest neighbor approach to lexical acquisition computes the distance between an unknown word and examples from the CiLin thesaurus based upon its morphological structure. In Chinese morphology, the two ways to generate new words are compounding and affixation. Orthographically, such compounding and affixation is represented by combinations of characters, and as a result, the character combinations and the morpho-syntactic relationship used to link them together can be clues for classification. Furthermore, his analysis of the Sinica Corpus indicates that only 49.68% monosyllabic3 words have one word class, but 91.67% multisyllabic words have one word class. Once characters merge together, only 8.33% words remain ambiguous. It implies that as characters are combined together, the degree of ambiguity tends to decrease.

(2) The similarity-based model. In [3], the author describes a similarity-based model to present the morphological rules for Chinese compound nouns. This representation model serves functions of 1) as the morphological rules of the compounds, 2) as a mean to evaluate the properness of a compound construction, and 3) as a mean to disambiguate the semantic ambiguity of the morphological head of a compound noun. The presupposition of automatic semantic classification for compounds is that the meaning of a compound is the semantic composition of its morphemic components and the head morpheme determines the major semantic class of this compound. The meaning of the input unknown word will be assigned with the moaning of the class with the most similarity morpho-semantic structures with this unknown word.

(3) The word information knowledge-based models. In [1] the 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. The models model the relationship between the semantic category of an unknown word and those of its component characters in different ways.(1) the overlapping-character 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. It is the most simple and most direct method. The results also show that using the POS category of the unknown word to filter out irrelevant candidate categories improves performance. (2) The character-category association model: The model uses statistical measures to compute

the strength of association between characters and categories and subsequently between words and categories. The results indicate that in the computation of character-category associations, it is useful to be sensitive to the positions in which the characters occur and the parts of speech of the words containing the characters. Assigning different weights to the associations between categories and characters in different positions improves performance, too. What is more, using the POS category of the unknown word to filter out irrelevant candidate categories improves performance. (3) The rule-based model: The rule-based model directly encodes the regularities observed in the relationship between the categories of a subset of unknown words and those of their component characters. (4) In [10], a three-layer backpropagation neural net (BPNN) was used to simulate the dependence between the semantic categories of a disyllabic word and those of its two component characters.

(4) The corpus-based model: in [1, 2] the corpus-based model uses the context in which the unknown word occurs to predict its category. This model performs better on more frequent words than on less frequent words, but the results are much lower than those of the knowledge-based models.

(5) The model combined contextual and structural information. In [5], context-based approach and structural information-based approach are treated as independent model with each other. (1) Structural information-based approach: includes two kinds of structural information-based models. (2) Structure-based approach is tightly coupled with context-based approach. Firstly, the structure-based approach generates a candidate synonym set. The context-based method would be used to measure the similarity between the test word and each of the word among the candidate synonym set. Secondly, the structure-based approach adjusts the similarity scores computed by context-based approach.

(6) The HowNet-based model. In [12], they propose a HowNet-based model of Chinese unknown words semantic analysis. This model treats the concept graph as representing method and treats the HowNet 2005 as the semantic knowledge resource. Firstly, the model bases on the HowNet dictionary to divide common unknown words. Second, comprehensively making use of the knowledge system of HowNet, the Chinese information structure is disambiguated by the part of speech sequence matching disambiguation method, concept map compatibility judgment disambiguation method, concept map compatibility disambiguation method and semantic similarity computation disambiguation method. Finally, the concept graph of the unknown words is generated based on the selected Chinese information structure. Then the semantic analysis of unknown words can be realized. During the semantic analysis process, on the one hand, this model determines the sense of the known words in an unknown word. On the other hand, this model structures the sense information of unknown words.

The first chapter of this paper introduces the data sets and the evaluate methods of unknown words. The second chapter elaborates the different models and algorithms, then analysis and compares them particularly. The third chapter summarizes the related work and error analysis. Finally, this paper analyzes the future work of Chinese unknown words and looks forward to the next research direction.

2. **Data Set and Evaluation Method.**

2.1. **Data Set.**

Recently, the research of the sense guessing of Chinese unknown words most frequently uses the dictionary, the electronic version of the CiLin and the extended CiLin thesaurus released by the Information Retrieval Lab at Harbin Institute of Technology in March, 2005. CiLin is written by Mei in 1983, including not only the synonyms but also the words with the same sense category which is the generalized relative words. Because of the CiLin was written many years ago, rarely updated later. So the Information Retrieval Lab at Harbin Institute of Technology made use of many related resources to finish an extended CiLin thesaurus.

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'.

