# Automatic Recognition of Chinese Unknown Word for Single-Character and Affix Models

Xin Jiang[1], Ling Wang[2], Yanjiao Cao[1], and Zhao Lu[1,*]

[1] Department of Computer Science and Technology, East China Normal University
200241, Shanghai, China
[2] Shanghai Interactive TV CO., LTD. Shanghai 200072, China
{xjiang,yjcao}@ica.stc.sh.cn, zlu@cs.ecnu.edu.cn

**Abstract.** This paper presents a novel method to recognize Chinese unknown word from short texts corpus, which is based our observation of both single-character and affix models of Chinese unknown word. In our approach, we collect some news titles of a news site and view these titles as short texts. There are three steps in our approach: First, all collected news titles are segmented with ICTCLAS, and statistics of potential unknown words are conducted. Second, all potential unknown words are classified into either single-character model or affix model based on structures of unknown word. Some filtration methods are used to filter garbage strings. Finally, unknown word is extracted according to the frequencies of word. We have got the excellent precision and the recalling rates, especially for the single-character model. The experiment results show that our approach is simple yet effective.

**Keywords:** Chinese unknown word, single-character model, affix model.

## 1 Introduction

As we all know that word segmentation is the key to Chinese information processing. Previous researches have made great progresses in word segmentation, but cases with unknown word are not satisfied, and any existing lexicon in Lexical analyzer is limited. What's more, it is difficult to cover all words in real texts or speeches. With the development of information technology, there are more and more Chinese unknown words emerging on the Internet, and they are difficult to be identified correctly for several reasons: (1) The meanings or usages of existing word can be changed; (2) There is neither a general role of the composition of unknown word, nor an unique regular pattern to distinguish these unknown words; (3) Unknown word which seldom appears in the corpus is extremely difficult to identify.

According to the definitions in linguistics, Chinese unknown word is not only a word that has not been recorded in popular dictionaries, but also an existed word that possess new meanings or new usages. In natural language processing, unknown word mainly refers to the word that has not been registered in the dictionary of Lexical analyzer, which mainly includes word with new morphology, named entities. Owning to the fuzzy definition of Chinese word, it is difficult to define a Chinese word explicitly.

* Corresponding author.

Y. Wang and T. Li (Eds.): Knowledge Engineering and Management, AISC 123, pp. 435–444.
springerlink.com 

The Chinese unknown word discussed in this study refers to a word which is not recorded in the dictionaries of ICTCLAS [1], that is developed by Institute of Computing Technology Chinese Academy of Sciences. After segmenting, a potential unknown word may be mixed with some single Chinese characters and always assembled by some single Chinese characters. It is a good idea to identify Chinese unknown word by combining some single characters with their adjacent words. It is discovered that a Chinese unknown word is always formed with two to four characters, and it is rare more than five. A Chinese unknown word is always formed as follows [2]:

1. Two-character new word, denoted as NW11 (two single-character word, "1+1");
2. Three-character new word, denoted as NW111 (three single-character word, "1+1+1"), NW12 (a single character followed with a bi-character word, "1+2"), NW21 (a bi-character word followed by a single character, "2+1");
3. Four-character new word, such as NW1111 (four single-character word, "1+1+1+1"), NW22 (a bi-character word followed by a bi-character word, "2+2"), NW211 (a bi-character word followed by two single character, "2+1+1"), NW121 (a single character followed by a bi-character word followed with a single character, "1+2+1"), NW112 (two single character followed by a bi-character word, "1+1+2"), NW13 (a single character followed by a tri-character word, "1+3"), NW31 (a tri-character word followed by a single character, "3+1");

In general, NW11, NW22, NW21 and NW12 cover more than 89% of all unknown Chinese words, they are 53%, 2%, 31% and 3%, respectively, and the others cover less than 11% [2]. In this study, we focus on unknown word of two surface patterns which account for more than 92% of Chinese unknown word, they are single-character model and affix model. The former model refers to NW11, NW111 and NW1111, the latter one refers to NW12, NW13, NW21 and NW31.

