National Conference on Information Technology and Computer Science (CITCS 2012)

The Research of the Maximum Length n-grams Priority Chinese Word Segmentation Method Based on Corpus Type Frequency Information

Lu Pengyu
School of Management
Harbin Institute of Technology
Harbin, China
lupengyu@hit.edu.cn

Jin Lijun
School of Management
Harbin Institute of Technology
Harbin, China
heyejunjun@163.com

Jiang Bin
School of Management
Harbin Institute of Technology
Harbin, China
956563910@qq.com

***Abstract*—In order to solve the difficulties to extract words in particular domain, we formulate a method of automatic word segmentation in Chinese based on corpus type frequency information. This method can effectively extract n-gram words that are not predefined in a lexicon by setting the maximum length (n) of the n-gram word we want to extract from a sentence and the minimum threshold frequency the n-gram word appears in corpus. When the real frequency the n-gram appears in corpus is above the threshold, the n-gram word will be extracted. If there are two or more n-grams have the same length, the higher frequency one will be chosen, and then the next higher frequency one if any of its characters are not in previous one.**

Keywords-word segmentation; word frequency; n-gram word; corpus type frequency information

I. INTRODUCTION

With the dramatically rising of user created contents (UGCs) on Internet, managers can improve marketing strategies and increase enterprise income by analysing UGCs; and consumers can make the right purchasing decisions according to the comments on commodities. And also, government departments are able to make policies and improve services by observing people's behaviours intentions based on UGCs. All these works need the help of Chinese text processing.

Chinese word segmentation is one of the key parts in Chinese text processing. The accuracy of word segmentation will directly affect the efficiency of text processing [1]. Different from English texts, Chinese texts have no distinct marks in the written or computer, which makes automatic segmentation more difficult. In recent years, the research of Chinese word segmentation method has made great progress, the precision and recall rate have also reached a high level. However, when texts involve a particular domain, the word segmentation accuracy rate is usually very low basing on traditional segmentation method. The main reason is that the granularity of the word itself is large, but they are segmented into smaller size, some long length n-gram words cannot be properly extract. Therefore, how to extract maximum length n-gram words is the key to improve the segmentation accuracy [2]. There are few researches on segmentation algorithm for the long length n-gram words in Chinese. This paper proposes a maximum length n-grams priority Chinese word segmentation method based on corpus type frequency information. This method can effectively get n-gram words by setting the maximum length of n-gram word and the minimum threshold frequency the n-gram word appears in corpus, and matching the input texts with the maximum length by word frequency in descending order, this method can extract long terms effectively. Compared to other segmentation methods, our method can improve the accuracy of n-gram word segmentation system.

This paper is organized as follows: Section 2 reviews some related researches of word segmentation methods based on word frequency. The maximum length n-grams priority Chinese word segmentation method based on corpus type frequency information is described in Section 3. A walk-through example of the method is presented in Section 4. Evaluation and experimental results are discussed in Section 5. Section 6 is summary and prospect, points out the innovation and deficiencies of this method and the potential applications of the proposed method.

II. RELATED WORK

Chinese word segmentation can be classified as dictionary-based method, statistics-based method, and hybrid method based on statistic and dictionary. The dictionary-based method is easy to segment words and its precision depends on the coverage of the dictionary; but it is difficult to identify unknown words and solve the ambiguity problem. Statistics-based method is able to reflect the credibility of words by the co-occurrence frequency or probability of adjacent words, such as mutual information based segmentation method and t-test based segmentation method [3]. Compared with dictionary-based method, statistics-based method is more effectively in identifying unknown words and getting rid of ambiguities, but its identifying precision is lower for common words. Up to now, a mature segmentation system generally combines different kinds of algorithms, in order to improve the precision and accelerates the speed of segmenting word at the same time [4, 5].

Heng Zhang, Wenzhao Yang et al. [4] proposed a Chinese word segmentation method based on the dictionary with word frequency. First, they carry out Chinese words auto-

Supported by National Natural Science Foundation of China (71071041)，Nature Scientific Research Innovation Foundation in Harbin Institute of Technology in China (HIT.NSRIF.2009109).

segmentation basing on Chinese dictionary. Then, they improve the precision by eliminating ambiguous word according to word frequency. Hongyan Cui[6] proposed an improved FMM algorithm with word frequency. According to word frequency, the algorithm extracts the 2-gram words at first, then segment the n-gram (n>2) words by putting them into the suffix set of 2-gram word root. Thus, the segmentation is speed up by transferring the global maximum matching into local variable length maximum matching in FMM. Xiaojuan Zhu, Tefang Chen et al. [7] proposed a Chinese word segmentation method based on word frequency and SVM. Through the system, continuous character bunch input can be segmented, and then the cut apart word bunch output can be gotten, the cut apart word bunch usually is two character words bunches, and one dictionary can be gotten. The dictionary stores word and the frequency that the word appears in these disposal tests. The segmentation system selects Mutual Information to statistic. Use SVM, the veracity of segmentation was better than the traditional method, and is of high stability. Qinyi Zhao et al. [8] proposed a method of string-scanning Chinese word segmentation based mutual information, this method can identify new words, eliminate ambiguity automatically, and have a good precision and recall rate. Huang Wei et al. [9] proposed a word segmentation method in feature selection in Chinese text categorization, this method segments word by calculating mutual information between two lexical entries and adjacent frequency of two or more lexical entries. Zhongjian Wang et al.[10] proposed a Word Segmentation Method Based on Inductive Learning and Segmentation Rule. The method extracts recursively a character string that occur frequently in text as word candidates, extracts segmentation rule with context information to deal with segmentation ambiguity. The method classifies those extracted word candidates to different ranking according to extraction situation, segments a text into words with extracted word candidates. In order to solve the problem that the longer words cannot be segmented correctly and be matched repeatedly, Ruilei Wang et al. [11] put forward idea for improving FMM algorithm that is to assign the maximum text-length to be treated dynamically based on the word length in Chinese word segmentation word bank. To fit this，they designed a word bank structure to enable the effective support on the improvement of FMM . Compared to normal FMM, the improved FMM sharply reduces matching times, and the speed and efficiency of Chinese Word segmentation algorithm have been obviously improved . Md. Aminul Islam et al. [12] proposed a generalized approach to word segmentation using maximum length descending frequency and entropy rate. The method uses corpus type frequency information to choose the type with maximum length and frequency from “desegmented” text. It also uses a modified forward-backward matching technique using maximum length frequency and entropy rate if any non-matching portions of the text exist. This method can extract n-grams effectively, but it does not consider the characteristic of Chinese texts, so the actual results are unsatisfied.

III. PROPOSED METHOD

Step1 Initialization：Establish table of the stop words, including copula, particle, modal particles and so on. Input the maximum length m of the character string word which is the character number contained in the word we want to extract. Meanwhile, according to Chinese characteristics, we set the minimum length of a word is 2. That is to say, the minimum segmenting word unit is 2-gram. Set up a set of the minimum word frequency threshold $T_m=\{t_m,t_{m-1},t_{m-2},\cdots,t_2\}$. In the set, t_m is the threshold of m-gram words. It means that if the occurrence frequency in the corpus of a m-gram word is more than the m-gram threshold t_m, the m-gram word will be extracted as a segmented word.

Step2 Input text S which need to be segmented (assume that it is pure Chinese text). Let the final set of the segmentation result is S'. Initially, $S'=\phi$.

Step3 Extract character string from the first Chinese character to the first Non-Chinese character (symbol, blank and so on) in S and make up a set of character string $C=\{c_1c_2c_3\cdots c_n\}$. In set C, n is the Chinese characters number contained in the character string. That is to say, c_n is an n-gram word. At the same time, get rid of C from S, put the first Non-Chinese character into the extracted word set C', and get rid of the first Non-Chinese character from S. Suppose the first Non-Chinese character is ‘，’, then $C'=\{，\}$.

Step4 Extract the stop words and put them into set C'. Whilst, the stop words in C are substituted as separator ‘/’. Suppose c_i is a stop word in C, then $C'=\{，c_i\}$, $C=\{c_1c_2c_3\cdots c_{i-1}/c_{i+1}\cdots c_n\}$.

Step5 Extract character string W from the first Chinese character to the first Non-Chinese character (symbol, blank and so on) in C, and get rid of W from C. Count the length of the character string to n'.

Step6 If $n'<m$, set $m=m-1$. If $n'\geq m$, extract the m-gram character strings from left to right in turn. It can be proved that the number of segmenting text stings that can be calculated by $n'-m+1$. Let set $W'=\{w_k|w_k=c_kc_{k+1}c_{k+2}\cdots c_{k+m-1}\}$, $k\in N\ and\ k\in[1,n'-m+1]$. Suppose c_i is a stop word in the original text, then the first character string extracted from C is $c_1c_2c_3\cdots c_{i-1}$, and $n'=i-1$. Suppose the maximum length of the segmenting word is m, then $W'=\{w_k|w_k=c_kc_{k+1}c_{k+2}\cdots c_{k+m-1},k\in N\ and\ k\in[1,i-m]\}$.

Step7 Acquire the occurrence frequency of w_k in corpus, and setup a set of word frequency f_w. let $f_w=\{f_{w_1},f_{w_2},\cdots,f_{w_k},\cdots,f_{w_{n-m}},\}$, f_{w_k} represents the occurrence frequency of w_k in corpus. If f_{w_j} is more than the threshold of m-gram word, that is $f_{w_j}\geq t_m$, put w_j into the extracted words set C'', and get rid of w_j from the set W'. Thus, $C''=\{c_i,c_jc_{j+1}c_{j+2}\cdots c_{j+m-1}\}$, and $W'=\{w_k|w_k=c_kc_{k+1}c_{k+2}\cdots c_{k+m-1},k\in N\ and\ k\in[1,j)\cup(j,n'-m+1]\}$. If $f_{w_j}\geq t_m$ and $f_{w_{j+1}}\geq t_m$ occurs at the same time, we will choose the segmenting word with larger occurrence frequency in corpus. Suppose $f_{w_j} > f_{w_{j+1}}$, Then w_j will be put into C'' , and it will also be deleted from W' .

Step8 After the m-gram words being extracted from W' , set $m = m - 1$. Loop step6-7 to extract new segmenting words until $m = 2$. Set $C'' = C'' \cup W', C' \cup C''$.

Step9 Loop step5-8 until $C = \phi$. Arrange all words in C' in the original text order, and separate these words from each other by "/". Set $S' = S' \cup C'$.

Step10 Loop step3-9 until $S = \phi$.

IV. A Walk-through Example

Step1 Initialization. Establish table of the stop word, assuming stop word containing '的','是','不是','了', and so on. Set $m = 6$, that is to say, the longest word contains six Chinese characters. $T_6 = \{5,20,100,500,2000\}$, that is, in corpus, the minimum occurrence frequency of the words containing six Chinese characters t_6 is 5, the minimum number of occurrence frequency of containing five Chinese characters t_5 is 20, ⋯, the minimum number of occurrence frequency of containing two Chinese characters t_2 is 2000.

Step2 Input text S which we need to segment. Suppose S ={针对信息管理与信息系统专业培养多学科交叉复合型人才的特点，设计了创新的人才培养模式。}. Initially, the final set of segmentation words $S' = 0$.

Step3 Extract character string from the first Chinese character to the first Non-Chinese character (symbol, blank and so on) in S . Then we get a set of character string C = {针对信息管理与信息系统专业培养多学科交叉复合型人才的特点}. Whilst, get rid of C from S . Put the first Non-Chinese character '，'into t extracted word set C' and get rid of it from S . Then $C' = \{,\}$， S ={设计了创新的人才培养模式。}.

Step4 Extract the stop word '的' from C and put it into C' . At the same time, the stop word in C is substituted as a separator '/'. Thus, C' = {，的}，and C = {针对信息管理与信息系统专业培养多学科交叉复合型人才/特点}.

Step5 Extract character string W in C from the first Chinese character to separator '/'(except '/') and get rid of W and '/' from C . Count the characters in W to n' as the length of the character string W . Then W = {针对信息管理与信息系统专业培养多学科交叉复合型人才}, C = {特点}, and $n' = 25$.

Step6 Suppose $m = 6$ ， thus $n' > m$. Segment character string from W in turn, and let the length of character string equal to m . These character strings make up the set W' = {针对信息管理/对信息管理与/信息管理与信/……/交叉复合型人/叉复合型人才}. The number of character strings in W' is 20, which can be calculated by $n' - m + 1$.

Step7 Get the occurrence frequency in the corpus of each character string in W' and make up the word frequency set f_w . Then $f_w = \{f_{w_1}, f_{w_2}, \cdots, f_{w_{20}},\}$. In f_w , $f_{w_6} = 1$ ， $f_{w_8} = 2$ and others all equal to 0. Because the threshold $t_6 = 5$, there is none f_{w_i} satisfies $f_{w_i} \geq t_6$, no segmented character string is extracted.

Step8 Set $m = m - 1 = 5$ and loop step6-7 to extract segmented words. Thus, W' = {针对信息管/对信息管理/信息管理与/……/与信息系统/信息系统专/息系统专业/……/培养多学科/养多学科交/多学科交叉/叉复合型人/复合型人才}, the number of character strings in W' is 21, and the word frequency set $f_w = \{f_{w_1}, f_{w_2}, \cdots, f_{w_{21}}\}$. In f_w , $f_{w_3} = 3$, $f_{w_6} = 1$, $f_{w_7} = 6$, $f_{w_8} = 2$, $f_{w_8} = 2$, $f_{w_9} = 2$, $f_{w_{14}} = 1$， $f_{w_{16}} = 38$， $f_{w_{21}} = 112$， and others all equal to 0. The threshold $t_5 = 20$， $f_{w_{16}} = 38 > t_5$， $f_{w21} = 112 > t_5$, as a result, w_{16} and w_{21} are extracted as two segmented words. Then, put w_{16} '多学科交叉' and w_{21} '复合型人才' into extracted words set C'' with separator '/', and get rid of them from the set W' at the same time. Thus, C'' = {多学科交叉/复合型人才}, and W' = {针对信息管/对信息管理/信息管理与/……/与信息系统/信息系统专/息系统专业/……/培养多学科/养多学科交/学科交叉复/科交叉复合/交叉复合型/叉复合型人}. Reduce the length of character string to be extracted in turn until that all 2-gram character strings are extracted. Thus, C'' = {多学科交叉/信息管理/信息系统/针对/专业/培养}, and W' = {与/复合型人才}. Set $C'' = C'' \cup W'$, then C'' = {多学科交叉/信息管理/信息系统/针对/专业/培养/与/复合型人才}. Set $C' = C' \cup C''$, then C' = {，/的/多学科交叉/信息管理/信息系统/针对/专业/培养/与/复合型人才}.

Step9 Loop step5-8 until $C = \phi$. Then, C' = {，/的/多学科交叉/信息管理/信息系统/针对/专业/培养/与/复合型人才/特点}. Arrange all words in C' in of the original text order. Thus, C' = {针对/信息管理/与/信息系统/专业/培养/多学科交叉/复合型人才/的/特点/，}. Set $S' = S' \cup C'$, then S' = {针对/信息管理/与/信息系统/专业/培养/多学科交叉/复合型人才/的/特点/，}.

Step10 Loop step3-9. Then, C = {设计了创新的人才培养模式}. The final extracted words set C' is {设计/了/创新/的/人才培养/模式/。}. Set $S' = S' \cup C'$, then S' = {针对/信息管理/与/信息系统/专业/培养/多学科交叉/复合型人才/的/特点/，/设计/了/创新/的/人才培养/模式/。}.Continue to loop step3-9 until $S = \phi$. Then, the final extracted words set S' is {针对/信息管理/与/信息系统/专业/培养/多学科交叉/复合型人才/的/特点/，/设计/了/创新/的/人才培养/模式/。}.

V. Evaluation and Experimental Result

The performance of word segmentation is usually measured using precision and recall rate. Recall rate is defined as the percentage of words in the manually segmented text identified by the segmentation algorithm, and precision is defined as the percentage of words returned by the algorithm that also occurred in the manually segmented text in the same position. In general, it is easy to obtain high performance for one of the two measures but relatively difficult to obtain high performance of both. F-measure (F) is the geometric mean of precision (P) and recall(R) rate and expresses a trade-off between those two measures. So we use precision, recall rate and F-measure to evaluate the accuracy of segmentation algorithm.

This experiment use news texts as testing data, totally 94.6M, involving 19 different areas, such as literature, art, education, philosophy, sports, transportation, economic, law, medical, and so on. We use CCL corpus (Online corpus established by Peking University Chinese Language Research Center,http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai) to count word frequency, the size of the corpus is 477 million characters (1.06GB). The maximum length of word in this experiment is 8-gram, the threshold set is {1500000, 2000, 800, 250, 80, 10, 6, 4}. That is, 1-gram can be extracted as an independent word only when its occurrence frequency in the corpus reaching 1.5 million. A 2-gram word will be extracted only when its occurrence frequency in the corpus reaching 2,000. In turn, when a 8-gram word is extracted, its occurrence frequency in the corpus should reach 3 times. The experiment results show that the average precision rate of this method is 77.7%, recall rate is 74.3%, and the F-measure value is 0.76. When texts involve a particular domain, this method can effectively get n-grams, the precision and recall rate can come up to 90% or more, the F-measure value can also reach 0.9.

Analysing the experiment results, the word segmentation errors is mainly caused by stop words, numerals, some typos, thresholds, and so on. For example, the errors related to numerals, such as “是一个”, “是一批”, are due to the different quantifiers which lead to the frequency of “是一” is much higher than the frequency of “一个” or “一批”.As a result, the segmentation consequence becomes “是一/个”, “是一/批”, and it influences the accuracy of word segmentation. In addition, the chosen stop words can also affect the segmentation result. For instance, if we extract “是” as a stop word, the previous example “是一个”, “是一批” will be segmented into “是/一个”, “是/一批” correctly. But the words related to “是”, such as “但是”, “可是”, “而是”, and so on, will be segmented into “但/是”, “可/是”, “而/是” incorrectly. In order to ensure the accuracy of word segmentation, all the word related to “是” should be put into stop word sets. But if the stop words are set too large, the efficiency of the segmentation procedure will be descended.

VI. CONCLUSION

The segmentation method proposed in this paper is the improvement of the original segmentation method based on word frequency. The corpus used by the algorithm records the number of occurrences of each word; this is the result of counting a large number of documents. Therefore, the word frequency of the corpus can become the basis for judging a word. This word segmentation method combines word frequency with long-term priority principle. It can effectively get n-grams and have a high identification of proper nouns and technical terms. This method is very practical. However, this algorithm in the experiment needs to use the network corpus; it affects the rate of word segmentation in some degree. Moreover, the content of the texts collected by the corpus also has influence on word precision and recall rate, especially texts in particular domains. If the corpus has few texts in some related domains, it will reduce the rate of identification of technical terms. In addition, stop words and threshold also affect the results of word segmentation. In order to improve the segmentation precision and recall rate, we will improve the algorithm by integrating our method with dictionary-based word segmentation method in the future.

REFERENCES

[1] Hanshi Wang, Jian Zhu, Shiping Tang and Xiaozhong Fan, “A New Unsupervised Approach to Word Segmentation,” Association for Computational Linguistics, 2011: 422-454.

[2] Zhuoming Liang and Juhuan Chen, “Fast Chinese Word Segmentation Based on Proper Nouns,” Computer Technology and Development(In Chinese) ,2008(18): 24-27

[3] Ping Chen, Xiaoxia Liu and Yajun Li, “Chinese Word Segmentation Based on Dictionary and Statistics,” Computer Engineering and Applications(In Chinese). 2008(44): 144-146

[4] Heng Zhang, “Chinese Word Segmentation Based on Dictionary and Frequency of the Words,” Microcomputer Information(In Chinese),2008(24): 239-240

[5] Qiyu Zhang, Ling Zhu and Yaping Zhang, “A Survey of Chinese Word Segmentation Algorithms,” Intelligence Research(In Chinese), 2008(11): 53-56

[6] Hongyan Cui, “Research on An Improved Chinese Segmentation Algorithm Based on Word Frequency Statistic,” Information Technology(In Chinese), 2008(4): 124-125

[7] Xiaojuan Zhu and Tefang Chen, “Study on Chinese Word Segmentation Based on Statistic and SVM,” Microcomputer Information(In Chinese), 2007(23): 205-207

[8] Qinyi Zhao and Lizhen Wang, “A Method of String-Scanning Chinese Word Segmentation Based on Mutual Information,” Journal of Intelligence (In Chinese), 2010(29): 161-162

[9] Huang Wei，Bing Gao et al. “Study on Method of Word Segmentation in Feature Selection in Chinese Text Categorization,” Third International Conference on Knowledge Discovery and Data Mining, 2010: 411-415

[10] Zhongjian Wang, Kenji Araki and Koji Tochinai, ”Word Segmentation Method Based on Inductive Learning and Segmentation Rule,” 2008 International Symposium on Computational Intelligence and Design, 2008: 96-98

[11] Ruilei Wang, Jing Luan, Xiaohua Pan and Xiupei Lu, “An Improved Forward Maximum Matching Algorithm For Chinese Word Segmentation,” Computer Applications and Software(In Chinese), 2011(28): 195-197

[12] Islam M, Inkpen D and Kiringa I, “A Generalized Approach to Word Segmentation Using Maximum Length Descending Frequency and Entropy Rate,” Computational Linguistics and Intelligent Text Processing, 2007: 175-85