Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (16): 3153-3160, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.3153.3160

Ambiguity Analysis Model of Word Segmentation Based on Word Group

Rongliang Luo, Hongxi Zhang and Minghui Wu
Department of Computer Science and Engineering, Zhejiang University City College,
Hangzhou, 310015, China

Abstract: We propose a word segment algorithm based on word group. In this model, word group is used for Ambiguity analysis. At first step, statistical information is used for build information base. In the process of dealing with sentence, a small step is triggered for counting information of adjoining situation and word frequency and calculates parameters of this model according to size of window. When get different word sequences, we use Analysis Tree to find the prime sequence. Because of short in decision distance, we get a low time complexity. Algorithm analyzing and result of experiment show that segmentation algorithm based on word group has higher efficiency and accuracy.

Key words: Analyses of ambiguity, analysis tree of ambiguity, word group, segmentation degree

INTRODUCTION

Word is the smallest unit of natural language Wang *et al*. (2011) it could be used independently and is also the basic unit of language information processing. Segmentation is an important task in the process of natural language processing on a language without separators (Wang *et al*., 2011), such as Chinese language Research in Chinese word segmentation has progressed tremendously in recent years (Li, 2011). Many excellent segmentation algorithms have been proposed in the field of segmentation. Practical results show that the performance of word segment system based on manual rules is not so good as system that based on statistic model (Huang and Zhao, 2007). According to whether dictionary is used and the way in which dictionary is used, word segmentation methods typically can divide into three kinds of methods: Dictionary model, statistical model and understanding model (Yue *et al*., 2012). Maximum Match (MM) is a algorithm adopted in the word segmentation method which is based on dictionary. This method has an advantage of higher efficiency while its accuracy is limited to the capacity of dictionary. Its limitation in grammar and semantics model leads to the incapability on the segmentation of ambiguity. Word segmentation method based on statistical model, such as that based on hmm, is with favorable ability to section ambiguity and recognize new words. However, it is with lower sensitivity to identify commonly used word; besides, more time and space spending are required. Meanwhile, some illegal words are recognized as word groups because of high-frequency co-occurrence. Word segmentation method based on semantics model works by imitating the way by which humans making sense of sentences. The process of segmentation in this method, is related with knowledge in Chinese language and artificial intelligence and so on; the method based on semantics model is regarded as the optimal one theoretically. But it is still at a testing stage needing exploration because a mass of language knowledge and information are required out of the complexity and flexibility of Chinese natural language.

Ambiguity resolution in word segmentation (AR) has been described as a task in which selects the appropriate word sequence in word segmentation when meet at list two way to split a sentence into a word sequence. AR is also an important part in the segment process and this is the main problem that we dealed with in this study. Different from hmm which makes decisions on the whole sentence wen deal with word segmentation, the algorithm introduced in this study carried out segmentation decision in the level of word group by setting segmentation window. With the shorter decision distance, the time complexity is decreased. At the same time, automatic segmentation algorithm has been taken into consideration to make segmentation with higher efficiency by utilizing dictionary. In the procedure of this algorithm of word segmentation, we making use of linguistic context, lexical statistical information and statistical law for word sequence. Dictionary, rules and statistics were applied in the algorithm. In the end, a superior segmentation was carried out with a time complexity of O (n).

Correspondong Author: Minghui Wu, Department of Computer Science and Engineering, Zhejiang University City College, Hangzhou, 310015, China

