A Character-net Based Chinese Text Segmentation Method

Lixin Zhou
zhoulx@ict.ac.cn

Qun Liu
Liuqun@ ict.ac.cn

Institute of Computing Technology, Chinese Academy of Science.
NO. 6 Kexueyuan South Road, Beijing, China, P.O. BOX 2704, 100080

Abstract

The segmentation of Chinese texts is a key process in Chinese information processing. The difficulties in segmentation are the process of ambiguous character string and unknown Chinese words. In order to obtain the correct result, the first is identification of all possible candidates of Chinese words in a text. In this paper, a data structure Chinese-character-net is put forward, then, based on this character-net, a new algorithm is presented to obtain all possible candidate of Chinese words in a text. This paper gives the experiment result. Finally the characteristics of the algorithm are analysed.

Keywords: segmentation, connection, character-net, ambiguity, unknown words.

1 Introduction

The segmentation of Chinese texts is a key problem in Chinese information processing. In the process of segmentation, the ambiguity processing, unknown Chinese words (not included in the lexicon) recognition (such as person names, organization names etc) are very difficult. As for those problems, many algorithms are put forward [Liu 2000]. But the existing algorithms haven't a universal data structure, each algorithm can resolve a problem, and correspond to a concrete data structure specifically. In process of the difficulties, the first step is identification of all possible candidates of Chinese words segmentation. For examples: "只有为人民工作才有真正价值", these words should be obtained: 只有、为人、人民、民工、工作、才、有、真正、价值。The ambiguous string is "为人民工作".There are some methods to resolve this problem: the one is the method forward maximum matching, backward maximum matching and minimum matching are used to find out the possible word strings from the character string [Guo 1997; Sproat et al. 1996; Gu and Mao 1994; Li et al. 1991; Wang et al. 1991b; Wang et al. 1990]. The second is The words finding automaton based on the Aho-Corasick Algorithm [Hong-I and Lua]. The former requires three scans of the input character string. In addition, during each scan, backtracking has to be performed in cases where a dictionary search fails. After that, the word recognition is built based on the candidates. The second requires building up a state chart, is difficult to combine with other algorithms.

In this paper, an algorithm is put forward to solve this problem, which uses the connection information between Chinese characters to recognize all possible candidates of segmentation words in a Chinese text. In the method, at first establish a Chinese character-net , try to establish

a universal data structure, which is easy to combine with other algorithms in Chinese text segmentation, and can use different kinds of information in a Chinese text, then identify all possible candidates of words segmentation easily.

2 Data Structure and Model

A Chinese character is considered as a node, a connection between characters considered as an edge. If a character is the final character of a Chinese word, the character is considered as a control node, and the formed edge weight is 1. The connection is defined as follows :

```
typedef struct CharConn
 {
    int        id;
    char     char1[5];
    char     char2[5];
    int        weight;
    int        wlen;
    char      wpos[20];
    char      bchar[5];
    int         route;
    CharConn   *next;
                }CharConn;
```

In the structure,

id is the sequence number of a connection edge,

char1 is the first character node,

char2 is the second character node;

weight is the weight of a edge, if char1 and char2 is in a Chinese word and char2 isn't the final character of a word, weight equal to 0; if char2 is the final character of a word(char2 is a control node), weight equal to 1.

wlen is the length of a word, if char2 isn't a control node, wlen is zero;

wpos is the part-of-speech of a word, if char2 isn't a control node, wpos is null;

bchar is the first character of a word, if char2 isn't a control node, bchar is null;

route is the former connection id, if the length of a word is greater to two characters.

For examples, as for these words : "号召", "号召力", "爱国", "爱国人士", "爱称", "人士", "士兵", "爱国心", "心情" the connection edge is in table 1.

Table 1.

id	Char1	char2	weight	wlen	wpos	bchar	route
1	号	召	1	4	v	号	0
2	号	召	0	0	null	null	0
3	召	力	1	6	n	号	1
4	爱	国	1	4	a	爱	0
5	爱	国	0	0	null	null	0
6	国	人	0	0	null	null	5
7	人	士	1	8	n	爱	6
8	爱	称	1	4	n	爱	0
9	人	士	1	4	n	人	0
10	士	兵	1	4	n	士	0
11	国	心	1	6	n	爱	5
12	心	情	1	4	n	心	0

