Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RESEARCH ON THE CHINESE SEGMENTATION TECHNOLOGY

[1,2] FENG HONGYU, [1]YAN SHITAO

[1] Henan Institute of Science and Technology, School of Information Technology, 453003,

[2] Wuhan University of Technology, School of Information Technology

feng_hongyu@126.com

ABSTRACT

Abstract: The Chinese segmentation technology is the basis and key of Chinese information processing. By researching on this technology, this paper indicates that the method based on dictionary and statistics are much mature, but the unregistered word recognition is the main problem in this field. Thus, this paper lays a foundation and sets the directional work for its later development.

Keywords: *Chinese Segmentation, Corpus, Evaluation*

1. INTRODUCTION

The Chinese segmentation technology is the basis and key of Chinese information processing, whose quality plays a crucial role in latter Chinese information processing. The machine translation, automatic abstact, information retrieval, question and answer system, speech recognition and other fields' development will inevitably need a good segmentation system. As we all know, the word is the smallest language unit that can be meaningful independently. In English, the space between words is the natural boundary, but in Chinese, the Chinese character is the basic writing unit without obvious distinguishable mark. Therefore, the Chinese segmentation problem becomes the basic and first work in Chinese information processing. To put it simple, the process of adding appropriate and clear boundary marks to make the formed string of words reflect the original meaning of the sentence is called Chinese segmentation.

At present, the more popular word segmentation method can be classified into three categories: the one based on the dictionary, also known as mechanical segmentation mainly uses one word table to do the pattern matching, thus to cut. It does not rely on lexical, syntactic and semantic knowledge. The advantages are the segmentation speed is quick, it is simple and accurate, and moreover, it is easy to be realized. So, it is widely applied in Chinese information processing. However, it has some disadvantages, the matching speed is slow, intersection and combination ambiguity segmentation problems can happen. Since the word itself doesn't have a standard definition, there is no unified standard word set, and different dictionaries will have different ambiguities. ②The other one is based on the statistical method. Based on the corpus supervised or unsupervised learning, the "language model" that describes one language can be got. It has the advantage that it is not limited by the processed text field, and it doesn't need a machine readable dictionary. Thus, the influence of unknown words can be reduced, as long as there is enough training text. Its defect is that it needs a lot of preliminary segmented corpus for support, and the space and time costs are great in training process. [1]③The Chinese segmentation based on the understanding of Chinese, also known as the knowledge segmentation, is to use the syntactic and semantic information of words and sentences, or from a large corpus to find out the combination characteristics to evaluate. The purpose is to find the closest segmentation results to the original semantic meaning. Although the knowledge segmentation overcomes the above two methods' faults, but the scheme algorithm is complex.

So Chinese segmentation based on the understanding Is still in the testing stage. Although in the experimental stage, but with further research, this method will become hot direction of the segmentation research. The judgment and identification of ambiguity Chinese is a breakthrough problem in Chinese segmentation.

And it is also an important standard to judge the quality of the system. [2]

This paper studies Chinese segmentation from the following aspects, algorithms, evaluation methods, problems and the future direction of development.

2. CHINESE SEGMENTATION BASED ON DICTIONARY

The Chinese segmentation based on dictionary method uses the dictionary to do the pattern matching, thus to cut the sentence. It mainly includes three basic algorithms: Forward Maximum Matching Method, Reverse Maximum Matching Method and Bi-direction Matching method. [3] This method is simple, convenient, but undoubtedly exists obvious defects. That is, it may cause ambiguity and can't identify new words. When dealing with the Chinese sentence "writers are in the study of life". Whether to cut this sentence into "postgraduate students | live" or "research | life", and this is the segmentation ambiguity. The segmentation method based on the string matching needs a dictionary. If encountered one word that does not exist in the dictionary, the process can't do the correct segmentation. Although most of dictionaries can cover most of the words, there are quite many words may not be included in the dictionary; these words are called unregistered words or new words. The unregistered words can cause the segmentation accuracy down to at least 5 times comparing with the ambiguity. [4] The actual used word segmentation system is to use the dictionary's mechanical segmentation as an initial step, and then utilizes other segmentation methods to further improve the segmentation accuracy, including the identification of unregistered words. [5]Take the positive maximum matching method as an example to present the basic algorithm thought. The Forward Maximum Matching Method is commonly referred to as the MM Method. The basic idea is: assume the longest word in the segmentation dictionary has I characters, and then use the first I words in the current document as the matching string, and then find the dictionary. If there exists such an I words in the dictionary, then the matching is successful, and the matching string is segmented as a word. Otherwise, the matching will be a failure; the last word in the matching string will be taken out. The above process will go on until the matching is successful. At this point, a round matching is finished, and then takes the next I string matching process until the document is scanned over. [6] The algorithm is shown in figure 1.

