Research on Longest Backward Segmentation for Context

Weichun Huang
East China Jiao Tong University
Nanchang, Jiangxi, China
e-mail: hwc1968@163.com

Jianjian Jiang
East China Jiao Tong University
Nanchang, Jiangxi, China
e-mail: 870798240@qq.com

***Abstract*—Segmentation is the basis of Chinese information processing. This paper presents us the algorithm about longest backward context for Chinese word segmentation, this method does not require labeling or semantic information, also not mechanically segment words, but automatically find word errors by real-time backtracking. Compared with simple matching algorithm, this method enhances the accuracy, adaptability and reduces some redundancy.**

Keywords-longest backward algorithm; Chinese word segmentation; matching algorithm

I. INTRODUCTION

In recent years, the research of intrusion detection system has gradually become one of the core research spots in the field[1]. With the widespread use of network technology, many emerging businesses on the Internet, such as online shopping and online banking as well as the construction of private networks, such as banking and financial networks etc, making network and information system's security and confidentiality issues seem increasingly important.

Chinese word segmentation is a very important and basic work in Chinese information processing. It is not only the basis of the automatic answering system, machine translation, automatic abstract system and semantic annotation systems based Chinese processing technology, but also the key of Chinese search engine technology. Chinese word segmentation technology has been studied for more than twenty years, but is still a bottleneck of the development of Chinese information processing theory. The reason is that Chinese is different from the Latin language. Between the word and the word, Latin language itself has a very clear dividing line (for example: spaces, punctuation marks, etc.), but for the Chinese that words have no dividing line between them. Written in a continuous manner coupled with the complexity of the Chinese itself, it has caused great difficulties.

This paper firstly analyzes the main technological in the field of Chinese word segmentation, that is maximum matching method based on string matching. On the basis of existing segmentation techniques, this paper adopts the unique advantage of dictionary matching algorithm that needs no corpus training in Chinese segmentation and constructs a system based on longest backward context segmentation. The experiment shows that the system improves its accuracy.

II. METHOD STUDY

A. Maximum Matching Method

Maximum matching Method is split from the left of the text, look for longest word included in the text and then segment them. For the text α, assuming that there is a string $ABCD \in W$, $AB \in W$ and $ABC \in W$. According to maximum matching method, the length of ABC is 3, the length of AB is 2, then we select ABC as a word in the text, the remaining D is selected as another word, so this text α is constituted by ABC and D. Maximum matching method has been described in detail below:

Step 1. Make $S = S$ as the initial state, the initial state is an empty set.

Step 2. Make $S = S \mid \alpha \mid \alpha\beta \mid \alpha\beta\delta$, $\alpha, \alpha\beta, \alpha\beta\delta$ in turn belong to adjacent characters in the text, these characters can be Chinese characters or foreign characters or number symbols and other characters. This step selects string segmented, to prepare for the next step.

Step 3. Look up $S = S \mid \alpha \mid \alpha\beta \mid \alpha\beta\delta$ in the dictionary, if searched successfully and the string is longer than the maximum word dictionary and mark it as a successful state ,then skip to Step 2 and continue. If searched unsuccessfully and it is shorter than the maximum word, then skip to Step 2 and continue. If searched unsuccessfully and the string is longer than the maximum word, then skip to Step 4.

Step 4. Make S backspace a string, that is $S = S^*$, S^* is previously successful state of S. Output S as the segmentation result, then S is the result of the text segmentation.

Step 5. Move the string pointer to the right, make S as an empty state, then skip to Step 1 and start a new search.

Figure 1 is the flow chart of Maximum Matching Method. Seen from the figure, the output is the final segmentation as it is not empty in step 4.When the algorithm starts from a successive addition of one word to form a new string, if the string is longer than the largest length in the dictionary, it indicates that the string is not an existing word, then we need make further analysis and processing.

