基于反序词典的中文分词技术研究

罗桂琼,费洪晓,戴 弋
(中南大学 信息科学与工程学院,湖南 长沙 410083)

摘 要:中文自动分词是计算机中文信息处理中的难题。在对中文分词的现有技术研究的基础上,对反序最大匹配分词方法进行了较深入的研究探讨,在此基础上对中文分词的词典结构和分词算法做了一部分改进,设计了基于反序词典的中文分词系统。实验表明,该改进算法加快了中文的分词速度,使得中文分词系统的效率有了显著提高。

关键词:中文分词;反序最大匹配;机械分词;反序词典
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2008)01-0080-04

Research of Chinese Segmentation Based on Converse Segmentation Dictionary

LUO Gui-qiong, FEI Hong-xiao, DAI Yi
(College of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: Chinese segmentation system is a difficulty in computer Chinese information handling. A deep discussion on methods of reverse ordering Chinese segmentation matching on the basis of existing technology in Chinese segmentation is made. On this basis of it, some improvements are made in the dictionary construction and segmentation arithmetic, designing a Chinese segmentation system based on reverse ordering dictionary. Experiment shows that the improving arithmetic accelerated the speed of Chinese segmentation.

Key words: Chinese segmentation; converse max matching; machine segmentation; converse segmentation dictionary

0 引 言

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[1]。在英文的行文中,单词之间是以空格作为自然分界符的,而中文只是字、句和段可以通过明显的分界符来简单划界,唯独词没有一个形式上的分界符,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得多、困难得多。

从实际应用上来说,中文分词又是实现计算机人工智能、智能搜索、人机对话、中文翻译等核心应用的关键技术[2~5]。

1 中文分词介绍

1.1 什么是中文分词

众所周知,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 I am a student,用中文则为:“我是一个学生”。计算机可以很简单通过空格知道 student 是一个单词,但是不能很容易明白“学”、“生”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。我是一个学生,分词的结果是:我 是 一个 学生。

1.2 中文分词技术难点

目前在中文分词过程中,有两大难题一直没有完全突破。

(1)歧义识别。

分词歧义的产生主要有两种情况:组合型歧义和交集型歧义。所谓组合型歧义是某个词的一小部分也是一个完整的词,如“中华人民共和国”,“中华”,“人民”,和“共和国”都是词,但是它们合起来也是一个词。而交集型歧义就是说两个相邻的词之间有重叠的部分,如“今天下午”,“天下”是一个词,“下午”也是一个词,它们重用了一个“下”字。研究表明,歧义的产生主要是后一种,它约占整个分词歧义的90%。所以,处理好交集歧义字段在很大程度上能保证一定的分词精

收稿日期:2007-04-03
基金项目:湖南省科技计划项目(2006JT1040)
作者简介:罗桂琼(1970-),女,硕士研究生,高级讲师,主要研究方向为计算机网络、信息提取;费洪晓,副教授,主要研究方向为网络管理与网络安全、信息过滤。

度。

另外还存在一种歧义是真歧义。真歧义是指给出一句话,人也无法判断哪个应该是词,哪个应该不是词。例如:"乒乓球拍卖完了",可以切分成"乒乓 球拍 卖 完 了",也可切分成"乒乓球 拍卖 完 了",如果没有上下文相关句子,将无法知道"拍卖"在这里是否是一个词。

(2)新词识别。

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,但又确实能称为词的那些词。最典型的是人名,人可以很容易理解句子"王军虎去广州了"中,"王军虎"是个词,因为是一个人的名字,但要是让计算机去识别就困难了。如果把"王军虎"做为一个词收录到字典中去,全世界有那么多名字,而且每时每刻都有新增的人名,收录这些人名本身就是一项巨大的工程。即使这项工作可以完成,还是会存在问题,例如:在句子"王军虎头虎脑的"中,"王军虎"还能不能算词?