Except for using the CiLin as the data set, some researches use HowNet, such as [12]. HowNet is a General Knowledge base which based on the concept represented by English-Chinese and the characteristic of the concept. HowNet describes the relation between concept with concept and between the characteristics of concept [13]. According to the difference of the knowledge, it can be classified to three parts, as knowledge dictionary, the file of sememe characteristic, Chinese information structure base. The knowledge dictionary describes the part of speech and concept term of each word through the record form. On the one hand, the file of sememe characteristic describes the hierarchical relationship of homogeneous sememe. On the other hand, the file describes the semantic relationship of different sememe. Chinese information structure base describes the dynamic role relationship and properties between each component of Chinese word, indicating Chinese language structure rule [14]. This knowledge can be as semantic resource of semantic analysis of unknown words, that is to say, treating knowledge dictionary as basic dictionary, treating Chinese information structure base as the collocation rule base. Refer to all files of sememe characteristic, different approaches are used to disambiguate for the rule in the information structure base. Then, the appropriate information structure can be chose for unknown words. The appropriate concept term can be chose for the known words in unknown words. Finally, the semantic structure of unknown words can be determined.

2.2. **Corpus Resource.**

In [4], unknown words are the Sinica Corpus lexicons that are not listed in the Chinese Electronic Dictionary of 80,000 lexicons and the CiLin. In the research of Chinese unknown words, the more use corpus is Sinica Corpus. The 5 million words Sinica Corpus contain 77,866 unknown words.

In [5] The Contemporary Chinese Corpus, which is segmented and POS-tagged, contains all the news articles (over 1.12 million tokens) of the People's Daily newspaper published in China from January to June 1998.

2.3. **Evaluation Method.** In the evaluation methods of Chinese unknown words, the most frequently used measure is accuracy, which is used to show the performance of the prediction. Accuracy reflects the capability of judgment of the classifier system about the whole samples-whether it can sentence the correct samples to correct and sentence the wrong samples to wrong. The accuracy in the sense guessing of Chinese unknown words denotes that the proportion of the unknown words which are assigned the correct semantic category in the all unknown words. Besides the accuracy, F1-score also is used to measure model performance. F1-score demonstrates the result of accuracy and recall. The higher F1-score gains, the more perfect the experimental approach performs. In [5], the author used accuracy and F1-score to evaluate the performance of Chinese unknown words.

3. **Models.**

3.1. **An example-based model.**

In [4, 11], morphological analysis is a primary step for predicting the syntactic and semantic categories of out-of-vocabulary (unknown) words. The designed Chinese morphological analyzer contains three major functions, 1) to segment a word into a sequence of morphemes, 2) to tag the part-of-speech of those morphemes, and 3) to identify the morpho-syntactic relation between morphemes, for example, modifier-head construction, verb-object construction. In the case of '舞蹈家', if it is an unknown word, it will be segment as '舞蹈' and '家', whose syntactic relation is modifier-head construction .

① The CiLin thesaurus is then searched for entries (examples) that are similar to the unknown word. A list of words sharing at least one morpheme with the unknown word, in the same position, is constructed. In the case of 舞蹈家/wudaojia, such a list would include 歌唱家/gechangjia SINGEXPERT 'singer', 回家/huijia GO-HOME 'go home', 富贵家/fuguijia RICH-FAMILY 'rich family' and so on.

② The examples that do not have the same morpho-syntactic relationships but shared morpheme belongs to the unknown word's modifier are pruned away. If no examples are found, the system falls back to the baseline classification method.

③ They assume that similarity of two semantic categories is the information content of their parents' node. For instance, the similarity of 哈密瓜/hamigua 'hami melon' (Bh07) and 番茄/fanqie 'tomato' (Bh06) is based on the information content of the node of their least common ancestor Bh. The CiLin thesaurus can be used as an information system, and the information content of each semantic category is defined as:

$$\mathit{Entropy(System)} - \mathit{Entropy(Semantic\ category)} \tag{1}$$

The similarity of two words is the least common ancestor information content(IC), and hence, the higher the information content is, the more similar two the words are. To simplify the computation, the probabilities of all leaf nodes are assumed equal. The similarity of two words is defined as:

$$Sim(W_1 \cap W_2) = \frac{IC(W_1 \cap W_2)}{Entropy(System)} = \frac{-\log_2(P(W_1 \cap W_2))}{Entropy(System)} \qquad (2)$$

The CiLin thesaurus can be used as an information system, and the information content of each semantic category is denoted as IC.

One problem for this algorithm is the insufficient coverage of the CiLin (CiLin may not cover all morphemes). The backup method is to run the classifier recursively to predict the possible categories of the unlisted morphemes. If a morpheme of an unknown word or of an unknown word's example is not listed in the CiLin, the similarity measurement will suspend measuring the similarity between the unknown word and the examples and run the classifier to predict he semantic category of the morpheme first. After the category of the morpheme is known, the classifier will continue to measure the similarity between the unknown word and its examples.

After the distances from the unknown word to each of the selected examples from the CiLin thesaurus are determined, the average distance to the K nearest neighbors from each semantic category is computed. The category with the lowest distance is assigned to the unknown word.