RESEARCH BACKGROUND

"Ten-year Review about Chinese Segmentation" of Huang and Zhao, mainly represents the achievements of Chinese segmentation in the past decade, displays issues and difficulties in the field of segmentation and summarizes the traditional methods and burgeoning ones. Statistical method based on label cut a figure gradually. Wang, has adopted a strategy of EAS to carry out segmentation in literature. For a segmentation sequence, there were two steps: evaluation and selection. The latter one was set to provide adjustment for the former one to improve the accuracy of segmentation. "Study of Chinese Segmentation Technology on Patent Literature" of Yue and Xu, adopts segmentation method based on the combination of field dictionary and statistics in order to solve the segmentation in Chinese patent literature more reasonably. Based on ictclas segmentation system, nc-value was used to make statistical sampling about domain-independent terminology with multi-word. However, the defect of this algorithm was obvious that this segmentation system was established on the foundation of ICTCLAS segmentation system of Chinese Academy of Sciences. Zhong built a normal form of Chinese segmentation in literature. According to the normal form defined, he applied tree structure to analyze and deal with the process of segmentation. Even though the results of segmentation in the experiment were satisfactory, the efficiency of the form still need to be verified. Wu and Lu design an algorithm for Chinese word segmentation in agriculture. They build a knowledge base as a central part of the agricultural information service platform and make their algorithm of dictionary search based on the Hash mechanism. But the algorithm is too special and limited. Ambiguity resolution in word segmentation is a important task in word segmentation. Although, the previous papers has their strategy for dealing with it, our algorithm is designed for solving this problem and making a model for word segmentation in another way.

Our study designed a segmentation model based on current context by making use of linguistic context, lexical statistical information and statistical law for word sequence, dictionary, rules and statistics to improve the accuracy and efficiency of segmentation. The structure of the study is as follows: Part two displayed the detail description about the algorithm. The first section presented the selection methods for different word sequences of segmentation and defined Analysis tree of ambiguity for the description of selection process. The second section introduced word sequence of segmentation, the variable used and calculation of the corresponding degree of segmentation MS_x to make judgment. The third section described the whole algorithm, including the background knowledge, the variable used and algorithm pseudo-code in detail. Furthermore, some examples were set to explain the process of its application. The third section also displayed partial data provided by SIGHAN Bakeoff-2 to analyze the results of segmentation.

ANALYSIS OF ALGORITHM

Before explain the detail, there are some defination and parameters need to introduced. This chapter includes three segments. In first section, two different in order to elaborate the analysis of ambiguity during the process of segmentation, Analysis Tree of ambiguity, as the model to analyze ambiguity in the process of segmentation, is defined. The second section introductoin degree of segmetation MS_x to judge word sequence of segmentation. The calculation of Ms_x was based on the analysis tree. The third sectoin described the whole algorithm, including pseudo-code description in important parts. Meanwhile, some example were set to explain the process of segmentaton algorithm as well.

Related definition of segmentation models

Definition 1: Combinatorial Ambiguity: If $W = a_1...a_ib_1...b_k$, $W_1 = a_1...a_i$, $W_2 = b_1...b_k$ are the words and the context is $<f_1, e_1>$, $<f_2, e_2>$; the words of $a_1...a_ib_1...b_k$ in $f_1a_1...a_ib_1...b_ke_1$ is W and results of $a_1...a_ib_1...b_k$ in $f_2a_1...a_ib_1...b_ke_2$ is the word sequence consists of W_1W_2, W, is named as the field of combinatorial ambiguity. There exists mutual inclusiveness between W_2 and W_1: The former includes the latter, while the latter is part of the former. Combinatorial ambiguity is one of the two kinds of analyses of Chinese word segmentation ambiguity.

Definition 2: Crossing ambiguity: If in character strings $W = a_1...a_ib_1...b_kc_1...c_j$, $W_1 = a_1...a_ib_1...b_k$, $W_2 = b_1...b_kc_1...c_j$, $W_a = a_1...a_i$, $W_c = c_1...c_j$, are the words W, is named as the field of W_1 and W_2. The length of ambiguity field is n = i+k+j. Crossing ambiguity is another important part of segmentation ambiguity.

Definition 3: Analysis tree of Ambiguity (analysis tree of ambiguity is set to present the selection process about the sequence out of ambiguity analysis).

