International Journal of Knowledge and Language Processing
Volume 2, Number 3, July **2011**

www.ijklp.org
KLP International ©2011 ISSN 2191-2734
pp.**1—13**

A Hybrid Model for Sense Guessing of Chinese Unknown Words

Likun Qiu[1,2], Kai Zhao[2] and Changjian Hu[2]
[1] Key Laboratory of Computational Linguistics Ministry of Education
Peking University
Beijing, 100871, China
qiulikun@pku.edu.cn

[2] NEC Laboratories
Beijing ,China
{zhao_kai, hu_changjian}@nec.cn

Received June 2011; revised July 2011

ABSTRACT. *This paper proposes a hybrid model to address the task of sense guessing for Chinese unknown words. Three types of similarity, i.e., positional, syntactic and semantic similarity, are analyzed; and three models are developed accordingly. Then the three models are combined to form a hybrid one (HPPS Model). To verify the effectiveness and consistency of HPPS, experiments were conducted on ten test sets which were collected from two popular Chinese thesauruses Cilin and HowNet. In addition, extra experiments were made on a test set of 2000 words which were collected from newspaper. The experiments show that HPPS Model consistently produces 4%~6% F-score improvement over the best results reported in previous researches.*
Keywords: Semantic category, Unknown word, LC Principle, Semantic classification

1. **Introduction.** In mathematics, semantics, and philosophy of language, the principle of compositionality has profound influence (Partee, 2004). It usually takes the following form: The meaning of a complex expression is the function of the meanings of its immediate sub-expressions and their mode of combination.

Since most Chinese words are composed of meaningful characters without rich inflections, the lexical structure of Chinese words might be considered as having similar property to the syntactic structure of phrases and sentences. Therefore, it is reasonable to transform the principle of compositionality to the following form: the sense of a Chinese word is the function of the syntactic and semantic properties of its immediate constituents and their mode of combination. Specifically, the principle might be transformed further to the following form: *The words formed by similar constituents in the same mode fall into the same semantic category*. This is referred to as *the principle of lexical compositionality* (LC Principle).

The task of sense guessing is to assign a semantic category to an unknown word. The assigned semantic category is chosen from a predefined set of semantic categories

(Figure 1). For example, a sense-guessing algorithm chooses *human*, which is one of the 1758 semantic categories defined by thesauruses *HowNet* (Dong and Dong, 2006), as semantic category of the unknown word 基民 *ji1min2 'stock fund investor'*.

This paper investigates sense guessing of Chinese unknown words based on the LC Principle. The word *similar* in the LC Principle has rich meaning. Three types of similarity can be defined here.

If two constituents (of two words) take the same position in the words, they have positional similarity.

If two constituents share the same POS tag, they have syntactic similarity.

If two constituents share the same semantic category, they have semantic similarity.

Three models are developed based on the three types of similarity accordingly. Firstly, a character-sense association model is developed based on the positional similarity. Given a group of words that start or end with the same character, the association between the character and word sense is computed.

Input:

(1) a set of existing words {w_1, w_2, …, w_n};

(2) a set of semantic categories {C_1, C_2, …, C_m};

(3) the relation between the words and the categories: {<w_i, C_j> | 1≤i≤n, 1≤j≤m }.

For an unknown word w,

Output: C_k, where 1≤k≤m.

FIGURE 1. The task of sense guessing

Secondly, a POS-sense association model is developed by a sequence-labeling method. All training words and testing words are segmented into constituents and tagged with POS. Then we solve the sense-guessing problem with algorithms such as CRFs and ME.

Thirdly, a sense-sense association model is developed based on the semantic similarity between the constituents of testing words and training words.

Finally, the three models are combined to form HPPS (a **H**ybrid model based on **P**osition, **P**OS and **S**ense**)**.

The remainder of this paper is organized as follows. Section 2 introduces previous work on sense guessing of unknown words. Method of (Lu, 2007) is described in Section 3 and is taken as baseline of this paper. Section 4 describes the HPPS Model. Section 5 gives the experiment results of the HPPS Model together with an error analysis. Section 6 presents conclusions.