The error rate of FMM is 1/169. The Reverse Maximum Matching Method is commonly referred to as the RMM Method. The basic principle of RMM method is the same with MM method, and the difference are that the direction of segmentation is reversed, and the used dictionary is also different. The error rate of RMM is 1/245.

To segment " in market state-owned enterprise development stability":

FMM results: market/China/have/enterprise/ development/stability

RMM results: in market /state-owned/enterprise/ development/stability

Bi-direction matching method combines the forward maximum matching method and the reverse maximum matching method. It firstly segment the document into several sentences according to the punctuation, then do the scanning segmentation to by the forward maximum matching method and reverse maximum matching method. If the results of the two segmentation methods are the same, the segmentation is right. Otherwise, it will be dealt with according to the minimum set processing.

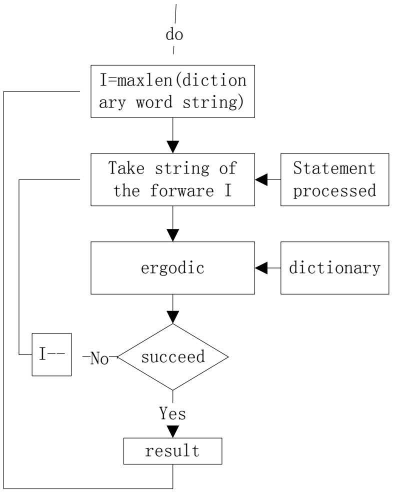
do
I=maxlen(dictionary word string)
Take string of the forware I
Statement processed
ergodic
dictionary
I--
No
succeed
Yes
result

Figure 1 The Process Figure Of FMM

3. CHINESE SEGMENTATION METHOD BASED ON STATISTICS

3.1 Introduction to Chinese Corpus

Corpus (corpus base) is database that deposits language material. [7] in 1991, China National Language Committee began to establish the national Chinese corpus to promote researches on Chinese lexical, syntactic, semantic and pragmatic. Its planned scale will reach to 70 million Chinese characters. Tsinghua University established a 100 million Chinese corpus in 1998, mainly for the studies on ambiguity segmentation problem. The computer language institute of Beijing University developed the multistage processing of modern Chinese corpus since 1992. It has achieved a lot in the construction of corpus; including the completion of the People Daily's tagging corpus with 26 million words in 1998, the textual level English-Chinese contrast bilingual corpus with 20 million words and more than 1000 English words, and the textual level information science and technology corpus with 80 million words. The balance Corpus Sinica Corpus of Taiwan University (http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/) is the world's first Chinese balance corpus with completed speech mark, and it includes 5.2 million words times (7.89 million characters). The corpus selects philosophy, art, science, life, society and literature texts published from 1990 to 1996. Its design thinking is: 1) follow the standards of Taiwan computer language association; 2) sample with natural paragraph, ignoring the article length; 3) use multiple classifications. In 2003, a Chinese-English parallel corpus was added, which includes 2373 Chinese-English parallel comparison texts. The scale of Peking University Modern Chinese Corpus is about 85 million Chinese characters.