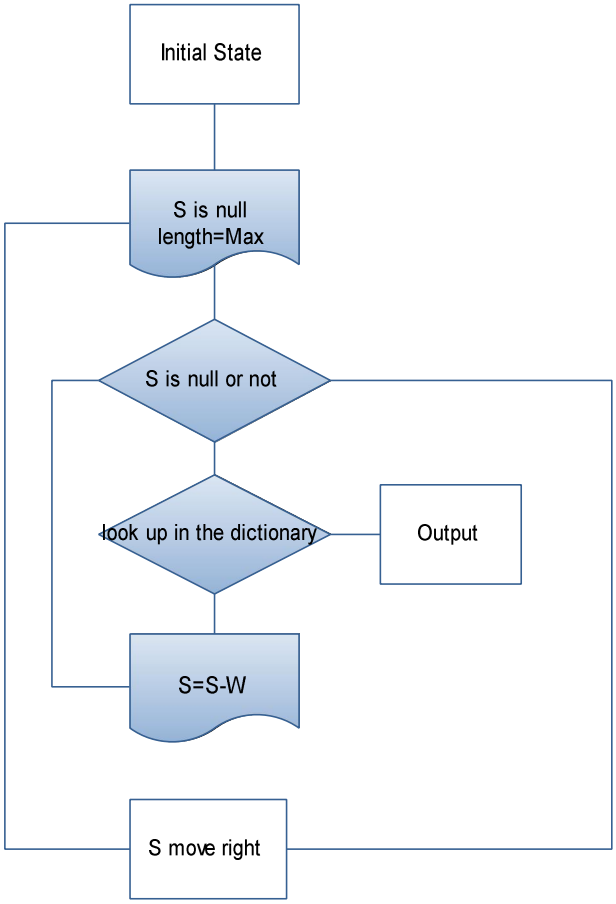

Figure. 1 Maximum Matching Method

B. *Dictionary Structure Design*

The dictionary used by algorithms is constituted by more than 100 million records. Faced with such a large number of data, string comparison method has low efficiency, thus adopt the method with hash to organize these words. It can be divided into two steps: hash storage and hash lookup. In the word hash storage, first make initial value $L = 0$, then get the number of bytes that the word takes up, according to the corresponding ASCII code for each byte to shift, finally obtained after the shift is the corresponding storage hash value. In the calculation of hash value, it will inevitably make conflict. when hash value of the two words make conflict, because chain address algorithm can resolve the conflict, post-order word after the conflict is linked to the back of the word.

The process of hash lookup is similar to that of hash storage. First according to the given hash function to calculate the corresponding hash value, and then following the given hash value to look up if the value exists or not. In order to improve search efficiency, we use bisearch method. From the principle of the method we can see, first arrange the above hash value, then let three pointers respectively point to the front, middle and end of hash sequence. By comparing with the value the middle point we can determine where the range hash value belongs to, then make sequential recursion until search successfully, we find that the word does not appear in the dictionary. Hash code is as follows:

```
for I := 1 to Length(Key) do
Result := ((Result shl 2) or (Result shr (SizeOf(Result) *
8 - 2))) xor Ord(Key[I]);
```

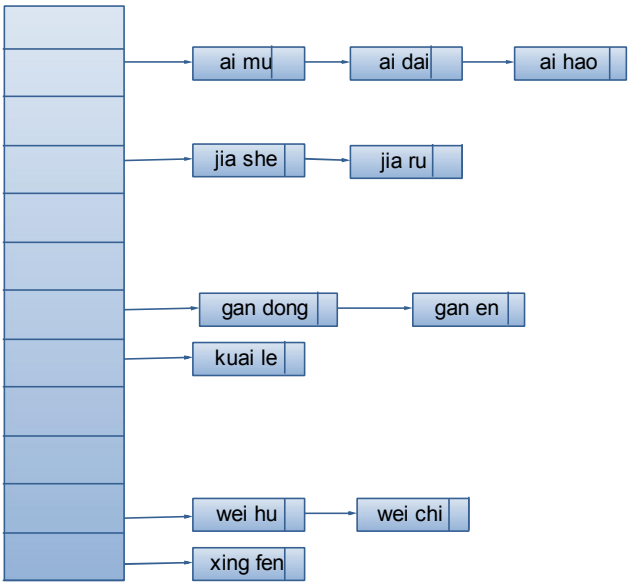

Figure. 2 The Scheme of Chain Address Algorithm

C. *Longest Backward Segmentation Model*

Longest backward segmentation algorithm integrates longest segmentation algorithm with Markov model. In the process of segmentation, it takes into account the probability of different combinations between words that will affect the segmentation for word, we call it segmentation factor referred to here as splitting factor. When there is a word can integrate either with the previous word or the back word, activate the factor, the factor can determine segmentation form.

