A Chinese New Word Detection Approach Based on Independence Testing

Dongchen Jiang[1(✉)], Xiaoyu Chen[2,3(✉)], and Xin Yang[1]

[1] School of Information Science and Technology,
Beijing Forestry University, Beijing 100083, China
jiangdongchen@bjfu.edu.cn
[2] Beijing Advanced Innovation Center for Big Data and Brain Computing,
Beihang University, Beijing 100191, China
[3] School of Mathematics and Systems Science,
Beihang University, Beijing 100191, China
chenxiaoyu@buaa.edu.cn

Abstract. New word detection is of great significance for Chinese text information processing, which directly affects the capabilities of word segmentation, information retrieval and automatic translation. Focusing on the problem of Chinese new word detection, this paper proposes an independence-testing-based detection approach with no need of prior information. The paper analyzes statistical characteristics of new words in Chinese texts, uses statistical hypothesis testing to infer the correlations between adjacent semantic units, and proposes an iterative algorithm to detect new words gradually. Our algorithm is evaluated on both large-scale corpus and short news texts. Experimental results show that this approach can effectively detect new words from all kinds of news.

Keywords: New word detection · Hypothesis testing
Test of independence · Semantic unit

1 Introduction

Words, the basic unit of a language, are important in information processing. In the fields of information retrieval, automatic translation, part-of-speech tagging and text semantic analysis, words are the basic symbolic units with particular meanings for processing. However, unlike English and other western languages, Chinese is based on characters without white spaces to mark word boundaries. Moreover, there are not unified definitions or rules to identify Chinese words. In order to process Chinese texts, it always needs a dictionary for word segmentation.

This work was supported by Fundamental Research Funds for Central Universities (No. BLX2015-17) and National Nature Science Foundation of China (No. 61702025).

J. Fleuriot et al. (Eds.): AISC 2018, LNAI 11110, pp. 227–236, 2018.
https://doi.org/10.1007/978-3-319-99957-9_17

However, the form of Chinese words is flexible and diverse. New words can be generated from existing words and characters through derivation, compounding, abbreviation, etc. The occurrence of new words makes it difficult to handle word segmentation with a fixed dictionary. Especially with the rapid development of the Internet, unknown names and places, new companies and expressions and other kinds of new words emerge frequently. The out-of-vocabulary problem becomes the most important factor that affects the accuracy of Chinese word segmentation [1]. Therefore, effective methods of new word detection are very important for Chinese language processing.

As words are not segmented by special symbols in Chinese and there are no morphological rules for Chinese word identification, it is not feasible to detect new words by syntax analysis or morphological analysis. Currently, frequently used methods for new word detection are based on statistics, semantical rules, or both. Candidates of new words are extracted according to basic statistical features, and they are filtered according to more complex statistical features or semantical rules. These methods are effective in practice, but they usually require prior knowledge or large-scale training corpus which leads to unexpected correlations between the detection results and the prior information.

To eliminate the dependencies on prior knowledge and training corpus, this paper studies the problem of Chinese new word detection from the perspective of statistical hypothesis testing, and proposes a new word detection approach with no need of prior information. The main contributions of this paper are as follows.

(1) Three criterias are proposed to describe the statistical characteristics of new words in Chinese texts;
(2) Statistical hypothesis testing techniques are used for Chinese new word detection, and an independence-testing-based detection approach is proposed.

2 Related Work

Currently, most effective methods of Chinese new word detection adopt a two-step approach. In the first step, all possible candidates are extracted from the target text. As the result may include many garbage strings, in the second step different strategies and methods are used to filter out garbage strings and improve the accuracy.

In the candidate extraction step, the most frequently used method is providing some kind of frequency threshold for character strings. Once the frequency of a string exceeds the threshold, the string will be extracted as a candidate. Zou et al. assumed that new words need to be repeated in a certain number of texts in a period of time. Thus, the total frequency of a string and the number of texts that contain the string are used as thresholds for candidate extraction [2]. Luo and Song proposed the concept of suffix array to handle candidate extraction. Strings that have the same prefix or suffix are indexed in right or left suffix array, and the ones that appear in both suffix arrays are treated as candidates

[3]. Li et al. set a fixed frequency threshold for n-gram candidate extraction [4]. Zhang et al. used hierarchical pruning to improve the n-gram based method, which reduces the number of garbage strings [5].