3.2 The algorithm based on statistics

The segmentation methods based on the statistic mainly uses statistics. The statistical model [8, 9] includes mutual information, N grammar model, and neural network model [10] the hidden Markov model (HMM) and so on. These statistical models make use of the probability of word joints as the condition of segmentation. Its principle is that from the perspective of word forms, the word is a stable word combination. So in context, more frequently the adjacent words occur at the same time, the more they are likely to form a word. Thus, the co-occur frequency or probability of word can better reflect the credibility of word formation. To summary the frequency of co-occurrence of each word adjacent in the corpus can calculate the mutual information. [11] There are many methods to test the combination degree of strings, such as the mutual information, chi-square test, t test, etc. [12]thinks the chi-square test is superior to mutual information. [13] adopts the marks to convert the Chinese segmentation problem to the sequence marking problem, and then uses the maximum entropy model to mark. The results are good. The first article based on word mark (Charac2ter2based Tagging) was published in 2002 in the first SIGHAN seminar. But it did not cause the attention of the academic circle. A year later, Xue's Maximum Entropy model based on the realization of word segmentation system took part in the Bake2off22003 assessments, and it won the second prize in the closed test project of AS corpus, yet its ROOV (0.729) was at the top. He also won the third place in the closed test of City corpus, and its ROOV (0.670) was still the highest. Since then, the machine learning method based on word obtained widely attention. The system with leading performance [14] almost adopted this similar marking method, and it becomes the mainstream in the field of segmentation research. CRFs can overcome the mark bias problem of maximum entropy model, and it is superior to the maximum entropy model based on word segmentation system. Thus, it becomes the first choice of word mark model [15-16].

The Conditional Random Fields, (CRF) in an undirected graph model that can calculate the conditional probability of the given input nodes [17]. Figure 2 is the basic process of word segmentation based on CRF algorithm.

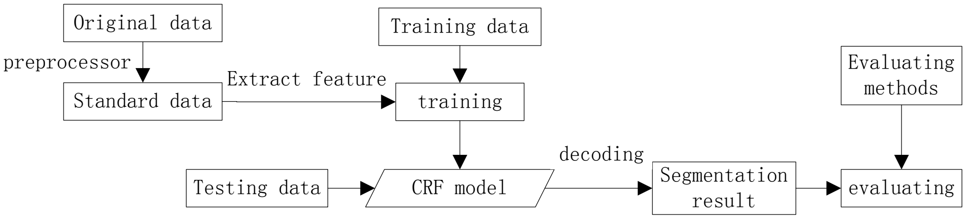

Figure 2 Basic Process Of Word Segmentation Based On CRF Algorithm.

The segmentation method based on word mark is actually the formation method. That is, regard the segmentation process as the word mark problem in word string. Since each word is occupying a certain formation position (i.e. lexeme) in constructing a specific word, if stipulates every word can only have four formation positions: namely B (initial), M (middle), E (postfix) and S (one word).

The advantages of word segmentation method based on statistical are: (1) it is not confined to the field of the texts; (2) it needs not a machine readable dictionary. While its defect is: (1) it needs a lot of training texts to establish the parameters of the model; (2) it requires a large amount of calculation; (3) the segmentation accuracy is relates to the text selected.

4. SEGMENTATION METHOD BASED ON UNDERSTANDING

The word segmentation method based on understanding is also called artificial intelligence segmentation method. Its basic idea is to simulate the language and sentence understanding of human's brain, thus to identify vocabulary unit. The basic model is to define the word by analyzing the content and the context of the information it provides based on syntax, grammar analysis, and semantic analysis. It usually includes three parts: segmentation subsystem, syntactic semantic subsystem and dispatching system. With the coordination of dispatching system, the segmentation subsystem can get the syntactic and semantic information to judge the ambiguity of segmentation. This method tries to make the machine have the human understanding ability, so it needs a large number of language knowledge and information. However, due to the generality and complexity of Chinese language, it is difficult to organization all kinds of language information into the machine directly readable form. Therefore, at present this method is still at the experimental stage.

5. EVALUATION STANDARDS

The evaluations on the Chinese automatic word segmentation and speech tagging system are important means to promote their development. The test index is mainly the correct ratio, recall rate and F - measure value. The calculation formula is as follows.

