A Dictionary Based Chinese Sentence Segmentation

Yi Pan (yipan@cs.wisc.edu)

Computer Science, University of Wisconsin Madison

Abstract

Chinese segmentation is to segment one Chinese text into words, the basic semantic to express complete meaning. Chinese segmentation is the base for voice identification, intelligent input and search engine. Most systems of Chinese word processing is also built up on the base of Chinese segmentation.

This paper researches the techniques of Chinese segmentation, designs one system for Chinese Automatic segmentation, which presents a better experimental testbed for future research on Chinese segmentation. First, the paper summarizes present development of the existing techniques for Chinese segmentation and their advantages and disadvantages. Then according to tasks to be fulfilled, the paper designs the theme of the system and specifies system functions. There are three modules in the system, dictionary building, find and match and post-processing. Based on these work, the paper designs and implements all these three modules in detail. Storage structure for the dictionary is efficient, the match speed is optimized and the post-processing removes most of the ambiguities with the effort of doing machine learning on different collection of topics corpus.

At last, the paper evaluates the correctness comparing with the solutions that does not use the auxiliary dictionary generated by the machine learning process.

There are several improvements of this segmentation system. A STL hash map is used for the dictionary storage. Segmentation ambiguity is effectively reduced.

Key word: Chinese Information Process; A Dictionary Match; Forward and Reverse Direction Longest Match; Statistic Based Auxiliary Dictionary; Ambiguity

1. Motivation

Communicating with the computer systems with the natural language is a long desired goal for computer intelligence. The task is so important because 1)People can communicate with the computer systems with their natural language so that they don't need to learn the machine language, which usually results in a boring and tedious feelings. 2) With the communication with the computer systems using the natural language, people actually will know better about how their natural language works and its deep inside scheme. This process will be a win-win situation.

The key for the natural language processing is how to understand the natural language for the computer systems, and one of the key issues in understanding the natural language is to segment the language. Segmentation is to divide sentences into small semantic units like words or characters. Word is the smallest meaningful component of a language. Chinese segmentation is to use the computers to process the pronunciation, shape and meaning of the Chinese sentences, yet Chinese language is very complicated due to the long history, diverted tribes and the evolution of the language is different. Comparing with the western language, Chinese 1)doesn't have space character, which is a natural language delimiter and 2) has many similar meaning words 3) has many virtual words. All these increase the difficulty of automatic segmentation.

In many field, like Chinese search engine, Chinese input method, Chinese speech recognition, they all require the segmentation. Not only Chinese, all those languages like Japanese, Korean which are represented by the UNICODE, have this segmentation issue also. So segmentation for these languages is an interesting and important topic in the natural language processing.

2. Related Work

There are several segmentation methods which are currently used in Chinese segmentation. They are dictionary based segmentation, statistic based segmentation and the hybrid of these two methods.

Dictionary based segmentation, also called the mechanic segmentation; it uses certain strategy to match the input sentences with a "large enough" dictionary. If it has a match in the dictionary, it will succeed. According to the different scanning direction, it can be divided into forward direction match and reverse direction match. What is usually used is the longest match algorithm. It will scan the input and tries to find the longest match word in the dictionary.

Statistic based segmentation's idea comes very naturally. From statistics point of view, if two or more characters are

at a high possibility that they appear together, it is very likely that they are actually a word. So the unigram or bigram model will be useful in analyzing the large set of collected language corpus. However, it is easy to see that with this method, it will treat some character series which appear in a high frequency as a word, which, may not be the truth.

The hybrid method is to try combining several methods together in order to get a better result. Like we can see in the mechanic segmentation, it can not solve the ambiguity problem and it can't take care of the unregistered word problem. So in the real practice, mechanic segmentation is usually used as the first step and some methods will be applied to refine the result.

3. My Design

In my system, I am going to use the dictionary based approach. For achieving this goal, there are several things I need to consider. First, how to build a dictionary. Second, how to store the dictionary so it can be efficiently searched. Third, how to process the ambiguity which is a big problem for the dictionary based segmentation.

For the first problem, I collected a large set of Chinese articles from website, manually divide them into words and also find some electronic version of Chinese dictionaries. Together I create a dictionary for my use, which is roughly 270,000 words in it.

For the second problem, I decided to use the STL hash map as the dictionary container. This will have a very efficient computation effort for the word look up.

For the third problem, I divide the problem into 2 sub problems, one is to identify the ambiguity and the other is to try to solve it. For the first goal, I use the 2-way scan for each sentence. I will do a forward direction longest match over the sentence and then do a reverse direction longest match. Then I will compare the output from these 2 methods. If they are identical, I consider there is no ambiguity. If not, there is an ambiguity here. For the second goal, I tried to collect one specific topic of articles, to be more precisely, the financial news. I tried to do statistics over these collected articles and create an auxiliary dictionary which will have the words and their frequency. Whenever an ambiguity is detected, the two different segmentation output will be checked against this auxiliary dictionary and the one with a higher frequency weight will be chosen as the final result.

