Journal of applied science and engineering innovation Vol.1 No.3 2014
ISSN (Print): 2331-9062
ISSN (Online): 2331-9070

Research of Reverse Backtracking Matching Algorithm for Chinese Word Segmentation

Yong Liu[1,a], Wei Li[2,b]
[1]Department of Computer Science and Technology, Qingdao University of Science & Technology, Qingdao, Shandong 266061,China
[2]Department of Computer Science and Technology, Qingdao University of Science & Technology, Qingdao, Shandong 266061,China
[a]liuyong0202@sohu.com, [b]liwei_58023@126.com

Key words: Chinese segmentation; overlapping ambiguity; reverse backtracking method

Abstract. With the rapid development of science and technology, we have entered the age of digital information. Search engines have become a preferred tool to find information, and Chinese word segmentation is an important part of the search engines. In order to improve the efficiency of the existing Chinese word segmentation and solve its existence problems of ambiguity, the reverse backtracking algorithm based on the reverse maximum matching method is designed. The algorithm pretreats the text to be split firstly, then use the forward and reverse maximum matching algorithm to discover ambiguities efficiently, finally use the backtracking mechanism to eliminate ambiguity. Experiments show that this method improves the speed and reduces the error rate of the overlapping ambiguity, improves the performance of the reverse maximum matching algorithm.

Introduction

Chinese information processing is one of the important research content of text mining, Chinese search engines, machine translation, and study copy detection are Chinese information processing, and the Chinese word segmentation is one of key technologies in Chinese information processing, especially it has a significant role in the massive information processing. Internet, search engines, text retrieval, computer and electronic products have higher requirements for the Chinese information processing technology, so the speed and accuracy of segmentation is extremely important, it has a greater impact on the efficiency of the system.

The processing of Chinese and English is different, words are separated by spaces in the English text; while Chinese text is a continuous string, in addition to punctuation, and there is no clear division flag between words. In the processing of natural language, the word is the smallest meaningful activity capable of independent language elements [1]. According to the characteristic of the behind partial statement of the Chinese central word, the accuracy of the reverse maximum matching algorithm is often higher than the forward matching algorithm. How to maximize the elimination of ambiguity and improve the accuracy of the segmentation will be a very important impact on the work after the entire text mining [2]. It can disambiguate by transforming the dictionary: hash structure tail word dictionary structure which record the word length [3], long-term priority [4], dual-dictionary mechanism [5]; It can also disambiguate by improving algorithm: the maximum matching algorithm by increasing word to find intersection of the word ambiguity [6].

The classification of Ambiguity

A large number of studies by experts and scholars have shown that ambiguous are primarily divided into two categories: overlapping ambiguity and combinational ambiguity [7].

Overlapping ambiguity. In a field of the ABC, if AB is an independent word, and BC is also an independent (where A, B, C is a string), so you can call the form ABC as overlapping ambiguity

fields. For example,"白天鹅"can cut into "白天/鹅" and "白/天鹅". where A = "白", B = "天", C = "鹅 ".

Combinational ambiguity. In a field of the AB, if AB is an independent word in vocabulary, A is an independent word, and B is also an independent word (where A, B is a string), so you can call the form AB as combinational ambiguity fields. For example: "现在/差/十/分/就/十一点/了" and "“她/十分/欣慰",the segmentation of “十分”in the two kinds of segmentation methods are different.

Literature [8] shows that 90% ambiguities is mainly overlapping ambiguity, so solving the overlapping ambiguity is the key to improve the quality of Chinese word segmentation.

Several algorithm analysis of Chinese word segmentation

In recent years, people have a certain research on Chinese word segmentation; they have made a variety of effective segmentation algorithms. These algorithms can be divided into three categories: segmentation algorithm based on string matching, segmentation algorithm based on statistics and segmentation algorithm based on understanding. We focus on segmentation algorithm based on string matching.