The semantic category is predicted as the category that gets the highest score in formula (3). The lexical similarity and frequency of examples of each category are considered as the most important features to decide a category.

Let W_1=unknown word, W_i =word whose semantic category defined in the CiLin, I=A…L (CiLin Taxonomy):

$$Rankscore\text{（}C_i) = \alpha * SS(C_i) + (1-\alpha) * FS(C_i) \qquad (3)$$

$$SS(C_i) = \underset{C(W_i)\in C_i}{\overset{i=A...L}{\arg\max}}\, Sim(W_1, W_i) \qquad (4)$$

$$FS(C_i) = \frac{Freq(C_i)}{\sum_{i=A}^{L} Freq(C_i)} \qquad (5)$$

The score of SS(C_i) is a lexical similarity score, which is from the maximum score of Similarity (W_1,W_2) in the category of W2. FS(C_i) is a frequency score to show how many examples there are in a category. α and (1-α) are respectively weights for the lexical similarity score and the frequency score.

For experiments, CiLin lexicons are divided into 2 sets: a training set of 80% CiLin words, a development set of 10% of CiLin words, and a test set of 10% CiLin words. On the test set, the baseline predicts 53.50% of adjectives, 70.84% of nouns and 47.19% of verbs correctly. The classifier reaches 64.20% in adjectives, 71.77% in nouns and 53.47%

in verbs, when α is 0.5 and K is five. The main contributions of the system are: first, it is the first attempt in adding semantic knowledge to Chinese unknown words. Since over 70% of unknown words are lexical words, the inability to resolve their meaning is a major obstacle to Chinese NLP such as semantic parsers. Second, without contextual information, the system can still successfully classify 65.76% of adjectives, 71.39% of nouns and 52.84% of verbs.

3.2. **Similarity-based Model.** In [3], different syntactic categories require different representational models and different fine-grained semantic classification methods. The presupposition of automatic semantic classification for compounds is that the meaning of a compound is the semantic composition of its morphemic components and the head morpheme determines the major semantic class of this compound. There are many polysyllabic words of which the property of semantic composition does not hold, for instances the transliteration words which should be listed in the lexicon. Since for the majority of compounds the presupposition hold, the design of our semantic classification algorithm will be based upon this presupposition. Therefore the process of identifying semantic class of a compound boils down to find and to determine the semantic class of its head morpheme. However ambiguous morphological structures cause the difficulties in finding head morpheme.

Example-based similarity measure supposed that a compound has the structure of XY where X and Y are morphemes and supposed without loss of generality Y is the head. In [3] supposed that each class of examples forms the following semantic relation rules. The rules show the possible semantic relations between prefix and suffix Y and their weight in term of the frequency distribution of each semantic category of the prefixes in the class. The rules are （Sem_1+Y, $Freq_1$）,（Sem_2+Y, $Freq_2$）,…,（Sem_k+Y, $Freq_k$）, where $Freq_i$ denotes the number of the words of the form Sem_i +Y).

The similarity is measured between the semantic class of the prefix X of the unknown compound and the prefix semantic types shown in the rule. One of the measurements proposed is, as follows, where Sem is the semantic class of X.

$$Similar(Sem, Rule) = \{\sum_{i=1,k} Information - Load(Sem \cap Sem_i) * Freq_i\} / Max - value \qquad (6)$$

Max-value is the maximal value of { $\sum_{i=1,k} Information - Load(S \cap Sem_i) * Freq_i$ }. The max-value normalizes the SIMILAR value to 0-1. $S \cap Sem_i$ denotes the least common ancestor of S and Sem_i.

The above similarity measure is called over-all-similarity measure, since it takes the equal weight on the similarity values of the input compound with every member in the class. Another similarity measure is called maximal-similarity, which is defined as follows. It takes the maximal value of the similarity between input compound and every member in the class as the output.

$$Similar(Word, Rule) = \underset{i=1,k}{Max}\{Information-Load(Sem \cap Sem_i)) / Max-value\} \quad (7)$$

3.3. **The structure information knowledge-based model.**

(1) In [1], the Baseline Model is the most simple and most direct 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.

For each semantic category in the CiLin thesaurus, the set of all unique component characters of its member words are extracted, and the total number of times each of these characters occurs in word-initial, word-middle, and word-final positions are recorded. With this information, the following three pairs of variants of the baseline model are proposed. In each pair, an a variant computes the score of a category for an unknown word by counting the number of overlapping characters between the category and the unknown word in a particular way, and its corresponding b variant computes a weighted or normalized counterpart of that score. In each of these variants, Score(Cat,w) denotes the score assigned to a category Cat for an unknown word w, n is the length of w (i.e., the number of characters in w), c_i is the ith character in w, p_i is the position of c_i in w, $p_i \in$ {word-initial, word-middle, word-final}, f (c_i) is the overall frequency of c_i in Cat, and f (c_i, p_i) is the frequency of c_i in position p_i in Cat, N is the total number of characters in Cat, Np_i is the total number of characters in position p_i in Cat, and N_w is the total number of words in Cat.

