Elimination Method Study of Ambiguous Words in Chinese Automatic Indexing

Wang Dan[1,2], Yang Xiaorong[1,2], and Zhang Jie [1,2]

[1] Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2] Key Laboratory of Agricultural Information Service Technology (2006-2010), Ministry of Agriculture, The People's Republic of China
{wangdan01,yangxiaorong,zhangjie02}@caas.cn

Abstract. Faced with huge amounts of information to realize the accurate retrieval under the network environment, the first step is indexing words cannot appear ambiguity word. Because Chinese's the basic unit is Chinese characters, Chinese characters form words, Word is divided into monosyllabic word and compound word, and there's no space between Chinese keywords and there are a lot of ambiguous concept. Therefore a lot of ambiguity in the indexing process will be produced. The result detected information of irrelevant or mistakenly identified. The paper focuses on a method to eliminating the crossed meanings ambiguous words in the automatic indexing. The paper puts forward a method to eliminating ambiguous words combined algorithm of exhaustive method and disambiguation rules. Experiments show that it can avoid a great lot segmenting ambiguities with better segmenting results.

Keywords: Chinese text, Automatic indexing, Keyword extraction, Ambiguous words, Elimination algorithm.

1 The Research Status

Chinese is one of the major languages in the world, and also spoken by the largest number of people in the world. It has become an important means that people get a lot of meaningful information from the voluminous network information by network technology. However, because Chinese's the basic unit is Chinese characters, Chinese characters form words, word is divided into monosyllabic word and compound word, and there's no space between Chinese keywords. Chinese words unlike western languages separated by spaces, Western processing (indexing) technology does not easily draw. This is the reason that Chinese information processing is difficult. Because of the difficulty of Chinese information processing, in order to obtain accurate information from the network information is harder. Chinese information processing, namely Chinese automatic indexing technology research began in the early 1980s, a lot of these skilled personnel on 1990s, a lot of papers published so far in the ascendant. After 30 years of research and practice, Chinese information

D. Li and Y. Chen (Eds.): CCTA 2013, Part II, IFIP AICT 420, pp. 79–88, 2014.

technology has been greatly developed, especially in the Chinese word segmentation techniques with some of the more mature approach. For example, forward sweep, reverse sweep, maximum matching algorithm for network information processing and retrieval provides a powerful tool. However, due to the complexity of Chinese information processing, the Chinese word segmentation process produces a lot of ambiguous words, resulting in retrieving information is not accurate, Especially in the vast network information, if given a search term, instead of not retrieve information, but to retrieve a large number of irrelevant information, readers need to get useful information after several rounds of screening. People prefer to go through a search operation will be able to get accurate information. In this article automatic indexing method to eliminate ambiguous words is to solve the problem of network information accurate retrieval. From the scope of information retrieval, the Chinese indexing ways have hand information indexing and automatic indexing by computer. The former is indexing staff through reading, analyzing literature, which precipitated a keyword and norms, and finally given this literature indexing terms, although indexing words good accuracy, retrieval efficiency, but the vast amounts of information need to be addressed in the current status, the artificial indexing alone is not possible, the need for computer technology to process vast amounts of information. Information processing in the network, especially when Chinese information processing automatically generate a lot of ambiguous words, for words to eliminate ambiguity generated methods of information retrieval is a hot technology, but also information retrieval technology's basic research, it is great significance to accurate network information retrieval.

2 Ask Questions

Chinese literature indexing are that indexers generally extract and record keywords or class number which have meaningful literature retrieval features from Chinese literature by analyzing the content of the literature. These keywords will be used as the basis for document retrieval. Generally to retrieve pertinent documents, First step, these keywords are subject indexing by the indexing staff. The second step, after these keywords are indexed by the computer processing, people can carry out precise searches, the prefix search operation and rear consistent retrieval operation. Chinese automatic indexing is the process of extracting keywords to achieve by computers. With the rapid development of information technology, Computer technology is increasingly used in Chinese text indexing. Keywords and class number are extracted from the Chinese text by computer technology according to some word segmentation algorithm and matching rules. For Chinese literature, these keywords for the expression of the concept of literature are contained in document titles, abstracts and text. But there are three problems which are no spaces between Chinese keywords, blurred boundaries of words and phrase, containing ambiguous words in words and phrase in Chinese literature. For example, "President Jiang Zemin" the words, has expressed a complete theme concept, which can be given as a keyword indexing. But

in Chinese it also contains "Jiang Zemin", "democracy" and "chairman" three words to express the three concepts. "Jiang Zemin" and "chairman" two words can also be given as a keyword indexing, but the word "democracy" is also given as a keyword indexing is produced ambiguous word indexing. Such an ambiguous word in the case of manual indexing does not appear, but with a computer indexing, if not treated, often appear. For another example, “the People's Republic of China” in the word extraction process can also put "Chinese" is extracted, can also cause ambiguity word indexing. The paper focuses on a method to eliminating the crossed meanings ambiguous words in the automatic indexing.

