中图分类号:TP391.1 文献标识码:A 文章编号:1009-2552(2009)07-0187-03

中文分词技术的研究现状与困难

孙铁利,刘延吉

(东北师范大学计算机学院,长春 130117)

摘 要:中文分词技术是中文信息处理领域的基础研究课题。而分词对于中文信息处理的诸多领域都是一个非常重要的基本组成部分。首先对中文分词的基本概念与应用,以及中文分词的基本方法进行了概述。然后分析了分词中存在的两个最大困难。最后指出了中文分词未来的研究方向。

关键词:中文分词;分词算法;歧义;未登录词

State of the art and difficulties in Chinese word segmentation technology

SUN Tie-li, LIU Yan-ji

(**School of Computer, Northeast Normal University, Changchun** 130117, **China**)

Abstract: Chinese word segmentation is a basic research issue on Chinese information processing tasks. And Chinese word segmentation is a very important component in many field of Chinese information process. The paper proposes an unsupervised training method for acquiring probability models that accurately segment Chinese character sequences into words. Then it presents a detailed analysis of the two great dificulties in word segmentation. And finally, it points out the research problems to be resolved on Chinese word segmentation.

Key words: Chinese word segmentation; segmentation algorithm; ambiguity; unlisted words

0 引言

随着计算机网络的飞速普及,人们已经进入了信息时代。在这个信息社会里,信息的重要性与日俱增,无论是个人,企业,乃至政府都需要获取大量有用的信息。谁掌握了信息,谁就能在竞争中处于有利位置。在这种环境下,搜索引擎技术逐渐成为技术人员的开发热点,而其中最为重要的技术就是分词技术。

分词技术属于自然语言理解技术的范畴,是语义理解的首要环节,它是能将语句中的词语正确切分开的一种技术。它是文本分类,信息检索,机器翻译,自动标引,文本的语音输入输出等领域的基础。而由于中文本身的复杂性及其书写习惯,使中文分词技术成为了分词技术中的难点[1-2]。

1 中文分词基本算法

近年来人们对中文分词技术有了一定的研究,提出了多种多样的中文分词算法。目前的中文分词算法主要分为三大类:基于词典的方法,基于统计的方法和基于规则的方法。

1.1 基于词典的分词算法

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。按照扫描方向的不同,该分词方法可以分为正向匹配和逆向匹配;按照长度的不同,可以分为最大匹配和最小匹配。常见的几种基于词典的分词方法思想如下。

1.1.1 正向最大匹配算法

正向最大匹配算法思想[3]:(1)从左往右取待切

收稿日期:2008-12-01

作者简介:孙铁利(1956-),男,东北师范大学计算机学院教授,博士生导师,主要研究方向为智能用户接口、信息处理、地理信息系统。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分汉语句的 m 个字符作为匹配字段,其中 m 为机器可读词典中最长词条的汉字个数。(2)查找机器可读词典并进行匹配。若匹配成功,则将这个匹配字段作为一个词切分出来;若匹配不成功,则将这个匹配字段的最后一个字去掉,剩下的字符串作为新的匹配字段,进行再次匹配。重复以上过程,直到切分出所有词为止。

该算法流程如图 1 如示。

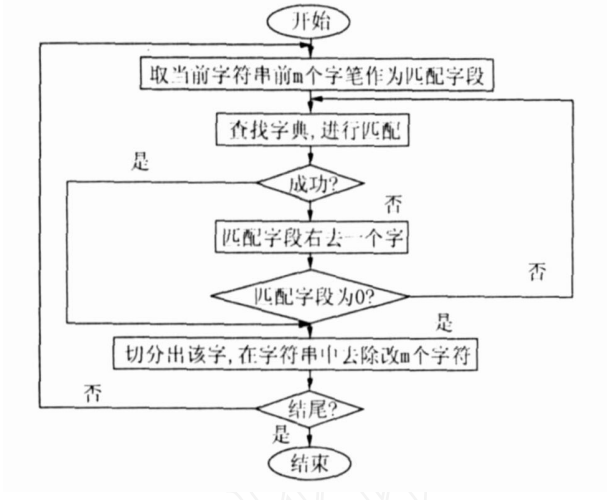

图 1　正向最大匹配法

1.1.2　逆向最大匹配算法

逆向最大匹配算法思想[4]:该算法是对正向最大匹配算法的逆向思维,主要是从右往左对字符串进行最大匹配。匹配成功,则将这个匹配字段作为一个词切分出来;若匹配不成功,则将这个匹配字段的最前一个字去掉,剩下的字符串作为新的匹配字段,进行再次匹配。重复以上过程,直到切分出所有词为止。实验表明逆向最大匹配算法效果要优于正向最大匹配算法。

1.1.3　全二分最大匹配算法

全二分最大匹配快速分词算法[5]:是一种基于 hash 表,每次的匹配操作都可以记忆,不需要任何的重复匹配操作,而且匹配操作都是使用二分法进行的,这样就最大限度地提高了分词的效率。