The proportion of correct number of segmentation words among the output test results. Assume that the number of output is N, and the correct number is N, thus the formula (1):

$$P = \frac{n}{N} * 100\% \qquad （1）$$

Recall ratio (R): the number of the right results accounts for the total number of standard answers. Assume that the number system outputs is N, and the number for correct results is n, and the number of standard answers is M, then the formula is(2) :

$$R = \frac{n}{M} * 100\% \qquad （2）$$

Two kind of marks: R_{OOV} refers to the recall rate of without the Set; R_{IV} refers to the recall rate of within the Set. F - Measure): the comprehensive value of accuracy and retrieve rate. The calculation formula is (3) :

$$F - measure = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P * R} * 100\% \qquad （3）$$

Generally, β=1, then（4）：

$$F1 = \frac{2 * P * R}{P + R} * 100\% \qquad （4）$$

Since the 1990s, the Chinese and Interface Technology Evaluation group of Chinese "863" plan has conducted many Chinese segmentation and speech tagging evaluations. In July, 2003, the SIGHAN held the first international Chinese segmentation evaluation in Sapporo, Japan called Bakeoff. [18] SIGHAN is the short name of the subordinate Chinese processing professional committee of ACL. In 2005 and 2006, it held the second and third session of Chinese segmentation evaluation in Jeju, South Korea and Sydney, Australia. [19-21] in 2007 and 2010, the fourth and the fifth Bakeoff were held. The segmentation system based on word mark caught the spotlight in the Bakeoff22005. The Low [22] system uses the maximum entropy model, and it won three titles (AS, CityU, PKU) and a second (MSRA) in the four open tests. The Tseng system used the condition random model, and it made two titles (CityU, MSRA), a second (PKU) and a third place (AS). In the Bakeoff22006, the segmentation system based on word has already blossomed everywhere [23 to 24]. The development of the Bakeoff has greatly promoted the progress of Chinese segmentation.

The Bakeoff adopts anther different evaluation system from 863, 973 segmentation. It posts four different standard training corpus (with marked material) on the Internet, a month later it announces and the four standard corresponding test corpus (original corpus). The contestant system can select any one or more standards to appraise his segmentation system. Each corpus can be divided into the closed and open tests. The closed test only be allowed to use the knowledge from the designated training corpus (such as vocabulary, N yuan grammar, etc.) to perform automatic segmentation study; while the Open test is not affected by such constraints. [4] Table 1 lists the evaluation results of CIPS SIGHAN 2010.

Table 1 The Evaluation Results Of CIPS SIGHAN 2010 [25 To 26]:

Domain	Mark	R	R_{OOV}	R_{IV}	P	F1
Literature	A	0.937	0.652	0.958	0.937	0.937
Computer	B	0.941	0.757	0.974	0.940	0.940
Medicine	C	0.930	0.674	0.961	0.917	0.923
Finance	D	0957	0.813	0.971	0.956	0.957

6. SUMMARY AND PROSPECTS

The Chinese segmentation is the first step of Chinese processing, whose quality plays a crucial role in latter Chinese information processing. This paper firstly studies the meaning of word segmentation, then focuses on the segmentation methods and the evaluation system, thus proves that the segmentation algorithm is more mature, but the unregistered word recognition is the main problem in this field. Segmentation algorithm is the core of the Chinese word, any mature segmentation system, impossible to use any single segmentation algorithms, typically needs to draw on the advantages of several different segmentation algorithms to achieve. For instance, it is possible to use a basic sub-word dictionary as a preliminary segmentation method for string matching, but also under the corpus context of the statistical results, the use of certain statistical theory or statistical model to treat ambiguity field and identify new words.

In computer simulating reasoning process, it needs to use a lot of knowledge of the language and information, coupled with the knowledge of Chinese general and complexity, it is difficult to organize of the various types of language knowledge into machine-readable form directly, so segmentation based on the understanding has many questions to deal with, but in the future, this method will become popular direction of the Chinese segmentation research. Also the segmentation in specific areas and identifying new words should be the breakthrough direction in segmentation research.

REFERENCES

[1] Fang Kui, Luo Wu, Wang Yujuan, Bu Weiqiong. On the Design and Implementation of Chinese Tokenizers in Agriculture. Agricultural Engineering, 2012 Vol． 2 No． 3

[2] Zheng Xiaogang, Han Lixin, Bai Shu, Zeng Xiaoqin. On the Combinational Chinese Segmentation Method. Computer Applications and Software, 2012 Vol.29 No． 7

[3] Zhou Chengyuan, Zhu Min, Yang Yun. Study on the Chinese Segmentation Algorithm. Computer & Digital Engineering, 2009 Vol. 37 No. 3