The methods for candidate-word filtering are mainly based on statistics or semantic rules. In statistics-based methods, different statistical features of new words are used to describe the internal association and external boundary of new words. He et al. proposed the concepts of inside word probability and double character coupling for filtering candidate words [6]. Luo and Zhao et al. used mutual information, left/right entropy and average entropy of left/right neighbor to filter candidate words [3,7]. Statistical models where various lexicons or statistical features are used, such as predication by partial matching (PPM) and conditional random fields (CRF), were also applied to new word detection [8,9].

Besides statistical features, manually constructed word-formation rules are also used for candidate words filtering. Zou et al. defined a set of rules by regular expressions for candidate-word filtering [2]. Cui et al. trained three garbage lexicons and one suffix lexicon using a large-scale corpus to remove garbage strings [10]. Zhang and Lin et al. integrated statistical features, semantical rules and other tactics to filter candidate words [5,11].

These methods are effective in new word detection. However, in most of the methods, the parameters of statistical features and semantic rules are obtained from training large-scale corpus or from prior knowledge. As there are dependent relations between the results and the prior knowledge or training corpus, once the prior knowledge is not proper for the target text or the training corpus has a different type with the target text, the accuracy of the results will be affected. To avoid the dependencies on prior information, this paper will study the problem of Chinese new word detection based on statistical characteristics of the target text.

3 Statistics-Based Modelling

3.1 Statistical Characteristics of New Words

Generally, people identify Chinese words according to their personal experiences and the meanings of the context. But for machines, context understanding is still a challenge in natural language processing, and the flexibility of Chinese makes the results of rule-based detection incomplete.

For a given Chinese text, we believe that a new word in the text can be identified if it satisfies the following characteristics: (1) the characters in the word are interrelated; (2) the word shows a certain independence and flexibility, i.e. it can be used as some specific component of a sentence and can be connected with different words or characters; (3) the word appears in the text at a certain frequency so that people can recognize it from the context. Based on the above analysis, the statistical characteristics of new words can be described by following three criterias.

1. The Chinese characters that compose a new word show strong correlations in a given text.

2. The occurrence of a new word and the occurrence of its adjacent characters (or words) are independent.
3. The frequency of a new word reaches a certain threshold in in a given text.

In order to obtain all new words of a given text automatically, all three criterias above need to be translated into mathematical methods, which can be processed on computer. As the correlations in Criteria 1 and the independency in Criteria 2 are both concepts in statistics, we can use hypothesis testing to infer whether Criteria 1 and 2 are satisfied. For example, for any two Chinese characters X and Y, the hypothesis can be set as for any two adjacent characters, the event of X being the first character and the event of Y being the second character are independent. According to statistics knowledge, this hypothesis can be tested by observation and computation. If the test result shows that the hypothesis is accepted, then the two characters come together at random. If the result shows the rejection of the hypothesis, then there is a correlation between the occurrences of the two characters. Moreover, it is possible that the two characters constitute a word (or part of a word). Therefore, both the internal correlations between characters in a word and the external independencies between different words or characters can be inferred by hypothesis testing. In addition, for the last criteria, it only needs to set a basic frequency threshold for filtering.

3.2 Basic Concepts and Modelling

The idea of the independence-testing-based new word detection is to combine all interrelated characters into the candidates of new words. In this paper, a Chinese character string that represent a relatively complete meaning is called a *semantic unit*. Without additional information, all Chinese characters are semantic units. For a given text, the string of two adjacent semantic units is called a *semantic pair*. In a semantic pair, the former semantic unit is called the *pre-unit* of the semantic pair, and the latter one is called the *post-unit* of the semantic pair.

For a given text T, let the number of all semantic pairs be n and all of these semantic pairs constitute a sample of hypothesis testing. For a semantic unit u, the number of all the semantic pairs which have u as pre-units is denoted as n_{u+}. Then the probability of u being a pre-unit of a semantic pair in T can be estimated by $p_{u+} = n_{u+}/n$. Similarly, n_{+u} denotes the number of all the semantic pairs which have u as post-units and p_{+u} is the relevant probability. Then, for any semantic units u and v, the independence hypothesis can be stated as follows:

> H_0: for any semantic pair in T, the event of u being its pre-unit and the event of v being its post-unit are independent.

Based on this hypothesis, the frequency $n_{u,v}$ of the semantic pair uv in T can be estimated by $np_{u+}p_{+v}$. Similarly, if we use $\tilde{u}$ to denote any semantic unit that is not u, then $n_{\tilde{u},v}$ can be estimated by $np_{\tilde{u}+}p_{+v}$ where $p_{u+} + p_{\tilde{u}+} = 1$.