2. **Related Work.** Methods involved in the sense-guessing process of unknown words might be classified into two types: structure-based methods and context-based methods. Most researches focusing on Chinese unknown words utilized structure-based methods. A hybrid model is proposed in (Lu, 2007). The accuracy of the hybrid model is 61.6% on *Cilin*. This is the best result in previous researches.

A similarity-based model is proposed in (Chen and Chen, 2000). The similarity of the modifiers of two words that share the same head is computed to represent the similarity of the two words. The F-score is 81%. However, the test set contains only 200 unknown nouns, which is too small to make a reliable evaluation.

By using a morphological analyzer, the morphosyntactic relationship between the morphemes of a word is detected in (Tseng, 2003). Before a most similar word of the test word is retrieved, the words with a different morphosyntactic relationship are filtered. However, the unknown words are only classified into the 12 major categories of *Cilin* (Mei *et al*., 1984), which is coarse-grained.

The method in (Chen, 2004) retrieves a word with the greatest association with the test word. The accuracy is 61.6% on disyllabic V-V compounds. However, the test words are included in the training data. This result is worse than the result of (Lu, 2007).

Meanwhile, we only found two researches that used context-based methods to processing Chinese unknown words. The experiments in (Lu, 2007; Chen and Lin, 2000) achieved 37% and 34.4% in terms of F-score respectively and show that the use of contextual information does not lead to performance enhancement. For English, context-based methods are used more popularly such as (Ciaramita and Johnson, 2003; Curran, 2006; Pekar and Staab, 2003). However, their results are similar to those analogous studies for Chinese unknown words. For instance, Pekar and Staab (2003) tried to classify nouns into 137 classes and only achieved a precision of 35.1%.

The idea of the LC Principle has been touched more or less by previous researches, such as (Chen, 2004; Lu, 2007). However, it has not been clearly stated and systematically studied.

3. **Baseline Model.** The method in (Lu, 2007) is taken as the baseline. It contains two separated models: a character-sense association model and a rule-based model.

3.1. **Character-Sense Association Model (CS Model).** The first model is the character-sense association model, which is used by both (Chen, 2004; Lu, 2007). It is referred to as the CS Model in this paper. To make the comparison reliable, we follow the designs of character-sense association model in (Lu, 2007).

This model uses x^2 to capture the relationship between the semantic category of an unknown word and that of its component characters. In (1), $Asso(c,t_j)$ denotes the association between a character c and a semantic category t_j, and $f(X)$ denotes the frequency of X.

$$Asso_{x^2}(c,t_j)=\frac{\alpha(c,t_j)}{\max_k \alpha(c,t_k)} \quad (1)$$

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

$$Asso(w,t_j)=\sum_{i=1}^{|w|}\lambda_i Asso(c_i,t_j) \quad (2)$$

Once the character-sense associations are calculated, the association between a word *w* and a category t_j, i.e., $Asso(w,t_j)$, is calculated in (2) as the sum of the weighted associations between each of the word's characters and the semantic category, where c_i denotes the *i*'th character of *w*, |*w*| denotes the length of *w*, and λ_i denotes the weights. All the λ s adds up to 1. The position-sensitive associations between a category and a character are computed in the initial, middle, and final positions of a word respectively.

3.2. **Rule-based Model.** There are two types of rules: Rules of type-1 and Rules of type-2. Rules of type-1 deal with coordinate multi-syllabic word. It presupposes that a coordinate multi-syllabic word and both of its components share the same category. In Rules of type-1, the unknown word *w* is divided into two parts A and B. Let f_A and f_B denote the number of times A and B occur in initial and final positions of word in C(*w*) respectively. Here, C(*w*) refers to the semantic category of word w. If C(A)=C(B) and both f_A and f_B surpass the predetermined thresholds, assign C(A) for AB.

Rules of type-2 guess the semantic category of a tri-syllabic or four-syllable word by finding a similar tri-syllabic or four-syllable word. A word w_1 is said similar to another word w_2, if their remaining parts have the same semantic category after the same characters at the same position are removed. By Rules of type-2, for an unknown word *w*, its similar words are collected from the thesaurus. The semantic categories of similar words are output as the categories of *w*. If there is no similar word, no result is output.