基于词典的分词算法优点是易于实现,在对精确率要求不高的系统中得到了很好的应用。其缺点在于由于词典是在分词之前准备的,其规模和内容受到了限制,对于未登录词的补充较难实现。

1.2　基于统计的分词算法

目前基于统计的分词算法有很多种,较为常见的算法是,基于互信息的概率统计算法,N-Gram 算法,基于组合度的汉语分词决策算法等等。

1.2.1　互信息的概率统计算法[6]

互信息是一种度量不同字符串之间相关性的统计量。对于字符串 X 和 Y,其互信息的计算公式如下:

$$MI(x,y) = \log 2 \frac{p(x,y)}{P(x)\,p(y)}$$

其中,$p(x,y)$ 为字符串 X 和 Y 共现的概率,$p(x)$,$p(y)$ 分别为字符串 X 和 Y 出现的概率。

互信息 $MI(x,y)$ 反映了字符串对之间结合关系的紧密程度:(1)互信息 $MI(x,y) > 0$,则 X,Y 之间具有可信的结合关系,并且 $MI(x,y)$ 越大,结合程度越强。(2) $MI(x,Y) = 0$,则 X,Y 之间的结合关系不明确。(3) $MI(x,y) < 0$,则 X,Y 之间基本没有结合关系,并且 $MI(x,y)$ 越小,结合程度越弱。

1.2.2　N-Gram 模型算法

N-Gram 模型思想[7]:一个单词的出现与其上下文环境中出现的单词序列密切相关,第 n 个词的出现只与前面 n-1 个词相关,而与其它任何词都不相关,设 W_1,W_2,…,W_n 是长度为 n 的字串,由于预测词 W_n 的出现概率,必须知道它前面所有词的出现概率,太过复杂。为了简化计算,规定任意词 W_i 只与其前两个相关,得到三元概率模型如下:

$$P(W) \approx P(W_1)\,P(W_2/W_1) \prod_{i=3\cdots n} P(W_i/IW_{i-2}\,W_{i-1})$$

以此类推,N 元模型就是假设当前词的出现概率只同它前面的 N-1 个词有关而得出的。

1.2.3　组合度的决策算法

组合度的算法思想[8]:在一篇文章中,如果汉字 B 紧跟在汉字 A 的后面,称 AB 为一个组合。运用组合度的数学公式,计算出每个词组的组合度,组合度越高,说明它是词组的可能性越大,组合度越低,说明它是词组的可能性越小。公式如下:

$$H_{AB} = -\ln\left(\frac{C_{n1}^{k} \cdot C_{n2}^{k} \cdot K! \cdot (N-K)!}{N!}\right)$$

其中,H_{AB} 为 AB 在文章中的组合度,N 为汉字个数,K 为 AB 组合的个数,$n1$ 是 A 的个数,$n2$ 是 B 的个数。

基于统计的分词方法优点在于它可以从已有的大量实例中进行归纳总结,分析语言内在的关联信息,将其加入到统计模型中去。简单的统计方法不需要词典,而是通过训练语料的迭代,建立统计模型。但统计方法本身也有一定的局限性,尤其是对常用词的识别精度很差。

1.3　基于规则的分词算法[9]

基于规则的分词方法是通过让计算机模拟人对

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

句子的理解,达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析,利用句法信息和语义信息来对文本进行分词。

这种分词方法优点在于它可以由实例中进行自动推理和证明,可以自动完成对未登录词的补充,但是由于它本身需要使用大量的语言知识。而汉语语言知识有其笼统、复杂性,难以将各种语言信息组织成机器可直接读取的形式,因此目前基于规则的分词方法还不是很成熟。这种方法目前总是和其他算法结合起来使用。

2 中文分词的混合算法

由于中文分词的三种基本算法,都有其各自的优缺点,所以为了能够达到更好的分词效果,人们开始有目的的把分词的几种基本算法以及其他知识结合起来,这就形成了现在日新月异的混合型分词算法。

(1) 吴建胜[10]等提出的基于自动机的分词方法,这种算法的基本思想是:在数据结构方面,把词典组织成自动机形式,在匹配算法上采用最大向前匹配算法,把二者有机的结合到一起,以达到更好的分词效果。

(2) 赵伟[11]等提出的一种规则与统计相结合的汉语分词方法,这种分词算法的基本思想是:基于一个标注好了的语料库,并且结合了规则和语料库统计两种分词方法。

(3) 张长利[12]等提出的一种基于后缀数组的无词典分词方法,这种分词算法的基本思想是:通过后缀数组和利用散列表获得汉字的结合模式,通过置信度筛选词,能够快速准确地抽取文档中的中、高频词,适用于对词条频度敏感、对计算速度要求高的中文信息处理.