3 Indexing Algorithm

There are two methods which extract the keywords or phrases from the Chinese literature in the current study. One is a method of rule-based segmentation[1], another is a method based on statistical analysis[2] of the sub-word. The former requires knowledge database for support, the latter does not need knowledge database and save part of the workload, but the search results are poor, for the accurate retrieval needs further filter the search results. I believe that the combination of the two methods used in automatic indexing will greatly improve the effectiveness of automatic indexing.

3.1 Automatic Indexing System Frame

This systematic framework for automatic indexing system and the elimination of ambiguous words is shown in Figure 1.

First step, the text is preprocessed by automatic indexing system, including some removing punctuation, extraction of feature words which are enclosed with a special symbol directly as indexing terms. The second step, the pretreatment of the text is processed and filtered by stop word list which include empty words and common words to get words or phrases, collection of short sentences. The third step, these words or phrases or short sentences are pumped word processing to obtain candidate keywords by the common vocabularies and professional vocabularies. Final step, Keywords are gave the corresponding weights based on the text of word frequency and occurrence of the word, which are sorted according to the statistics, and then according to presetting threshold value of keywords, keywords are selected by automatic indexing system.

Segmentation processing of acquired longer keywords by matching vocabulary words will produce ambiguous words. The removingmethods of ambiguous words are discussed in the next section.

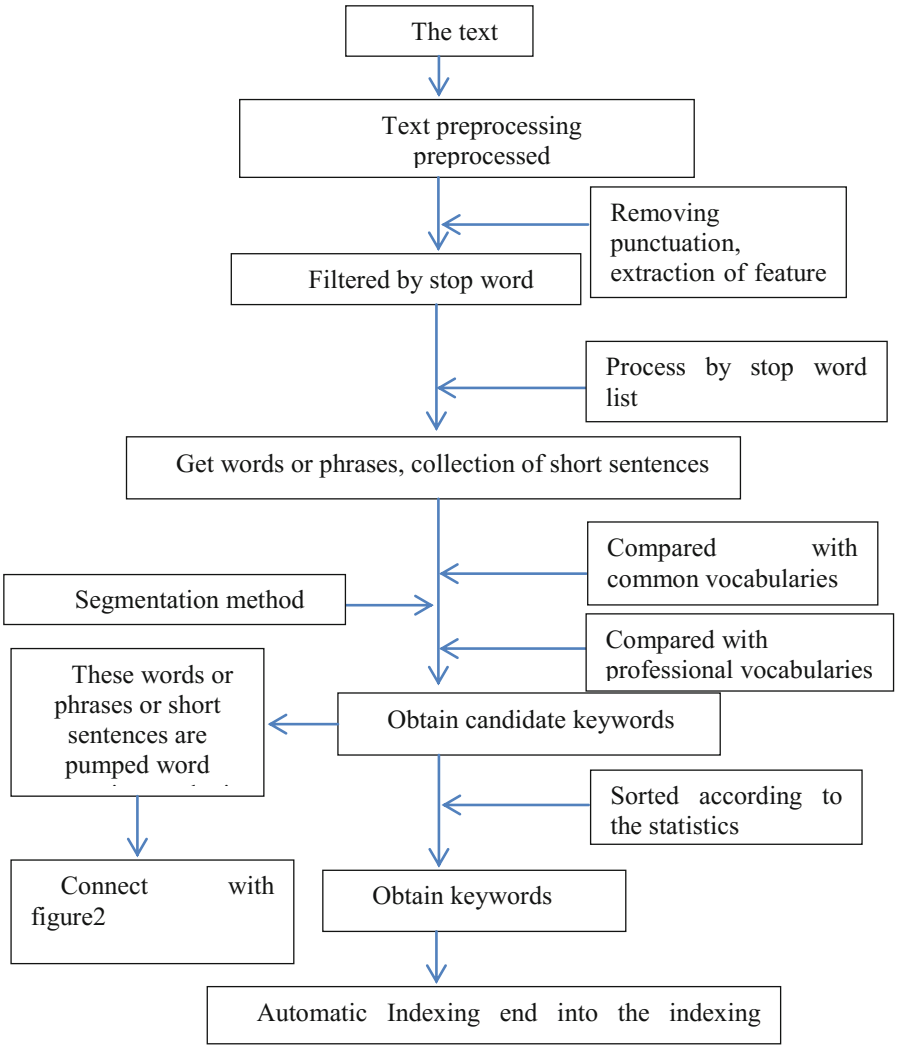

Fig. 1. Automatic indexing system frame

3.2 Text Preprocessing

First of all, information with a special symbol is extracted by automatic indexing system, including the title with quotation marks to cause, or place names and the person's name with a special symbol to cause. And then preprocessing information is preliminary segmentation processed by stop word list and removing punctuation to get words or phrases, collection of short sentences.