Formally, for a tri-syllabic word ABC, if there is a word XYC such that C(AB)=C(XY), then C(ABC)=C(XYC); if there is a word XBC such that C(A)=C(X), then C(ABC)=C(XBC). For instance, for a test word 推销商 *tui1xiao1shang1 'salesman'*, collect its similar word 销售商 *xiao1shou4shang1 'salesman'* from the thesaurus, i.e., C(推销)=C(销售). Then C(销售商) is assigned to 推销商 as its semantic category.

3.3. Combination. The former two models are combined together (see Figure 2).

For an unknown word *w*, the rule-based model is applied. Denote the output as {C_1, C_2, …, C_n}.

If n=1, then C_1 is output.

If n>1, rank all C_i, where i=1, …, n, according to their association with w (apply CS Model to achieve the association). Then the top-ranked one is output.

If n=0, the character-sense association model is applied. Denote its output as

{C_1, C_2, …, C_m}.

If m=1, then C_1 is output.

If m>1, rank all C_i according to their association, and output the top-ranked one.

If m=0, nothing is output.

FIGURE 2. The baseline method

4. Proposed Method: HPPS Model. Three models are developed based on the three types of similarity of the LC Principle. The first model is inherited from the CS model without modification. The second model uses a sequence-labeling method to guess sense based on constituents of words and POS tags of the constituents. The third model automatically generates mapping rules from the semantic category of constituents to the semantic category of the whole word. Then the three models are combined to form the HPPS model.

4.1. **Sequence-Labeling Model (SL Model).** The second model considers the sense-guessing task as a sequence-labeling problem. This model builds mapping from the constituents of a word and their POS tags, to the semantic category of the word. This is referred to as the SL Model. Since many studies have shown that CRFs (Conditional Random Fields) are the best model for sequence-labeling problem (Lafferty et al., 2001; Vail et al., 2007), CRFs are adopted as method in this model.

For any unknown word *w*, it is not necessary to infer its semantic category from all words in the thesaurus, because most words in the thesaurus have no relation with *w*. Generally, only those words sharing the same character with *w* may possibly share the same semantic category with *w*. Therefore, only this kind of words is selected to form the training set of *w*. In detail, if *w* is a noun, the words sharing the same final character with *w* are chosen. If *w* is a verb or adjective, the words sharing either the initial character or the final character are chosen.

Two types of features are employed: the constituent characters of a word and the POS tags of those constituents. In both training and testing process, the internal constituent structure of the words is analyzed and POS-tags are attached to constituents. For example, 文化部门 *wen2hua4-bu4men2 'branch of culture'* has the following characters: 文, 化, 部, 门, and is segmented and POS-tagged as "文/N 化/V 部/N 门/N", in which "文/N" means that "文" is a noun and "化/V" means that "化" is a verb.

Particularly, twelve n-gram templates are selected as features for CRFs: C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1, $C_{-1}C_1$, P_{-1}, P_0, P_1, $P_{-1}P_0$, P_0P_1, $P_{-1}P_1$, where C stands for a character, P for the POS of a character, and the subscripts -1, 0 and 1 for the previous, current and next position respectively.

In the training process, firstly, each training word is segmented and POS tagged by a standard tool. That is, for word *w*, the following form is achieved: $<A_1, P_1>$, $<A_2, P_2>$, …$<A_{n-1}, P_{n-1}>$, $<A_n, P_n>$ where each A_i is a constituent after segmentation, and P_i is

its POS tag. Secondly, each constituent is attached with a category label. Given $C(w)=C_1$, the following form is achieved $<A_1, P_1/C_1\_I>$, $<A_2, P_2/C_1\_M>$, … $<A_{n-1}, P_{n-1}/C_1\_M>$, $<A_n, P_n/C_1\_F>$ where C_1_I, C_1_M, C_1_F denotes the *Initial*, *Middle*, and *Final* part of C_1 respectively. For instance, for *w*=文化部门, the following form is achieved: <文, N/Di09_I>, <化, V/Di09_M>, <部, N/Di09_M>, <门, N/Di09_F> in which Di09 is the semantic category of *w* in *Cilin*. Thirdly, the feature templates are used to extract features. Fourthly, CRFs are applied on the training sets to obtain a model.