Analysis tree of Ambiguity is a graph G. Graph G is consisted of tow sets: Set V and set E, denoted by:

$$G = (V, E)$$

V, the word sequence produced in analysis of ambiguity, is called as vertex for short in the following parts; E is the finite of edges; edges are the order dual pairs to determine the vertex; V meet the following two conditions:

- If $v \in V$, with $v_1 \in V$ and $(v_1, v) \in E$, then $v \in VE$, where the set $VE = \{x | x \in V \text{ and } (x, v) \in E\}$
- $\forall v \in V$, take $S = v$, if $v_e \in V$ and (v, v_e), let $S = v_e$; Repeat this action until there exist no and the same vertex could be concluded

Descriptive graph about the analysis process of ambiguity is accordance with the above definition of Analysis Tree of ambiguity. There are two steps during the process from word sequences of ambiguity to completing resolution. First, repeat detection about the word sequence of combinatorial ambiguity and calculate degree of segmentation with the analytic method of combinatorial ambiguity. Then, find the better one in the ambiguity sequences.

$$sg = \underset{cb(sg_1, sg_2)}{MAX}(MS(sg_1), MS(sg_2)) \quad (1)$$

After the better one finded, if sg_1 is the optimal among the selected parts, a new vertex sg is added into the Analysis Tree of ambiguity V with initial value sg_1 and edge (sg_1, sg) and edge (sg_2, sg) are added into E. By these steps, the elementary Analysis tree of ambiguity is set, in which MS is calculated based on the value of cb (sg_1, sg_2).

Second, calculating degree of segmentation about the rest word sequence based on crossing ambiguity method respectively and selecting the word sequence with the largest degree of segmentation as the result. This step could be defined as:

$$sg = \underset{x \in left}{MAX}(MS(x)) \quad (2)$$

Evaluation of word sequence: Evaluation phase is a process in which calculating segmentation degree about word group according to static information, context information and information about word group is needed. Therefore, three problems need to be solved in this step:

- Where do word sequences come from that need the calculation of segmentation degree?
- Which data are necessary and how do we get these data?
- How to calculate the segmentation degree?

We will discuss these problems gradually in following text:

- **The first question:** Calculating object the word sequences needing calculation are produced in the analysis of ambiguity during segmentation process. Discovery process of ambiguous word sequences has been explained in detail in Section 2.3.3. word sequences that produce in the analysis of ambiguity are the sequences needed to calculating. Each act of segmentation will produce several word sequences with corresponding segmentation degree MS_x, which is regarded as a value to measure the quality of segmentation. The normal form of word sequence producing MS_x is ABC, maxLen(D). B is the next forecasting word needing segmentation; A, as the previous word adjacent to B, has been analyzed; C, adjacent to B, is an auxiliary forecasting word; D is the third forecasting word; no integrated word needs to be provided except the length of the longest word adjacent to B
- **Information needed:** Two parts of information are needed in calculating SV: Dictionary information; context information, including context word frequency information and word group information. Word sequence selected here is named as word group to present the information source. Dictionary information come from statistic analysis of training sample:
  - Word frequency FRL, the frequency of the word among vast training samples. After the statistic analysis, proper smoothing is needed
  - t BU, the times of two words adjacent to each other in training samples. Proper smoothing is needed, too

Context information is as follows:

- FR in above paragraph, frequency of word
- single character condemnatory factor (SPF). SPF represents the sum of the number of single character and the number of adjacent words. For example, in the length of A, B, C and D in ABC, maxLen(D) is 1, respectively, then the number of single character is 4 and the number of adjacent time is 3, thus the value of SPF is 7
- Condemnatory factor of rarely used character Records the number of rarely used character and serves as punishment of rarely used characters in word sequence
- Calculation of segmentation degree. Although, it is very difficult to score partial words, properly when they are compared with full words, however,

statistical parsing associate scores with each decision in the parsing process, selecting the parse which is built by the highest scoring sequence of decisions. Assuming ABC maxLen(D), stored in X, the definition of segmentation degree produced by X is:

$$MS_X=(FR_X+B_X)^{1/K_X} \quad (3)$$

MS_x is the metric of segmentation degree about the whole word sequence. In this definition, the index in the above equation is the reciprocal of condemnatory factor; the radix is the sum of word frequency metric and adjoin metric. The index 1/k(x) stands for condemnatory factor. In the bracket, X stands for ABC, maxLen(D). FR stands for word frequency metric of word group. FR is used as comprehensive metric of counting word frequency in lexicon and that in above context of segmentation process. B stands for adjoin metric of adjacent words in lexicon.