There are many approaches in Chinese unknown word recognition. These methods can be classified into three categories: (1) Rule-based methods, such as two methods suggested in [4, 5], they have advantages of high accuracy and strong pertinence. However it is difficult to define and evolve rules. The rules are always domain-specific which result in bad portability and flexibility. (2) Statistical methods, such as the method based on a SVM classifier in [2, 9] focus on two surface patterns, NW11 and NW21, the model of conditional random field in [6, 8] and the method based on Independent Word Probability (IWP) [7]. These methods need training through large-scale corpus and produce sparse data owing to few countable structure laws, which will lead to low accuracy rate. (3) Hybrid methods, which combine the two kinds of method. The author focus on bi-gram, tri-gram, quad-gram word and other common model, garbage string dictionary is used by self-learning method to filter [11].

Although these practicable methods achieve reasonable precision or recalling rates in some special cases, they have inherent deficiencies: (1) It is difficult to define and evolve the rules. And the rules are always domain-specific which result in bad portability and flexibility; (2) For statistical method, it needs training through large-scale corpus and owing to few countable structure laws sparse data will be produced; those finally lead to a low accuracy rate. Furthermore it's time-consuming, expensive and inflexible. What's more many approaches focus on the unknown word on specific areas or specific categories of unknown word. But the categories of unknown word

are diverse and the amount of such word is huge. With the rapid development of the Internet, this situation is becoming more and more serious.

This paper presents a novel approach considering features of unknown word, they are single-character model and affix model. The remainder of the paper is structured as follows: Section 2 details our method; Section 3 shows experiments and evaluations; Section 4, concludes this paper and future work.

## 2 Our Approach

### 2.1 Single-Character Model and Affix Model

In this section, several definitions of the single-character model and the affix string model are given.

**Definition 1 Single-character string.** A Chinese word containing two or more two single Chinese characters is called a single-character string.

For example: *上海/ns公布/v购买/v经/n适/n房/n细则/n.* Here, *"经适房", "经适"* , *"适房"* are all single-character strings.

**Definition 2 Parental string and substring.** For three single-character strings, *A, B* and *C,* if *A=B+C* and both lengths of *B* and *C* are large than 1 or equal to 1, then *A* is viewed as the parental string of *B* and *C*, *B* and *C* are viewed as two substrings of *A*.

**Definition 3 Potential Unknown Word (PUW) and Longest Potential Unknown Word (LPUW).** Given a string *T*, $T = \{X_1X_2 \ldots X_i \ldots X_n\}$, $(1 \le i \le n)$ , and each $X_i$ in the string *T* is a single Chinese character. The string *NW*, $NW(i.j) = \{X_jX_{j+1} \ldots X_k\}$, $(1 \le j, k \le n)$, is viewed as a potential unknown word if the string *NW* meets:

1. *NW* is a substring of the string *T*.
2. The length of the string *NW* is equal to 2 or larger than 2.

and more, the string *NW* is viewed as a longest potential unknown word if it meets:

1. $j = 0$ or $X_{j-1}$ is not a single Chinese character,
2. $k = n$ or $X_{k+1}$ is not a single Chinese character.

In above example, the string *NW "经适房"* is the longest potential unknown word, and the string *NW "经适"* and *"经适房"* are all potential unknown words.

For unknown words of the affix string model, they can be classified two forms, the postfix string such as NW21 and NW31, the prefix string such as NW12 and NW13.

**Definition 4 Affix string model.** Given a string T, that is $T = \{X_1X_2\}$, $X_1$ is an existing word that consists of two or three single Chinese characters, and $X_2$ is a single Chinese character. Then the sting *T* is viewed as a postfix string. If $X_1$ is a single Chinese character and $X_2$ is an existing word, the string *T* is viewed as a prefix string.

For NW21, "2+1", a postfix string example: ***国土/n部/q*****拟/v全国/n推广/vn土地/n问/v责/ng连/d坐/v制/v**

For NW31, "3+1", a postfix string: *中医/n治疗/vn****扁桃体/n炎/n*****常用/a的/ude1四/n种/q方法/n**

For NW12, "1+2", a prefix string: *广东/ns 廉江/ns公安局/n**副/b局长/n**迁/v豪/ag 宅/ng收/v红包/n*.

## 2.2 The High-Level Structure of Our Approach

For Chinese unknown word of the two modes, the single character model and the affix mode, our idea is to construct a candidate word list based on statistics of a large-scale corpus of short text. We collected the news titles from Sina.com, and these titles are parsed as a corpus and segmented by ICTLACS. Based on our observations about the structure features of Chinese unknown word, our aim is to identify Chinese unknown word of the single-character model and the affix model.