With above analysis, the statistic $Q^2_{u,v}$ can be constructed to characterize the total frequency errors of semantic pairs associated with u and v.

$$Q^2_{u,v} = \frac{(n_{u,v} - np_{u+}p_{+v})^2}{np_{u+}p_{+v}} + \frac{(n_{u,\tilde{v}} - np_{u+}p_{+\tilde{v}})^2}{np_{u+}p_{+\tilde{v}}} + \frac{(n_{\tilde{u},v} - np_{\tilde{u}+}p_{+v})^2}{np_{\tilde{u}+}p_{+v}} + \frac{(n_{\tilde{u},\tilde{v}} - np_{\tilde{u}+}p_{+\tilde{v}})^2}{np_{\tilde{u}+}p_{+\tilde{v}}} \tag{1}$$

If we assume that in Eq. (1) each error indicating the difference between the actual frequency and the relevant estimated frequency fits a normal distribution, then $Q^2_{u,v}$ fits the chi-squared distribution with 1 degree of freedom, i.e. $Q^2_{u,v} \sim \chi(1)^2$.

For any semantic pair uv in T, the 2 × 2 contingency table for uv can be constructed accordingly, i.e.

Table 1. 2 × 2 Contingency table of uv

Pre-unit	Post-unit		
	$+v$	$+\tilde{v}$	Row total
$u+$	a	b	$a+b$
$\tilde{u}+$	c	d	$c+d$
Column total	$a+c$	$b+d$	n

In Table 1, a is the frequency of the semantic pair uv in T, b, c and d are the frequencies of $u\tilde{v}$, $\tilde{u}v$ and $\tilde{u}\tilde{v}$, respectively, and $n = a+b+c+d$ holds. With this table, Eq. (1) can be simplified, and we have

$$Q^2_{u,v} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}. \tag{2}$$

Accordingly, the correlation between the occurrences of u and v in one semantic pair can be inferred by independence testing: given a significance level α, if $Q^2_{u,v}$ is in the critical region, H_0 is rejected, which means that the occurrences of u and v in one semantic pair is correlated; otherwise, H_0 should be accepted, which indicates that u and v occur in the semantic pair uv independently.

4 Algorithm Based on Independence Testing

4.1 Problem Analysis

Independence testing can be used to infer the correlation between the occurrences of two semantic units in one semantic pair. However, even if the testing result is rejection of the hypothesis, it is still problematic to identify the semantic pair of the two units as one new word.

Firstly, the rejection of H_0 means that the event of u being a pre-unit of one semantic pair and the event of v being the post-unit of the same semantic pair are correlated. However, the correlation may not be determined by the co-occurrence of u and v. It may be determined by the co-occurrence of u and $\tilde{v}$ or the co-occurrence of $\tilde{u}$ and v. Therefore, to show that the correlation is indeed determined by the co-occurrence of u and v, it is still necessary to demonstrate $n_{u,v} \geq \delta n p_{u+} p_{+v}$ where the coefficient δ is significant larger than 1.

Secondly, two interrelated semantic pairs often share one semantic unit, i.e. the post-unit of one semantic pair is the pre-unit of the other semantic pair. In this case, improper word identification tactic may result in inaccuracy or irrationality.

If we identify one of these overlapped interrelated semantic pairs as a word, the correlation of the other semantic pair is affected, which may cause inaccurate word identification. Take the name “特朗普” (Trump) as an example, the consecutive semantic units “特朗普” contains two overlapped semantic pairs “特朗” and “朗普”, but neither of them is a word. Thus, it is inappropriate to select one of the overlapped semantic pairs as a new word.

However, if we identify all the consecutive semantic units that constitute several overlapped semantic pairs as one word, it may also cause irrationality. For example, “人民” (people) and “民办” (civilian-run) are two interrelated semantic pairs sharing the common character “民”. And it is unwise to identify “人民办” as one new word in the sentence “为人民办实事” (do practical work for the people).

The main reason for this problem is that the test of independence is used to infer the correlation between two semantic units and it is insufficient to analyze the correlations between multiple consecutive semantic units. Therefore, it needs other tactic to detect words formed by multiple semantic units.

4.2 Algorithm Description

In order to solve the problem of interference between overlapped interrelated semantic pairs, we apply an iterative approach to merge the most interrelated semantic pair as a new semantic unit for each iteration and gain new words gradually.