Segmentation algorithm based on statistics: This algorithm only need statistic the information of the word without dividing dictionary, so it is also called non-dictionary algorithms. Segmentation algorithm based on statistics must establish digital statistical model firstly, and calculate the parameter values in the statistical model by training the established corpus. Statistical model has a strong robustness and generality. This method is mainly making use of the joint probability between word and word as segmentation information. The basic idea is that: input each of the string in the corpus, full cut of the string and list all the possible results of segmentation, calculate the occurrence probability of the statistical data which can reflect the characteristics of the language, then select the largest probability from the segmentation results. Currently more common algorithm is the probability and statistics algorithm based on mutual information, combined degree algorithm and N-Gram model algorithm. The advantage of this algorithm is not required to establish artificial rule base, obtain the data required from the corpus through training by the machine automatically; automatic disambiguation effectively; identify the unknown word; solve the limitation of mechanical word segmentation algorithm. Improve the accuracy. The disadvantage is that the sensitivity is low, it often takes some useless phrases; large amount of calculation and the accuracy of segmentation is associated with the choice of training text.

Segmentation algorithm based on understanding: this algorithm is to achieve the effect of the word through the computer simulation of people’s understanding of the sentence. The basic idea is that the segmentation, the syntactic analysis and the semantic analysis are done at the same time, and make use of syntactic and semantic information to handle ambiguity. It usually consists of three parts: the segmentation subsystem, syntactic and semantic subsystems and total control section. The advantage of this algorithm is that it can be carried by the instance of reasoning and proof, and can add unknown word automatically. The disadvantage is that it requires a lot of Chinese knowledge to make the algorithm difficult to achieve.

Segmentation algorithm based on string matching is also called mechanical segmentation method, which matches the Chinese string stay with the entry in a “big” machine dictionary for distribution according to a certain strategy, if a string is found in the dictionary, then the match is successful (identify a word).Scanning the string from front to back is called forward match , scanning the string from back to front is called reverse matching; matching following long-term priority is called the maximum matching, otherwise called the minimum match; it also can be combined with the process of speech tagging in the matching process, which is called integrated approach combined with segmentation and labeling. If the length of matching and the direction of scanning combined, there are four kinds of matching methods: forward maximum matching, forward minimum matching, reverse maximum matching and reverse minimum matching.

Because of the characteristic of the Chinese, it usually use the maximum matching, namely

forward maximum matching and reverse maximum matching. As a result of long-term priorities and maximum matching algorithm, it splits the text forward or reverse in accordance with sub-word dictionary, if somewhere there can be a word in two cases, it will not be able to more accurately identify ambiguity and handle it, that is to say, it will not be able to identify and split ambiguous strings effectively. The error rate of using the forward maximum matching merely is 1/169, the error rate of using the reverse maximum matching is 1/245, so in order to solve the situation of the low accuracy of ambiguity of the current segmentation based on string matching algorithm, and we propose a reverse with backtracking Chinese word segmentation algorithm. It adds a backtracking mechanism to the reverse maximum matching segmentation method, which takes bidirectional consideration in the course of reverse matching, not only matches the word backward, but also analysis the next step to ensure that the result of the last matching process does not affect this matching.

The design of reverse backtracking matching algorithm

Pretreatment for segmentation of text. Chinese word segmentation is a complex process, the text to deal with the segmentation may be relatively large, we can not use word segmentation algorithm directly to split without any pretreatment conditions, because it is very easy to produce mismatch error. Since we are using the word segmentation based on string matching, the longer the sub-word string is, the greater the probability of error is. If a long string is divided into a relatively short word strings to segmentation in no error conditions, then the error rate will be greatly reduced, and it can improve the accuracy and efficiency of matching. After the analysis, if you first cut out all the segmentation of the text and some vocabulary words before using a algorithm based on string matching to cut the text, then cut the rest of the relatively short strings for segmentation, we can ensure that we will improve the accuracy of the segmentation.

The design of reverse backtracking matching algorithm. Reverse maximum matching algorithm idea is: assume that the maximum length of the words in the dictionary is N, which has up to N characters, so take the first N characters in the content as an entry, then search the word dictionary, if there is such a dictionary entry, then the match is successful, this entry must be segmented; if you have not found such an entry when you scan the whole word dictionaries, the match fails. You need to reduce the length of matches, remove the top of the character in this entry, and then follow the above method to match, so go on and on until the match is successful.