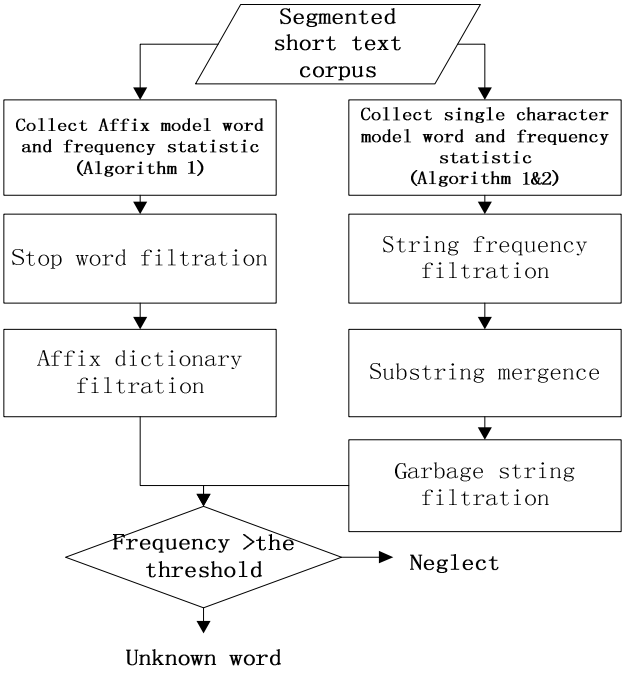


**Fig. 1.** The high-level structure of our approach

As shown in Fig.1, all potential unknown words of the two models are extracted from the corpus and their frequencies are also collected. All unknown words are added into a set of potential unknown word. For the single-character model, we adopt the string frequency filtration algorithm and the substring mergence algorithm to filter a word which is low frequency or is a redundant string. And we do the garbage string filtration. For the affix model, the stop word filtration algorithm and the affix dictionary filtration algorithm are adopted to filter a potential unknown word which is based on stop words and an affix dictionary respectively. In the end, we confirm an unknown word according to its frequency.

### 2.3 Potential Unknown Word Recognition

We introduce the PUW recognition algorithm to extract all potential unknown words. We pay our attention on the single-character model and the affix model: (1) For the Single-character model, the longest potential unknown word is extracted and its frequency is collected. All substrings of the longest potential unknown word are extracted as a potential unknown word, and their frequencies are collected also. (2) For the Affix model, postfix strings such as NW21 and NW31, and prefix strings, NW12 and NW13, are extracted and their frequencies are collected.

Two algorithms of potential unknown word recognition: (1) The Potential Unknown word Detection (PUD) algorithm is used to detect longest potential unknown words for the single-character model and potential unknown words for the affix model. All frequencies are collected; (2) The Sliding Window Algorithm (SW) is used to get all substrings of longest potential unknown words as potential unknown words for the single-character model, and get their frequencies

Here are some important variables used in the algorithm and their meanings: *SSTC* is the set of segmented short text corpus; *a* is one of segmented short text in *SSTC*, and *a*={*w[0]w[1]… w[k]*},*w* is a segmented fragment of *a*; *N(w)* is the frequency of *w*; *length(w)* is the length of *w*.

**Algorithm 1:** The Potential Unknown word Detection Algorithm (PUD)

```
Input:  set SSTC
Output:  A set slpuw of  LPUW, and a set spuw of  the affix model PUW
1.  for each a_i in SSTC
2.      get  a_i ={w[0]w[1]… w[k]}
3.      for each w[j]w[j+1] in a_i
4.          if just one length of w[j] and w[j+1] is equal to 1
5.              if  w[j]w[j+1]  not in spuw
6.                  add to spuw
7.              else
8.                  N(w[j]w[j+1])++
9.          end if
10.     end for
11.     Set temp to null
12.     for each w[j] in a_i
13.         if length(w[j])==1
14.             w[j] appended to temp
15.         end if
16.         else if length(temp)>1
17.             if temp not in slpuw
18.                 add to slpuw, temp set to null
19.             else
20.                 the N(temp)++, temp set to null
21.         end else
22.     end for
23. end for
```

The main steps of the PUD algorithm are: Get a short text from the segmented corpus, traverse the segmented fragments, extract all longest potential unknown words and all potential unknown words of the affix model, and their frequencies are recorded. After using the PUD algorithm, all longest potential unknown words are recorded in the set *slpuw*. Then the Sliding Window (SW) algorithm is used to detect all substrings of the longest potential unknown word, their frequencies are recorded.