新词中除了人名以外,还有机构名、地名、产品名、商标名、简称、省略语等,这都是很难处理的问题,而且又正好是人们经常使用的词,因此对于分词系统来说,对新词的识别也是十分重要的。

1.3 中文分词应用领域

(1)搜索引擎。

搜索引擎通常由信息收集和信息检索两部分组成。中文的词与词之间是没有分隔符的,因此若想建立基于词的索引,就需要专门的技术,这种技术就是本文所研究的"汉语词语切分技术"[6]。

(2)中文校对系统。

电子信息的形成可通过多种途径,最通常的方法即是用键盘输入,因而不免造成一些输入错误,由此产生了用计算机进行文本校错的研究。中文自动校错系统可应用于报刊及出版社、打字业等需要进行文本校错的行业[7]。

(3)中译英系统。

机器翻译是研究如何利用计算机进行语言之间翻译的一门边缘学科,它的发展取决于计算机科学、语言学、数学、人工智能、心理学等一系列学科的发展。随着信息的急剧增加,国际交流的日趋频繁,机器翻译的潜在需求越来越大[8]。

(4)语音输出系统。

语音分词的目的是根据人们在表达上的习惯和语流的停顿及强弱变化,在每个词之间插入长度不等的空语音符号(停顿),提高语流的节奏和自然度,以利于听者的理解。

2 中文分词技术

中文分词技术属于自然语言处理技术范畴,对于一句话,人可以通过自己的知识来明白哪些是词,哪些不是词,但如何让计算机也能理解?其处理过程就是分词算法。

现有的分词算法可分为三大类:基于规则的分词方法、基于统计的分词方法和基于理解的分词方法。

2.1 基于规则分词

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。

常用的几种机械分词方法如下:

1)正向最大匹配法(由左到右的方向);

2)逆向最大匹配法(由右到左的方向);

3)最少切分(使每一句中切出的词数最小)。

目前机械式分词占主流地位的是正向最大匹配法和逆向最大匹配法[9]。在匹配过程中,又可分为增字和减字匹配两种。还可以将上述各种方法相互组合,例如,可以将正向最大匹配方法和逆向最大匹配方法结合起来构成双向匹配法。由于汉语单字成词的特点,正向最小匹配和逆向最小匹配一般很少使用。一般说来,逆向匹配的切分精度略高于正向匹配,遇到的歧义现象也较少。统计结果表明,单纯使用正向最大匹配的错误率为1/169,单纯使用逆向最大匹配的错误率为1/245。但这种精度还远远不能满足实际的需要。实际使用的分词系统,都是把机械分词作为一种初分手段,还需通过利用各种其它的语言信息来进一步提高切分的准确率。

一种方法是改进扫描方式,称为特征扫描或标志切分,优先在待分析字符串中识别和切分出一些带有明显特征的词,以这些词作为断点,可将原字符串分为较小的串再来进机械分词,从而减少匹配的错误率。另一种方法是将分词和词类标注结合起来,利用丰富的词类信息对分词决策提供帮助,并且在标注过程中又反过来对分词结果进行检验、调整,从而极大地提高切分的准确率。

基于词典的机械分词法,实现简单,实用性强,但机械分词法的最大的缺点就是词典的完备性不能得到保证。据文献统计,用一个含有70000个词的词典去

切分含有 15000 个词的语料库，仍然有 30% 以上的词条没有被分出来，也就是说有 4500 个词没有在词典中登录。

2.2 基于统计分词

基于统计的方法是基于(两个或多个)汉字同时出现的概率，通过对语料库有监督或无监督的学习，得到描述一种语言的“语言模型”(常用一阶隐马尔可夫模型(1stHMM))。基于统计的方法有许多优点：未登录词的影响降低了，只要有足够的训练文本就易于创建和使用[10]。

2.3 基于理解分词

基于理解的中文分词又称之为知识分词，知识分词是一种理想的分词方法，它不存在上面的问题，但这类分词方案的算法复杂度高，其有效性与可行性尚需在实际工作中得到进一步的验证。知识分词利用有关词、句子等的句法和语义信息或者从大量语料中找出汉字组词的结合特点来进行评价，以期找到最贴近于原句语义的分词结果。

3 改进的反序最大匹配算法

3.1 反序分词词典结构

分词词典结构设计的好坏直接影响分词系统的效率，好的词典结构使得分词算法易于实现且呈现出高效率，而一个设计不合理的分词词典结构不仅使分词算法难于实现且将严重影响分词速度。