More specifically, for a text T, the statistic Q^2 of each semantic pair can be calculated under the independence hypothesis H_0 according to Eq. (2). For any semantic pair whose Q^2 is in the critical region, the larger Q^2 is the stronger the correlation between its internal semantic units is. Therefore, we can select the internal related semantic pair with the largest Q^2 as a semantic unit. Based on this idea, we merge the most related semantic units and obtain new words gradually. The independence-testing-based Chinese word detection algorithm is proposed as follows:

In the algorithm, U and V are used to record the adjacent semantic units which have the largest Q^2. $freq$ is used to record the frequency of semantic pair UV. For each iteration, UV will be merged into one semantic unit and

Algorithm 1. $IHT-WD$

```
Input: A string of Chinese text $T$
Output: A list $wordlist$ of new detected words
foreach $character\ c\ of\ T$ do
    if $c$ is a Chinese character then
        insert $(c, pos_c)$ into $semanticUnitList$;
$Q^2 = \chi(1)^2_\alpha$, $freq = threshold$;
set $U$ and $V$ to be empty;
while $Q^2 \geq \chi(1)^2_\alpha$ do
    forall the semantic pair $UV \in semanticUnitList$ do
        set $UV$ as one semantic unit and update $semanticUnitList$;
    update all frequency information of $semanticUnitList$;
    set $Q^2 = 0$, $freq = threshold$;
    foreach semantic pair $uv \in semanticUnitList$ do
        if $n_{u,v} \geq threshold \wedge n_{u,v} \geq \delta n p_{u+} p_{+v}$ then
            calculate $Q^2_{u,v}$ according to Equation (2);
            if $Q^2_{u,v} > Q^2 \vee Q^2_{u,v} = Q^2 \wedge n_{u,v} > freq$ then
                update $Q^2$, $freq$, $U$ and $V$ according to $Q^2_{u,v}$, $n_{u,v}$, $u$ and $v$;
foreach semantic unit $u \in semanticUnitList$ do
    if $u$ is not a character $\wedge\ n_u \geq threshold$ then
        insert $u$ into $wordlist$;
return $wordlist$;
```

semanticUnitList will be updated accordingly (the two operations will not be performed in the first loop as U and V are empty at the beginning). After these operations, the algorithm will find the most interrelated semantic pair in *semanticUnitList*. If its error statistic is in the critical region and its actual frequency exceeds the frequency threshold *threshold*, the merging and the updating operations will be performed in next iteration.

In this algorithm, *threshold* is a frequency threshold which is determined by two factors. Firstly, for any text, there should be a basic frequency threshold *basic_freq* for all new words according to Criteria 1 of Sect. 3. *threshold* must not be smaller than *basic_freq*. Secondly, as the number n of all semantic units may vary from text to text, if *threshold* is fixed, the noise will increase as the text grows. To avoid this, the threshold should grow as the length of the text increases. In this paper, the average word frequency is used for the threshold. As the number of words are unknown before processing, Heaps' law is used for estimation, i.e. if there are N words in the text, the text contains KN^β different words. In this paper, N is estimated by half of the initial semantic pairs n. Then, the frequency threshold *threshold* can be calculated as follows:

$$threshold = max(basic\_freq, (\frac{n}{2})^{1-\beta}/K).$$

This iterative merging algorithm can solve the problems mentioned above. Take the name " 特朗普 " as an example: if " 特朗 " instead of " 朗普 " is merged in one iteration, then " 特朗 " will be merged with " 普 " in some following iteration. For the example of " 人民 " and " 民办 ", if " 人民 " is merged first, then " 人民办 " and " 民办 " will become two different semantic pairs. " 民 " and " 办 " still can be merged, but the merge of " 人民 " and " 办 " is rare in practice.

5 Experiments

The detection algorithm is estimated by precision. As the three criterias for new words in Sect. 3 also accord with the characteristics of some commonly used Chinese words, the detection results may contain existing words, new words and garbage strings. If we use R, E, N and G to denote the set of result strings, the set of existing words, the set of new words and the set of garbage strings, respectively, then word detection precision P_w and new word detection precision P_n can be used for evaluation of our approach.

$$P_w = \frac{|E| + |N|}{|R|}, \quad P_n = \frac{|N|}{|N| + |G|}.$$

In the experiments, parameters in the detection algorithm are set as follows: the significant level α is set to 0.5%; the basic frequency $basic\_freq$ is set to 4; in Heaps' law, we set $K = 1$, $\beta = 0.75$, which is suitable for news and other short texts; the coefficient δ for co-occurrence judgement is set to 4.