**Algorithm 2:** The Sliding Window Algorithm (SW)

```
Input:  The set slpuw of LPUW
Output: A set subset of substring
1.  for each c_k in slpuw
2.      let s=c_k, j=2, substring is null
3.      for ( ; j< length(s); j++)
4.          for( i=0; i+j-1<length(s); i++)
5.              substring=s.sub(i,i+j) //substring of s the index is between i and
    i+j;
6.              if substring not in subset
7.                  added to subset; N(substring)=N(s);
8.              else
9.                  N(substring)++
10.         end for
11.     end for
12. end for
```

## 2.4 Filtration of PUW

If a new word appears in the network, its frequency may be high in a certain time. But in the set of our collecting potential unknown word, some strings will be neglected either for their low frequencies or their massive redundant information. For above two kinds of situations, the string frequency filtration and the substring merging are used to filter single-character string.

**Filtration of Single-Character Model**

Word is the smallest linguistic unit which could be used independently. Once a new word or expression is accepted by people, they would be used repeatedly. Based on this observation, we mainly study these single-character strings repeatedly used in a corpus. That is to say, strings whose frequencies are larger than a threshold in the potential unknown word set would be our objects.

A potential unknown word of the single-character model is not only a longest potential unknown word but also substrings whose length is at least 2. These substrings are parts of the longest potential unknown word. Such as, if "*经适房*" is a longest potential unknown word, both "*经适*" and "*适房*" are potential unknown words, and the substring merging is used to filter potential unknown word [10]:

If $N\{C_iC_{i+1}\dots C_{i+j}\} = N\{C_{i+1}C_{i+2}\dots C_{i+j+1}\} = N\{C_iC_{i+1}\dots C_{i+j}C_{i+j+1}\}$, then delete $N\{C_iC_{i+1}\dots C_{i+j}\}$ and $N\{C_{i+1}C_{i+2}\dots C_{i+j+1}\}$

If $N\{C_iC_{i+1}\ldots C_{i+j}\} > N\{C_iC_{i+1}\ldots C_{i+j}C_{i+j+1}\}$ or $\{C_{i+1}C_{i+2}\ldots C_{i+j+1}\} > N\{C_iC_{i+1}\ldots C_{i+j}C_{i+j+1}\}$, then delete $N\{C_iC_{i+1}\ldots C_{i+j}C_{i+j+1}\}$.

Here, $C_iC_{i+1}\ldots C_{i+j}$ represents a potential unknown word, and $C_i$ represents a single Chinese character, $N\{C_iC_{i+1}\ldots C_{i+j}\}$ represents the frequency of $C_iC_{i+1}\ldots C_{i+j}$.

However, for single-character model, there are no significant rules to follow. To any combination of single Chinese characters, if the meaning of the combination is clear and widely used, it can be considered as a Chinese unknown word. Garbage strings have obvious linguistic features also. So the problem of Chinese unknown word recognition can be turned into garbage string filtering from the set of potential unknown words. The garbage strings of single-character model mainly refer to noise strings generated by prepositions, adverbs and other functional words.

In this study, we extract news titles as a corpus to extract unknown word. It is obviously that there are a lot of Arabic numbers and some special characters such as Roman alphabets or English letters whose frequencies are considerably high. Strings with these characters could not be viewed as words, such as *"20年"* (20 years), *"32级"* (32 level) and *"300余"* (a little more than 300). Strings with these characters will affect recognition accuracy, so they must to be neglected.

**Filtration of Affix Model**

Potential unknown word of the affix model is consisted of a Chinese word and a single Chinese character. It means that a single Chinese character is suffixed or prefixed to a word. For the prefix form, NW12 and NW13 are two types to be mainly identified. They always appear with some prefixes such as *"副, 近, 新, 原, 创, 反, 亚, 非…"* . For these prefixes, the corpus is trained to get them and add them into a prefix dictionary. The training steps are:

1. For a well segmented corpus, look for all potential unknown words which are in the form 1+2 or 1+3.
2. Figure out every first word that appears in this mode.
3. Count out a number of the highest *N* words and add them to the prefix word dictionary.

As well, for the postfix form, NW21 and NW31 are two types to be mainly identified. Word with this kind of form always possesses obvious linguistic features. The last character can always generate a large number of three-character words and a few four-character words. We also call them postfix, such as *"门, 热, 控, 秀, 局, 案, 部, 者…"* .. For these postfixes, a training corpus is adopted to extract them and add them into the postfix dictionary. The training steps are shown as follows:

1. For a well segmented corpus, look for all potential unknown words which are in two forms of 2+1 and 3+1.
2. Figure out every last word that appears in this mode.
3. Count out the highest *N* words and add them to the postfix word dictionary.