Function definition of FR is as follows. FRC_{wi} in the equation stands for frequency of w_i in lexicon and w_i length stands for the length of w_i. AFRC stands for the mean frequency of words in lexicon. FRL_{wi} stands for the frequency of in above context. And AFRL stands for the mean frequency of words in above context. FR_x is a logarithmic equation, including two logarithmic parameters mean relatively length of word group in lexicon and that in above context. These two logarithmic parameters stand for the beneficial improvement about metric of segmentation degree. The logarithm formula should be a natural logarithm function.

Equation 4 stands for measurement of word frequency which is beneficial to the improvement of segmentation degree:

$$FR_x = \ln\left(\frac{\sum_{i=1}^{n}(FRC_{w_i}\times w_i.length)}{AFRC\times n}\right)+\ln\left(\frac{\sum_{i=1}^{n}(FRL_{w_i}\times w_i.length)}{AFRL\times n}\right) \quad (4)$$

B stands for the adjacent situation among word sequence and the calculative method is shown as B (X). BU (i, i+1) stands for the adjacent frequency of w_i and w_{i+1} in word group of the lexicon. w_i.length and w_{i+1} stand for their own length, respectively. ABU stands for mean adjacent frequency which is greater than 0 in lexicon. B_x stands for the measurement of neighborhood degree among word sequences. B_x is in direct proportion to two successive calculation of adjacent times and lengths of two words, while it is inversely proportional to mean adjacent times. Thus, with the increase of statistical data, the balance of the equation could also be kept:

$$B_x = \sum_{i=1}^{n}\frac{BU(i,i+1)\times(w_i.length+w_{i+1}.length)}{ABU\times(n-1)} \quad (5)$$

K_x in Eq. 5 stands for the value index penalty resulted from condemnatory factor. In definition equation, SPF stands for penalty coefficient function. If it is single character, the calculation returns to 1, or it will return to 0. Similarly, RPF stands for condemnatory factor function of rarely used character. If it is rarely used character, the calculation returns to 1, or it will return to 0. In this equation, C is a constant, determined according to the training result about training samples. K_x is a exponential function. Its radix is the sum of condemnatory factors of single character and rarely used character. Its index is the constant of (0,1), set according to training in training set. The bigger the index, the smaller segmentation degree:

$$K_x = pow\left(\sum_{i=1}^{n}SPF(Len(W_i))+RPF(Len(W_i), Len(W_{i+1})), C\right) \quad (6)$$

Review of algorithm: In this section, ambiguity analysis model of segmentation algorithm based on word group will be introduced as a whole. Besides, the description of adding pseudo-code is also presented in detail and some examples are listed to show the segmentation process of algorithm.

Brief introduction of algorithm: To solve AR problem in word segmentation and making a model for word segmentation, we designed our algorithm into four steps:

- **Step 1:** Pretreatment for word sequences in which word sequences are produced
- **Step 2:** Discovering and calculating segmentation degree about sequence of combinatorial ambiguity circularly, then analyzing the results
- **Step 3:** Dealing with the analysis of crossing ambiguity
- **Step 4:** Updating temporary lexicon

Code Tagging: P stands for the position of the ending position of last segmentation; E stands for the next position for interrupt of character sequence; the array of target[] stores the target forecasting words; assistFir[][] stands for auxiliary forecasting word I; assistFir[i][] stands for the corresponding array of auxiliary forecasting word to target[i]. AssistSec[][] stands for the maxlength of auxiliary forecasting word II; assistSec[i][j] stands for the maxlength of forecasting word following assisFir[i][j].

```
01:  If    P<E then
02:        seg(P, target[ ]);
03:        For I0 to Len(target)
04:                seg(target[I].p, assistFir [i] [ ])
05:        End for
06:        For i0 to Len(target)
07:         For j0 to Len(assistFir[i])
08:         segLen( i , j , assistSec)
09:                End for
10:        End for
11:        While combMean()
12:                resolveComb()
13:        P = resolveMix()
14:  end if
```

Pretreatment was consisted of code segments from 02 to 10, including three steps:

- **Step 1:** querying the dictionary to segment all target forecasting words. This step was carried out by the code: "seg()"
- **Step 2:** According to the position of every target forecasting word, the corresponding auxiliary forecasting word I was forecasted. This step was carried out by the code "seg()"
- **Step 3:** According to the position of every auxiliary forecasting word I, the maxlength of corresponding auxiliary forecasting word II was forecasted. This step was carried out by the code "segLen()"

For example: 听说大学生活像白纸 the results after pretreatment were shown in Fig. 1.