In the testing process, the unknown word is segmented and POS-tagged by the same tool first. Then features are extracted. Finally the sequence is input to the obtained model to acquire a semantic category. For instance, given an unknown word 花费 *hua1fei4 'expend'*, it would be analyzed as 花/V 费/V for feature extraction. Then the model gives an output: <花, V/He13_I> <费, V/He13_F> . That is, the semantic category He13 is assigned to the word 花费 (in *Cilin* He13 referring to expending or storing).

4.2. **Sense-Sense Association Model (SS Model).** The third model simulates three ways of word forming in Chinese based on the semantic similarity. This is referred to as the SS Model. In detail, the first way is the same as Rules of type-1 of Lu (2007) and is called coordinate analogy. The other two are called double parallel analogy and paired parallel analogy respectively. Compared with the Rules of type-2 of Lu (2007), the two newly proposed analogies have three advantages. The first, a pattern is given instead of rules. That is, rules will be automatically generated from a thesaurus based on the patterns. The second, the pattern is in probabilistic form, which extends the coverage. The third, restriction on word length in the Rules of type-2 is removed, which covers more cases.

Double Parallel Analogy

In linguistics, a group of words is said *parallel* if they share the same character(s) at the same position, i.e., $\{D_1A, D_2A, \ldots, D_nA\}$, where each D_iA is a word, and D and A are constituents containing one or more characters. In many cases, parallel words also share the same semantic category, i.e., $C(D_1A)=C(D_2A)=\ldots=C(D_nA)$. That gives a hint for sense guessing: it is probably correct to guess an unknown word as $C(D_1A)$ if it takes a similar structure $D_{n+1}A$.

However, there are also many violations, especially when A is polysemous. To filter those violations, an extra limitation may be set on the semantic categories of the different part of the parallel words. Particularly, the semantic categories of the different part are required to be the same, i.e., $C(D_1)=C(D_2)=\ldots=C(D_n)$. This limitation helps filter many violations. Since the semantic categories of both part of and the whole words are required to be the same, it is called *double*.

If a group of words in the thesaurus are found to be double parallel, then it is confident to guess a similar-structure unknown word $D_{n+1}A$ as $C(D_1A)$. In real cases, one or more negative examples may occur. Here, a negative example refers to a word E_1A satisfying $C(E_1)=C(D_1)=\ldots=C(D_n)$ but $C(E_1A)\neq C(D_iA)$, where $1\leq i\leq n$. Less negative examples, more possible a guess is correct. Therefore a threshold T is introduced. In addition, to ensure the correctness of guessing, a limitation is added to the number of parallel words, i.e., $\{D_1A, D_2A, \ldots, D_nA\}$. n must be not less than a threshold N.

Denote the thesaurus as S. Given two thresholds N and T. Double Parallel Analogy gives a pattern as follows. For a constituent A that contains one or more characters, collect parallel word set PS={D_iA| $D_iA \in S$}, where D_i contains one or more characters. On PS, if $|\{DA|\mathrm{C}(D)=CM_1\}|\geq N$ and $\mathrm{P}(\mathrm{C}(DA)=CM_2|\mathrm{C}(D)=CM_1)>T$, where CM_1 and CM_2 are two semantic categories, then a rule is generated: For an unknown word w=BA, C(BA)=CM_2 if C(B)=CM_1.

For example, for A=人 *ren2 'person'*, collect parallel word set *PS*={D_iA} from *Cilin*. *PS* contains more than 300 words. Among them, four words (Table 1) satisfies C(D_i)=Ed03, where 1≤i≤4. Given N=3 and T=0.5. Since $|\{DA|\mathrm{C}(D)=\mathrm{Ed03}\}|=4>N$ and $\mathrm{P}(\mathrm{C}(DA)=\mathrm{Ak03}|\mathrm{C}(D)=\mathrm{Ed03})=\frac{3}{4}>T$, a rule is generated: *C(D 人)=Ak03 if C(D)=Ed03*. Then, for an unknown word 圣人 *sheng4ren2 'sage'*, since C(圣)=Ed03, this rule assigns Ak03 to 圣人.

