Artificial Intelligence Technologies and Applications
C. Chen (Ed.)

doi:10.3233/FAIA231409

A New Method for Improving the Accuracy of Word Segmentation in Modern Chinese Texts

Yan ZHANG [1]
College of Chinese Language and Culture Jinan University Guangzhou, China

Abstract. Chinese does not adopt the form of word segmentation in writing like English. The role of words in NLP is enormous and can have a significant impact on subsequent tasks. Presently, NLP research in China mainly focuses on modern Chinese, leaving a gap in studying the pre modern Chinese stage during the late Qing Dynasty to early Republic. Many texts from this time were in traditional paper books with varying preservation conditions, making it hard to obtain a representative dataset, limiting in-depth research. Most texts haven't undergone digital processing, lacking segmentation or annotation. This poses great challenges for researchers. Researchers must design word segmentation algorithms from scratch, increasing difficulty and workload. By using N-gram to extract bigram and trigram fragments, and combining with 100 Year Chinese New Words and Phrases Dictionary and other books, this paper constructs a Chinese wordlist at the end of the Qing Dynasty to the beginning of the Republic of China. 10000 sentences were randomly selected from the corpus constructed in this article and manually segmented and annotated. The experimental results show that adding the wordlist constructed in this article has indeed improved the accuracy of word segmentation for most of these 10000 sentences. Specifically, the three software (CoreNLP, FudanNLP, and Jieba) show a certain degree of performance improvement when using the wordlist. For HanLP, the use of wordlist in this set of data did not bring significant performance improvement, or even a slight decline to some extent.

Keywords. modern Chinese; segmentation; method; accuracy

1.Introduction

Natural Language Processing (NLP) is an important direction in the field of computer science and artificial intelligence, studying various theories and methods of communication between computers and human beings in natural languages (such as English and French) [1-2]. Natural language is used to interact with computers, which is a manifestation of social and economic development and progress, as well as an inevitable historical mission of NLP technology [3-4]. In terms of broad categories, NLP can be divided into two categories focusing on basic research and applied research. Basic research is manifested in basic subjects, such as English, mathematics, statistics and other subjects, and in subdivision content, such as identifying errors in wrong

[1] Corresponding author: Yan ZHANG, College of Chinese Language and Culture Jinan University Guangzhou, China; email: zyahsf@126.com

sentences, eliminating wrong words, and classifying different types of statistics. Applied research is manifested in many aspects of practical applications, such as information query, word segmentation, text translation, part of speech classification, etc.

In recent years, with the rapid development of artificial intelligence technology, NLP has become an important branch of artificial intelligence. China's achievements in NLP are reflected in many aspects. First, China has made important breakthroughs in core technologies such as Machine translation, text generation and speech recognition, which has promoted the development of applications such as automated translation and intelligent speech assistant. Secondly, China has advantages in the construction of large-scale corpora and data mining, providing strong data support for the training and optimization of deep learning models. In addition, China's NLP research institutions and enterprises actively participate in international competitions and academic cooperation, maintain close cooperation with world-class research teams, and constantly improve their research level. Chinese researchers and enterprises have invested a lot of energy and resources in this field and made significant progress. For example, the big model of Spark cognition of Iflytek (https://xinghuo.xfyun.cn/), ERNIE Bot of Baidu (https://yiyan.baidu.com/), Tongyi Qianwen of Alibaba (https://tongyi.aliyun.com/). With the continuous innovation and progress of NLP technology, China will continue to play an important role in this field and make greater contributions to promoting the application and development of artificial intelligence technology.

Generally speaking, in the past, Chinese NLP focused on Chinese texts in general fields, such as automatic news summarization, sentiment analysis, and grammar correction in Chinese writing. However, there is still a lack of Chinese language research in certain specific fields, which either have a limited audience or limited research materials. In short, conducting research on them requires more effort than conducting research on general modern Chinese, but the benefits achieved are not as good as the former. Cognitive scientists believe that language and cognition are closely linked, and through language, we can glimpse human understanding of the world [5, 6, 7]. Therefore, with the development of modern NLP, we will conduct research on the large number of ancient and modern Chinese texts that exist in China, and explore the scientific laws behind them.