If the word group with long-word is segmented inaccurately, a lot of single characters are produced, such as the results "听说/大/学/生/活/像/白/纸" of MM. In this example, the penalty coefficient set was 4. For example, the penalty coefficient of "听说/大学生/活/像/白/纸" was 6. The penalty coefficient was designed for the segmentation of single character. The detail processes were as follows.

Data declaration: FR_x is a logarithmic formula, including two logarithmic parameters mean relatively length of word group in lexicon and that in above context. These two logarithmic parameters stand for the beneficial improvement about metric of segmentation degree. Thelogarithm equation should be natural logarithm function. This equation stands for measurement of word frequency which is beneficial to the improvement of segmentation degree.

As to the calculation of B_x and K_x, analysis method of combinatorial ambiguity and that of crossing ambiguity is the same, while that of FR_X is different form them. On basis of combinatorial ambiguity, $FRC_X = FR_Y$. During this equation, there is mutual inclusiveness between Y and X. For the example displayed above, there were two sequences: X = "听说/大学生/活/", Y = "听说/大/学生/活".

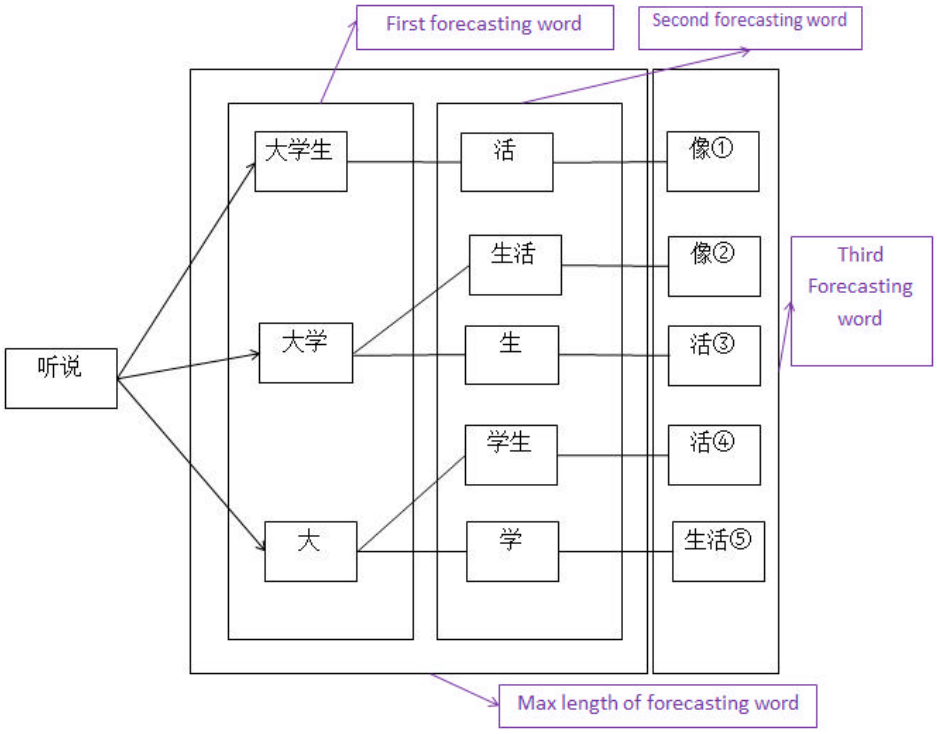

Fig. 1: Example of segmentation

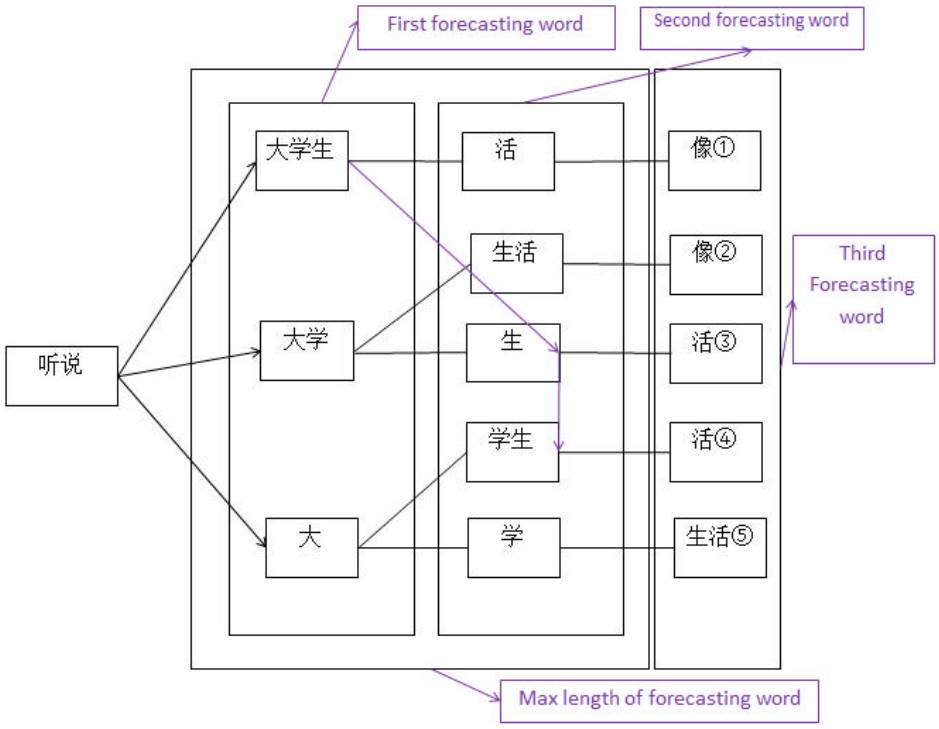

Fig. 2: Example of segmentation