An input text file will be given by the user and three output files will be generated, they are the result from forward direction segmentation, reverse direction segmentation and the ambiguity-removal process output respectively. By comparing these three outputs, we can find by using this ambiguity-removal process, the segmentation result is much better than either the forward direction longest match or the reverse direction one.

4. Implementation Detail

For the time constraint, I only did the collection and manual processing for about 10 financial articles from google news. It is easier to see that with a larger corpus, the auxiliary dictionary will play a more important role in removing the ambiguity. However, for this project, it still gives the positive effect in ambiguity-removal.

For storing the dictionary, I use the STL hash map and the pair data structure. Each word in the text file will be formatted as a pair and inserted into the hash map.

Non-Chinese character processing is taken care of in this project, if there are spaces, Chinese/English punctuations and other special characters, they will be segmented separately.

When implementing the forward and reverse direction longest match, I set the longest word length to be recognized to be 5 characters. A longer maximum matching length will result in a great increase of processing time for the system as it will generate much more searching sub tasks for a given sentence. I found 5 Chinese characters (10 bytes) to be a reasonable number.

I implemented a scheme to evaluate the weight from two different segmentations for an identical sentence. With the help of the auxiliary dictionary, I can check out the frequency of each word. The idea is simple, if a word is having a higher chance appear in one specific kind of articles, and the sentence is from an article in that specific topic, the segmentation which segments the sentence in this way is more likely the correct segmentation for this particular sentence.

5. Evaluation

After the dictionary is loaded into the system, the segmentation process for the sentences is very fast and the correctness is acceptable. Here is an example output:

With the forward direction longest match:

据|彭|博|社|5|月|2|日报|道|巴|菲|特|表示|，|接替|他|担任|伯克|希尔|公司|首席执行官|的|人选|目前|已经在|公司|工作|。|…

…

…

而不|幸|的|是|这次|轮到|了|我|。|”|…

With the reverse direction longest match:

据|彭|博|社|5|月|2|日|报道|巴|菲|特|表示|，|接|替他|担任|伯克|希尔|公司|首席执行官|的|人选|目前|已经在|公司|工作|。|…

…

…

而|不幸|的|是|这次|轮|到了|我|。|”|…

After the ambiguity-removal process:

据|彭|博|社|5|月|2|日|报道|巴|菲|特|表示|，|接替|他|担任|伯克|希尔|公司|首席执行官|的|人选|目前|已经在|公司|工作|。|…

…

…

而|不幸|的|是|这次|轮到|了|我|。|”|…

It is easy to see that in the reverse direction longest match, |接|替他| is not correct while the segmentation in forward direction longest match is right with this one. However for |而不|幸| is not correct for forward direction longest match while the one in reverse direction |而|不幸| is correct with this sentence. In the final result, which finds these discrepancies in the two methods and consults further with an auxiliary topic specific machine learning dictionary, it actually finds the correct solution. So it is better than either one.

Note that the people's name and other specific words are not recognized. This is due to the so called unregistration problem. These words are not collected in the dictionary so they can not be identified here.

6. Conclusion and future work

The dictionary based word segmentation is fast to work and with the introduction of the auxiliary dictionaries which are generated by learning each specific topics of article collection; this method will have an acceptable correctness and reasonable segmentation speed. If we can make a preprocessing to identify which topics an article belongs to, we can use the desired auxiliary dictionary for that topic. This will have a very good result for removing ambiguity.

The dictionary itself will have a big impact on the segmentation result. It is ideal to have an "all in it" dictionary, although not possible. But the larger the dictionary is, the better segmentation result we will have.

References

1. Thomas Emerson, March 2000. *Segmentation Chinese in Unicode*: 16th international Unicode Conference.
2. Chao Yang, 2005, *Research in Segmentation technique*: www.lw86.com.
3. Dexi Zhu, 1982. *Chinese Grammar*: Commercial Press.
4. 1993, *GB/T13715-92 Modern Chinese Segmentation Standard for Information Processing*: China Standard Press.
5. Maosong Sun, Jiayan Zou. *Research in Chinese Automatic Segmentation*
6. Maosong Sun, Zhengping Zuo, Changning Huang, 1999. *Research for the Dictionary for Chinese Automatic Segmantation:* Qinghua University.
7. Fei Zou, 2002. *Web based Chinese Segmentation Technique* :Nanjing University.
8. Halpern, Jack and Jouni Kerman, 1999. *The Pitfalls and Complexities of Chinese to Chinese Conversion*: Proceeding of the 14th international Unicode Conference.
9. Jian-yuan Nie, Martin Brisebois, Xiaobo Ren, 1996. *On Chinese Text Retrieval*: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval.
10. KWOK, K.L., 1997. *Comparing representations in Chinese information retrieval*: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval.