Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **6**, November **2018**

A New Chinese Word Segmentation Method Based on Maximum Matching

Yue Zhao[1], Hang Li[1,*], Shoulin Yin[1] and Yang Sun[1]

[1]Software College
Shenyang Normal University
No.253, HuangHe Bei Street, HuangGu District, Shenyang, P.C 110034 - China
1151971241@qq.com;lihangsoft@163.com;ysl352720214@163.com;17247613@qq.com
*Corresponding author: lihangsoft@163.com

Received April, 2018; revised November 2018

Abstract. *Automatic Chinese word segmentation is a hot issue in information extraction, machine translation, information retrieval, automatic text categorization, speech recognition, and the voice conversion, natural language understanding. English uses a space as a nature delimiter, but because Chinese inherits from the ancient Chinese tradition, there is no space between words. Only in Chinese words, sentences and paragraphs, it can be demarcated by obvious delimiter, only words have no formalized delimiter, although the English phrases exist this problem. However, Chinese unique composition determines the Chinese is far more complicated than English. So in this paper, we propose a new Chinese word segmentation method based on maximum matching. Experiments show that the new method has a better segmentation effectiveness than other methods.*
Keywords: Chinese word segmentation, Information extraction, Maximum matching

1. **Introduction.** Chinese word segmentation technology [1] has a widely applications, such as search engine Chinese proofreading, post-processing of Chinese speech recognition, machine translation, full text retrieval of Chinese document library, Chinese text retrieval, Chinese character recognition, Chinese character keyboard input and traditional Chinese characters conversion [2-4]. Chinese word segmentation technology is also on its importance. For any applications, Chinese word segmentation technology is the most basic technology, its accuracy for each application domain has the decisive influence of considerable. But if the segmentation speed is too slow, even if the segmentation accuracy is very high, it cannot be accepted. So the accuracy and speed of the Chinese word segmentation algorithm should achieve the high demands.

Because of the complexity of the Chinese grammar and semantics, even if Chinese word segmentation algorithm is more accurate, it still cannot avoid the appearance of some ambiguity fields. And with the development of the Chinese language, all areas of professional vocabulary and new words appear in life can cause identification error problem of the new words in the process of Chinese word segmentation. Chinese word segmentation, i.e., a Chinese character sequence is segmented into words according to certain rules. For example, 'I am a student', the Chinese word it can be segmented for 'I/am/a/student'. Chinese word segmentation's difficulty mainly is embodied in two aspects: linguistics and computer science.

Difficulties in linguistics. 1) There is no uniform definition of the word. At present, there is no formal definition of words in the language field for the common recognition and strict

unification. Further research is needed in this regard. 2) The Chinese word segmentation has not yet formed an accepted participle standard. Computer, just like persons, is also facing the same problem. The same text may vary from person to person, and it is very likely that the result of different participles will be segmented. 3) Word's specific methods remain difficulty though now they had proposed some word segmentation rules, but the reality is not so simple, true text tends to be diverse and complex to obtain participle reason. To achieve better effect, it should consider participle specification participle word structure and the key factors such as true real corpus. Difficulties in computer science. 1) It is difficult to find a reasonable model of natural language situation. 2) How to effectively use and represent the grammar and semantic knowledge required for participles. 3) How to understand and formalize semantics.

Currently, in terms of Chinese word segmentation algorithm, computer science and language academic experts have invested a lot of research works, and taken breakthrough in many ways. For example, Chang [5] demonstrated that optimizing segmentation for an existing segmentation standard did not always yield better MT performance. He found that other factors such as segmentation consistency and granularity of Chinese words could be more important for machine translation. Cai [6] proposed a novel neural framework which thoroughly eliminated context windows and could utilize complete segmentation history. This model employed a gated combination neural network over characters to produce distributed representations of word candidates, which were then given to a long short-term memory (LSTM) language scoring model. Chen [7] proposed adversarial multi-criteria learning for Chinese word segmentation by integrating shared knowledge from multiple heterogeneous segmentation criteria. Qian [8] proposed a new psychometric-inspired evaluation metric for Chinese word segmentation, which addressed to balance the very skewed word distribution at different levels of difficulty. The performance on a real evaluation showed that the proposed metric gave more reasonable and distinguishable scores and correlated well with human judgement.

But the existing methods on the segmentation results are not ideal. In the specification of Chinese, automatic segmentation algorithm, segmentation ambiguity manage, natural language understanding and artificial intelligence,there many difficulties need to overcome. As a result, word segmentation rate and word segmentation algorithm is easy to implement, which becomes quite critical. So we propose our new segmentation method in this paper. The structure of this paper is as follows. Section 2 provides the preliminaries, such as maximum matching, forward maximum matching and reverse maximum matching. The proposed method is presented in section 3. To illustrate the effectiveness of our new method, experiments are shown in section4, and results analysis are given in this section. Finally, a conclusion of this paper is given in section 5.

2. **Preliminaries.**

2.1. **Maximum Matching.**

Maximum matching algorithm mainly includes: forward maximum matching algorithm, reverse maximum matching algorithm, and the bidirectional matching algorithm [9-11]. Their main principles are to cut and separate words, then compare with the word library. If it is a word, then it is cut out, otherwise, by increasing or reducing a word, then going again, until leave a word. If the word string cannot be cut, it can be as the unknown word to process.

2.2. **Forward Maximum Matching.**

In a computer, it stores a known word list, this word list is also called below table. Assuming that the longest word in the word segmentation dictionary has i Chinese characters, the top i characters in current word character are as matching fields, and it looks for dictionary. If in the dictionary, there is a such i

word, it is matching. as a word segmentation If the dictionary can't find such a word, it will match fails.Then the last word in the character will be deleted, until the match is successful. The rest of the string length is zero. Figure1 is the process of forward maximum matching.

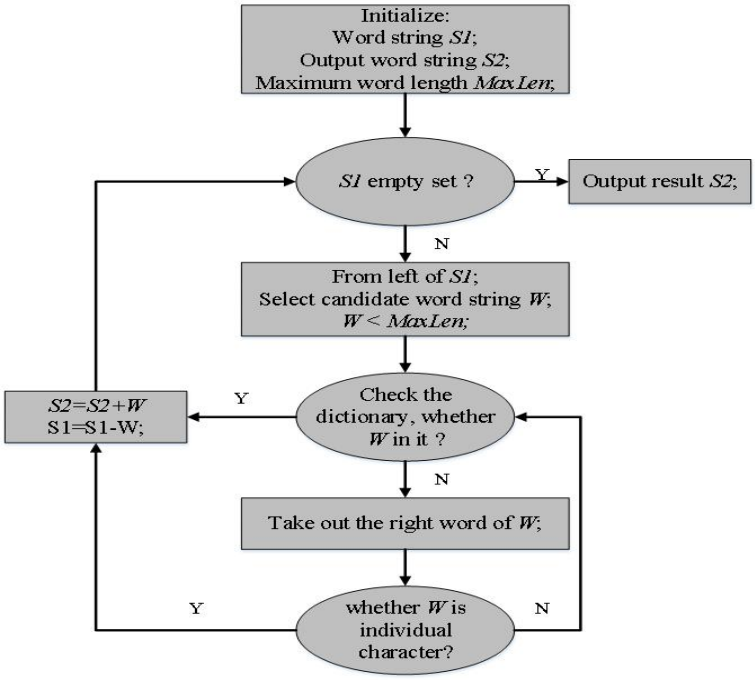

FIGURE 1. The process of forward maximum matching.

2.3. **Reverse Maximum Matching.** This matching method is likely to forward maximum matching. However, the difference is that the direction of cut word is reverse. Reverse maximum matching's cut word direction is from left to right and to match character word.

3. New Maximum Matching for word segmentation.

3.1. **Maximum Matching Analysis.** Although this method is easily to implement. There are still some disadvantage. Error rate of forward maximum matching is about 0.58% and that of reverse maximum matching is about 0.41%. Although the results show that the error rate is relatively low, the error cannot be ignored, it will directly affect the lexical analysis, the final segmentation result is inaccurate. And second, the reverse method of cutting word accuracy is higher than forward method. However, the reverse method requires the configuration of the reverse diction dictionary, which is not compatible with people's language habits, and it is not convenient to modify and maintain.

Forward maximum matching algorithm is from left to right to match characters in the text, if the match is successful, it will split out a word, but there is a problem: to achieve maximum matching, matching is not the first time to cut. The length of big words longer is difficult to determine, if it is too long, then the match will take amount of time, the time complexity of the algorithm is increased obviously. If it is too short, it cannot correctly segment the longer word, which results in the decrease of segmentation accuracy.

Maximum matching algorithm belongs to the mechanical word segmentation algorithm, which cannot solve the problem of two kinds of ambiguity, one is intersection ambiguity, the other is combination ambiguity. They are the main reasons for the effect the precision

of mechanical word segmentation system. Another drawback of mechanical participles is the inability to synthesise new words that do not exist in a dictionary.

3.2. **Dictionary structure design.** The word segmentation method based on dictionary is simple, easily to implement and cut. Because it largely design a better dictionary structure. A dictionary is the basis for each word, and the number of words that can be divided depends on the size of the dictionary. The structure of the dictionary is designed by combining the word segmentation algorithm.

Definition 1. A set of terms that begin with the same word and have the same length (the same number of Chinese characters in the entry) are called phrases.

Definition 2. A dictionary is made up of many phrases. A string always match with the same Chinese characters at the beginning of the longest words. If matching is not successful, then it will match long term. If the string match all entries with the same acronym, it will waste a lot of time. When the match is not successful, a word is deleted from the end of the string to match the dictionary, and the decline length is not always 1, or it may waste time. So it can save the entire dictionary in an ordered list by saving the same first word and the same length to a group. When a string is matched to a dictionary, it only matches the entries in a group with the same initial and the same length, and the number of matches is greatly reduced. Therefore, this paper designs a dictionary storage structure suitable for the new algorithm. First, the entries are arranged in alphabetical order; second, the words that begin with the same word are arranged in length from long to short according to the length of each word. It reads the dictionary into memory before the word segmentation. The structure of the dictionary in memory is shown in figure 2.

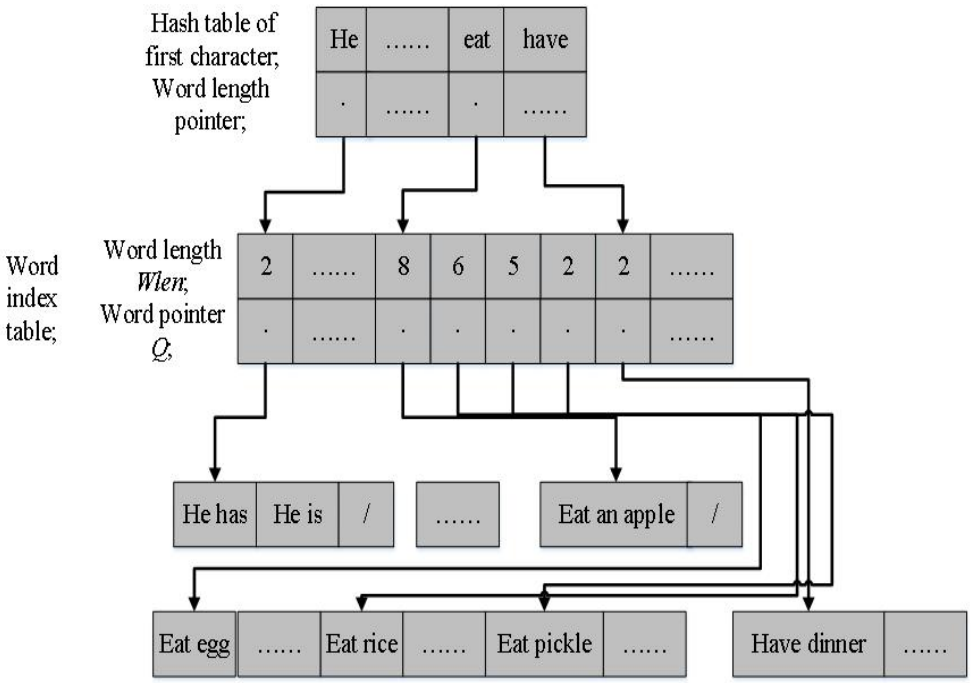

FIGURE 2. Data structure of dictionary.

The dictionary is composed of three aspects:

- Hash table of first character. Internal code is the existence form Chinese characters in computer. The internal code will be transferred as zone bit code, then the zone bit code is converted to a decimal number, this number is the serial number in a

Hash table. The serial number indicates the entry address of the word length list with the word as the first word.

Calculation of this entry address is as follows:

$$offset = (ch1 - 0xB0) \times 94 + (ch2 - 0xA1). \quad (1)$$

Where $offset$ is the serial number in Hash table. $ch1$ and $ch2$ are the high byte and low byte of this word's machine code, respectively.

- Word index table. Word length records the item length with the same first word. For example, word "eat" has four words length of 8, 6, 5, 2. The word pointer indicates the first address of an entry with a certain word.
- Dictionary main text. It stores entries in groups.

3.3. **Process of New Maximum Matching.** In this paper, we adopt the ideas of group to modify maximum matching algorithm. Based on the binary word by word segmentation dictionary mechanism, it puts all same acronym in one group. In the process of search queries, it can query word by word to narrow range. Finally, it looks for the word only within the scope of the same acronym. To further achieve query, it can also narrow the range of the group. This method is mainly to group the same set of all the entries in the acronym with the different word. In this case, the group has the same acronym and the same entry word length. Then the search range will become small.

Group puts the same acronym in the dictionary text, and the same long term word into one group. In a big group, the same acronym to line up with the word long from short to long, will begin with a long word to display in the the Hash table. Each word index table can be tagged separately according to the different length, thus it will further separate the same acronym dictionary of the text of the entries according to the different word long into n groups.

When cut the word, according to the head characters, it finds the word begin with the head word. And then by long term according to the order of the long and short retrieval thesaurus, and then it stays words again other contrast, if the match is successful, it is a word segmentation, then continue to match the next word; If the match fails, then the matching string reduces a word from the tail to rematch; The word segmentation process until the Chinese character string match is completed.

Detail processes of this new method.

- While(there is next word) // if there is another word, it continues to cut words
- Search(the first Chinese character in current word)
- // search the first word group with the first Chinese character
- For ($i := lengthofcurrentword$; $i > 0; i--$)
- // search the word library according to the length of word from long to short
- If match(current word)// successful
- Split(current word);// cut a word
- Get(next term);// cut next term
- Break; Else
- Length of current word-1;

4. **Experiments and analysis.** On the basis of word segmentation algorithm, this paper develops a matching participle procedure based on Delphi platform. Using the web spiders crawling to catch nearly 50 piece of network news as a participle text, and using the annotation of the People's Daily corpus by Peking University in January 1998. They are for reference according to the method described in this article structure thesaurus to construct a new dictionary, and as the article word segmentation dictionary. At the

same time, the segment word of People's Daily corpus by person in January 1998 as a result of the test corpus. First of all, through the participle program running, it handles participle text segmentation. Then word segmentation results will be made a comparison with prepared test corpus to check word segmentation accuracy results. The results show that it reaches to 91.7%, word segmentation is faster too. But the results have nearly 8.3% of the error. The main reason is that the text participle exists a certain amount of new words and dictionaries is not perfect enough. The word segmentation algorithm is based on the dictionary and corpus to match word segmentation, which were not included in dictionary of new words, which creates a mismatch and identification, visible on the thesaurus construction precision.

Operation efficiency and the accuracy of results are the important measurement for word segmentation algorithm. With the same other conditions, if operation efficiency and accuracy are higher, it shows that the performance of this segmentation algorithm is better, and the word segmentation is faster too. Figure3 and figure4 are the comparison results of segmentation efficiency and accuracy with other other three methods PAS [12], FENM [13] and SLR[14].

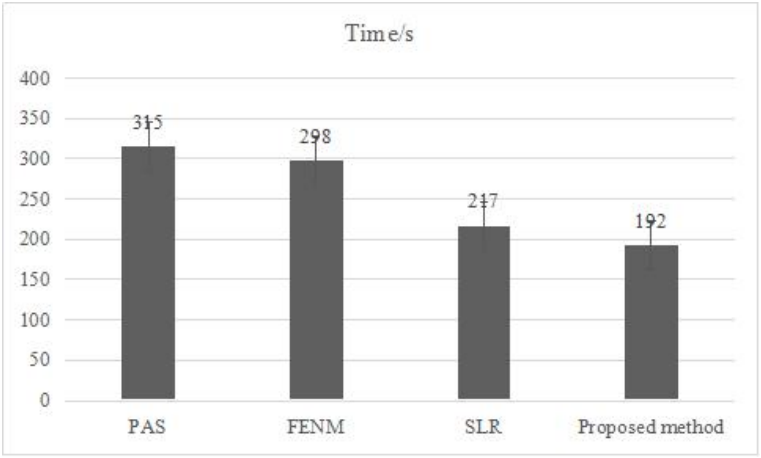

FIGURE 3. Time comparison of different methods.

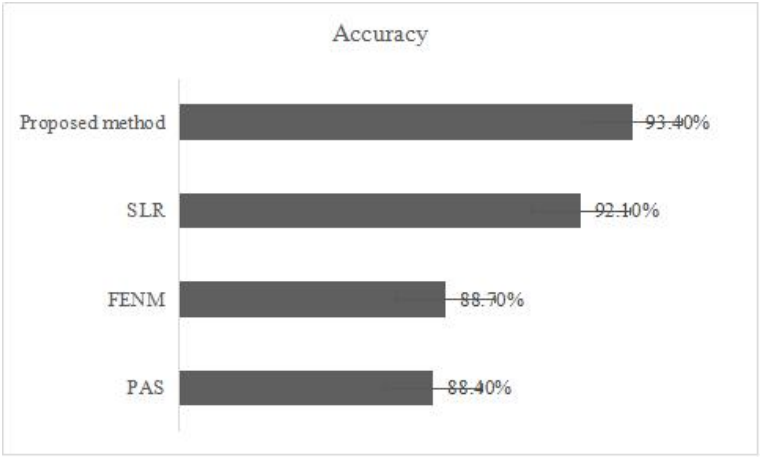

FIGURE 4. Accuracy comparison of different methods.

According to the word segmentation dictionary design in this paper, when it looks for a word, it first detects the starting position on word table based on first word Hash table.

And then according to the corresponding to the first pointer of this word, it finds the starting position of the word in the word index table. According to the maximum word length, it needs to find the word's position in dictionary main text based on the dictionary main text pointer of corresponding word length, then it can locate the required word. if it appears to finding failure in any of these steps, namely, it does not find the matching word. Then it stops the next steps and executes the next search. This can reduce the searching time, increase the speed of segmentation, and then the efficiency of word segmentation is improved.

5. **Conclusions.** In order to improve cutting word efficiency, and segmentation ability of long term, this paper proposes a specifications established by the thesaurus, and on the basis of word segmentation algorithm. The experiments show that the retrieval efficiency is improved obviously. The ability of cutting long words is stronger. At the same time, the algorithm also has shortcomings. Chinese is a more complicated language, its structure is complex and diverse, its usage is flexible and changeable, and its application is ubiquitous, which determines its huge vocabulary and demands efficiency. In addition, with the rapid development of social economy, a lot of new words often random emerge, these new words must not be included in the thesaurus. It gives a thesaurus updates and maintenance, which has brought the new challenge. Therefore, the establishment of a thesaurus is a huge and difficult project, word library maintenance updates and new words recognition remains to be further research.

Acknowledgment. This study was supported by the Natural Science Fund Project Guidance Plan in Liaoning Province of China (No. 20180520024). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] H. Yang, M. Chen, Z. Zhen, Analysis on Applicability of Common Chinese Word Segmentation Software in Literature Study of Traditional Chinese Medicine Text, *Journal of Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology*, 2017.

[2] Y. Sun, Y. Zhang , Y. Zhang, Chinese Text Proofreading Model of Integration of Error Detection and Error Correction, *Chinese Lexical Semantics. Springer International Publishing*, pp. 376-386, 2016.

[3] S.Liu, Ideographical member identification and extraction method and machine-translation and manual-correction interactive translation method based on ideographical members, 2017.

[4] X. U. Qinghua ,H. Zhao, Q. U. Junyan , et al., Thinking and practice of scientific journals' editing,proofreading and typesetting process, *Journal of Chinese Journal of Scientific & Technical Periodicals* 2016.

[5] P. C. Chang, M. Galley, C. D. Manning Optimizing Chinese Word Segmentation for Machine Translation Performance, *Journal of Philosophy*, pp. 224-232, 2016.

[6] D. Cai, H. Zha, Neural Word Segmentation Learning for Chinese, pp. 409-420 , 2016.

[7] X. Chen, Z. Shi , X. Qiu, et al. Adversarial Multi-Criteria Learning for Chinese Word Segmentation, *arXiv:1704.07556* , pp. 1193-1203, 2017.

[8] P. Qian, X. Qiu, X. Huang, A New Psychometric-inspired Evaluation Metric for Chinese Word Segmentation, *Meeting of the Association for Computational Linguistics.,* pp. 2185-2194, 2016.

[9] K. Abugharbieh, A. Balabanyan, A. Durgaryan , et al. Line-impedance matching and signal conditioning capabilities for high-speed feed-forward voltage-mode transmit drivers, *Journal of Microelectronics Journal*, vol. 55, pp. 26-39, 2016.

[10] L. Zhang , Y. Li, J. Meng, Design of Chinese Word Segmentation System Based on Improved Chinese Converse Dictionary and Reverse Maximum Matching Algorithm, *Web Information Systems - WISE 2006 Workshops, WISE 2006 International Workshops, Wuhan, China, October 23-26, 2006, Proceedings. DBLP*, pp. 171-181, 2006.

[11] I. Hussain, M. Zubair, J. Ahmed, et al. Bidirectional exact pattern matching algorithm, *International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science. IEEE*, pp. 293-293, 2010.

[12] N. Ma, Y. Li, X. He, et al., Practical approach of word segmentation in poor resource situation, *Journal of Application Research of Computers*, 2016.

[13] X. Chen, X. Qiu,X. Huang A Feature-Enriched Neural Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 3960-3966, 2016.

[14] T. Daikoku, Y. Yatomi, M. Yumoto, Statistical learning of an auditory sequence and reorganization of acquired knowledge: A time course of word segmentation and ordering, *Journal of Neuropsychologia*, vol. 95, pp. 1-10, 2016.