Universe Scientific Publishing

ORIGINAL ARTICLE

Study on Chinese Word Segmentation

Gaga Mao

Southwest University of Nationality

Abstract: Search engine technology is widely applied currently, which gradually deepens the research of full-text retrieval technology and Chinese word segmentation technology. Chinese word segmentation is one of the key technologies of Chinese languages information, the quality of which directly affects the information processing efficiency of Chinese languages.

Keywords: Chinese Languages; Word Segmentation; Research

In the early 1980s, the first automated Chinese word segmentation system appeared in China. The research work on Chinese word segmentation is mainly in China, and research institutions are mainly domestic university laboratories and research offices, such as Institute of Computing Technology, Chinese Lexical Analysis System(ICTCLAS) of digital research laboratory of Institute of Computing Technology in Chinese Academy of Sciences, word segmentation system of Institute of Computational Linguistics, Microsoft Research SYNAC(The Syntactic Analyzer for Chinese) of Natural Language Institute of Microsoft Research Lab, word segmentation system of colleges and universities like Beijing Normal University and Nanjing University, and currently outstanding commercial software - Hylanda Intelligent Word Segmentation of Hylanda Technology Company. With the deepening of the study on Chinese word segmentation, the accuracy as well as speed of word segmentation are gradually increasing. The accuracy has been increased from 80% to about 99%, and the speed has been increased from hundreds of words per second to hundreds of thousands of words per second. At present, search engines have become an indispensable tool in our daily life, work and study. Common search engines include Baidu search, Google search, Sogou search, 360 search etc. A search engine refers to indexing the collected documents and web pages, establishing an index database where the user can perform full-text retrieval operations by querying keywords. Search engines, as one of the most technologically useful applications on the Internet today, involve very complex technologys, mainly include "word segmentation - index - search". Thus, word segmentation plays a very significant role in search engines. Currently, there are many algorithms for Chinese word segmentation, which is one of the most popular research directions. Above all, this paper studys and analyses the Chinese word segmentation algorithm.

1. Review of the Chinese word segmentation technologies

1.1 Full-text search technology

The so-called full-text search means that the computer indexing program creates an index for each word by scanning each word in the article, to indicate numbers and locations of the word appearing in it. When the user queries, the retrieval program will search them according to the index established in advance, and feed back the results of the lookup to the user. The index item can be a character, a word or a phrase, regarding to whether the word segmentation

Copyright © 2019 Gaga Mao
doi: 10.18686/ahe.v3i3.1393
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

technology is adopted in a document of Chinese words, and thus, the index item can be divided into a character-based full-text index and a word-based full-text one. Character-based full-text indexing is the indexing of each character in an article, and the word is broken down into characters combination when searched. For different languages, characters have different meanings. For example, characters and words in English are actually one, however, there are big differences between characters and words in Chinese. Thus, with this method, it could show a more comprehensive search results, but follows a lower precision. Sometimes there are even ridiculous search results, such as with a feedback of “Marx” when searching for the currency unit “mark”..

Word-based full-text indexing refers to indexing words, that is, semantic units, in the article. It searches by word, and is able to process synonymous items. English and other western languages are similar to word processing because they are segmented according to blanks, and they are also easy to add synonyms. For Chinese languages, syncopated to achieve the purpose of indexing by words are needed. Cutting words in the Chinese characters documents, improving accuracy of the word segmentation, extracting keywords as the index item, and realizing the index by word can greatly improve search accuracy .

1.2 Chinese word segmentation technology

Chinese word segmentation is very different from English word segmentation. For English, a word is a word, while Chinese is a word-based writing unit. There is no obvious label between words, so it needs to be syncopated artificially. The Chinese word segmentation system is a system that uses computers to automatically recognize words in its texts. There are many achievements in its research and many algorithms are presented. According to its characteristics, the existing word segmentation algorithms can be divided into four categories: word segmentation based on string matching, word segmentation based on understanding, word segmentation based on statistics, and word segmentation based on semantics.

2. Chinese word segmentation method

2.1 A word segmentation method based on string matching

This method is also called mechanical word segmentation method, and the dictionary-based word segmentation method. It matches the Chinese character string to be analyzed with the term in a "sufficiently large" machine dictionary according to a certain strategy. If a string is found in the dictionary, the match is successful (a word is recognized). The method has three elements: the word segmentation dictionary, the text scanning order, and the matching principle. The scanning order of the text contains forward scan, reverse scan, and two-way scan. The matching principle mainly are the Maximum match, the minimumt match, the word-by-word match and the best match.

Maximum matching method (MM). The basic idea is: if the number of Chinese characters contained in the longest entry in the automatic word segmentation dictionary is i, then the first i characters in the current string sequence of the processed material are taken as matching fields to find the word segment dictionary. The i word, the match is successful, the matching field is segmented as a word. If the i word is not found in the dictionary, the match fails. Matching field removes the last Chinese character, and remaining characters are used as new matching fields, and then the matching is performed, and then continues until the matching is successful. Statistical result shows that the error rate of this method is 1/169.

2) Reverse maximum matching (RMM).The process of word segmentation in this method is the same as that in MM method. The difference is that it starts from the end of the sentence (or article) and removes the Chinese character in front every time when the matching is not successful. The statistical result shows that the error rate of this method is 1/245.

3) Word by word. The word in the dictionary is searched word by word for the whole material to be processed in decreasing order from long to short until all the words are separated. Whatever the capacity of the segmentation dictionary and how small the material being processed is, the segmentation dictionary must be matched.