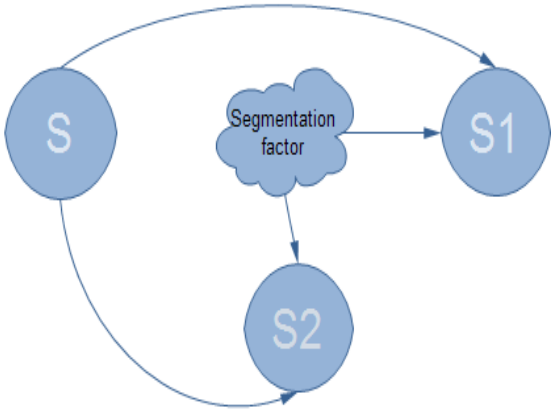

Figure. 3 Longest Backward Segmentation Model

In the figure, S is the original segmentation state, S1, S2 is two different states, for the stability of two states is judged by the segmentation factor.

Longest backward segmentation algorithm makes longest segmentation for word, when faced with the words having two more meanings, then judge them by Markov. For example, there is a text "zu he cheng gong fen qi." Possible segmentation results are as follows:

Result 1: zu he/cheng/gong fen qi

Result 2: zu he/cheng gong/fen qi

Two segmentation results of the text are given above, they describe completely different semantics, we can find that segmentation for the "功" determines the orientation of text . Here introduces two definitions:

Algorithm first pre-process the given text, by removing some stopping-words from the text to reduce the noise to the text thus to prepare for the next segmentation. If the segmentation for words is located in the starting position of the text, then there is no need to go back but directly make maximum segmentation. Between the first segmentation and the second segmentation, choose the base word to make reverse segmentation, we call it backward-process. If no match with this word is found, suffix frequency is zero, so you only need to continue making maximum segmentation forward; if matched successfully and suffix frequency is no longer a zero-that is $p(E) \neq 0$, calculate the Markov values.

$$p(E) = p(e, W_i) w_k / p(W_1) \quad (3\text{-}6)$$

Which is the word as a suffix corresponding to the probability for the corresponding term weights. Access to the frequency of its corresponding ending after the reverse for maximum word-based segmentation to identify its most positive word segmentation is needed to determine at this time, the word is given first frequency, to calculate the corresponding Markov value, obtain the following equation.

$$p(B) = p(b, W_i) w_k / p(W_i) \quad (3\text{-}7)$$

It can give segmentation arbitration results based on the corresponding Markov state transition probability. Markov state transition probability regards segmentation as the transition from one state to another state of the transfer. Combined with context to determine a steady state, It is the final segmentation results. In this paper, classification dictionary will be divided into two categories, general-purpose dictionary and professional dictionary. In accordance with specific major professional dictionary establishes their own dictionaries, general-purpose dictionary is the language used by the people, including some common words, idioms and so on. In segmentation algorithm, two types of dictionaries correspond to different weights. In accordance with the principle of professional thesaurus priority, if there is professional vocabulary, allocate it to professional dictionary, so its weight is greater than that in the general-purpose dictionary.

There is an existing text "yi chuan chuan lian de cai deng", according to the segmentation algorithm given in this section , after the first segmentation, the segmentation result is as follows: "yi chuan chuan/lian de cai deng". At this point, "lian" is regarded as the base word, reversely back to get the word "chuan lian", and maximum segmentation is empty, then the first segmentation result needs to be corrected, getting a new result "yi chuan/chuan lian/de cai deng".

Figure 4 shows the segmentation arbitration map of the sentence" yi chuan chuan lian de cai deng ". This algorithm corrects the segmentation result as " yi chuan/chuan lian/de cai deng ", reducing the possibility of ambiguity .