According to the explanation of combinatorial ambiguity, in the field of 大学生, X included Y. In the process of calculating segmentation degree, $FRC_X = FR_Y$. According to the explanation of crossing ambiguity, $FRM_X = FR_X$ and there was no mutual inclusiveness exists.

According to the algorithm described above, the second step and third step of the example were as follows:

- **Step 2:** Word sequences produced by pretreatment was examined to find whether there is combinatorial ambiguity exists. ③④ is find to be a combinatorial ambiguity case. Choose the one with higher segmentation degree According to computing method of segmentation degree based on combinatorial ambiguity, the partially optimal sequence ④ has been selected; Then detection has been implemented again and again until that there was no combinatorial ambiguity case exists in the remanent word sequences
- **Step 3:** Calculating the segmentation degree of the rest sequences ①②④⑤ across to analysis algorithm of crossing ambiguity and choosing the largest segmentation degree. Than, the optimal result ② has been concluded

The Analysis tree defined was as follows:

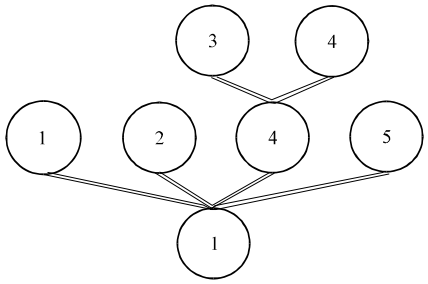

Fig. 3: Analysis tree of example about segmentation

As the definition of analysis tree shows, analytical algorithm of crossing ambiguity only was used in the last step (decision of ①②④⑤) and that of combinatorial ambiguity has been used in the first and second steps (decision of ③④).