The complexity of the algorithm is relatively small; technology is relatively easy, it only need to build vocabulary. But its ambiguity identification is relatively poor; the segmentation accuracy is not high. So the backtracking reverse segmentation algorithm is proposed.

Reverse backtracking algorithm pre-processes the word of the text in the first place; replace the punctuation of the text with spaces, the text is divided into shorter string. It no longer need to cut the string that the length is less than 2.The design idea of the algorithm is:

(1) Set the maximum length of the dictionary is N, firstly split the first N characters(less then removed all) as a string str1, and finds a match in the dictionary that the length is N. If successful, turn to (3); if the match is failure, turn to (2).

(2) Remove the left character of a word to continue match the dictionary which length is (N-1), and if unsuccessful, repeat the above steps until the match is successful.

(3)Assume that the match string length is n, if n <2 then print str1 directly; Otherwise, take the (n-1) characters of str1 to match the dictionary again. If the match fails, cut the str1 and continue to operate from the beginning; if the match is successful, and then turn to (4).

(4)A Chinese characters str3 contain the string str1 and its preceding character match the dictionary, if the match is successful then cut str2 out, continue from the beginning; if the match fails, cut str1 out, continue from the beginning, repeat the above steps until the segmentation is finished.

Reverse backtracking matching algorithm flow chart is shown in Fig.1:

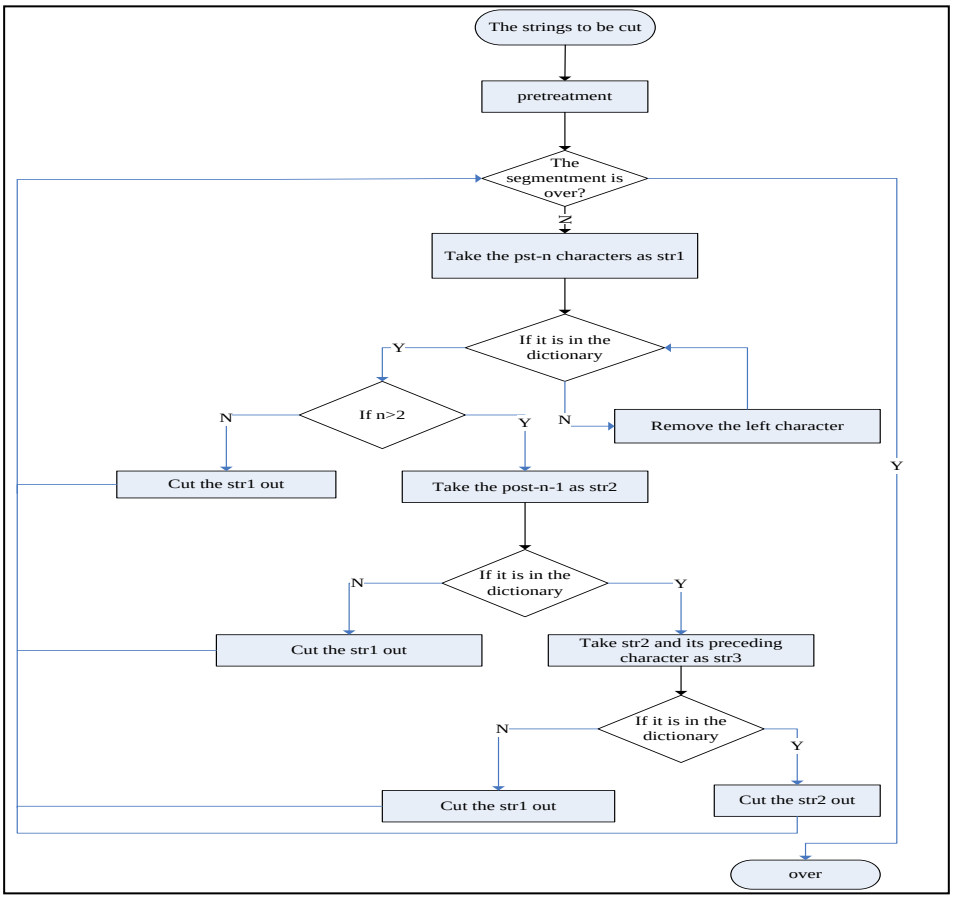

Fig.1: Reverse backtracking matching algorithm flow chart

Analysis of experimental results

The system mainly consists of three modules: the main interface module segmentation module, comparison module.

The main interface of the system shown in Fig.2:

The segmentation module system shown in Fig.3:

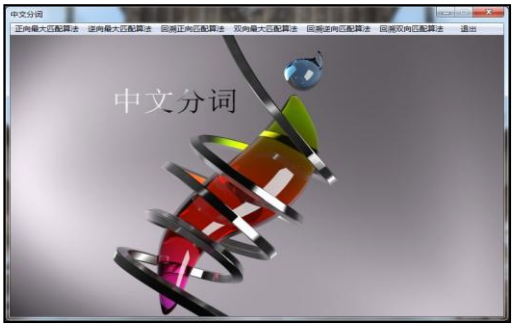

Fig.2: The main interface of the system

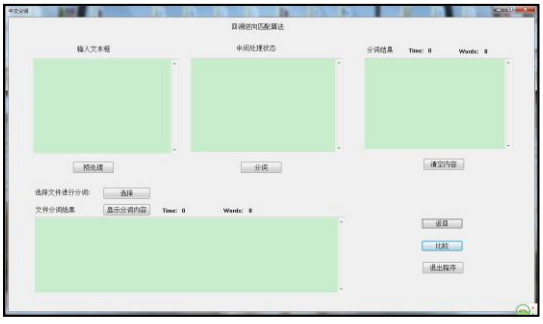

Fig.3: The segmentation module system

Systematic comparison module shown in Fig.4:

Reverse maximum matching algorithm segmentation results shown in Fig.5:

A total of contained words: 265

The total time of segmentation: 858ms

Journal of applied science and engineering innovation Vol.1 No.3 2014
ISSN (Print): 2331-9062
ISSN (Online): 2331-9070

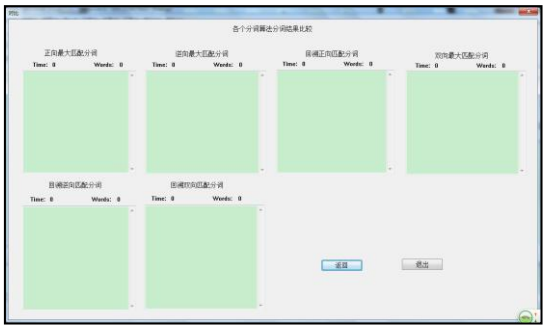

Fig.4: Systematic comparison module

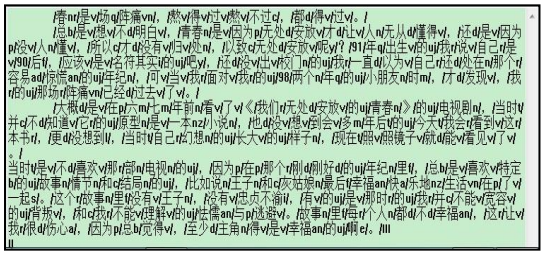

Fig.5: Reverse maximum matching algorithm segmentation results

Reverse backtracking maximum matching algorithm segmentation results shown in Fig.6:

A total of contained words: 265

The total time of segmentation: 796ms

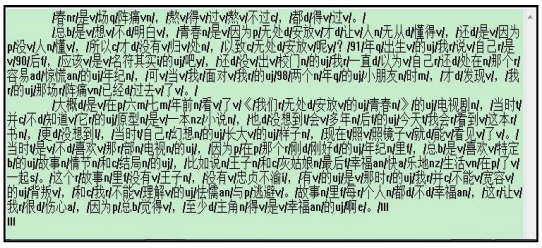

Fig.6: Reverse backtracking maximum matching algorithm segmentation results

Reverse maximum matching method complies with long term priority rule, it is a segmentation based on dictionary, it is characterized by simple and easy to implement, in most cases it can be a good segmentation, but it often can not carry out effective segmentation for the string of ambiguous. Such as"幸福快乐地生活",the "快乐" and "乐地"are ambiguity, reverse maximum matching algorithm for the segmentation results:幸福/快/乐地/生活, this is obviously wrong, while the reverse backtracking maximum matching algorithm can get correct results:幸福/快乐/地/生活.Select randomly real corpus, use the reverse maximum matching algorithm and the reverse backtracking maximum matching algorithm to cut the text, compare some of different parts of the strings, the results are shown in table.1:

Table.1 Reverse maximum matching algorithm and reverse backtracking segmentation results contrast

The results of reverse maximum matching algorithm	**The results reverse backing maximum matching algorithm**
/也 d/没 v/想 v/到会 v/多 m/年后 t/	也 d/没想到 l/会 v/多年 n/后 n/
我会 r/看到 v/这 r/本书 r/	我会 r/看到 v/这本 r/书 n/
在 p/那个 r/刚 d/刚好 d/的 uj/年纪 n/	在 p/那个 r/刚刚 d/好 d/的 uj/年纪 n/
幸福 an/快 a/乐地 nz/生活 vn/	幸福 an/快乐 an/地 an/生活 vn/
故事 n/里 f/每 r/个人 n/都 d/不 d/幸福 an/	故事 n/里 f/每个 r/人 r/都 d/不幸 n/福 n/

Experiments show that a backtracking mechanism is introduced into the reverse backtracking maximum matching segmentation algorithm, so it can effectively eliminate most overlapping ambiguity which may lead by reverse maximum matching algorithm. But also because of the

mechanism, it appears a new error which does not appear in the reverse maximum matching algorithm, from the experimental results of the whole, the error rate of reverse backtracking maximum matching algorithm is much lower than reverse maximum matching algorithm, it effectively improves the accuracy of the segmentation.

Conclusion

This study studies the algorithm of Chinese word segmentation and ambiguity algorithm, it focus on the reverse backing segmentation algorithm, and introduce the backtracking mechanism. So in the entry of the matching process, it not only focuses on the current word matching, but also considers the effect of the match on the follow-match, it is better to analysis and process the ambiguity, it can reduce the possibility of ambiguity.

References

[1]Ling Fang.The key arithmetic for Chinese word segmentation. Master thesis, Beijing University of Posts and Telecommunications.2005.

[2]Zhenguo Ding, Zhuo Zhang, Jing Li. Improvement on reverse directional maximum matching Method based on hash structure for Chinese word segmentation. Computer Engineering and Design.2008, 29: 3208-3211, 3265.

[3]Zhen Liang, YuSheng Li. Reverse backtracking research of Chinese segmentation based on dictionary of Hash structure. Computer Engineering and Design, 2010, 31: 5158-5160.

[4]Zhanxiao Tian, Xian-Zhong Han, Ke-Jian Wang. An improved ambiguity resolution of RMM based on long-term priority, Journal of Agricultural University of Hebei, 2009, 32: 100-102,107.

[5]Ling Li.Design of Chinese Word Segmentation System Based on Dual-dictionary Mechanism. Mechanical Engineering & Automation, 2013, 17-19.

[6]A-ming Hu, Wei-dong Wang. Optimization of Chinese words' ambiguity recognition algorithm. Modern Electronics Technique, 2012, 35: 107-109.

[7]Suzhi Zhang, Fangmei Liu. Research on Chinese Word Segmentation Based on Matrix Restraint. Computer Engineering, 2007, 33: 98-100.

[8]Yong-gang Cao, Yuzhong Cao, Maozhong Jin, Chao Liu. Information Retrieval Oriented Adaptive Chinese Word Segmentation System. Journal of Software, 2006, 17: 356-363.