Pair One: In a variant, the score of a category is the sum of the number of occurrences of each component character of the unknown word in the category. In the b variant, each number is weighted by the total number of characters in the category.

$$\text{Variant 1a: } Score(Cat,w) = \sum_{i=1}^{n} f(c_i) \quad (8)$$

$$\text{Variant 1b: } Score(Cat,w) = \sum_{i=1}^{n} \frac{f(c_i)}{N} \quad (9)$$

Pair Two: In a variant, the score of a category is the sum of the number of occurrences of each component character of the unknown word in the category in its corresponding position. In the b variant, each number is weighted by the total number of characters in the corresponding position in the category.

$$\text{Variant 2a: } Score(Cat,w) = \sum_{i=1}^{n} f(c_i, p_i) \quad (10)$$

$$\text{Variant 2b: } Score(Cat,w) = \sum_{i=1}^{n} \frac{f(c_i, p_i)}{N_{p_i}} \quad (11)$$

Pair Three In a variant, the score of a category is the total number of occurrences of the final character c_n of the unknown word in the word-final position p_n in the category. In the b variant, the score is weighted by the total number of words in the category.

Variant 3a: $Score(Cat,w) = f(c_n, p_n)$ （12）

Variant 3b: $$Score(Cat,w) = \frac{f(c_n, p_n)}{N_w}$$ （13）

The first pairs of variants capture the overlapping-character hypothesis in the most straightforward manner. The second pairs are sensitive to the positions in which the component characters occur in the unknown word and in the member words of the category. The last pairs look at the final character of the unknown word and the final characters of the member words in each category only. In each of these variants, the category with the maximum score is proposed as the category for the unknown word.

(2) The Character-Category Association Model

The following two statistical measures are used to compute character category associations: mutual information and X^2.

$$Asso_{MI}(Char,Cat_j) = log\frac{P(Char,Cat_j)}{P(Char)P(Cat_j)}$$ （14）

$$Asso_{\chi^2}(Char,Cat_j) = \frac{\alpha(Char,Cat_j)}{Max_k\alpha(Char,Cat_j)}$$ （15）

$$\alpha(Char,Cat_j) = \sqrt{\frac{[f(Char,Cat_j)]^2}{f(Char)+f(Cat_j)}}$$ （16）

The category of word can be calculated as the sum of the weighted associations between a category and each of the word's component characters.

$$Asso(W,Cat_j) = \sum_{i=1}^{\backslash W\backslash}\lambda_i Asso(Char,Cat_j)$$ (17)

The character-category association model can also be made sensitive to the positions in which the characters occur in the words. In general, if a list of candidate categories is desired, an ordered list of candidate categories whose associations with the word surpass an empirically determined threshold can be proposed.

(3) A Rule-Based Model

The rule-based model uses a few rules to predict the semantic categories of a subset of unknown words based on the syntactic and semantic categories of their components. The rule-based approaches have not been used for sense guessing of unknown words in previous studies. However, there are several regularities in the data that can be captured in a more direct and effective way by rules than by statistical models. A separate set of rules are developed for words that are two, three, and four characters long.

1) For disyllabic unknown words AB:
   For a disyllabic word AB, if A and B share a common category c, 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 are both greater than or equal to the predetermined

thresholds, propose c for AB, e.g., tan-ta ‘collapse’, where tan and ta both mean ‘collapse’ and share the category Id05. The thresholds for f_A and f_B are determined empirically in the development stage. They are both set to 1 if the unknown word AB is a noun and to 0 and 3 respectively otherwise.

2) For a trisyllabic word ABC:
a. If BC equals xuéji¯a ‘-ist’, propose the category Al02 for ABC.
b. If A in {dà ‘big’, xiˇao ‘little’, lˇao ‘old’}, propose the category of AB for ABC, if C is the diminutive suffix er or the category of BC for ABC otherwise.

3) For a trisyllabic word ABC:
a. If A and BC share a common category c, propose c for ABC.
b. If AB and C share a common category c, propose c for ABC.
c. If there is a word XYC where XY shares a common category with AB, propose the category of XYC for ABC.
d. If there is a word XBC where X shares a common category with A, propose the category of XBC for ABC.

4) For a four-character word ABCD:
a. If CD equals xuéji¯a ‘-ist’ or dàxué ‘University’, propose the category Al02 or Dm05 for ABCD respectively.
b. If there is a word XYZD/YZD where XYZ/YZ shares a common category with ABC, propose the category of XYZ/YZ for ABCD.
c. If there is a word ABCX where X shares a common category with D, propose the category of ABCX for ABCD.
d. If there is a word XYCD where XY shares a common category with AB, propose the category of XYCD for ABCD.
e. If there is a word XBCD/BCD, propose the category of XBCD/BCD for ABCD.