The symmetrical form of the analogy also applies. That is, if the word set takes AD form, then the rule takes AB form. If each word is restricted to 3 or 4 characters, N=1 and T=0, then this analogy regresses to Rules of type-2 of Lu (2007). That is, the double parallel analogy covers Rules of type-2.

TABLE 1. Words of parallel set {DiA} satisfying C(Di)=Ed03

D_iA	**Word**	**C(D_iA)**
D_1A	坏人 *huai4ren2 'bad person'*	Ak03
D_2A	歹人 *dai3ren2 'gangster'*	Ak03
D_3A	好人 *hao3ren2 'good person'*	Ak03
D_4A	美人 *mei3ren2 'beautiful person'*	Ac03

TABLE 2. Parallel sets of character pair A=峰 and B=头

D_iA/B	**Word**	**C(D_iA/B)**
D_1A	上峰 *shang4feng1 'leader'*	Aj08
D_1B	上头 *shang4tou5 'leader'*	Aj08
D_2A	山峰 *shan1feng1 'peak'*	Be04
D_2B	山头 *shan1tou2 'peak'*	Be04
D_3A	尖峰 *jian1feng1 'high-point'*	Dd13
D_3B	尖头 *jian1tou2 'sharp-end'*	Bc01
D_4A	洪峰 *hong2feng1 'flood'*	Bg01
D_5B	木头 *mu4tou5 'wood'*	Bm03

Paired Parallel Analogy

Many pairs of characters have the ability to form words with the same semantic category, if the pair of words has the same semantic category itself. That is, a pair of characters A and B has the ability to form words DA and DB with C(DA)=C(DB), if C(A)=C(B) holds. Denote the thesaurus as *S*. Given a threshold T, a probabilistic pattern is given as follows.

For a pair of characters A and B with C(A)=C(B), combine their own parallel word sets as $PS=\{D_iA|D_iA\in S\}\cup\{D_iB|D_iB\in S\}$. If $\mathrm{P}(\mathrm{C}(DA)=\mathrm{C}(DB)|(DA\in PS, DB\in PS))>T$, then a rule is generated: For an unknown word w=EA, C(EA)=C(EB) if EB∈S.
For example, A=峰 *feng1 'peak'*, B=头 *tou2 'top'*. In *Cilin*, C(A)=C(B)=Bc01. From PS

(Table 2), three words-pairs are found: {{D_1A, D_1B}, {D_2A, D_2B}, {D_3A, D_3B}}. Given T=0.5. Since $P(C(DA)=C(DB) \mid (DA \in PS, DB \in PS)) = \frac{2}{3} > T$, a rule is generated: C(D峰)=C(D头) if the word D头 ∈ S. Then, for an unknown word 眉峰 *mei2feng1 'eyebrow'*, the above rule is applicable because DB=眉头 *mei2tou2 'eyebrow'* exists in the thesaurus, with semantic category Bk12. Then Bk12 is assigned to 眉峰. The symmetrical form of the analogy also applies.

4.3. **HPPS Model.** HPPS is a hybrid method of SS, CS, and SL models. About the three models, the SS Model is most credible. That is, if it gives a guess, the guess is always correct. But it cannot give guess in many cases, because of its strict constrains. The CS and SL Model have similar credibility and coverage. However, CS Model is more credible in Case-1, while SL Model more credible on Case-2.

For a Case-1 word w=AB, in the training set, there exist at least two words w_1=A* and w_2=*B, satisfying C(w_1)=C(w_2). Here, * means any character. For example, for w=包间 *bao1jian1 'compartment'*, in *Cilin*, there exist two words 包厢 *bao1xiang1 'balcony'* and 房间 *fang2jian1 'room'*, satisfying C(包厢)=C(房间). Other words are Case-2 words. For example, in *Cilin*, for w=半径 *ban4jing4 'radius'*, w_2=直径 *zhi2jing4 'diameter'* can be found, but there is no w_1=半* satisfying C(w_1)= C(w_2).