3.3 The Structure of Knowledge Database

Knowledge databases consists of stop word vocabularies, general and professional thesaurus, which are the basis for automatic indexing, its quality directly affects the effectiveness of automatic indexing words.

The text with removing punctuation is slicing processed by stop word vocabularies which include the commonly used function words in Chinese, for example, prepositions, conjunctions, auxiliary, and also include generic words. The use of stop words vocabularies can filter out unwanted words and rough cut of the text, in order to speed up text processing speed.

Word vocabulary is an important basis for containment of the keywords in the process of word segmentation. In order to accelerate the matching speed and accuracy of word segmentation, and general vocabulary can be divided into general vocabulary and professional vocabulary. "Classified Chinese Thesaurus" and all kinds of professional thesaurus have been expanded and modified to do common vocabularies and specialized vocabulary.

3.4 Word Segmentation Method

3.4.1 Forward Longest Matching Method

The strings obtainedby coarse segmentationhave been verbatim scanned from left to right and match with Thesaurus,and the keywords of thesauri maximum matching as the primary keywords. For example, in thesaurus in the "cadres tenure" in Chinese,and also included "cadres"、"office"、"age". Longest matching method is that "A short length is not taken" the word extraction rules, only extracting "cadres tenure" is used.

3.4.2 Reverse Longest Matching Method

The stringsobtained by coarse segmentation have been verbatim scanned from right to left and match with Thesaurus, and the keywords of thesauri maximum matching as the primary keywords.

First of all, according to the longest matching rule,keywords which have been matched with the thesauri have been extracted as primary keywordsin this article. Secondly,when strings length of primary keywordsare greater than or equal to four Chinese words, and then slicing process, may produce ambiguous words which this article discusses.The smallest unit of slicing process is two Chinese characters.

3.5 The Frequency and Weight of Keywords

According to the important degree of each part in the text, divided the parts identified, given the size of the contribution to content weight of the text before the text is pretreated.

3.5.1 Text Area Value

The weights of title, abstract and keywords in the text should be different, the former is big and the latter is small. Actually, the weights of the keywords from the title should be absolutely great in order to ensure that the word appears in the indexing words.

3.5.2 Important Statement

Article subtitled or weight of keyword in each of the first paragraph or the end of the paragraph statement is greater than weight of keywords in the body.

3.5.3 Word Frequency Statistics

According to the frequency and the weight of the keywords, the keywords to obtain by segmentation have been counted and sorted. According to the indexing depth which is maximum number of index words, the final text keywords of text have been gave. In accordance with the literature reports, the average depth of manual indexing are 7,usually the average depth of automatic indexing are 10 to 15.

3.6 Long Keyword Processing

After the keywords obtained by two-way scan the maximum matching have already been independent retrieval concepts, which directly access to a retrieval system for indexing processing, and providing search services in automatic indexing. But some of these words are very long term, word in the more length of the keywords not only contains the independence concept, but also has search significance, if not for slicing process, will be lost, causing leakage marked on the automatic indexing system. In general, the keywords of the more length should be re-carved process, slicing process may produce above-mentioned ambiguous words.

4 Produce Ambiguous Words

4.1 Type of Ambiguous Words

Ambiguous word is defined by different segmentation methods produce non-paper meaning of the word. Ambiguous words have two types of cross-type ambiguous words (cross ambiguity) and the combination ambiguous words (covering ambiguity). According to statistics, the cross ambiguity words accounted for 86% of the total, so to solve cross ambiguity words is the key of the segmentation of words. To exclude ambiguous words in automatic indexing algorithm is cross ambiguous words.

4.2 Methods to Disambiguate Words[4]

Currently the typical method of disambiguate words are:

4.2.1 Exhaustive Method

In general, exhaustive method is to find all possible words in the Chinese string to be split. Most matching algorithm is used in the forward or reverse matching algorithm method of exhaustion, or combination of forward or reverse matching algorithm exhaustive methods. When segmentation word is not correct, this method will produce ambiguous words.