The NLP achievements on modern Chinese are very rich, but there is still a lack of research on the early stage of modern Chinese in the late Qing Dynasty to early Republic of China, which is between ancient Chinese and modern Chinese. The Chinese language at this stage is different from ancient Chinese and also from modern Chinese. Overall, its characteristics are more inclined towards the modern Chinese language we currently use. Therefore, we refer to this stage of Chinese as pre modern Chinese. At this stage, the number of Chinese texts is not large, and high-quality research materials are even more scarce. Due to the decline of China's national strength during this period, a large number of literati advocated for a comprehensive study of everything in the West, including the languages of western countries. Therefore, Chinese language at this stage is often full of western language grammar, vocabulary, and other characteristics, making it completely different from ancient Chinese and Mandarin. Therefore, various Chinese NLP tools based on general field research in China are not suitable for processing text during this period. Due to the fact that Chinese is different from phonography writing in English, word segmentation is a prerequisite for subsequent NLP tasks. However, due to the characteristics of Chinese at this stage, there are currently many errors in the mainstream Chinese NLP when segmenting them. For example: "心裏就未免設疑，故此禀報領事官，函致稅務司。" The sentence comes from *An Enchiridion of*

Mandarin (官话指南). The book was compiled by the students as Interpreters of the Japanese legation in China in Qing Dy-nasty, who sorted out their daily schoolbooks, with the help of their Chinese teachers. We used four mainstream word segmentation software to segment this sentence, and the results are as follows:

心/裏/就/未免/設疑/，/故此/禀報領/事官/，/函致稅/務司/。(CoreNLP)

心裏就/未免/設疑/，/故/此禀報/領事官/，/函致/稅務司/。(FudanNLP)

心/裏/就/未免/設疑/，/故此/禀報/領事官/，/函致/稅務司/。(Jieba)

心/裏/就/未免/設/疑/，/故此/禀/報/領/事/官/，/函/致/稅/務/司/。(HanLP)

The accurate result of expert manual word segmentation should be:

心裏/就/未免/設疑/，/故此/禀報/領事官/，/函致/稅務司/。

The four major word segmentation software did not correctly divide words, the main reasons are as follows: sentences are mixed with traditional Chinese characters and simplified Chinese characters, and domestic word segmentation software can't handle this situation well; There are many elegant archaic words in sentences, such as "函致". Mainstream word segmentation software is generally based on wordlists and combined with various algorithms for word segmentation. However, they are based on wordlists that are specific to modern Chinese in the general field. For Chinese texts from the late Qing Dynasty to the early Republic of China, many distinctive words were not included in mainstream word segmentation software. Therefore, this article explores the development of a new method to construct a characteristic wordlist of Chinese at the end of the Qing Dynasty to the beginning of the Republic of China, in order to optimize the accuracy of mainstream word segmentation software for text segmentation during that period. At the same time, the improvement of word segmentation accuracy in word segmentation software is of great significance for subsequent NLP tasks.

2.Literature Review

2.1 Modern Chinese

Chinese is the language with the highest number of speakers in China and also the language with the highest number of speakers as the first language in the world. Modern Chinese can be divided into standard language and dialects. Mandarin is the standard language of modern Chinese, with Beijing pronunciation as the standard pronunciation, northern dialect as the basic dialect, and typical modern vernacular writings as grammatical norms.[8]

Modern Chinese was finally formed during the May 4th Movement of 1919. Modern Chinese has two sides; One aspect is its written form, namely vernacular ("白话"), and the other aspect is its oral form, namely Mandarin ("普通话") [9].

There are many differences in pronunciation, vocabulary, and grammar between modern Chinese and ancient Chinese. At present, the NLP of Chinese language mainly focuses on the study of modern Chinese, such as FudanNLP of Fudan University (https://github.com/FudanNLP/fnlp), Hanlp of Han He (https://github.com/hankcs/HanLP), CoreNLP of Stanford University (https://github.com/stanfordnlp/CoreNLP), etc. The study of ancient Chinese also involves many aspects, such as the sikuGPT

developed by Nanjing Agricultural University and Nanjing Normal University on GPT2 for the generation and translation of ancient Chinese texts.

However, the research on Chinese NLP in the transitional stage lacks corresponding research results. For example, the Chinese texts in the late Qing Dynasty to the early Republic of China are obviously different from the ancient Chinese in the Pre Qin and Han Dynasties, and also from the modern Chinese after the May Fourth Movement of 1919. It seems to be a transitional stage between the two, but its language features are more inclined towards modern Chinese. Therefore, traditional modern Chinese and ancient Chinese NLP tools do not seem suitable for processing it.