According to the above observations, HPPS Model is designed as shown in Figure 3. SS Model is running first. For words which SS Model gives no guess, give them to CS or SL Model. CS and SL Model have their own advantages: CS Model is more credible when both initial and final positive examples are found while SL works better when only one positive example is found. Therefore, they are used to process Case-1 and Case-2 words respectively.

For an unknown word w,

Apply the SS Model. Denote the output as {C_1, C_2, …, Cn}.

If n=1, then C_1 is output.

If n>1, rank all C_i, where 1≤i≤n, according to their association with w (apply CS Model to achieve the association). Then the top-ranked one is output.

If n=0:

Apply CS in Case-1;

Apply SL in Case-2.

FIGURE 3. The HPPS Model

5. **Experiments.**

5.1. **Data Preparation.** Three types of test sets were constructed. The first two are based on popular Chinese thesauruses *Cilin* and *HowNet*. *Cilin* contains over 70,000 words, which are classified into 1428 semantic categories. *HowNet* contains over 90,000 words and 1758 semantic categories. To compare with previous work fairly, the test sets were constructed following the procedure of Lu (2007): (1) select the January/1998 part of the *Contemporary Chinese Corpus* from Peking University (Yu et al. 2002). That corpus contains all the articles published in People's Daily, a major newspaper in China; (2) remove words that are not in *Cilin*; (3) remove words that are not 2-4 characters length; (4) remove words that are not noun, verb, or adjective. Then 35151 words were left. (5) construct ten test sets, each of which contains 3,000 words. Basically the words were randomly selected from the 35151 words, but with a frequency control: in each test set, 1/3 of words occurring 1-3 times, 3-6 times, and 7 or more times in the corpus respectively. The ten sets are referred to as IV (in-vocabulary) sets, because all the words are currently included in *Cilin*. The ten IV sets of *HowNet* were constructed in the same way.

The third type of test set was constructed by simulating the real unknown words identification process: words occurring in February-June/1998 period of the *Contemporary Chinese Corpus*, but not in the January/1998 period of the corpus, *Cilin* and *HowNet*, were collected. It seems that those words are unknown words in January 1998. Then these words were filter by length, POS and frequency like above. From the left words, 2000 were randomly chosen, which forms the test set. It is referred to as OOV (out-of-vocabulary) set. Compared with those IV sets, the OOV set is more close to the real case of unknown words. Only 2000 words were chosen because of the high cost of human tagging. Two annotators performed the tagging task. Each word was asked to assign a semantic category in *Cilin* and *HowNet* respectively. There was about 15% disagreement initially between the annotators. Then they discussed the disagreement and solved it. Only one category was remained for one word (in fact, one category in *Cilin* and one category in *HowNet*).

5.2. **Baseline: Results of Method of Lu (2007) on *Cilin* and *HowNet*.** The method of Lu (2007) includes the CS Model and a rule-based model. For the CS Model, a training process is needed. When the training set is constructed, a *remove-one* policy is used. That is, for a test word *w*, all other words in the thesaurus are included in the training set except *w* (i.e., remove one word *w* from the thesaurus). That policy is a little confusing for polysemous words.

A polysemous word has more than one token in the thesaurus. For example, the word 老爷爷 *lao3ye2ye2 'grandpa'* appears twice in *Cilin*, corresponding to two semantic categories *old-man* and *grandpa*. The remove-one policy may have two meanings, when a polysemous word *w* is taken as the test word:

Remove all tokens of *w*. For example, if *w*=老爷爷, then the two tokens in *old-man* and *grandpa* are removed from the training set. We call it *all-token-removing* policy.

Remove one token of *w*. For example, if 老爷爷 of category *old-man* is selected as test word, the token of category *old-man* is removed, but the token of category *grandpa* is still remained in the training set. We call it *one-token-removing* policy.

It is not clear which policy was adopted in Lu (2007). Therefore we implemented and tested both policies, with parameters the same as Lu (2007). Table 3 summarizes the test results.