TIME COMPLEXITY

- In the segmentation of a single word. In the period of atomic segmentation, there are at most 4 predicted words of each word. Twice predictions are needed. Therefore, the number of predicted words is 4*4 = 16. Then, to judge whether there is a word of a length

greater than 1 is determined. The number of each word' segmentation needing calculation is 16. As each calculation is worked out in the time complexity of constant, time complexity of a word is 16*O(1), also known as O (1)

- If the length of an article is "n", there are at most "n" words. Therefore, if the segmentation length of an article is "n", the time complexity of this article is n*O (1), known as O (n). That is:

$$\sum_{i=1}^{n} 4*4=16*n \in O(n) \quad (7)$$

The test result shown that the segmentation speed of this algorithm was 27 million characters per second. Segmentation speed of the main segmentation tool, such as that of massive Chinese intelligent segmentation is 20 million per second. And the segmentation speed of latest tool 2012.8 Pan gu segmentation is 19.4 million characters per second.

EXPERIMENT AND ANALYSIS

The environment of this test shows as below: Windows 7 with sp1, cpu E5400, RAM 2g, language java.

Parameter setting: In the test, the value of pns was very essential to the result of the test. With the change of pns, the result of segmentation would be different, including the accuracy and recall rate. By experiment, pns was selected as 0.07. The results were as following:

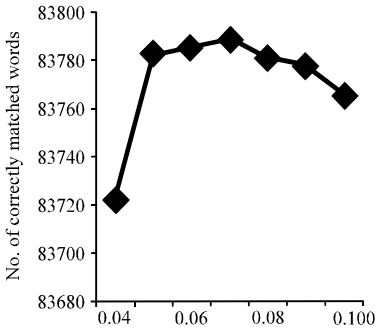

Fig. 4: Changes of segmentation results resulting 1

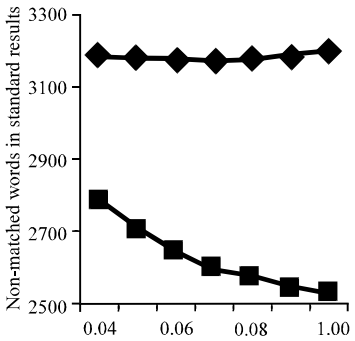

Fig. 5: Changes of segmentation results resulting

During the procedure, pns was an important parameter to single characters in the optimum sequence. It was set to compute the influence produced by condemnatory factors k; the bigger value, the less number of single characters in segmentation results. That was the reason why Graph 4 was a decreasing function. Meanwhile, when the number exceeds a specific value, the accuracy would be decreased. There was an optimal point in this function. The number of correctly matched words and that of non-matched words in standard results were regarded as the main reference standard for effects of segmentation. According to the graph, when was 0.07, the number of correctly matched words reached its peak while that of non-matched words in standard results was the least value.

RESULTS

First we make compare to a famous competition of Chinese word participle. Data used in this test were downloaded from the website of SIGHAN (webset: www.sighan.org), published by SIGHAN Bakeoff-2. SIGHAN Bakeoff is a popular contest for language processing. As said by Changning Huang and Hai Zhao in REFERENCES document (Huang and Zhao, 2007). We have conducted test against Peking University test data in SIGHAN Bakeoff contest. This part will introduce the steps and results of the test.

The corresponding lexicon was concluded by the training samples. Smooth processing about the lexicographic data has been made later. Input the selected data, The results of Peking University's test data shows in Table 1.

The evaluation criteria adopted in this test were accuracy P, recall rate R and measure-valued F. The computational formulas were as follows:

$$p\ (\%) = \frac{\text{The No. of words that are correctly segmented}}{\text{The No. of words that are segmented}} \times 100 \quad (8)$$

$$p\ (\%) = \frac{\text{The No. of words that are correctly segmented}}{\text{The No. of words in the reference}} \times 100 \quad (9)$$

$$F = \frac{2 \times P \times R}{P+R} \quad (10)$$

According to the results of the test, accuracy R, recall rate R and measure-valued F of Our model were the highest among those of attendances in SIGHAN Bakeoff-2. In the Bakeoff-2 contest, most participant use statistic model can beat the ones that based on manual