4.2.2 Lenovo – Backtracking

Li Guochen et al [5] proposed Lenovo – Backtracking. First of all, Chinese string to be split according to feature words have been divided into several sub-strings, each sub-strings is either single word or word group, and then word group is subdivided into words by the entity thesaurus and rule base, when word Segmentation, a certain grammar knowledge is used.

4.2.3 Phrase Matching and Semantic Rules Law

Yao Jiwei, Zhao Dongfan[6] proposed a combination disambiguation method of a local single phrase matching and semantic rules based on the phrase structure grammar.

4.2.4 Part-of-speech Tagging

BaiShuanhu[7] eliminated ambiguity words by the combining method of Markov chain tagging technology and word segmentation algorithm

The paper puts forward a method to eliminating ambiguous words combined algorithm of exhaustive method and disambiguation rules.

5 Ambiguous Words Elimination Algorithm

The elimination algorithm of words ambiguous this article discusses is reprocessed the candidate words have already been cut a longer keyword or word group. Longer word is a meaningful indexing terms which appears in the dictionary as indexing words.

If no longer word segmentation processing, the leakage phenomenon of keywords may occur. In order to no-produce leakage phenomenon, we need word segmentation processing again in order to increase the literature retrieval point. Cut out of the word may appear ambiguous word, such as in the word "Jiang Zemin" the ambiguous word of "democracy".

5.1 The Disambiguation Process

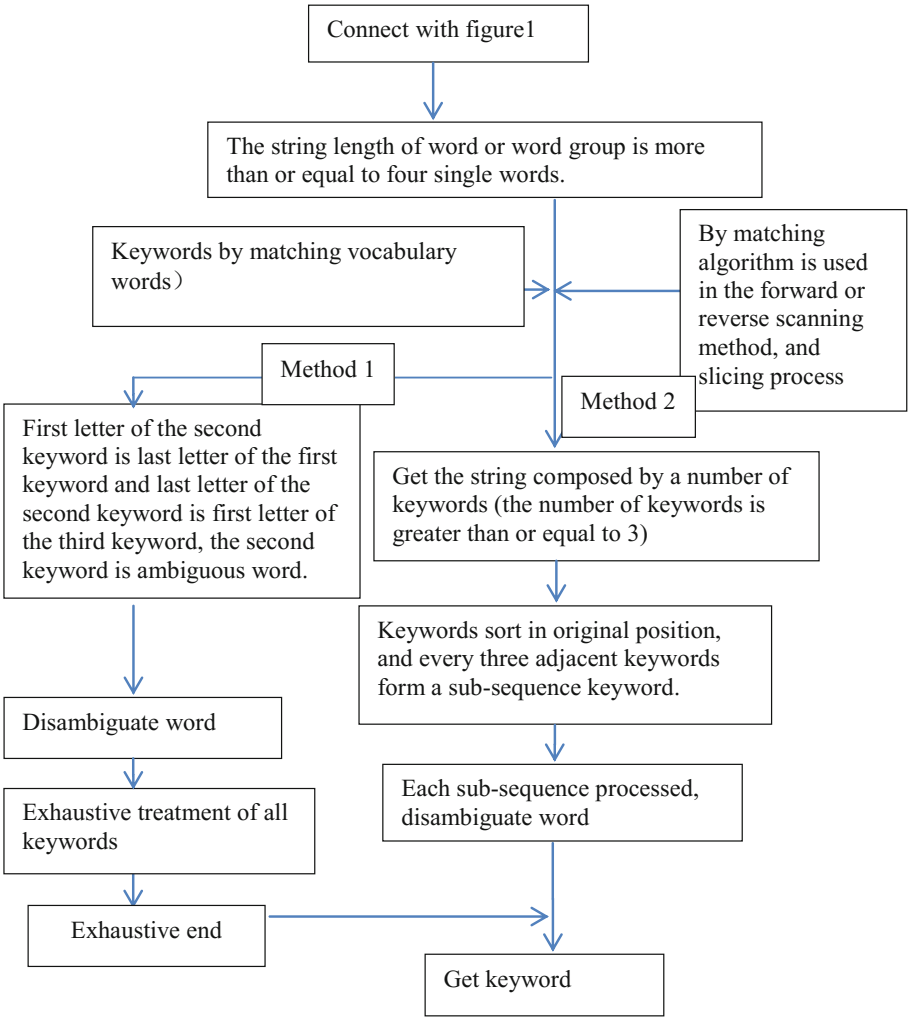

Fig. 2. The Disambiguation process

5.2 Disambiguation Algorithm

The longer Chinese words is set, it's the string length of are greater than or equal tofour single words. Without considering the single case of Chinese characters, maybe 3 or more than 3 of keywords have been cut with different methods of separation. After matching with keyword thesaurus, an N (N≥3) consisting of a sequence of keyword strings which referred to a large sequence string has been obtained. At this point, disambiguation in two ways:

Method 1: The sequence of each keyword in accordance with the Chinese characters string (longer keyword) sort order, in turn the adjacent three keyword consisting sub-sequence string, thus combined into several sub-sequence of strings, and then disambiguation process separately for each sub-sequence strings. That is, first letter of the second keyword is last letter of the first keyword and last letter of the second keyword is first letter of the third keyword, the second keyword is ambiguous word. Thus, after processing for each sub-sequence string, we eliminate all ambiguity word.