Stop characters are firstly used to filter potential unknown word during the process of affix model filtration. Here, stop characters are ones which often appear in news titles

but cannot be viewed as a part of a word with other words, such as *"被"*, *"致"*, *"在"*, *"将"*, *"为"*, *"称"* …. Therefore, potential unknown word with stop characters as their postfix or prefix would be neglected.

In addition, further filtration for potential unknown word is conducted based on composition of the affix model. If the prefix of a potential unknown word with the form of NW12 or NW13 is the character that is in the postfix dictionary, the potential unknown word must be neglected. For example, the *"案"* in *"案抓获"* (a potential unknown word) always exists in the postfix dictionary. However, it appears as a prefix here, so *"案抓获"* would be neglected from the set of unknown word. Similarly, if the last character of a potential unknown word with the form of NW21 or NW31 is the one that is in the prefix dictionary, the potential unknown word must be neglected.

## 3 Experiment Evaluations

In this paper, the news titles are taken from sina.com as a short text corpus, because the news titles are always in real times and we can get the latest information from news titles. In addition the words in news titles are always simple and formal; it will be performance better in the experience, especially in the result of filtration. In our experience 100,000 news titles are selected randomly. The news titles are in two years period, from April, 2009 to April, 2011, and we collected in April, 2011.The titles are segmented, potential unknown words and their frequencies are collected.

In this study, two experiments are conducted to recognize Chinese unknown word, one is for the single-character model and the other is for the affix model. Both two experiments choose an optimal threshold mentioned in Fig.1. The programs of two experiments are developed by JAVA and MySql.

The performance of two experiences is measured by precision, recall and F of unknown words identification, which are defined as follows:

Precision= (number of correct identification )/(total number of identification made)*100% (1)

Recall= (number of correct identification)/(total number of identification made+number of flitted ones)*100% (2)

F= (2*Precision*Recall)/(Precision+Recall)*100% (3)

### 3.1 Experiments for Single-Character Model

In the experiment, we extract domestic news titles from Sina.com with the size of 1,000,000. They are used to identify unknown word after segmented by ICTCLAS. After a word of single-character model is identified, its frequency filtration and substring merging are used to filter single-character string, and we do the garbage string filtration. Fig.2 shows the PRF values when frequencies of all words of the single-character model are higher than a specific threshold to determine an optimal threshold.

All unknown words of the single-character model are extracted, these extracted words are with either forms of NW11, NW111 and NW1111. As shown in Fig.2, the best performance when the threshold is set to 75. In this experiment, there are 112,154

words of the single-character model extracted in total. The number of low frequency words that their frequencies are below 20 is 111,310, and the result in Fig.2 is based on the words that their frequencies are larger than 20.

### 3.2 Experiments for Affix Model

We use the same dataset to extract words of the Affix model. Stop characters and the prefix (or the postfix dictionary) are used to filter garbage strings from all potential unknown words. And their frequencies are viewed as the threshold to identify unknown word. Fig.3 shows the PRF when the frequencies of unknown words of the affix model are more than a threshold.

As shown in Fig.3, the best performance is that the frequency of unknown word is larger than 100. However, the number of identified unknown word is little in that case. So on the whole, the threshold should be set to 75. In this experiment, there are 145,900 words of the affix model extracted in all. 144,872 words can be viewed as the low frequency words, that is to say, their frequencies are lower than 20, and the result in Fig.3 is based on the left 1,028 words that their frequencies are larger than 20.

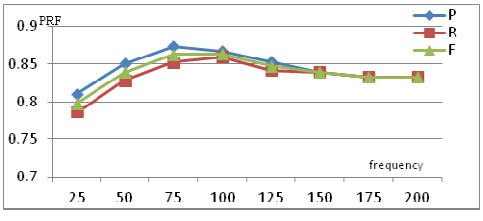


**Fig. 2.** The curve graph of PRF of the single-character model

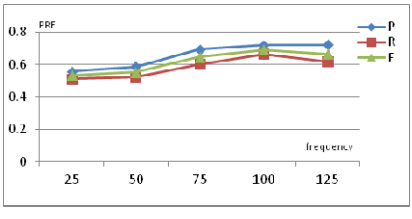


**Fig. 3.** The curve graph of PRF of the affix model

## 4 Conclusions and Future Work

Nowadays, there are many researches on Chinese unknown word recognition. However, most of them are not suitable or flexible to all types of words. The unknown word identification method discussed in this study is based on a large-scale corpus of short texts. In our approach, we collect news titles of sina.com and view these titles as short texts corpus. The single-character model and the affix model are two models we focus on. Potential unknown words are extracted and their frequencies are collected. The different methods are used to filter junk strings of the two models. Finally unknown word is extracted based on their frequencies.