文中在反序最大匹配分词方法的基础上对中文分词的词典结构和分词算法做了一部分改进，设计了一个中文分词系统 CSS(Chinese Segmentation System)。反序机械分词词典每一项包括三个元素：词、以这个词为前缀的词的最大长度和以这个词为前缀且词长最大的词在反序词典中的索引号。分词算法通过一个 HASH 映射表直接从以该词为前缀的最大词长的词开始匹配，这符合最大匹配分词算法的特点，从而加速了分词的速度。CSS 中文分词系统的反序词典结构的设计充分利用了“最大”这个关键之处，结构如下所示[11]：

```
typedef struct_SReverseDictItem
  { char * pReverseWord; //反序词条
int nMaxLength; //包含这个词的词条最大长度
int nIndex; //在 Dict 中以该词为前缀且词长最大的词条在反序词典中的索引
  } SReverseDictItem;
  typedef struct -SReverseDict
{ int nItemcount;
SReverseDictItem * pSReverseDictItem
} SReverseDict; //指向反序分词词典的一个结构体
```

3.2 改进的反序最大匹配算法

分词算法必须与分词词典结构配合，它与分词词典的结构具有同样重要的作用。CSS 中文分词系统分词算法的简单描述如下：

装载 ReverseDict 以及反序词典词根 HASH 表，如果 ReverseDict 还没有生成，则生成 ReverseDict 以及反序词典词根 HASH 表，并将生成的 ReverseDict 以及反序词典词根 HASH 表以 txt 文件的形式保存(这样以后系统重起的时候就不需要重做生成 ReverseDict 以及反序词典词根 HASH 表的工作，只需要从相应的 txt 文件装载)。

预处理，利用特殊的标记(0Xa1)将输入的中文文本分割成较短的汉字串，汉字串以中文空格分开，这些标记包括所有的标点符号，例如：“，”、“。”等等。

从指向要进行分词的中文字符串的串尾的指针取两个字节即一个汉字，查看这个汉字对应的反序词典词根 HASH 表 m_nIndexDict[i][j]，如果这个汉字对应的 m_nIndexDict[i][j]的值是 -1，表示当前的词典里没有以这个汉字为词根的词，指向串尾的指针向前移两个字节，开始下一次的匹配；如果这个汉字对应的 m_nIndexDict[i][j]的值大于 0，则进行下一个步骤。

在反序机械词典 ReverseDict 里从 m_nIndexDict[i][j]位置开始查找这个字是否是其他词的前缀，如果是则进行下一步，否则，从指向要进行分词的中文字符串的串尾的指针取 4 个字节，即二个汉字查看是否是其他词的前缀，直到找到。

从反序机械词典 ReverseDict 里查看以这个词为前缀的最大词长 n，然后从指向串尾的指针向前切分 2 * n 个字节，在 ReverseDict 里从这个词条 nIndex 所记载的索引值位置开始进行，在词典中进行从后往前的匹配，如果匹配成功，则将该词附加到一个字符串变量 string 中(该字符串变量用于存储分词结果并显示该结果)，并在 string 后插入一规定符号“/”作为切分标记(该符号由系统定义)，同时指向待匹配的字符串串尾的指针向前移 2 * (n + 1)个字节；否则从指向串尾的指针向前切分 2 * (n - 1)个字节，继续匹配，直到匹配成功或只剩下 2 个字节。

循环以上三个步骤，直到整个中文字符串处理完为止。

4 测试结果

文中设计的这个中文分词系统 CSS，在反序最大匹配分词方法的基础上对中文分词的词典结构和分词算法做了一部分改进，经过测试，CSS 系统对几个主要类别的文档分别进行统计的结果如表 1 所示。

表1　分词速度统计表

文档类型	分词速度(字/秒)
政治	2300
经济	1980
文化	2820
体育	2032
科技	2400

经过多次测试,如果使用传统的逆向最大机械分词算法,MAXL取14个字节,即7个汉字,平均分词速度为280字/秒;使用改进后的分词词典,逆向最大机械分词算法分词速度达到2530字/秒。