Table 1: Comparison of test results

Attendance	R Score	P Score	F Score
Our's	0.97	0.964	0.967
Aitao Chen	0.953	0.946	0.95
Yahoo! Inc\Huihsin Tseng	0.946	0.954	0.95
Stanford NLP Group			
Huipeng Zhang	0.952	0.945	0.949
Information Retrieval Lab, Harbin Institute of Technology			
Huipeng Zhang	0.952	0.943	0.947
Information Retrieval Lab, Harbin Institute of Technology			
Chen Jiajun	0.941	0.95	0.946
Nanjing University			

Table 2: Speed of this model and some commonly used tools

Tools	Speed (thousand word per second)
Paoding	100
Imdict	30
IK analyzer	60
Je-anlysis	30
This model	2700

rules and this is mainly because manual rules can hardly summarize all regulation among word in sentences, but statistical model can distinguish details gradually. As we combine semantics model , dictionary model and manual rules, we get superiority in this aspect. The results shown that our model had a good p performance on precision in word segment.

As time needed of algorithm in the competition above was not recorded, we make another simple compare to some popular Chinese word participle tools. Paoding is a familiar tool used for Chinese word segment. imdict, IKAnalyzer and je-anlysis are also common tools used in luc lucene or other projects. Their speed of word segment show in the form Table 2.

Imdict is based on HMM model, of which the time complexity is State×L^2. In the equation, state representate the State number of HMM model and L means the length of the sentence to be segmented. Obviously, it is not coincidence that imdict has a low speed. Paoding split sentence into any possible words, than choose the best on among them and it is also not comparable to O (n) in time complexity. IK Analyzer has the time complexity of Exponential growth. Because of the excellence in time complexity, our model has a good representation in speed.

An appropriate pns make contribution to precision of "Ambiguity analysis Model of Word Segmentation Based on Word Group". In the test on data provided by SIGHAN Bakeoff-2 ahead, we get a good accuracy in Chinese word segment. The speed data presented in the Table 2 accord with Time complexity analysis section.

CONCLUSION

Words are the basic units to process for most NLP tasks (Zhang and Clark, 2010), therefore segmentation is a significant procedure for dealing with Natural Language Processing in the language where there is no separator between words (Zhang *et al*., 2010). Inappropriate word segmentation, would bring about great impact on the data processing, such as Chinese search engine (Song *et al*., 2012). In this study, a type of Chinese segmentation model has been described. When get different word sequences, we use Analysis Tree to find the prime sequence for Ambiguity analysis. Three important merits of this model described were as follows: First, this algorithm has made full advantage of context information to provide more segmentation information for ambiguity analysis processing and neologism discovering when making segmentation; second, the combination of statistical segmentation and dictionary segmentation has resulted in the quick speed of segmentation and relatively high accuracy; third, this algorithm was accompanied with certain learning ability and the accuracy could be improved in a better training formwork. In subsequent work, some useful characteristics of words could be added in the field of segmentation and the learning ability and efficiency of the algorithm could be improved as well.

ACKNOWLEDGMENT

This study is partly supported by the Science Foundation of Zhejiang Province under Grant No. 2010R50009, People's Republic of China.

REFERENCES

Huang, C.N. and H. Zhao, 2007. Ten-year review about Chinese segmentation. J. Chin. Inform. Process., 21: 2-16.

Li, Z., 2011. Parsing the internal structure of words: A new paradigm for chinese word segmentation. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, June 19-24, 2011, Portland, ON., USA., pp: 1405-1414.

Song, C., S.Y. Zhao and X.Z. Zhou, 2012. Algorithm research of segmentation technology in vertical search engine. Comput. Technol. Dev., Vol. 2.

Wang, H.S., J. Zhu, S. Tang and X. Fan, 2011. A new unsupervised approach to word segmentation. Comput. Linguist., 37: 421-454.

Yue, J.Y., J.A. Xu and Y.Y. Zhang, 2012. Chinese word segmentation technology for patent documents. Peking University Press, Beijing.

Zhang, C.P., L.L. Zhao and C.M. Wu, 2010. Method of Chinese word segmentation based on character-word classification. J. Comput. Appl., 30: 2034-2037.

Zhang, Y. and S. Clark, 2010. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. Proceedings of the Conference on Empirical Methods in Natural Language Processing, October 9-11, 2010, MIT Stata Center, Massachusetts, USA., pp: 843-852.