ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.5,No.1,January 2009, pp. 192-193

E-mail: eduf@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

中文分词算法解析

张磊,张代远

(南京邮电大学 计算机学院,江苏 南京 210003)

摘要:中文分词是计算机中文信息处理中的难题,而中文分词算法是其中的核心,但由于中英文环境中语素的不同特点,使得中文必须要解决分词的问题。这篇文章较为深刻的阐述了中分分词的算法,基于对分词算法的理解和对匹配法分词的分析,对最大匹配分词方法进行了较深入的研究探讨,提出了什么算法是解决分词效率的最佳方法以及各种方法的比较优劣等问题,及可能产生的歧义,对不同的算法给予了充分的解释,通过对各种算法的比较,总结出了比较常用和效率较高的算法。

关键词:中文分词;最大匹配算法;最大概率算法;算法;系统

中图分类号:TP391 **文献标识码**:A **文章编号**:1009-3044(2009)01-0192-02

Chinese Lexical Analysis Algorithm

ZHANG Lei, ZHANG Dai-yuan

(Nanjing University of Post & Telecommunications Computer College, Nanjing 210003, China)

Abstract: Chinese Lexical Analysis is a difficult problem in the Chinese information processing ,and the algorithm is the core of it,but there are some different factors between Chinese and English, Chinese Lexical Analysis should be solved completely. This paper present some kinds of algorithms,and analyzing the advantages and Disadvantages of these algorithms to find the best one .At the same time,it is very easy to readers to understand the paper,and using the plot to express the meaning of algorithm.

Key words: chinese lexical analysis; forward maximum matching method; maximum probability method; algorithm; system

1 引言

何为分词? 中文分词与其他的分词又有什么不同呢? 分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。英文是以词为单位的,词和词之间是靠空格隔开,而中文是以字为单位,句子中所有的字连起来才能描述一个意思。例如,英文句子有 take me home,用中文则为:"带我回家"。计算机可以很简单通过空格知道 home 是一个单词,但是不能很容易明白"回"、"家"两个字合起来才表示一个词。把中文的汉字序列切分成有意义的词,就是中文分词,有些人也称为切词。虽然英文也同样存在短语的划分问题,但是在词这一层上,中文比之英文要复杂的多、困难的多。 具体到计算机科学,中文分词则是在计算机中通过人为的规则,编写一个计算机应用程序来对中文文本进行处理,得到词的序列的过程。中文分词是为方便处理中文信息而产生,属于中文信息处理技术的范畴。

2 中文分词算法

根据不同的分词算法:常用的分词算法有最大匹配,最大概率分词法等等。

最大匹配法(Forward Maximum Matching method, FMM 法)

2.2.1 最大匹配法剖析

选取包含 6-8 个汉字的符号串作为最大符号串, 把最大符号串与词典中的单词条目相匹配,如果不能匹配就削掉一个汉字继续匹配,直到在词典中找到相应的单词为止。匹配的方向是从右向左。例如图 1:最大分词算法。

1) S2="";S1 不为空,从 S1 左边取出候选子串 W="计算语言学";

2) 查词表,"计算语言学"在词表中,将 W 加入到 S2 中,S2="计算语言学/",并将 W 从 S1 中去掉,此时 S1="课程是三个课时";

3) S1 不为空,于是从 S1 左边取出候选子串 W="课程是三个";

4) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到 W="课程是三";

5) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到W="课程是";

6) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到W="课程";

7) 查词表,W 在词表中,将 W 加入到 S2 中,S2="计算语言学/课程/",将 W 从 S1 中去掉,此时 S1="是三个课时";

8) S1 不为空,于是从 S1 左边取出候选子串 W="是三个课时";

9) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到 W="是三个课";

10) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到 W="是三个";

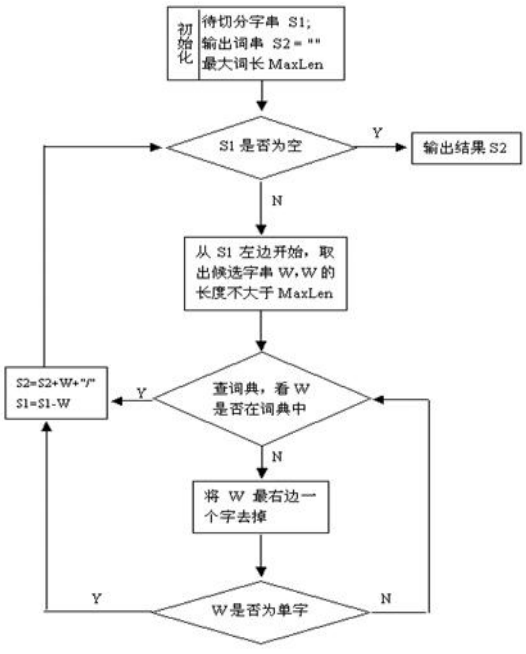

图 1 最大分词算法

收稿日期:2008-11-10

11) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到 W="是三";

12) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到 W="是",这时 W 是单字,将 W 加入到 S2 中,S2="计算语言学/课程/是/",并将从 S1 中去掉,此时 S1="三个课时";

13) S1 不为空,从 S1 左边取出候选子串 W="三个课时";

14) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到 W="三个课";

15) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到 W="三个";