TABLE 3. F-score of CS Model on IV set of Cilin

	Development	**Test**
All-token-removing	0.561	0.545
One-token-removing	0.591	0.578
Lu (2007)	0.586	0.582

From Table 3, we can see that the one-token-removing policy achieved much more similar performance to Lu (2007) than all-token-removing policy. Therefore, we guess that one-token-removing policy was adopted in Lu (2007). However, all-token-removing is more reasonable than one-token-removing, because when people say that one word is an unknown word, they mean that the word did not occur before, and this is the first time it appears. Therefore, an unknown word surely has no token included in the whole thesaurus. According to the above analysis, the all-token-removing policy is adopted in the following experiments. Among the ten test tests (of *Cilin* or *HowNet*), one set is used for development, while the other nine sets are tested then based on the parameters achieved in development process (test process).

TABLE 4. Results of Rule-based Model of Lu (2007) and SS Model on IV set of Cilin and Hownet

Model	**Thesaurus**	**Development**			**Test**		
		P	R	F	P	R	F
Rule-based Model of Lu	*Cilin*	0.819	0.154	0.259	0.778	0.152	0.254
	HowNet	0.751	0.175	0.284	0.726	0.171	0.277
SS Model	*Cilin*	0.814	0.249	0.381	0.787	0.253	0.383
	HowNet	0.769	0.311	0.442	0.763	0.311	0.442

Table 4 summarizes the results of rule-based model of Lu (2007) in terms of precision, recall and F-measure on IV sets of *Cilin*. The model achieves an overall 77.8% precision and 15.2% recall. Combined the CS Model and the rule-based model together, Lu (2007)'s hybrid model achieves 56.5% F-score (see Table 5).

5.3. **Results of Proposed Methods on *Cilin* and *HowNet*.** Table 4 also shows that on *Cilin*, the SS Model improves 0.9% in precision and 10.1% improvement in recall over rule-based model of Lu; on *Hownet* the SS Model improves 3.7% in precision and 14% in

recall over rules-based model of Lu. The improvement verifies that the probabilistic-pattern based method can cover more words than manually crafted rules.

After development process, the following parameters were achieved: the two T thresholds of SS Model are both 0.65; the threshold N in double parallel analogy is 3 for disyllabic words and 1 for other words; thresholds for f_A and f_B in coordinate analogy are 1 and 1 for nouns, and 0 and 3 for other words. In SL Model, ICTCLAS 3.0 (Zhang, 2002) was used as word segmentation and POS tagging tool, while "CRF++, Yet another CRF" toolkit (Kudo, 2005) was used as the implementation of CRFs.

TABLE 5: Results of proposed methods on IV sets and OOV set of Cilin and HowNet

Data Type	**Thesaurus**	**Model**	**Development**			**Test**		
			P	R	F	P	R	F
IV	*Cilin*	baseline	0.581	0.580	0.580	0.566	0.565	0.565
		HPPS	0.619	0.618	0.619	0.610	0.609	0.610
	HowNet	baseline	0.525	0.525	0.525	0.510	0.510	0.510
		HPPS	0.576	0.575	0.575	0.564	0.564	0.564
OOV	*Cilin*	baseline	/	/	/	0.569	0.569	0.569
		HPPS	/	/	/	0.630	0.630	0.630
	HowNet	baseline	/	/	/	0.557	0.557	0.557
		HPPS	/	/	/	0.604	0.604	0.604

Table 5 shows that the HPPS Model improves 4.5% on IV sets of *Cilin* and 5.4% on *HowNet* in F-score over the baseline model. It also summarizes the results on OOV set of *Cilin* and *HowNet*. The HPPS Model achieves improvements of 6.1% F-score on OOV set of *Cilin* and 4.7% F-score on OOV set of *HowNet* over the baseline. Compared with IV sets, the improvement on OOV set is a little bigger on *Cilin*, but a little smaller on *HowNet*. Generally speaking, the performance improvement over the baseline is consistent on the three types of test set. The average improvement is 5.2%.