通过分析以上的测试结果,可以看出该中文分词系统将来所要继续改进的方向是词典识别专有名词的能力和其动态性增容性方面。可以建立不同的分类词典,对专门的领域使用专门的分词词典,并且词典容量也可以动态变化以适应不同领域中出现的新词汇。

5　结束语

自动分词是汉语自然语言处理的第一步。目前,汉语自然语言处理的应用系统处理对象越来越多的是大规模语料(如Internet信息搜索引擎,各种全文检索系统等),因此分词的速度和分词算法的易实现性变得相当关键。在多种分词算法中,正向最大匹配分词算法简洁、易于实现,在实际工程中应用最为广泛。但基于统计的分词算法和基于理解的分词算法都是对基于规则分词算法扩充和完善,一般的分词系统都是将其中几种结合起来一起使用,很少单纯使用一种分词算法。基于理解的分词算法实现起来复杂,但其分词精度相当高,适合于要求分词精度高的场合;而基于统计分词算法对识别未登录词和专有名词有着自己的优势。三者的有机结合将是未来的发展方向。

参考文献:

[1] 冯书晓,徐　新,杨春梅.国内中文分词技术研究新进展[J].情报杂志,2002(11):29-30.

[2] JieeSoft.OFBiz简单介绍[EB/OL].2004-04-16/2004-06-05.http://www.jieesoft.com/modules.php.

[3] 文庭孝,邱均平,侯经川.汉语自动分词研究展望[J].现代图书情报技术,2004,112(7):6-10.

[4] 湛　燕,陈　昊,袁　方,等.基于中文文本分类的分词方法研究[J].计算机工程与应用,2003(23):87-91.

[5] 文庭孝.汉语自动分词研究进展[J].图书情报,2005(5):54-63.

[6] 邹海山,吴　勇,吴月珠,等.中文搜索引擎中的中文信息处理技术[J].计算机应用研究,2000(12):21-24.

[7] 吴　岩,李秀坤,刘　挺,等.中文自动校对系统的研究与实现[J].哈尔滨工业大学学报,2001,33(1):60-64.

[8] 吕学强.机器翻译概述[J].辽宁师专学报,2002,4(1):8-11.

[9] 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.

[10] 李家福,张亚非.一种基于概率模型的分词系统[J].系统仿真学报,2002,14(5):544-546.

[11] 彭希鸿.基于WEB内容挖掘的网页分类与过滤研究与应用[D].长沙:中南大学,2003.

(上接第79页)

到服务的查找都利用基于本体的匹配算法,因此同传统的关键字匹配机制比较可以获得更准确的语义信息,进而能更精确地定位服务,提高查准率并改善Web服务发现性能。

3　结束语

基于P2P的底层架构,文中提出一个语义Web服务发布和发现模型。该模型的底层采用两层结构:第一层节点采用非结构化的方式连接,通过一定的路由机制保证通信的畅通,避免回路,同时有效支持系统的可扩展性;第二层采用集中式的结构组织节点,有效地提高查询效率。在此模型的基础上,设计具体高效的匹配算法将是下一步研究的目标。

参考文献:

[1] 宋　炜,张　铭.语义网简明教程[M].北京:高等教育出版社,2004.

[2] 方馨馨,熊齐邦.基于P2P网络的语义Web服务发现机制[J].计算机工程,2005,31(17):115-117.

[3] 尹晓璐,李广军.基于语义的Web服务查询[J].实验科学与技术,2005,24(1):31-34.

[4] Cerami E.Web服务精髓[M].陈　逸译.北京:中国电力出版社,2003.

[5] Raman R,Solomon L M.MatchMaking:an extensible Framework for distributed resource management[J].Cluster Computing,1999(2):129-138.

[6] Li Lei,Horrocks L.A Software Framework for Matchmaking Based on Semantic Web Technology[C]//Proceedings International WWW Conference.Budapest,Hungary:[s.n.],2003:20-24.

[7] Le-Hung Vu,Hauswirth M,Aberer K.Towards P2P-based Semantic Web Service Discovery with QoS Support[D].Lausanne,Switzerland:School of Computer and Communication Sections,Ecole Polytechnique Federale de Lausanne(EPFL),2005.