16) 查词表,W 不在词表中,将 W 最右边一个字去掉,得到 W="三",这时 W 是单字,将 W 加入到 S2 中,S2="计算语言学/课程/是/三/",并将 W 从 S1 中去掉,此时 S1="个课时"。

2.2.2 最大匹配法分词的问题

1) 长度限制;

2) 效率低;

3) 掩盖分词歧义;

4) 最大匹配的并不一定是想要的分词方式。

2.2 逆向最大匹配法(Backward Maximum Matching method, BMM 法)

匹配方向与 MM 法相反,是从左向右。实验表明:对于汉语来说,逆向最大匹配法比最大匹配法更有效。 逆向最大匹配的分词原理和过程与正向最大匹配相似,区别在于前者从文章或者句子(字串)的末尾开始切分,若不成功则减去最前面的一个字。比如对于字符串"处理机器发生的故障",第一步,从字串的右边取长度以步长为单位的字段"发生的故障"在词典中进行匹配,匹配不成功,再取字段"生的故障"进行匹配,依次匹配,直到分出"故障"一词,最终使用 BMM 方法切分的结果为:故障、发生、机器、处理。该方法要求配备逆序词典。

2.3 双向匹配法(Bi-direction Matching method, BM 法)

比较 MM 法与 BMM 法的切分结果,从而决定正确的切分,双向匹配法属于最大匹配算法的一种增强算法。这种算法有它的优点:可以兼顾汉语句法规律的多样性(即以正向优先为主,逆向优先仍然存在的情况),但是需要一种评估机制来评估两种方向的优劣,比如在我们的系统中,采用了结合度来对不同的分词结果进行分析。但是,正向匹配和逆向匹配都固有的缺点—不能有效处理歧义字段。因此,要提高双向分词的正确率,除了建立对两个方向进行评估的有效机制外,更为根本的是要对正向匹配和反向匹配本身进行改进,以提高其准确性。

由于以上四种方法都是基于字符串匹配的分词方法,这种方法又叫做机械分词方法,它是按照一定的策略将待分析的汉字串与一个"充分大的"机器词典中的词条进行匹配,若在词典中找到某个字符串,则匹配成功(识别出一个词)。按照扫描方向的不同,串匹配分词方法可以分为正向匹配和逆向匹配;按照不同长度优先匹配的情况,可以分为最大(最长)匹配和最小(最短)匹配;按照是否与词性标注过程相结合,又可以分为单纯分词方法和分词与标注相结合的一体化方法。所以后面不做过多陈述。

2.4 最大概率法

基本思想是:一个待切分的汉字串可能包含多种分词结果;将其中概率最大的那个作为该字串的分词结果。如图 2 所示为最大概率法分词。

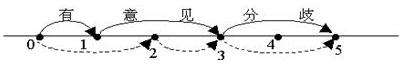

图 2 最大概率法分词

1) 对一个待分词的字串 S,按照从左到右的顺序取出全部候选词 $w_1,w_2,\cdots,w_i,\cdots,w_n$;

2) 到词典中查出每个候选词 的概率值 $P(w_i)$,并记录每个候选词的全部左邻词;

3) 按照公式 1 计算每个候选词的累计概率, 同时比较得到每个候选词的最佳左邻词;

4) 如果当前词 w_n 是字串 S 的尾词,且累计概率 $P'(w_n)$最大,则 w_n 就是 S 的终点词;

5) 从 w_n 开始,按照从右到左顺序,依次将每个词的最佳左邻词输出,即为 S 的分词结果。

最大概率法问题:

1) 并不能解决所有的交集型歧义问题;

2) 无法解决组合型歧义问题。

3 总结

以上分析了各种算法的优缺点,就目前而言,分词算法已经算是比较成熟了,有简单的有复杂的,比如正向最大匹配,反向最大匹配,双向最大匹配,语言模型方法,最短路径算法等等。这里就不展开说了。但是要记住一点的是:判断一个分词算法好不好,关键看两点,一个是消除歧义能力;一个是词典未登录词的识别比如人名、地名、机构名等,比如著名的 ICTCLAS 分词系统就是采用最大匹配算法。

参考文献:

[1] 张华平,刘群.基于 N-最短路径方法的中文词语粗分模型[J].中文信息处理学报,2002,16(5):77-83.

[2] 黄昌宁.中文信息处理中的分词问题[J].语言文字应用,1997,6(1):72-78.

[3] 孙宏林.现代汉语语料库分词中的若干问题[M].北京:清华大学出版社,1997.

[4] 朱珣.中文自动分词系统的研究[D].武汉:华中师范大学,2004.

[5] 张利,张立勇,张晓淼,等.基于改进 BP 网络的中文歧义字段分词方法研究[J].大连理工大学学报,2007,41(1):131-135.

[6] 张卫; 中文词性标注的研究与实现[D].南京师范大学,2007.

[7] 钱揖丽,郑家恒. 文本切分知识获取及其应用[J].计算机工程与应用,2003,39(2):63-64.

[8] 孙茂松,卢红娜,邹嘉彦.基于隐 Markov 模型的汉语词类自动标注的实验研究[J].清华大学学报:自然科学版,2000,40(9):58-61.

张磊(1982-),男,辽宁阜新人,硕士,主要研究方向:中文信息处理。