5.4. **Error Analysis.** The result of HPPS on one IV set of *Cilin* is selected to do error analysis. There are mainly four types of error.

The first type of error is caused by the ambiguity of constituents. For instance, the words ended with character 头 *tou2 'head'* are among several semantic categories. It is difficult to identify that 丫头 *ya1tou5 'girl'* is *girl* while 白头 *bai2tou2 'white-head'* is *head*. The second type of error is caused by the defect of the thesaurus. For instance, HPPS assigned *amount* to 库存量 *ku4cun2liang4 'the quantity of goods in stock'*. However, it is assigned *artifact* in *HowNet*, the same category as 库存 *ku4cun2 'inventory'*. The third type of error is caused by some characters that have no ability of forming new words. For instance, the character 拓 has two meanings. One is *ta4 'rub'*, and the other is *tuo4 'develop'*. However, the meaning *ta4 'rub'* comes from Archaic Chinese and rarely used in modern Chinese. Therefore unknown words like 拓展 *tuo4zhan3 'develop'* must have the 'develop' meaning. The fourth type of error is caused by metaphors, idioms, domain specific terms, transliterations, abbreviations and so on.

For instance, 二恶英 *e4er4ying1 'dioxin'* is a domain specific term and 夸克 *kua4ke4 'quark'* is a transliteration.
The ratio of the four types of error *is* 45%, 25%, 5% and 25% respectively.

6. **Conclusions.** This paper contributes to the research of sense guessing for Chinese unknown words. Specifically, we (1) propose a method for generating rules for sense guessing (Sense-Sense Association Model), (2) consider the sense guessing task as a sequence-labeling process and tackle it with CRFs (Sequence Labeling Model), and (3) combine the two models with Character-Sense Association Model together as HPPS Model. Experiments conducted both on IV set and OOV set show the effectiveness of the HPPS Model.

Acknowledgement
The work is supported by the National Natural Science Foundation of China (Grant No. 61103089) and China Postdoctoral Science Foundation (Grant No. 20100480169). The earlier version of this paper has been presented in the 23nd PACLIC conference. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Chen, C.-J. 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*, pp. 33-40.

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

[3] Chen, K.-J. and C.-J. Chen. 2000. Automatic semantic classification for Chinese unknown compound nouns. In *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 173-179.

[4] Ciaramita, M. and M. Johnson. 2003. Supersense Tagging of Unknown Nouns in WordNet. In *Proceedings of the 2003 Conference on Empirical Methods on Natural Language Processing*.

[5] Curran, J. R. 2005. Supersense Tagging of Unknown Nouns using Semantic Similarity. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 26-33.

[6] Dong, Z. D. and Q. Dong. 2006. *HowNet And the Computation of Meaning*. World Scientific Publishing Co., Inc. River Edge, NJ, USA.

[7] Kudo, T. 2005. CRF++: Yet Another CRF toolkit. http://chasen.org/~taku/software/CRF++.

[8] Lafferty, J., A. McCallum and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of International Conference on Machine Learning*, pp. 282-289.

[9] Lu X. F. 2007. Hybrid Models for Semantic Classification of Chinese Unknown Words. In *Proceedings of North American Chapter of the Association for Computational Linguistics - Human*

Language Technologies 2007 conference, pp. 188–195.

[10] Mei, J. J., Y. M. Zhu, Y. Q. Gao and H. X. Yin. (eds.) 1984. *Tongyici Cilin*. Commercial Press, Hong Kong.

[11] Partee, B. H. 2004. *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*. Oxford: Blackwell Publishing, pp. 153-181.

[12] Pekar, V. and S. Staab. 2003. Word classification based on combined measures of distributional and semantic similarity. In *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics*.

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

[14] Vail, D. L., M. M. Veloso and J. D. Lafferty. 2007. Conditional Random Fields for Activity Recognition. In *Proceedings of 2007 International Joint Conference on Autonomous Agents and Multi-agent Systems*.

[15] Yarowsky, D. 1992. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 454-460.

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

[17] Zhang, K. ICTCLAS1.0. http://www.nlp.org.cn/ project/project.php?proj_id=6.