3.4. **The corpus-based model.** The corpus-based model first uses 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.

① **Context Extraction and Representation**

For each semantic category, a generalized context is built from the contexts of its member words. Comparing the context of an unknown word with the generalized context of a category instead of the context of a particular word also helps alleviate the data-sparseness problem.

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, they 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

They experiment with topical context and micro-context with window sizes of 100 and 6 respectively.

Context Representation

They represent the context of a category as a vector. They then compute the weight of a context word w in context c, as follows, where N denotes the size of the corpus.

$$W_{PWI}(w,c) = log\frac{P(w,c)}{P(w)P(c)} \tag{18}$$

$$W_t(w,c) = \frac{P(w,c) - P(w)P(c)}{\sqrt{P(w,c)/N}} \tag{19}$$

② **Contextual Similarity Measurement**

They 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 $\vec{y}$ is calculated, where x_i and y_i denote the weight of the ith context word in $\vec{x}$ and $\vec{y}$, as follows:

$$\cos(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{20}$$

3.5. **The hybrid model.**

In [1], the hybrid model combines structure information and contextual information.

In [5], the main part of the hybrid model is structure-based approach. The context-based approach is used to sort the candidate sense category received by the structure approach. It is a loose combination, and the experience performance is not improved more. The hybrid approach included two steps. Firstly, the two knowledge-based models provide some candidate supersense for the unknown words. Secondly, the context information is computed and compared to determine which supersense should win.

Particularly, given an unknown word w and a thesaurus T, the following three steps are executed:

a) Candidate Synonym Set Building. A character filtering process and a POS (Part Of Speech) filtering process are completed to acquire some words sharing character and POS with w from T. Those words compose a set, which is called candidate synonym set and denoted by CS(w).
b) Contextual Similarity Computation. Collect context of w and words in CS(w) from a corpus. Denote context of a word {SS（w_i）}, where w_1=w or $w_1 \in$ CS(w). Define contextual similarity of w and $w_i \in$ CS(w) as sim(w, w_i)=λ(w,w_i) * CTS(w, w_i), where CTS(w, w_i) denotes the pure contextual similarity, while weighted by λ(w, w_i) (structural similarity between a test word and corresponding training word).
c) KNN Classification. A KNN classifier is employed to assign w a supersense s, which is

selected from $\{SS(w_i)\}$. Here, $SS(w_i)$ denotes the supersense of $w_i \in CS(w)$.

A KNN classifier takes the following steps:

(1) Rank $sim(w, w_i)$ from big to small.

(2) Keep $sim(w, w_i)$ unchanged for the first K words and set $sim(w, w_i)=0$ for the remaining.

(3) Collect supersenses of all $w_i \in CS(w)$ to form a supersense set $\{SS(w_i)\}$. For a supersense $s_j \in \{SS(w_i)\}$, denote its similarity sum as $sim(s_j)$, and compute it as $sim(s_j)=\Sigma sim(w, w_i)$, where w_i satisfies $SS(w_i)= s_j$.

(4) Assign w the supersense having the biggest similarity sum, i.e., $Argmax(sim(s_j))$.

3.6. **The HowNet-based model.** Concept graph is a knowledge presentation based on linguistics, psychology, philosophy, logistics and mathematics [15]. The concept graph consists of concept node and relation node. Concept node stands for entity, attribute, status and event. Relation node stands for the relationship between each concept. Because the concept graph establishes on predicate logic, it can translate completely natural language with each other and represent the sense of natural language [16]. HowNet mainly describes the relation between concept with concept and between the characteristics of concept, so the knowledge included by HowNet can be represented and analyzed sense by concept graph [17]. The concept is different from general ones. When the semantic analysis of unknown words model based on HowNet uses concept graph as the knowledge representation, the nodes of concept graph only include concept nodes and the knowledge represented by relation nodes integrate into arcs of concept graph.

Unknown words are the input of the semantic analysis of unknown words. The output in this task is the semantic information of the unknown words represented by concept graph. The basic idea of the task is to integrate the semantic information of unknown words through the semantic information of known words, according to the knowledge provided by HowNet. So there are four problems as follows:

(1) The problem of participle of unknown words intends to make unknown words consist of several known words.

(2) The problem of the representation of Chinese information structure intends to formalize Chinese information structure, which can be represented and can be done semantic analysis by concept graph.

(3) The problem of disambiguation includes the disambiguation to Chinese information structure and the sense disambiguation. It is designed to make sure the matching relationship of known words and the sense of polysemant.

(4) The problem of process of semantic analysis of unknown words intends to arrange reasonably the process of semantic analysis, to make unknown words gain efficient analysis.