(4) 孙晓[13]等提出的基于动态规划的最小代价路径汉语自动分词方法,这种分词算法的基本思想是:基于最长次长匹配的方法建立汉语切分路径有向图,将汉语自动分词转换为在有向图中选择正确的切分路径,其中有向图中的节点代价对应单词频度,而边代价对应所连接的两个单词的接续频度;运用改进后 Dijkstra 最小代价路径算法,求出有向图中路径代价最小的切分路径作为切分结果。

混合型分词算法多种多样,所结合的知识点也有很多,可以结合数据结构知识来形成新的词典机制;也可以结合标记语料库的方法更好的完善分词算法。显而易见,混合型分词算法在大多数方面要优于基本型分词算法。它将成为今后分词算法研究中的一个热点。

3 中文分词目前的困难

由于中文词与词之间不象西文那样有明显的分隔符,所以构成了中文在自动切分上的极大困难。在现有的中文自动分词方法中,基于词典的分词方法占有主导地位。而中文分词的主要困难不在于词典中词条的匹配,而是在于切分歧义消解和未登录词语的识别。在中文分词过程中,这两大难题一直没有完全突破。

3.1 歧义处理[14-15]

歧义是指同样的一句话,可能有两种或者更多的切分方法。目前主要分为交集型歧义、组合型歧义和真歧义三种。其中交集型歧义字段数量庞大,处理方法多样;组合型歧义字段数量较少,处理起来相对较难;而真歧义字段数量更为稀少,且很难处理。

分词歧义处理之所以是中文分词的困难之一,原因在于歧义分为多种类型,针对不同的歧义类型应采取不同的解决方法。除了需要依靠上、下文语义信息;增加语义、语用知识等外部条件外,还存在难以消解的真歧义,增加了歧义切分的难度。同时未登录词中也存在着歧义切分的问题,这也增加了歧义切分的难度。所以歧义处理是影响分词系统切分精度的重要因素,是自动分词系统设计中的一个最困难也是最核心的问题。

3.2 未登录词识别[16]

新词,专业术语称为未登录词。也就是那些在字典中都没有收录过词。未登录词可以分为专名和非专名两大类。其中专名包括中国人名、外国译名、地名等,而非专名包括新词、简称、方言词语、文言词语、行业用词等。

无论是专名还是非专名的未登录词都很难处理,因为其数量庞大,又没有相应的规范,而且随着社会生活的变迁,使未登录词的数量大大增加,这又为未登录词的识别增加了难度。因此,未登录词识别是中文分词的另一大难点。

4 结束语

本文主要是对中文各类分词算法做出了系统的介绍,分析了每类分词算法各自的优缺点。提出了将中文分词算法分为基本分词算法和混合型分词算法两大类型,得出了混合型算法往往要优于基本型算法的结果。同时分析了中文分词的两大难点——歧义处理和未登录词的识别,指出了它们的困难所在,这就为以后的中文分词研究工作奠定了基础。

(下转第 192 页)

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(上接第189页)

参考文献:

[1] 张春霞,郝天永.汉语分词的研究现状与困难[J].系统仿真学报,2005,17(1):138-147.

[2] 赵川,杜玲,岳鹏,等.基于中文的自然语言理解初探[J].现代电子技术,2007(6):82-85.

[3] 郭辉,苏中义,王文,等.一种改进的MM分词算法[J].微型电脑应用,2002,18(1):13-15.

[4] 陈耀东,王挺.基于有向图的双向匹配分词算法及实现[J].计算机应用,2005,25(6):1442-1444.

[5] 李振星,徐泽平,唐卫清,等.全二分最大匹配快速分词算法[J].计算机工程与应用,2002(11):106-109.

[6] 费洪晓,康松林,朱小娟,等.基于词频统计的中文分词的研究[J].计算机工程与应用,2005(7):67-68.

[7] 吴应良,韦岗,李海.一种基于N-gram模型和机器学习的汉语分词算法[J].电子与信息学报,2001,23(11):1148-1153.

[8] 刘利东.基于组合度的汉语分词决策算法研究[J].德州学院学报,2003,19(2):65-70.

[9] 张江.基于规则的分词方法[J].计算机与现代化,2005(4):18-20.

[10] 吴建胜,战学刚,迟呈英.一种基于自动机的分词方法[J].计算机工程与应用,2005,41(8).

[11] 赵伟,戴新宇,尹存燕,等.一种规则与统计相结合的汉语分词方法[J].计算机应用研究,2004(3):23-25.

[12] 张长利,赫枫龄,左万利.一种基于后缀数组的无词典分词方法[J].吉林大学学报:理学版,2004,42(4):548-553.

[13] 孙晓,黄德根.基于动态规划的最小代价路径汉语自动分词[J].小型微型计算机系统,2006,27(3):516-519.

[14] 谭琼,史忠植.分词中的歧义处理[J].计算机工程与应用,2002,38(11):125-127.

[15] 翟凤文,赫枫龄,左万利.基于统计规则的交集型歧义处理方法[J].吉林大学学报:理学版,2006,44(2):223-228.

[16] 陈小荷.自动分词中未登录词问题的一揽子解决方案[J].语言文字应用,1999(3):103-109.

责任编辑:李光辉

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net