2.2 Chinese Segmentation

Word segmentation is a very important step in Chinese Natural language processing, which refers to the segmentation of continuous Chinese text into meaningful words. This process may seem simple, but it is actually a complex and critical step in Chinese processing, with the following importance:

- Chinese words do not have clear boundaries: compared to languages such as English separated by spaces, Chinese words do not have clear boundaries between them, which requires word segmentation when processing Chinese text in order to conduct subsequent semantic analysis and understanding. In each task of Chinese NLP, word segmentation is often a pre step, such as machine translation, text classification, keyword extraction, etc. Only through correct segmentation results can effective subsequent processing be carried out.
- Improvement of processing efficiency: Chinese word segmentation can divide long texts into several smaller words, thereby reducing the complexity and computational cost of subsequent processing and improving processing efficiency. The quality of word segmentation directly affects the accuracy and efficiency of subsequent processing. If the word segmentation is not accurate, it will lead to semantic understanding errors and affect the emotional analysis, Information extraction and other tasks of the text.

The effect of word segmentation involves grammatical results, parts of speech classification and word segmentation processing. The effect of word segmentation must be maintained with high accuracy. Of course, it will affect the later part of language text classification. If the subject of the previous sentence is divided into wrong words, it will affect the word segmentation of the following words. If continuous mistakes will lead to the complete distortion of the original meaning of the sentence, and even the meaning of the sentence may become the opposite. For example, the original intention of the sentence is to express “不高兴 bu gao xing (unhappy)”. If there is a mistake in the word segmentation, the word “bu (not)” is separated, and then the sentence becomes the meaning of happy. Chinese word segmentation generally has two ways:

- Choosing the thesaurus as the basic reference, the algorithm adopts the matching method: forward / inverse maximum matching segmentation method.
- Choosing a way that does not depend on the thesaurus. For example, use statistical methods to segment words. In context, the more adjacent characters appear at the same time, the more likely they are to form a word.

The above two segmentation methods can not completely solve the problems of ambiguity recognition and unknown word recognition. But there are many difficulties in Chinese word segmentation, this section took the segmentation of ambiguous characters

as an example. Segmentation ambiguity can be divided into three types: intersected type, combined type and mixed type.

1) Intersected type.

A Chinese character string AXB. If AX and XB are words at the same time, where A, X and B are Chinese character strings respectively. Therefor there are different segmentation methods for character X, which is called intersection ambiguity. For example, "从中学(cong zhong xue)". It can integrate "cong" and "zhong" into "congzhong (from among)", and "zhong" and "xue" into "zhongxue (middle school)", then there will be intersection ambiguity.

2) Combined type.

For a language M, such as "W_1W_2", there are two words "W_1" and "W_2" , or one word "W_1W_2". For example, "中国家和(zhong guo jia he)", then the string can be divided into "Zhongguo (China)" and "Jiahe (a place name)", or the other is the whole "zhongguojiahe".

3)Mixed type.

It refers to the synthesis of intersection and combination. For example, "哪里的人才很受欢迎? (Na li de ren cai hen shou huan ying?)" can be divided into "Nali de ren / cai / hen shou huanying? (Where are the people popular?)" and "Nali de / rencai / hen shou huanying? (Where are the talents popular?)".

Therefore, scientifically segmenting Chinese text is a challenging NLP task.

3. Experimental Design and Validation

This section will explain how to optimize the efficiency and accuracy of modern Chinese segmentation. The experimental process is shown in the figure 1. Each process of the following experiments will be explained one by one in the following text.