Compared with other existing methods, this method has some advantages. First, the method gives the detailed forms and characteristics of unknown word, and we divide unknown word into the single-character model or the affix model. Different method is used to identify unknown word of two models to improve accuracy. In addition, the filtration method based on stop words is used in the affix model, and it performs a good effect. Second, in the experiments we get the latest news titles as the short text corpus, we can get the latest unknown word timely. And the word in news title is always formal. We can also get the unknown words however Li, H.Q [2] and Qin,

H.W [9] cannot, for example, NW111, NW13 and NW31. The experiment results show that our approach is effective.

This study focus on two models of Chinese unknown words, the approach cannot detect unknown words which are not belonging to the two modes, such as NW22. The experiment in this study does not take some real-time news titles, and our approach does not extract the latest high frequency of unknown words. For future work, we will try to design independent modules for other models of unknown words. And there is a need to expand the corpus to identify more words. Moreover, it is necessary to find some efficient methods used in filtration of junk strings.

**Acknowledgments.** This work is supported by an Opening Project of Shanghai Key Laboratory of Integrate Administration Technologies for Information Security (No. AGK2010004).

## References

1. http://ictclas.org/ (September 2011)
2. Li, H., Huang, C.-N., Gao, J., Fan, X.-z.: The Use of SVM for Chinese New Word Identification. In: Su, K.-Y., Tsujii, J., Lee, J.-H., Kwong, O.Y. (eds.) IJCNLP 2004. LNCS (LNAI), vol. 3248, pp. 723–732. Springer, Heidelberg (2005)
3. Zhang, H.P., Liu, Q.: Automatic Recognition of Chinese Unknown Words Based on Roles Tagging. Chinese Journal of Computers, 85–91 (January 2004)
4. Wu, A., Jiang, Z.X.: Statistically-Enhanced New Word Identification in a Rule-Based Chinese System. In: Proceedings of the Second Chinese Language Processing Workshop, pp. 46–51 (2000)
5. Isozaki, H.: Japanese named entity recognition based on a simple rule generator and decision tree learning. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 314–321 (2001)
6. Chen, K.J., Ma, W.Y.: Unknown word extraction for Chinese documents. In: The 19th International Conference on Computational Linguistics, pp. 169–175 (2002)
7. Meng, Y., Yu, H., Nishino, F.: Chinese new word identification based on character parsing model. In: Proceedings of First International Joint Conference on Natural Language Proceeding Sanya, Hainan Island, China, pp. 489–496 (2004)
8. Xu, Y.S., Wang, X., Tang, B.Z.: Chinese Unknown Word Recognition using improved Conditional Random Fields. In: Eighth International Conference on Intelligent Systems Design and Applications, pp. 363–367 (2008)
9. Qin, H.W., Bu, F.L.: Research on a Feature of Chinese New word Identification. Computer Engineering (2004)
10. Lv, H.L.: Chinese New Word Identification Based on Large-scale Corpus. Dalian University of Technology (2008)
11. Cui, S.Q., Liu, Q., Meng, Y., Yu, H., Nishino, F.: New Word Detection Based on Large-Scale Corpus. Journal of Computer Research and Development (2006)
12. Ding, J.L., Ci, X., Huang, J.X.: Approach of Internet New Word Identification Based on Immune Genetic Algorithm. Computer Science. 240–245 (Janruary 2011)
13. Zhang, Y., Sun, M., Zhang, Y.: Chinese New Word Detection from Query Logs. In: Cao, L., Zhong, J., Feng, Y. (eds.) ADMA 2010, Part II. LNCS, vol. 6441, pp. 233–243. Springer, Heidelberg (2010)
14. Zhu, Q., Cheng, X.Y., Gao, Z.J.: The Recognition Method of Unknown Chinese Words in Fragments Based On Mutual Information. Journal of Convergence Information Technology (2010)