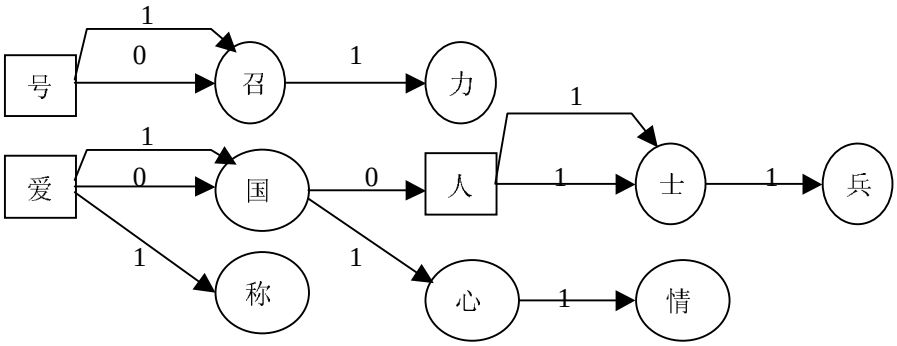

Fig. 1

3 Algorithm

Based on the Chinese character net which is described in section 2, the algorithm finding all the possible candidates of words segmented in a Chinese text is as follows:

```
Begin the algorithm
  Variable
    CString strSrc;//the source string
    CString strRes; //the result of all
                    possible word candidates
    int i; //the current character in the source
           string
    int iFind; //the position-number of the
           final-character of the last formed
           word
    int len; //the length of the source string
    Char str1[5]; //the current first character
    Char str2[5]; //the current second character
    BOOL  Find=0; // a tag points the current
                  routes are or aren't in
                  words
    int  Frec=0; //a tag points the route is or
                  isn't in a word
  while(i < len-1) {
      get the first current character into str1
          from the source string;
      get the second current character into str2
          from the source string;
    select the connection between str1 and str2
    from the Chinese-character-net;
  if(Find the connections) {
    for( each connection obtained )
     if(weight == 0) {
        add the current route to route list;
     }
     else if(weight == 1) {
       j = the position-number of the
           first character of the current
           formed word;
       if(its route matches the former right
           route) then {
         if(iFind > j)
           process the middle characters
           ( between iFind and j ) as single
           characters;
         add the candidate word to the result
         string strRes;
         set iFind to current value;
       } else set Frec = -1;
       reduce the current route from
         the route list ;
     }
```

```
        if (each of the connections isn't in a
            word) then
            Find = false;
        End for
        If ( Find == false) then {
            process the current character as
            single character;
            set iFind += 2;
        } else if(not find connection) {
              process the current character as single
              character;
              set iFind += 2;
            }
        i = i + 1; //forward a Chinese
                   character
    }
End    the algorithm
```

By the algorithm, the following Chinese character string "主持人的话：暑假期间，我们又一次看到了久别的《成长的烦恼》，它就像一本家庭教育百科全书，艺术地展现着现实生活中的种种问题，也提供给我们看问题的一个新视角。" is segmented into "主持 人 的话 ： 暑假 假期 期间 ， 我们 又 一 次 看到 了 久别 别的 《 成长 的 烦恼 》 ， 它 就 像 一本 本家 家庭 教育 百科全书 全书 ， 艺术地 展现 着 现实 生活 中 的 种 种 问题 ， 也 提供 供给 我们 看 问题 的 一个 新 视角 。".There are "暑假期间", "久别的",and "本家庭" ambiguous segmentation strings. An other example is the following Chinese character string "一家乡镇企业的负责人参加会议" is segmented into "一 家 家乡 乡镇 企业 的 负责 负责人 人参 参加 会议". In the text, "家乡镇", and "负责人参加" are the ambiguous segmentation strings.

4 Experiment

Based on a basic Chinese word dictation obtained from Beijing University, which has 61135 Chinese words, we obtain the connections between each two characters, establish a Chinese character net which has 76259 connections. The records increase 24.7% ((76259-61135)/ 61135). In the character net, there are 2857 connections which have same char1 and same char2. In a general Chinese machine readable lexicon, there are about only 12% of words whose length are longer than three Chinese characters, about 70% of words whose length equal 4, and about 15% of words whose length equal 6. So, in the algorithm in this paper, the structure of the character-net is fine and the confliction may be processed seldom in the selection of the connections between same char1 and same char2. About 1500 Chinese characters can be processed per second.

5 Analysis of the Algorithm

In Chinese, the meaning of a character is atomic and based, the meaning of most of the Chinese words can be derived from the characters in the word, as is to say, the meaning of a Chinese word is compound or derived. This paper resolves the difficulties in segmentation of Chinese texts by the thought. The information in a Chinese text are divided into three kinds: (1) about characters, (2) about connections between characters, and (3) about Chinese words. As is expressed in Fig. 2.