The algorithm is firstly evaluated by using the ICTCLAS testing corpus - People's Daily of Jan 1998, which is provided by Peking University. All news are input as one single text, and the output is a list of semantic units. The existing words and the new words are obtained in different methods. All words in the segmentation result of the ICTCLAS corpus are set as the existing words. After filtering out the existing words from the result, new words are manually annotated according to the following criteria: (1) a new word should be an existing entry in Wikipedia or Baidu encyclopedia; (2) a new word should represent a clear concept in real life; (3) a new word should be the abbreviation of an existing word. If a semantic unit meets one of the criteria, it is identified as a new word (Table 2).

Table 2. Result on ICTCLAS corpus

Total	Existing	New	P_w
4471	3308	350	81.82%

The algorithm has detected 4471 different semantic units from the ICTCLAS corpus, 3308 of them are existing words, and 350 of them are new words. The precision of word detection is 81.82%. As the word segmentation result of the ICTCLAS corpus has already included some of the new words, it is not appropriate to show the precision of new word detection by this experiment. To illustrate the effect of the algorithm on new word detection, we randomly download 100 pieces of news from www.gmw.cn, which cover politics, international news, economy, life, sport, education, etc (Table 3).

Table 3. Result on news from www.gmw.cn

Total	Existing	New	P_w	P_n
1217	767	368	93.26%	81.78%

The algorithm obtains 1217 semantic units in total, which includes 767 existing words and 367 new words. The new word detection precision is 81.78% while the total word detection precision is 93.26%. Because there is no standard test set for Chinese new word detection, it is improper to compare our method with existing ones directly. But in terms of the new word detection precision, our approach is competitive. Some examples of new words are listed as follows:

- Entries in Baidu encyclopedia: 公办学校 (the public school), 智能芯片 (intelligent chip), 基本医疗保险 (basic medical insurance), 非洲裔美国人 (African American), etc.
- Semantic units representing a clear concept: 一带一路 (the Belt and Road), 安倍政府 (Abe administration), 星巴克 (Starbucks), 积极废人 (an active loser), etc.
- Abbreviations: 上合组织 (Shanghai Cooperation Organization), 外研社 (Foreign Language Teaching and Research Press), 世卫组织 (World Health Organization), etc.

6 Conclusion and Future Work

This paper presents three statistical criterias for new word detection and proposes an independence-testing-based approach for Chinese new word detection. Compared with the existing methods, our approach does not need prior knowledge or large scale training corpus and it is more suitable for detecting new words from news texts. The experiment on randomly selected news shows that the new word detection precision of the algorithm is over 80%, which is competitive compared with the existing methods. As the method only use some statistical characteristics of new words, it is recommended to combine this method with semantic-rule-based methods to improve the accuracy.

In the future, we will construct a test set of Chinese new word detection for comparison and calculate the recall rate of our approach. Furthermore, since both the coefficient δ and the frequency threshold *threshold* can be determined by statistical methods, the methods of setting these parameters in Algorithm 1 and the impact of these parameters will be studied in a late stage.

References

1. Huang, C.N., Hai, Z.: Chinese word segmentation: a decade review. J. Chin. Inf. Process. **21**(3), 8–19 (2007)
2. Zou, G., Liu, Y., Liu, Q.: Internet-oriented Chinese new words detection. J. Chin. Inf. Process. **18**(6), 1–9 (2004)
3. Luo, Z., Song, R.: An integrated method for Chinese unknown word extraction. In: Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing, pp. 148–154. Association for Computational Linguistics (2004)
4. Li, D., Tu, W., Shi, L.: Chinese new word identification algorithm based on context-aware. Comput. Eng. Des. **33**(10), 4022–4027 (2012)
5. Zhang, H., Yong, L.I., Yan, Q.: Method of new Chinese words identification from large scale network corpora. Comput. Eng. Appl. **51**(5), 208–213 (2015)
6. He, M., Gong, C., Zhang, H., Cheng, X.: Method of new word identification based on lager-scale corpus. Comput. Eng. Appl. **43**(21), 157–159 (2007)
7. Zhao, X., Zhang, H.: New words identification based on iterative algorithm. Comput. Eng. **40**(7), 154–158 (2014)
8. Zeng, H.L., Zhou, C.L., Shi, X.D., et al.: New word detection algorithm for Chinese based on extraction of local context information. In: Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, pp. 797–801. IEEE Xplore (2008)
9. Peng, F., Feng, F., Mccallum, A.: Chinese segmentation and new word detection using conditional random fields. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, pp. 562–568 (2004)
10. Cui, S.: New word detection based on large-scale corpus. J. Comput. Res. Dev. **43**(5), 927–932 (2006)
11. Zhang, H., Luan, J., Li, Y., Qi, X.: Method of new Chinese word detection based on statistical learning framework. Comput. Sci. **39**(2), 232–235 (2012)