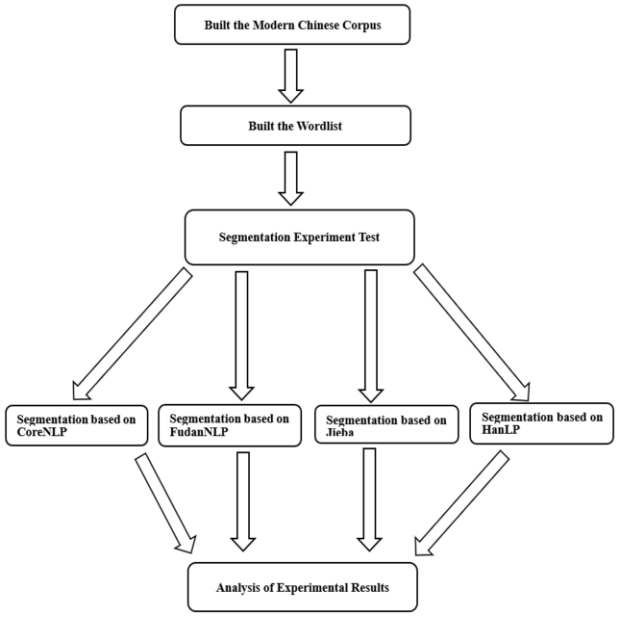

Figure 1. The experimental flowchart.

3.1Built the Modern Chinese Corpus

We have selected a large number of Chinese texts from the late Qing and early Republic of China to fully reflect the Chinese language of that era, such as T. F. Wade's *YU-YEN TZU-ERH CHI* ("语言自迩集"), Calvin Wilson Mateer's *A Course of Mandarin Lessons, based on Idiom* ("官话类编"), and Wu Qitai's *Guide to Mandarin* ("官话指南"), etc. Most of the texts come from the CCL corpus of Peking University (http://ccl.pku.edu.cn:8080/ccl_corpus/). In the end, we obtained a corpus of Chinese texts from the late Qing Dynasty to the early Republic of China, covering nearly a hundred years from the early 19th century to the early 20th century.

3.2Built the Wordlist

The basic idea of the statistical based word segmentation method is that the more connected words appear in Chinese sentences, the more they are used as words, the higher the reliability of sentence splitting, and the higher the accuracy of word segmentation. The basic principle is to count the number of times words appear, and words that appear high enough are retained as separate words. The statistical based word segmentation method can effectively handle the problem of unregistered words and ambiguous words, without the need for manual wordlist. However, it relies too much on a corpus, which has a large computational load and average word segmentation speed. Common statistical based word segmentation methods include N-Gram Model and Hidden Markov Model (HMM).

In order to build a Chinese wordlist suitable for the end of the Qing Dynasty to the beginning of the Republic of China, we chose the bigram and trigram of N-Gram Model to screen out the unique words at this stage. The word sequence s is $\omega_1, \omega_2, \dots, \omega_m$ .The specific calculation formula is as follows:

When N=2, the model is called a binary grammar model, where the probability of each word appearing is related to the first word, including:

$$P(\omega_i \mid \omega_1, \omega_2, \dots, \omega_{i-1}) = P(\omega_i \mid \omega_{i-1}) \tag{1}$$

Then, the probability of the word sequence s is:

$$P(s)=P(\omega_1,\omega_2,\dots,\omega_m)=P(\omega_1)P(\omega_2|\omega_1)\dots P(\omega_m|\omega_{m-1}) \tag{2}$$

When N=3, the model is called the trigram model, where the probability of each word appearing is related to the first two words, including:

$$P(\omega_i|\omega_1,\omega_2,\dots,\omega_{i-1})=P(\omega_i|\omega_{i-2}\omega_{i-1}) \tag{3}$$

Then, the probability of the word sequence s is:

$$P(s)=P(\omega_1,\omega_2,\dots,\omega_m)=P(\omega_1)P(\omega_2|\omega_1)P(\omega_3|\omega_1\omega_2)\dots P(\omega_m|\omega_{m-2}\omega_{m-1}) \tag{4}$$

In order to avoid the situation where certain words do not appear in the corpus, resulting in zero numerator or denominator, this article adopts a data smoothing method

with an addition of 1. Through calculation, we can obtain many bigram and trigram fragments with higher frequency of use.

At the same time, we also referred to books such as *Research on New Words and Phrases during the Republic of China Period* (“民国时期新词语研究”) [10], *100 Year Chinese New Words and Phrases Dictionary* (“100 年汉语新词新语大辞典”) [11], *The Great Chinese Dictionary* (“汉语大词典”) [12] to determine whether the selected bigram and trigram text fragments are the same word. New words and phrases mainly come from words borrowed from foreign languages and words created by Chinese people. Many words are unique at this stage, and with the development of modern Chinese, they are no longer used, such as the English word—telephone, which was originally called “德律风”. Therefore, words unique at this stage are often misclassified by mainstream Chinese word segmentation software such as CoreNLP.