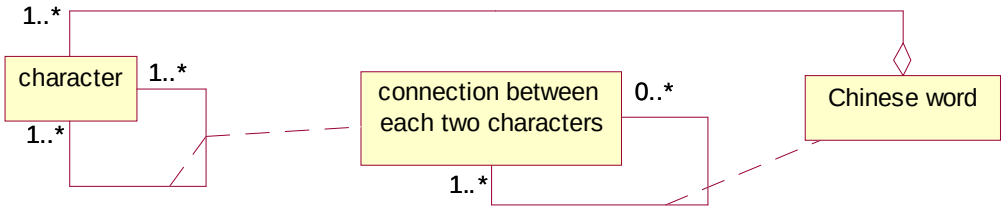

Fig. 2

In fig.2, a character and another character, which have relation between them, can compose into a connection. A connection and zero or several connections compose into a Chinese word. A Chinese word is composed of one or several Chinese characters.

About a character, there are following information: (1) the probability used in a person name, (2) if it is a single character word etc.

About a connection, there are information as described in section 2 and 3.

About a word, there are following information: (1) if it is used as a prefix or a suffix(such as “班长”, “部长”, “山脉”, “新村”, “先生”); (2) mutual information between words, etc.

In the process of segmentation of Chinese texts, we make the segmentation character by character. At first, the information of a character is processed, for example in this step we can obtain the possible person names; the second is obtaining and processing the information of connections between each two characters by the Chinese-character-net described in this paper; then we obtain all the possible candidate of segmentation words in a Chinese text. The third, we use the information of words and between words, resolve the ambiguity of segmentation words, identification of unknown words such as person names, place names and organization names.

So the algorithm in this paper is easy combined with other existing algorithms.

6 Conclusion

In this paper, an algorithm of finding all possible candidate words in segmentation of a Chinese text has been presented. The algorithm is based on a Chinese-character-net, which is established by the information of connections between each two Chinese characters. The algorithm has some characteristics as follows:

(1) the character net is a basic data structure, makes the use of all information in segmentation consistently and easy.

(2) The scan of a text need only once.

(3) The algorithm is easy combined with other existing algorithms.

(4) The algorithm is effective.

(5) The algorithm is easy extensible. After all possible candidate segmentation words are obtained, we can obtain the result of FMM by FMM thought, obtain the result of BMM by BMM thought, and can process ambiguity and unknown words by probability grammar or HMM method.

Based on the result obtained by the algorithm, different tactics of processing the possible candidates words segmented can be adopted according to the needs of different kinds of

applications such as search engine [Zhou 2001], text classification, machine translation, information extraction, retrieval or filter etc.

Acknowledgements

Our thanks go to the Institute of Computational Language, Peking University for the Lexicon. The paper is supported by the ICT Youth Science Foundation.

References

Liu Kaiying. Automatic Segmentation and part-of speech Tagging for Chinese Text. ShangWu Press, Beijing, May 2000.

Gu Ping and Mao Yu-Hang. The adjacent matching algorithm of Chinese automatic word segmentation and its implementation in the QHFY Chinese-English system, Proceedings of the 1994 International Conference on Chinese Computing, Singapore, ICCC-94, 1994.

Guo Jin. 1997. Critical tokenization and its properties, Computational Linguistics. 23, 4, 569-596.

Li B.Y., Lin S., Sun C.F. and Sun M.S. A maximal matching automatic Chinese word segmentation algorithm using corpus tagging for ambiguity resolution, Proceedings of R. O. C. Computational Linguistics Conference IV, Taiwan, ROCLING-IV, 1991. 135-146.

Hong-I Ng, Kim-Teng Lua. A Word-Finding Automaton for Chinese Sentence Tokenization. National University of Singapore, http://www.comp.nus.edu.sg/ %7Erpir/ members.html# nghongi

Sproat Richard, Shih Chilin, Gale William, and Chang Nancy. 1996. A stochastic finite-state word segmentation algorithm for Chinese, Computational Linguistics. 22, 3, 377-404.

Wang Xiao-Long, Wang Kai-Zhu and Bai Xiao-Hua. 1991. Separating syllables and characters into words in natural language understanding, Journal of Chinese Information Processing. 5, 3, 48-58.

Wang Yong-Heng, Su Hai-Ju and Mo Yan. 1990. Automatic processing of Chinese words, Journal of Chinese Information Processing. 4, 4, 1-11.

Zhou Lixin. Research of Segmentation of Chinese Texts in Chinese Search Engine. Proceeding of IEEE International Conference on Systems, Man, and Cybernetics SMC' 2001,Tucson, Arizona, USA, October 7-10, 2001.