Moving Integrated Product Development to Service Clouds in the Global Economy
J. Cha et al. (Eds.)

doi:10.3233/978-1-61499-440-4-233

Word Segmentation Algorithm on Procedure Blueprint

Jianbin LIU[1], Duo YAO and LiRui YAO
Computer School, Beijing Information Science & Technology University, Beijing, China

Abstract. Word segmentation plays a most important role in the semi-auto conversation process from logic structure to implementation structure. This paper is intended to design a segmentation method for Procedure Blueprint. And this design will be illustrated by the point of long words priority, reverses matching principle and characteristic of Chinese language center rear and the reduction of cross-shaped ambiguity. Meantime, based on the Procedure Blueprint and computer implementation, we are planning to improve the dictionary organization structure, the Segmentation method and the participle efficiency.

Keywords. Procedure Blueprint; Segmentation; Dictionary organization; reverse maximal matching

Introduction

According to Professor Jianbin Liu, Procedure Blueprint (PB) is a kind of visualized modeling language composed of three-layer external view, two-level mapping and unity internal structure. These views are Abstract Concept Structure Diagram (ACSD), Abstract Logic Structure Diagram (ALSD) and Abstract Implementation Structure Diagram (AISD). In ACSD and ALSD, PB uses natural and limited natural languages to describe concept algorithm, behavior process and program structure of programming language [1]. However, the different i ， AISD uses programming language operation expression. The semi-auto conversation mainly concludes three steps, natural language segmentation, semantics analysis and conversation. Its detailed process is shown in figure 1.

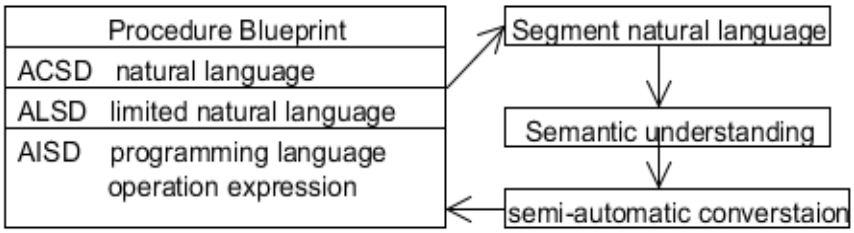

Figure 1. Semi-automatic conversation flow diagram

[1] Corresponding Author

The main concept and algorithm of these segmentation methods is the basis and key of semantics analysis, by which we can split the natural language [2]. PB field words are relatively concentrated. Meanwhile the reverse matching word segmentation method can be completely realized.

1. Related concepts

The Chinese word segmentation is that computer cuts Chinese sentences into a series of independent and meaningful words in accordance with certain words segmentation algorithm. However, the PB Word segmentation is a practical application of Chinese word segmentation technology. Till now, this technology has been deeply researched many times, certainly, the fruitful results has been greatly achieved, which is a very important guidance to our studies.

At present, the used word segmentation methods are mechanical segmentation method and segmentation method based on rules and statistics [3]. The mechanical segmentation method algorithm can be simply implemented, and the participle efficiency is high, but it lacks of disambiguation process. The Segmentation method based on rules is difficult to establish and tedious to deal with conflicts and short of flexibility. However, the Segmentation method based on statistics is easier to reach a higher accuracy with a large number of training corpus and segmentation rates related to algorithm and search space.

The basic idea of forward maximum matching algorithm is that: assuming the largest words containing in Chinese characters is n, taking the Chinese characters from the pending sentences before n and then [4] [5] matching with the words in the segmentation dictionary. If the dictionary still contains the n, the match is successful and those Chinese characters are cut into a word; picking up n from n+1, and then matching. If these are no such word, the last Chinese character of the n should be removed; and the operation should be repeated until the matching is successful. The basic idea of the reverse maximal matching is similar with forward matching, but their different is that the segmentation of the reverse maximal matching is begun from the end of sentence; if the match is failure, then the first Chinese character of the n should be removed.

2. Segmentation method in PB

Word segmentation algorithm based on the statistics is complex and requires of training corpus. Considering there is no such kind of trained corpus, this paper collects the source files of PB in recent years. One of the linguistics experts state that most of PB users is difficult to describe their thinking, and is lack of training corpus, knowledge reserves and professional level. However, the mechanical word segmentation method, which is broadly used in certain domain, is easily implemented and achieved its high efficiency. This paper adopts developed reverse maximal matching mean in segmentation.

2.1. Dictionary design

The structure of word dictionary is very important to the word segmentation efficiency [6]. As for the mechanical segmentation method, the word dictionary is the most important basis for word segmentation and the number of the segmentation words depends on the dictionary vocabulary.

The traditional dictionary is built by text, and its data is simply displayed without effective organization [7]. We define the O(n) (n is the number of entries in the dictionary) as the time complexity of search. If the dictionary is well –organized, and then the comparisons times and match times will be reduced in a great extent. This paper adopts the reverse matching method and is accord with the last letter of characters to make the sequence and create the index. For example, the index of "variable(▪量)", "constant(常量)", "weight(重量)" is created as "amount(量)". Assuming it has N entries, M is the number of index, and Qi is the number of entries under i word of index. Usually, one time maximum matching times in traditional dictionary is N, and in the worst case, the matching number become M + Qi (N =Q1+Q2+ …+ QM) after creating index. Make the sequence of the entries according to the length of the vocabulary ensures the reduction of matching times of pending vocabulary and dictionary. The dictionary organization is shown in figure 2.

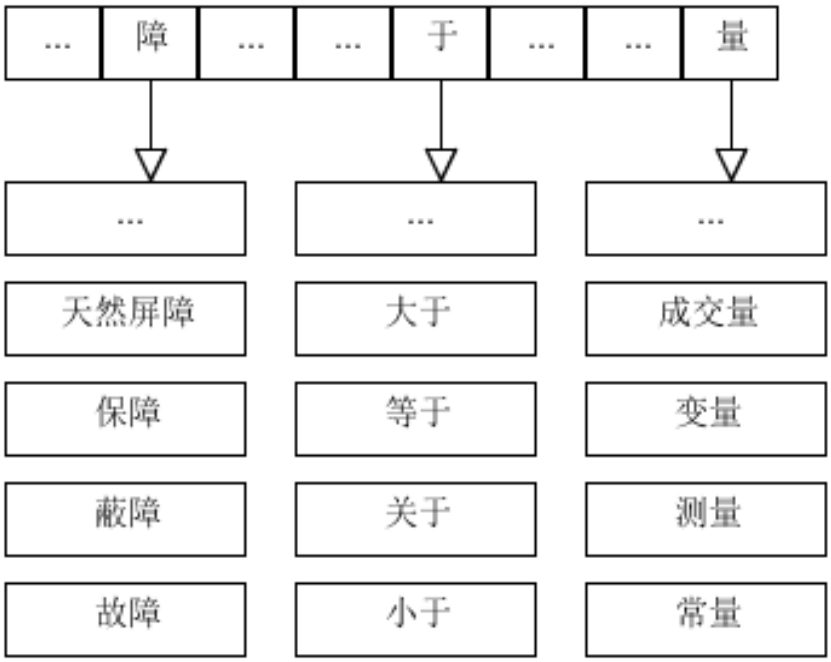

Figure 2. Dictionary Organization.

2.2. Words segmentation algorithm

The above viewpoint expresses the usage of dictionary structure in word segmentation algorithm and the relationship of Organizational structure with Mechanical word segmentation algorithm. The following point will describe the algorithm combined with the above dictionary organization.

Defining the Limited natural statement is "$W_1W_2W_3\ldots W_n$", we show the segmentation steps as following.

(1) Obtaining those sentences which is needed to be analyzed and judging its word strings .If the word strings is NULL, match over ,otherwise ,turns into (2);

(2) Obtaining the length of the sentence and naming it as 'sentenceLenth', assuming that 'dictionaryLen' is number of Chinese characters of the largest entry in dictionary, then taking the smaller one as the length of sentence will be of participle, named as 'Len';

(3) Initializing the location sign i=n, and getting Wn (that represents the last Chinese character) and assigning 'W_n' to W_{pos};

(4) Judging whether W_{pos} is a Chinese character, if the answer is true, then (6) will be carried out, otherwise (5) will be carried out;

(5) Storing the letters or digits into temporary phrase, i minus one and assign Wi to W_{pos};

(6) Cutting out 'Len' from the last character in the pending sentence, naming as pending 'initial Word'. Then 'initialWord' = $W_{(i\text{-}length+1)}W_{(i\text{-}length)}\ldots W_{(i)}$.

(7) Matching Wpos index with those words ended with Wpos and then using index to quickly locate. If the 'initialWord' can be searched in the dictionary, then it is divided into a word, and setting i=i-Len, $W_{pos}=W_i$, (4) will be carried out. If the 'initialWord' can't be searched in the dictionary, then the first Chinese character in 'initialWord' is removed and matching Wpos index with those words ended with Wpos. The option has to be repeated until the correct segmentation is achieved or length of the pending sentence becomes one.

(8) Cutting words out, the remaining part of the sentence is regarded as a new one to be processed. If the length of sentence is one, then the word segmentation is finished. Otherwise (1) will be carried out.

2.3. Word segmentation algorithm advantages

- Participle efficiency

  Statistical Word segmentation method needs training corpus support. Its segmentation speed is impacted by the algorithm complexity of time and the overhead expense of space. Mechanical word segmentation method is easily implemented, its segmentation speed is fast, and it is appropriate for certain participles in specific areas. In this paper, we organize the participles dictionary, create index with those vocabularies of the same last character and gather them into a virtual mini-dictionary, and finally push this dictionary become a mini dictionary of unity of N. Comparing with the entire dictionary, this method narrows matching range and time.
- Ambiguity automatic digestion

  According to statistical analysis of the typical corpus, there are 6% ambiguity fields. As long as the word segmentation algorithm eliminates the false

ambiguous, the segmentation will be improved more accuracy. The algorithm refers to this paper is a good way to dispel the crossing ambiguity. Taking " her hair and clothing is very special(她今天的发饰和服装很特别)" as an example of segment process. "Kimono(和服)" and "Closing(服装)" constitute the crossing ambiguity. In reverse matching word segmentation method, there is no entry to match "and closing(和服装)" in dictionary. Then "and(和)" is removed, "Closing(服装)" is successfully matched with entry, so taking "Closing(服装)" as a new word. At the same time, the method eliminates overlap type ambiguity.

3. The realization of word segmentation method

The article states that the dictionary organization and the thought of word segmentation are based on semi-automatic conversion system. Finally, this paper will realize the method by combining with the system.

Considering the segmentation method is applied in conversation system, vocabulary library should contain the PB and programming words. Meanwhile, we took the process of collecting and analyzing the PB programming corpus, and we found that the common corpuses in PB are added into segmentation dictionary, and it can avoid necessary vocabulary becoming an unknown word. Those vocabularies contain digits and variable names, which are not in vocabulary library. Hence, in the segmentation, the character of classes should be judged. It isn't processed until the character is Chinese character. Continuous non-Chinese characters are processed as a word.

As for the implementation, the text adopts hash map storage dictionary to improve matching speed. Realizing that the algorithm on the computer is with java programming language, the class structure of segmentation system is shown in figure 3.

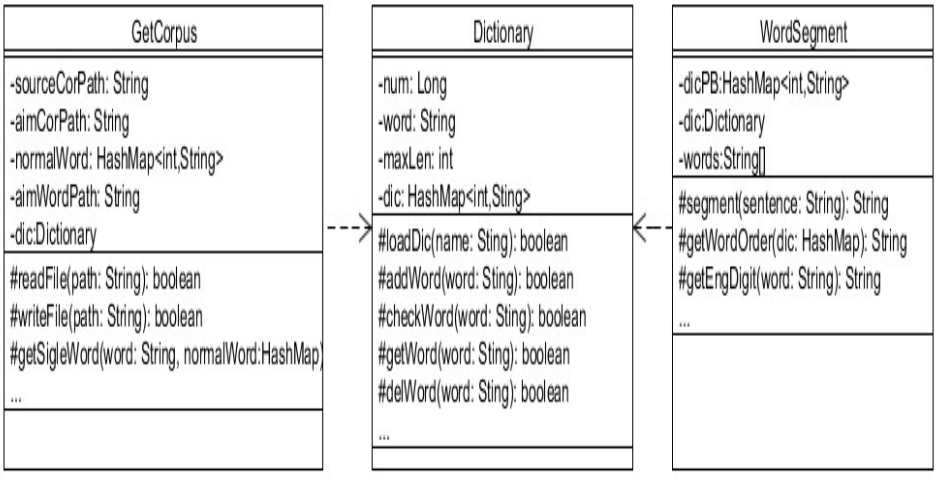

Figure 3. Word segmentation system classes figure schemes

4. Summary

According to Chinese language center rear characteristic, this article draws the reverse matching method to digest crossing ambiguity. Formatting the dictionary in rear Chinese character can narrow the range of searching and improve segmentation speed. But there is no special treatment for ambiguity resolution. At the same time, improving words segmentation methods for library design is very important and they all need further analysis and improvement.

Acknowledgment

This research is sponsored by the projects as follows：

The Funding Project for Beijing talents training mode innovation experimental zone-Software Enguneering;

The Beijing Characteristic Specialty Construction Project for Software Engineering.

References

[1] Liu Jianbin, *Procedure Blueprint designing methodology*, Beijing: Science Press, 1,2005
[2] Liu Yaofeng, Wang Zhiliang, Wang Chuanjing. Model of Chinese Words Segmentation and Part-of-Word Tagging, *Computer Engineering*, 36(4) (2010) 17-19.
[3] Liu Hongzhi, Research on Chinese Word Segmentation Techniques, *Computer Development and Application*, 23(3) (2010) 173-175.
[4] Wu Tao, ZahngMaodi, Zhang Chuanbo, Research of Chinese Word Segmentation Algorithms Based on Statistics and Reverse Maximum Match, *Computer Engineering & Science*, 30(8) (2008) 79-82.
[5] Wang Ruilei, Luan Jing, Pan Xiaohua.An Improved Forward Maximum Matching Algorithm for Chinese Word Segmentation, *Computer Applications and Software*, 28(3) (2011) 195-197.
[6] Zhang CaiQin，Yuan Jian, Improved forward maximum matching word segmentation algorithm, *Computer Engineering and Design*, (11) (2010) 2595-2597, 2633.
[7] Wu Jing，Cai Di，Wang Zheng, Word processing and application of GIS in natural language queries, *Geo-Information Science*，7(3): (2005) 67-71.