With the help of N-gram and various dictionaries, we finally got wordlist of Chinese at the end of the Qing Dynasty to the beginning of the Republic of China.

3.3Segmentation Experiment Test

In order to determine which segmentation result is better, there are usually two ways. One is to call various NLP software interfaces to segment specific sentences and compare the segmentation results by feeling, but this segmentation result has a significant subjective color. Another method is to analyze the segmentation results and standard segmentation results through a test set, and obtain accuracy, recall, etc. Obviously, the second method is more scientific, but it requires a large amount of manually annotated and accurately segmented datasets. The currently publicly available segmentation datasets include *the People’s Daily* of 1998 segmentation dataset, the Second International Chinese Word Segmentation Bakeoff Data, etc.

Obviously, manually annotated Chinese segmentation datasets require a lot of manpower and resources, and there are not many publicly available datasets. As for the Chinese word segmentation dataset from the late Qing Dynasty to the early Republic of China, it does not exist. Therefore, due to limited energy, we adopted three sets of books: T F. Wade’s *YU-YEN TZU-ERH CHI*, Calvin Wilson Mateer’s *A Course of Mandarin Lessons, based on Idiom*, and Wu Qitai’s *Guide to Mandarin* randomly selected 10000 sentences and manually segmented them as the segmentation data standard set.

This article tested the accuracy of four segmentation software, CoreNLP, FudanNLP, Jieba, and HanLP, and tested their word segmentation performance without and using our constructed word list. We selected 10000 sentences from our manual word segmentation as the baseline for comparison. The performance indicator used in this article is accuracy. Comparing the results of software segmentation with those of manual segmentation, it is completely consistent that the sentence segmentation is accurate. Finally, calculate the ratio between the number of accurate words for word segmentation and the total number of words for 10000 sentences, and use this ratio as the accuracy value for the final word segmentation software. The formula is as follows:

$$\text{Accuracy} = \text{accurate word segmentation} / \text{total number of words} \tag{5}$$

In order to know the accuracy of these four software, we ran it through the Second International Chinese Word Segmentation Bakeoff Dataset (SICWSBD, http://sighan.cs.uchicago.edu/bakeoff2005/). It is worth noting that the four software all

have many word segmentation method interfaces, and this article uses their default interfaces. This article tested simplified Chinese data from Peking University in SICWSBD. The following figure 2 shows the accuracy of four software.

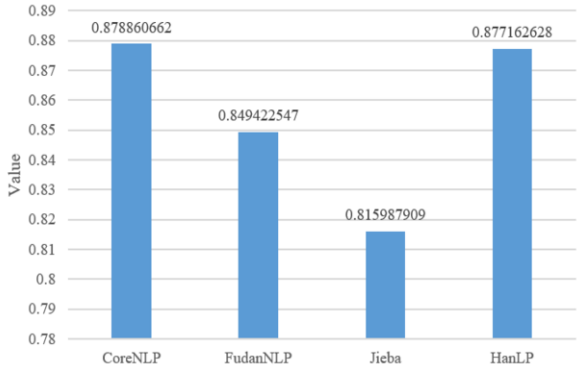

Figure 2. The accuracy of four word segmentation software.

From figure 2, it can be observed that CoreNLP and HanLP have the highest accuracy, and the numerical values are very close, about 87.80%. And Jieba's accuracy is the lowest, only about 81.60%. The paper did not use their other complex word segmentation algorithms, but still achieved excellent word segmentation performance. Overall, their performance is relatively excellent.

After understanding the performance of four software segmentation methods, we tested their segmentation performance on the 10000 sentences we selected. The results are as follows:

Table 1 Experimental Test Results Table

Software	***Without Wordlist***	***With Wordlist***	***increase or decrease***
CoreNLP	0.850613550	0.864768947	increase
FudanNLP	0.825861927	0.830571504	increase
Jieba	0.804382982	0.819233572	increase
HanLP	0.865304394	0.861596828	decrease

Drew the data from table 1 into figure 3.