Method 2: Premise condition: Without permutations for large sequence string and do not form several sub-sequence string. That is, the first and the last letter of each keyword (referred to as A word) of large sequences check with the first and the last letter of the rest of the keywords. If the above conditions are satisfied simultaneously, namely, the first letter of A word is the last letter of one of the keywords and the last letter of A word is the first letter of another keywords, then A word is ambiguous word. After the above process for each keyword, all ambiguity word have been eliminated.

5.3 Sample and Analysis

The preselected words which had been split match with the common vocabulary or professional vocabulary. Such as “President Jiang Zemin”, “the People's Republic of China” ,“Major general”, “East China Sea Fisheries distribution”, and after segmentation produced ambiguous words “Democracy”, “Chinese”, “General” and “Seawater” are common vocabulary words. After the above disambiguation algorithm processing, we can eliminate these ambiguous words in turn.

In addition, specialized vocabularies such as Agricultural Thesaurus [8]often have a class of words. For example, for the following words through the segmentation processing, hyponym of "microbial fertilizer" (CLC as S114):anti-bacteria fertilizer、rhizobium fertilizer、azotobacter fertilizer, the "fertilizer" word (hypernym) are also extracted as keywords. If "fertilizer" is used as indexing terms, the upper word indexing error occurs. According indexing rules, the upper word can’t serve as index terms. Because of the massive literature retrieval operation, with the upper word as a search condition, the detection result is often a large number of irrelevant documents, the upper word indexing is the main reason. Strictly speaking, though these words are not ambiguous word, but it is ambiguous word indexing and should be eliminated.

Through the above-mentioned two types of ambiguous word indexing analysis and preliminary disambiguation experiments, this method to eliminate the ambiguous word is an effective method and provide readers reference.

6 Conclusion

Disambiguation and unknown word identification are the difficult problems in the Chinese word segmentation field. There are a lot of types and cause of ambiguous word, different word processing method produces different methods to eliminate ambiguous words. This article is related to the ambiguous word appears under certain

circumstances, although conditions require to eliminate such ambiguous words, but it is still an effective way to eliminate ambiguous words. I also hope that the majority of researchers study algorithms in a wide range of methodological and propose innovative solutions to design a common method of clearing ambiguous words and improve the accuracy and speed of segmentation. In addition, the extensive literature research focuses on statistical segmentation,which also focuses on based on combination with statistical segmentation and other methods todisambiguate, they will give the Chinese word segmentation technology to bring a substantive breakthrough.

References

1. Li, D., Cao, Y., Wan, Y.: New Security Feature Extraction Method Based on Association Rules. Computer Engineering and Applications (S1), 105–107 (2006)
2. Xiao, H., Xu, S.-H.: A Method of Automatic Keyword Extraction based on Co-occurrence Model. Transactions of Shenyang Ligong University (5), 38–41 (2009)
3. Su, X., Liu, X., Shao, P.: The Word-indexand Position Retrievalforthe Document TitlesIn Chinese. Journal of Nanjing University(Natural Sciences Edition) (2), 329–333 (1990)
4. Weng, H.: Comparison Studies on Inconsistencies and Ambiguity Automatic Identification Method in Chinese Information Processing. Language Applied Research (12), 93–94 (2006)
5. Li, G., Liu, K., Zhang, Y.: Segmentating Chinese Word and Processing Different Meanings Structure. Journal of Chinese Information Processing (3), 27–32 (1988)
6. Yao, J.-W., Zhao, D.: Disambiguation Method in Chinese Word Segmentation Based on Phrase Match. Journal of Jilin University(Science Edition) 48(3), 427–432 (2010)
7. Bai, S.: Chinese word segmentation and POS integrated approach to automatic annotation. In: Advances in Computational Linguistics and Applied, Beijing, pp. 56–61. Tsinghua University Press (1995)
8. Cai, J.: "Chinese Library Classification" professional classification "Agricultural Professional Classification". Beijing. Library Press (October 1999)