In order to simplify calculation, it orders the unknown words have the same information structure with the example which has better similarity. So the calculation disambiguation of semantic similarity is gained through calculating the semantic similarity between known words and words in information structure base. Then the information structure of the example which has the maximum similarity is chose as the information structure of unknown words. The disambiguation of information structure can come true. The key is how to calculate the similarity between unknown words and example. If the unknown words choose one information structure, the semantic structure of them is consistent with the information structure and the complete binary tree, SYN_TREE, constructed by SYN_S. Therefore, the integral similarity can be get, according to integrate the similarity between the words in unknown word and the words in example based on SYN_TREE.

Assuming that unword denotes as unknown words and example denotes as word example. The complete binary tree constructed according to information structure, as SYN_TREE. The calculation of semantic similarity can divide into two steps.

The first step is to calculate the similarity of each word in unword and word in example, then to put in array simPart, refer to [14].

The second step is to integrate the integral similarity, according to the structure of SYN_TREE. The algorithm is simTree(SYN_TREE, simPart, a, b), where a and b denote parameters. a and b denote the weight of head word and modifier in the process of calculating integral similarity. a + b = 1,a ＞ =0.5, the basic idea of simTree is as follow.

When the left sub tree and right sub tree of SYN_ TREE exist, the simTree is called recursively, as (21):

$$sim = simTree(TREElchild, simPart, a, b) * a + simTree(TREErchild, simPart, a, b) * b \quad (21)$$

, where TREE.lChild and TREE.rChild denote the left sub tree and the right sub tree of SYN_TREE respectively.

When SYN_TREE is the leaf node, sim = (the similarity of word graph the location of the corresponding of the TREE).

The process of semantic analysis of unknown words is as follows:

Unword denotes the input unknown words. The semantic analysis of unword divided three steps:

The first step is to participle and to save the result into synList.

The second step is to do something to every participle result as follow:

(1) For the words in unword, searching the HowNet knowledge dictionary, saving the part of speech and corresponding concept term into wordlist. So if unword has n words, there are n-wordlist.

(2) For n wordlist, calculating the Cartesian product. So the Cartesian product of wordlist includes all the combination of the concept terms of the word in unword.

(3) For every line of Cartesian product, the first and second semantic structure can be disambiguated and constructed the information structure set by the part of speech sequence matching disambiguation method and concept map compatibility judgment disambiguation method. If the better information structure set only has one information structure, the concept of the word can be determined by the information structure, and turn to step three. Otherwise, using concept map compatibility calculates the disambiguation to cut down the better information structure. Then using the part of speech sequence matching

disambiguation method and concept map compatibility judgment disambiguation method disambiguates the third information structure and add the third information structure after cutting down into better information structure set and disambiguate again the better information structure set by semantic similarity computation disambiguation and write down the maximum similarity, the tab of Cartesian product and the tab of corresponding information structure.

(4) After dealing with every line of Cartesian product, they choose the maximum similarity in the similarity set as the similarity between information structure and semantic similarity of unknown words, then treat the corresponding information structure as the information structure of unknown words and treat the corresponding concept term of the corresponding Cartesian product as the concept term of words in unknown words.

The third step is as follow. If there is only one kind of participle result of unword, they construct the concept graph of unknown words based on the information structure selected. After that, the semantic analysis of unword stops. If there are several kinds of participle results of unword and information structure is disambiguated by calculating the semantic similarity for every result, they calculate the similarity between unknown word and information structure firstly, then choose the corresponding word segmentation of maximum similarity as the final participle result of unknown word. Finally, they construct the concept graph of unknown word based on the information structure selected.

4. **Related work and error analysis.** The most of previous research approaches classified the unknown words to the category of other words whose sense category is known by calculating their similarity or association. In the later researches, new approaches are proposed, including knowledge-based model, rule-based model, corpus-based model, the hybrid model which combined the knowledge-based model and contextual information. As result of some of the papers using different data set and evaluation approaches, this paper overviews the experience results of different papers.

In [18], the author used the WordNet test set to experiment with different parameter settings. The experiments were performed by considering all possible configurations of the parameters. The approach need to decide when to use the similarity measure and when to fall-back to the guessing rules. This is determined by looking at the frequency and number of attributes for the unknown word. In the WORDNET test set, the best configuration of the experience gets the accuracy of 68%.

In the knowledge-based models in [1], the experience result of the overlapping-character baseline model is the best accuracy score of 51.7%. The experience result of the character-category association model is the best accuracy score of 58.2%, 6.5% higher than the baseline model. The experience result of the rule-based model is the best accuracy score of 80.3%, considerably higher than all other models.

In the hybrid model [1] combined structure-based methods and context-based method, structure-based methods are the main part and a context-based method is only used to rank the candidate supersense provided by the structure-based methods. That is a loose-coupled combination. The experiments show that the use of contextual information doesn't further improve performance on the basis of structural information.