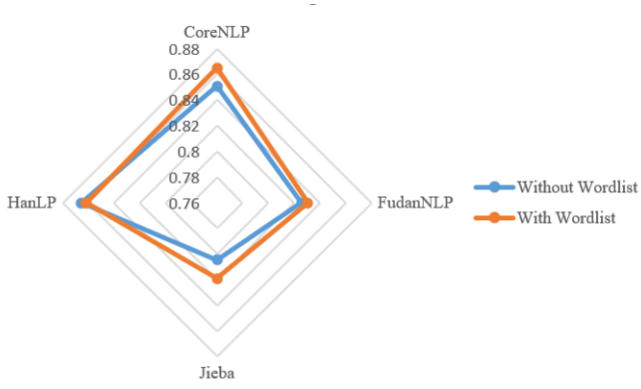

Figure 3. Segmentation accuracy for 10000 sentences.

3.4Analyze the Experimental Results

Based on the values in table 1 and figure 3, the following conclusions can be drawn:

- When the wordlist is used, the performance is relatively good, and most of its values are higher than the corresponding performance when the wordlist is not used.
- Some software may not get significant performance improvement after using the wordlist (for example, HanLP), while some software may have some performance improvement (for example, CoreNLP and Jieba).

4. Conclusions

Chinese does not adopt the form of word segmentation in writing like English. The role of words in NLP is enormous and can have a significant impact on subsequent tasks. Presently, NLP research in China mainly focuses on modern Chinese, leaving a gap in studying the pre modern Chinese stage during the late Qing Dynasty to early Republic. Many texts from this time were in traditional paper books with varying preservation conditions, making it hard to obtain a representative dataset, limiting in-depth research. Most texts haven't undergone digital processing, lacking segmentation or annotation. This poses great challenges for researchers. Researchers must design word segmentation algorithms from scratch, increasing difficulty and workload.

By using N-gram to extract bigram and trigram fragments, and combining with 100 Year Chinese New Words and Phrases Dictionary and other books, this paper constructs a Chinese wordlist at the end of the Qing Dynasty to the beginning of the Republic of China. 10000 sentences were randomly selected from the corpus constructed in this article and manually segmented and annotated. The experimental results show that adding the wordlist constructed in this article has indeed improved the accuracy of word segmentation for most of these 10000 sentences. Specifically, the three software (CoreNLP, FudanNLP, and Jieba) show a certain degree of performance improvement when using the wordlist. For HanLP, the use of wordlist in this set of data did not bring significant performance improvement, or even a slight decline to some extent.

Acknowledgment

Thank you to my supervisors Professor J. Li and Professor Y. J. Wang for their help in the article. Sincere gratitude to anonymous reviewers for their criticism and correction of the article.

References

[1]A. Wong, J. M. Plasek, S. P. Montecalvo, et al. Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 2018, 38(8): pp. 822-841.

[2]W. H. Weng, K. B. Wagholikar, A. T. Mccray, et al. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Medical Informatics & Decision Making, 2017, 17(1): pp. 155.

[3]H. S. Chase, L. R. Mitrani L, G. G. Lu, et al. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Medical Informatics & Decision Making, 2017, 17(1), pp. 1-8.
[4]J. Downs, S. Velupillai, G. George, et al. Detection of Suicidality in Adolescents with Autism Spectrum Disorders: Developing a Natural Language Processing Approach for Use in Electronic Health Records. AMIA. Annual Symposium proceedings. AMIA Symposium, vol. 2017, 2018. pp. 641-649.
[5]Perlovsky L. Language and cognition[J]. Neural Networks, 2009, 22(3), pp. 247-257.
[6]D. Seleskovitch, Language and cognition. Language interpretation and communication. Boston, MA: Springer US, 1978, pp. 333-341.
[7]G. J. Whitehurst, and B. J. Zimmerman, The functions of language and cognition. NewYork: USA, Academic Press, 2014.
[8]Y. S. Hu, Modern Chinese.Shanghai: Shanghai Educational Publishing House, 1981, pp. 4-9.
[9]S. K. Zhang, Chinese Language Learning Series. Jinan, Shandong: Shandong Education Press, 1983, pp. 30.
[10]N. Li, Research on New Words and Phrases during the Republic of China Period. Jinan, Shandong: Shandong University Press, 2014.
[11]Z. R. Song, 100 Year Chinese New Words and Phrases Dictionary, Shanghai: Shanghai Lexicographical Publishing House, 2014.
[12]Z. F. Luo, The Great Chinese Dictionary, Shanghai: Chinese Dictionary Publishing House, 1992.