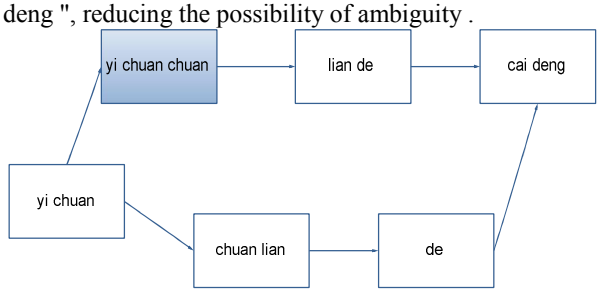

Figure.4 The Scheme of Segmentation Arbitration

D. The realization of Longest Backward Segmentation

The algorithm first makes the initial segmentation with the given text, and then use reverse-back method, begin from the base word tentatively to find words reversely while continuing positive segmentation. When there is a conflict, it gives steady-state on the basis of Markov state transition probability, the arbitration result is given.

To achieve the segmentation with a document, the following four stages is as follow:

1)Pre-processing stage. Before segmenting with a text, you need to preprocess the text that refers to the pre-filter, filter out some disturbing sub-factor. The purpose of preprocessing is to provide a system containing only Chinese characters. When the system segments with the information, segmentation is required in the original text. Sentence will usually contain a number of particle, punctuation, all these need to be handled in the pre-processing stage.

2)Dictionary structure stage. Loaded the dictionary into memory, put entries in the hash table and sort it, so use binary search method according to the dictionary, greatly improve the efficiency of the algorithm.

3)Segmentation stage. If there are English words in the document, classifies these English words into corresponding strings, so are figures. There are the only Chinese strings in the whole document, we make segmentation adopting the longest backward segmentation algorithm.

4)Post-processing stage. After the segmentation, the algorithm will find some words that appear not in the dictionary, these are unknown words. At the end, this algorithm adds these unknown words automatically to the dictionary.

III. Application Study and Experimental Evaluation

TABLE I. Contrast of precision and recall by longest backward segmentation and maximum matching method

Method	Precision(%)	Recall(%)
Maximum Matching Method	87.6%	87.2%
Longest Backward Segmentation	89.6%	89%

From the above experiments, it show that the precision and recall of Longest Backward Segmentation is higher that of MMM and recall rate improves by nearly 2%.

The reason is that the longest backtracking model make retrospective segmentation, and the maximum matching method doesn't. For example: under the premise of failing to recall the correct results, recall of maximum matching method 87.2% while longest backward segmentation for context is 89%.

IV. Summary and Future Prospects

This paper proposes longest backward segmentation algorithm. In accordance with the principle of -specific dictionary priority, it arbitrates to the word with ambiguity by Markov state transition to further support for the arbitration. The significance of this research is to conduct basic research for Jiangxi technology project detection system, the algorithm can deal with unknown words. There is still much room for improvement in dealing with ambiguity.

References

[1] T.F. Smith and M.S. Waterman, Identification of common molecular sub-sequences[J]. Journal of Molecular Biology, 147:195-197, 1981.

[2] Bao, J. P., J. Y. Shen, X. D. Liu, and H. Y. Liu, A fast document copydetection model[J]. Soft Computing 10, 41–46, 2006.

[3] C. J. Neill and G. Shanmuganthan, A Web-Enabled Plagiarism Detection Tool[C]. IT Professional, pp, 19-23, September, 2004.

[4] Juan MET, Pedro G T, and Jesus EDVl, "Anomaly detection methods in wired networks", A survey and taxonomy, Computer Communication, 2004, pp.1569-1584.

[5] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval[M]. McGraw-Hill Book Company, 1983.

[6] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, A Large-Scale Study of the Evolution of Web Pages[C]. Proceedings of the 12nd International World Wide Web Conference, pp, 669-678, May, 2003

[7] Arslan Abdullah N.An algorithm for string edit distance allowing substring reversals[C]. Sixth IEEE Symposium on BioInformatics and BioEngineering, 2006, Washington D.C., USA, pp, 220-226, 2006.

[8] Li Yujian Pliu Bo.A Normalized Levenshtein Distance Metric[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,pp 29(6):1091-1095, 2007.

[9] Xia Tian,An Edit Distance Algorithm with Block Swap[C]. The 9th International Conference

[10] T.F. Smith and M.S. Waterman, Identification of common molecular sub-sequences[J] Journal of Molecular Biology, 147:195-197, 1981.