The similar example approach adopts a K-nearest neighbor approach such that the distance between an unknown word and examples from the CiLin thesaurus is computed based upon its morphological structure [4]. Without contextual information, the system can still successfully classify 65.76% of adjectives, 71.39% of nouns and 52.84% of verbs. As Tseng reported results on adjectives, nouns, and verbs individually, the overall results are computed based on the proportion of words of the three different parts of speech in the test data. The final result of test data is 66.8%. The result of the hybrid model is 61.6%, comparing with it.

While [4] reported higher overall results than [1], there are three issues that deserve closer examination.

First, [4] adopted a head-determination baseline model that assigns the category of the morphological head to each unknown word, and reported very high results for [1]. In [4], the 70.8% accuracy reported for classifying nouns is especially high. However, [7] reported much lower results (18.8%) for a similar baseline model, which assigns one of the categories of the two component characters to each disyllabic V-V compound.

Second, whereas it is clear that the test words are not considered as their own neighbors in the testing phase, it is unclear whether the words in the development and test data are included or excluded in training the model, i.e., in calculating the entropy of the system and the information content of individual categories.

Finally, a distinct feature of [4] is that, similar to [8], it used a morphological analyzer to determine the head and analyze the structure of each unknown word. Given that the exemplar-based model does not improve the results of the head-determination baseline model by a large margin, especially for nouns, it seems to be the case that the difference between the results reported in [4] and those reported in [1] and other studies could to a large extent be attributed to the use of the morphological analyzer.

In [3], the author proposed a similar example-based method. The similarity of the modifiers of two words that share the same head is computed to represent the similarity of the two words. It is hard to directly compare the results in this study with the results reported in [3], as they only evaluated their model on a small test set of 200 unknown nouns.

In [7], the author argued that this approach does not work well for exocentric words, i.e., words whose categories are different from those of their heads and other component morphemes. In [7], the author proposed a different similarity-based model that retrieves synonyms of unknown words, where the nearest synonym of an unknown word is defined as the example that has the greatest word-word association with the unknown word. In [7], the results shew that the accuracy reaches 61.6% on 500 disyllabic verb-verb compounds.

For an unknown word [5], the proposed method analyzes its structure to collect some known words from a thesaurus to work as its candidate synonyms. Then the contextual similarity is computed between the unknown word and the candidate synonyms. Finally, the method selects some most similar candidate synonyms, and assigns the dominant supersense among those synonyms to the unknown word. In [5], they choose the Extended

CiLin as data set. Two sets were constructed following the procedure in [1], which selects words from CiLin, and then filters and groups them with the help of Contemporary Chinese Corpus. Two sets were constructed which contain 3,000 test words respectively. The result of the experience is 67.1%.

In [3], the classification errors are caused by a) some of the testing examples have no semantic composition property, b) some semantic classifications are too much fine-grained. There is no clear cut difference between some classes, even human judge cannot make a right classification, c) there are not enough samples that causes the similarity-based model does not work on the suffixes with few or no sample data. In [5], the cause of error about the hybrid model which combines the contextual information and structure information is as follows:

a) First, for an unknown word w, there is no word in thesaurus T that shares character with w. In such a case, no supersense is assigned to w.

b) Second, although some words do share characters with w, all of them do not share supersenses with w.

c) Third, w does not appear in the Raw-Corpus. In this case, the proposed method degrades to the baseline model.

d) Fourth, although one or more candidate synonyms share supersense with w, the proportion of those words is too small to help that supersense win among all the candidate supersenses.

In [12], the concept graph of the sense of unknown words is get after analyzing the unknown words. Because the correctness evaluation of the concept graph of the sense can't be evaluated by automatic system, the experience judges the correctness by manual method. It means that unknown words gain a correct semantic analysis, if the sense which the concept graph of the sense of unknown words expresses is consistent with the comprehension of people. Otherwise, the experience considers the semantic analysis not correct. After judging by manual method, there are 17666 words which are correct in 22278 unknown words. The accuracy is 79.3%. There are three influence factors effects the accuracy as follows.

(1) For the verb unknown words, because verb has high dynamics, the accuracy of analysis is influenced, such as '砍价'.

(2) For thumbnail-type unknown words, because of being short of the connection between acronym and lexicon, the accuracy of analysis is low, such as '党代会'.

(3) For the unknown words which are idiom, it also can't get high accuracy, such as '大器晚成'.

5. **Conclusions.** In essence, the sense guessing of unknown words is the task to classify the unknown words into the fine-grained category of Chinese dictionary, such as CiLin. The research approaches of this task mainly include knowledge-based approach, contextual information-based approach and the hybrid model combined the former two models. The previous research mainly focused on knowledge-based model, then joined the research of contextual information. However, the result is not very ideal. Later, it is not very obvious that the hybrid model combining loosely the knowledge-based model and contextual

information-based model performs. In recent research, the better result is achieved by combining closely the knowledge and contextual information of unknown words to get a hybrid model.

The next research direction of the sense guessing of unknown words includes, as follows.