[4] Huang Changning, Zhao Hai. A Ten years Overview of Chinese Segmentation. JOURNAL OF CHINESE INFORMA TION PROCESSING, 2007, Vol. 21 , No. 3

[5] Chen Ping, Liu Xiaoxia, Li Yajun. Chinese segmentation Method based on Dictionary and Statistics. Computer Engineering and Applications, 2008,44(10)

[6] Zhang Dan, Literature Review on the Chinese Segmentation Method. Science and cultural,2006

[7]Zong Chengqin, Statistics Natural Language Processing [M]. Tsinghua University Press, 2008

[8]Wang Xiaolong, Guan Yi． Computer Natural Language Processing [M]． Beijing: Tsinghua University Press, 2005．

[9] James·Preiss, Bayesian Statistics Principle, Model and Application[M]. Beijing: China statistical Publishing House,1992．

[10]Jurafsky D,Martin J H. Speech and Language Processing:An Intro-duction to Natural Language Processing,Computational Linguisticsand Speech Recognition [M].USA: Prentice Hall, 2000

[11] Zhang Feng, Xu Yun, Hou Yan, Fan Xiaozhong. Chinese Term Extraction System Based on Mutual Information. Application Research of Computers, 2007,5

[12]Ye Na, Chen Xiaofang, Cai Dongfeng[J]. Journal of Shenyang Institute of Aeronautical Engineering, 2010, 27(4):32-36．

[13]Nianwen Xue. Chinese Word Segmentation as Charac2ter Tagging [J] .Computational Linguistics and Chinese Language Processing, 2003, 8 (1):29-48.

[14]Hai Zhao ,Chang Ning Huang and Mu Li. An Improved Chinese Word Segmentation System with Conditional Random Field [C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney, Australia: 2006: 1082117.

[15]John Lafferty, Andrew McCallum, Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]//Proc. of ICML218. Williams College, USA, 2001:282-289.

[16] Fuchun Peng, Fangfang Feng and Andrew McCallum. Chinese Segmentation and New Word Detection using Conditional Random Fields [C]// COLING 2004. Geneva, Switzerland , 2004: 562-568.

[17]Zhou Bo, Cai Dongfeng. Research on Chinese Organisations names' Recognition Based on Conditional Random Fields[J] . Journal of Shenyang Institute of Aeronautical Engineering, 2009, 26(1):49-52.

[18]Sproat, R. and Emerson, T. The First International Chinese Word Segmentation Bakeoff[A]. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C]. Sapporo, Japan: July, 2003, 133-143.

[19]Emerson , T. The Second International Chinese Word Segmentation Bakeoff [A] . In : Proceedings of the Fourth SIGHAN Workshop on Chinese Language Pro2cessing[C] . J eju Island , Korea : 2005 ,1232133.

[20] Levow , G. The Third International Chinese Lan2guage Processing Bakeoff : Word segmentation and named entity recognition[A] . In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Pro2cessing[C]. Sydney: July 2006, 1082117.

[21] Chengjie Sun , Chang2Ning Huang et al. Detecting segmentation errors in Chinese annotated corpus[A] .In : Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing[C] . Jeju Island , Kore2a : 2005 , 128.

[22J in Kiat Low , Hwee Tou Ng and Wenyuan Guo. A maximument ropy approach to Chinese words Seg2mentation [A] . In : Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing[C] . J eju Island , Korea : 2005 , 1612164.

[23] Huihsin Tseng , Pichuan Chang et al. A conditional random field word segmenter for SIGHAN Bakeoff2005[A] . In : Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing[C] . Jeju Island, Korea : 2005 , 1682171.

[24] Hai Zhao , Changning Huang et al. Effective tag set selection in Chinese word segmentation via conditional random field modeling [A] . In : Proceedings of PA2CL IC220[C] . Wuhan, China : November 123 , 2006 ,87294

[25]K. Wang, C. Zong and K. Su. A Character-Based Joint Model for Chinese Word Segmentation. Proc. COLING 2010, Aug. 23-27, 2010, pp. 1173-1181

[26]K. Wang, C. Zong and K. Su. A Character-Based Joint Model for CIPSSIGHAN Word Segmentation Bakeoff 2010. Proc. CLP2010, Aug. 28-29, 2010 245-248