4) Establish the segmentation marking method. Segmentation marks have natural and unnatural categories. Natural segmentation marks refer to the non-literal signs appearing in articles, such as punctuation marks. Unnatural marks are the use of affixes and words that do not form words (including monophonic words, polyphonic words and onomatopoeia, etc.). Setting up the segmentation mark method, the first step is to collect a lot of segmentation mark. When making word segmentation, find out the segmentation mark first to divide the sentence into some shorter fields, and then use MM, RMM or other methods for fine processing. This method is not really a word segmentation method, but a pre-processing method of automatic word segmentation, which requires extra time to scan segmentation marks and increase storage space to store those unnatural segmentation marks.

5) Optimum matching method(OM). This method can be divided into forward best matching method and reverse best matching method, which is based on the following point: the words are arranged in the order of word frequency in dictionary. In order to shorten the retrieval time of the word dictionary and achieve the best effect, thid method thereby reducing the time complexity of word segmentation and speeding up the segmentation. In essence, this method is not a purely word segmentation method, and it is just a way of organizing the word segmentation dictionary. The OM method's word segmentation dictionary must have a data item of a specified length in front of each word, so its spatial complexity is increased without having an effect on improving word segmentation accuracy, and time complexity of word segmentation processing is reduced.

2.2 Word segmentation method based on understanding

This method is also called artificial intelligence-based word segmentation method. The basic idea is to perform syntactic and semantic analysis at the same time as word segmentation, and use syntactic information and semantic information to deal with ambiguity. It usually consists of three parts: the word segmentation subsystem, the syntactic and semantic subsystem and the general control part. Under the coordination of the general control part, the word segmentation subsystem can obtain the syntactic and semantic information about words, sentences, etc. to judge the participle ambiguity., which means it simulates the process of human understanding of the sentence. This method of word segmentation requires a large amount of linguistic knowledge and information. At present, the word segmentation methods based on understanding mainly include expert system segmentation and neural network segmentation. Due to the generality and complexity of knowledge of Chinese language, it is difficult to organize various language information into a form that can be directly read by machines. Therefore, the word segmentation system based on understanding is still in the experimental stage.

1) Expert system word segmentation. From the perspective of expert system, the knowledge of word segmentation (including knowledge of common sense segmentation and the ambiguous segmentation rule of disambiguation segmentation), which is the ambiguous segmentation rule, is independent from the inference engine that implements the word segmentation process, so that the maintenance of the knowledge base and the realization of the inference engine do not interfere with each other, making the knowledge base easy to maintain and manage. It also has the ability to discover intersection ambiguity fields, polysemy combination ambiguity fields and certain self-learning capabilities.

2) Neural network word segmentation. This method is to simulate human brain parallel, distributed processing and the establishment of numerical calculation models. It stores the implicit methods of word segmentation knowledge into the neural network, and modifies the internal weights through self-learning and training to achieve the correct word segmentation results. Finally, the neural network automatic segmentation results are given.

3) Neural network expert system integrated word segmentation. This method first starts the neural network to perform word segmentation. When the neural network cannot give accurate segmentation to the newly appearing words, the expert system is activated to analyze and judge, refering to the knowledge base for reasoning. The preliminary analysis is obtained, and the learning mechanism is started to train the neural network. The method can take the advantages of both the neural network and the expert system in a more sufficient way, and further improve the word segmentation efficiency.

2.3 Segmentation method based on statistics

The main idea of the method is that the words are stable combinations. So, in the context, the more the adjacent words appear at the same time, the more likely they are to constitute a word. Therefore, the probability or frequency of occurrence of words adjacent to a word can be better reflected in the reliability of the word. The frequency of the combination of adjacent words appearing in the training text can be counted, and the mutual information between them can be calculated too. The mutual information reflects the closeness of the relationship between Chinese characters. When the tightness degree is above a certain threshold, it can be considered that the block may constitute a word. This method is also known as no-dictionary participle.

The main statistical models used in this method are: N-ary grammar model, Hidden Markov model and Maximum entropy model, *et al*.. In practical application, it is generally combined with the dictionary based segmentation method, which not only gives full play to the characteristics of fast and efficient matching segmentation, but also takes the advantages of dictionary free segmentation combined with context to identify new words and automatically eliminate ambiguity.

3. Conclusion

Due to the uniqueness of Chinese characters, there is no perfect algorithm for Chinese word segmentation. The further improvement of the Chinese word segmentation algorithm should be based on the achievements and comprehensive use of a variety of methods, as well as applications of new models and methods. Through these continuous exploration, Chinese word segmentation algorithm is becoming more and more perfect.

Acknowledgment

The fund project of the research zone on word segmentation of ethnic languages :fund project:The special fund project of central college basic scientific research operating expenses of Southwest Minzu University (2019NQN30). Funded results of the construction project of "national language and character information processing laboratory" of key laboratories of Southwest Minzu University.

References

1. Lin Shen. Research on word segmentation algorithm and improvement of national characters [J]. Computer knowledge and technology 2017; 13(31): 199-200
2. Qiuyun Gan. An overview of word segmentation algorithm for ethnic characters [J]. Journal of Tangshan Normal University 2013; 35(05): 55-57
3. Hongzhi Liu. Research on word segmentation technology of national characters [J]. Computer development and application 2010; 23(03): 1-3