Firstly, additional resources, such as Chinese-English dictionaries, parallel corpora, and larger corpora with richer linguistic annotation may prove useful for improving the knowledge-based and/or corpus-based models.

Second, future work needs to explore alternative ways to combine the two more effectively.

In addition, in recent research based on HowNet, the sense classification of the common unknown words is in the position of beginning stage. In the next work, it is necessary to refer to the existing research achievement and add the module of semantic parsing. In order to improve the analysis efficiency of verb unknown words, it may be useful to add verb syntactic parsing module which based on statistical analysis. About the semantic parsing of the thumbnail-type unknown words, it may be useful to append the mapping table of common expression and acronym, design the semantic analysis module of the thumbnail-type unknown words. About generating concept graph, recently, the approach compounds each word graph by the structure information. It is necessary to add the pruning module of concept graph.

Acknowledgment. This work is partially supported by Chinese National Fund of Natural Science under Grant 61272221 and 61073119, and Jiangsu Province Fund of Social Science under Grant 12YYA002.

REFERENCES

[1] Xiaofei Lu. Hybrid Model for Chinese Unknown Word Resolution. Ph.D Dissertation, the Ohio State University, 2006.

[2] Xiaofei Lu. Hybrid Model for Semantic Classification of Chinese Unknown Words. In: *Proceedings of North American Chapter of the Association for Computational Linguistics -Human Language Technologies' 07*. Rochester: NAACL，188-195, 2007.

[3] Chen,K., Chen C. Automatic Semantic Classification for Chinese Unknown Compound Nouns. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*. Saarbrücken, Germany, COLING, 173–179, 2000.

[4] Tseng,H. Semantic Classification of Chinese Unknown Words. In: *Proceedings of the Student Research Workshop at the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan: ACL, 72–79, 2003.

[5] Likun Qiu, Yunfang Wu, and Yanqiu Shao. Combining Contextual and Structural Information for Supersense Tagging of Chinese Unknown Words. In: *A. Gelbukh, ed. CICLing'11*, Part I, LNCS 6608. Berlin: Springer-Verlag, 15–28, 2011.

[6] Jurafsky, D., & Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Englewood Cliffs: Prentice Hall, 2008.

[7] Chen, C. Character-sense Association and Compounding Template Similarity: Automatic Semantic Classification of Chinese Compounds. In: *Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing*. Barcelona, 33-40, 2004.

[8] Chen,H., Lin, C. Sense-tagging Chinese Corpus. In: *Proceedings of ACL-2000 workshop on Chinese Language*. Hong Kong: ACL, 7-14, 2000.

[9] Chen, K.J., C.R. Huang, L. P. Chang & H.L. Hsu. SINICA CORPUS: Design Methodology for Balanced Corpora. In: *Proceedings of the 11th Pacific Asia Conference on Language, Information, and Communication.* Seoul: PACLIC, 167-176, 1996.

[10] Lua, K.-T. Prediction of Meaning of Bi-syllabic Chinese Compound Words using Back Propagation Neural Network. Computational Processing of Oriental Languages, 11(2): 133–144, 1997.

[11] Tseng H, Chen K J. Design of Chinese Morphological Analyzer. In: *Proceedings of the first SIGHAN workshop on Chinese language processing-Volume 18*. Stroudsburg: ACL, 1-7, 2002.

[12] Ruixia Zhang, Guozeng Yang, Qingxin Yan. The HowNet-based Semantic Analysis Model of the Common Chinese Unknown Words. *Computer Applications and Software*, 29(8), 126-130, 2012.

[13] Zhendong Dong, Qiang Dong. The Introduction of HowNet. http: / /www.keenage.com. 2006.

[14] Zhendong Dong, Qiang Dong. About HowNet- Chinese Message Structure Database. http: / /www.keenage.com.

[15] Sowa J F． Conceptual Graphs［R/OL］． http: / /www. jfsowa.com/cg /index.htm．

[16] Lei Zhang, Xueliang Li. Concept Structure and the Application. Ph.D Dissertation, Northwestern Polytechnical University, 2001.

[17] Ruixia Zhang, Han Xiao. The HowNet-based Word Graph Construction. *Journal of North China Institute of Water Conservancy and Hydroelectric Power*, 29(3) : 53-56, 2008．

[18] James R. Curran. Supersense Tagging of Unknown Nouns using Semantic Similarity. In: *Proceedings of the 43rd Annual Meeting of the ACL*, 26–33, 2005.

[19] Hearst and Hinrich Schutze. Customizing a Lexicon to Better Suit a Computational Task. In: *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text.* Columbus, 55–69, 1993.

[20] Dominic Widdows. Unsupervised Methods for Developing Taxonomies by Combining Syntactic and Statistical Information. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Edmonton, 276–283, 2003.

[21] Massimiliano Ciaramita and Mark Johnson. Supersense Tagging of Unknown Nouns in WordNet. In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing.* Sapporo: EMNLP, 168–175, 2003.