文章编号:1001 - 9081(2000)02 - 0008 - 04

文语转换系统中基于语料的汉语自动分词研究

应志伟 柴佩琪 陈其晖 同济大学计算机系 上海(200092)

摘 要:基于一个实际的文语转换系统,介绍了它的一些处理方法,采用了一种改进的最大匹配法,可以切分出所有的交集歧义,提出了一种基于统计模型的算法来处理其中的多交集歧义字段,并用穷举法和一些简单的规则相组合的方法从实用角度解决多音字的异读问题以及中文姓名的自动识别方法,解决了汉语切分歧义、多音词处理、中文姓名的自动识别问题,达到实现文语转换的目的。

关键词:文语转换;汉语分词;最大匹配法;多交集歧义;多音词;姓名识别
中图分类号:TP317.2 **文献标识码**:A

RESEARCH OF CHINESE WORD SEGMENTATION IN TTS SYSTEM

YING Zhi - wei CHAI Pei - qi CHEN Qi - hui
Department of Computer, Shanghai Tongji University, Shanghai 200092 *China*

Abstract: To obtain a high natural speech in Chinese Text-To-Speech System, Chinese word segmentation is an important problem, which must be solved firstly. To clear up different meanings of word segmentation, get correct pronunciation of polyphonic word, and identify Chinese name are several important and difficult problems in Chinese word segmentation. In this paper, we introduce some methods to solve those problems based on a practical TTS system. A revised MM method of Chinese word automatic segmentation is put forward to obtain all different meanings of intersection, and accordingly an algorithm, which is based on statistic model, is also given to clear up different meanings. Moreover, An algorithm of correcting pronunciation of polyphonic word and an algorithm of identifying Chinese name are also presented.

Key words: TTS; chinese word segmentation; polyphonic word; chinese name identification

1 引言

汉语自动分词,是实现高自然度文语转换系统的第一步,是汉语信息处理领域中的一门基础课题。虽然,几十年来,国内提出了数十种分词的方法,但都没有一个最后认定的切分标准,仍存在一些理论问题尚未完全解决。

汉语自动分词的基本方法可分为三类:形式分词方法,语法分词方法和语义分词方法。我们在文语转换系统的研究开发过程中采用了形式分词方法,辅之以一些分词规则,使之满足一般文语转换系统中对分词精度的要求,而又比较容易实现。本文的目的是试图在综合前人研究基础上,结合实际的分词系统,对分词方法作进一步的探索、尝试。

2 一种改进的最大匹配分词算法设计

所谓形式分词是指不直接进行语法、语义分析而只是借助于分词词典,基于一些统计信息进行分词的一种方法。一般来说,形式分词的方法有正向最大匹配法、逆向最大匹配法、逐词遍历匹配法、设立切分标记法、最佳匹配法等等。

所谓语法分词方法指对文章进行语法分析后,根据一定的语法规则对句子进行分词的一种分词方法。

形式分词方法是基于字符串匹配的原理上进行的:字符串匹配可分成最大匹配法和最小匹配法。对于汉语这样一种语言,由于它的每一个字几乎都可以单独成词,采用最小匹配法显然是不可行的,所以在形式分词方法中采用最多的是最大匹配法。最大匹配法可分成增字最大匹配法和减字最大匹配法。

增字最大匹配法,即对当前匹配的字符串从一个字开始,匹配中不断增字,直到匹配不下去或满足预定条件为止。而每一轮匹配的结果则取匹配成功

收稿日期:1999-08-12
作者简介: 应志伟 硕士研究生,研究方向:语音识别与语言处理; 柴佩琪 教授,博士生导师。

的最大的字符串。

减字最大匹配法,每一轮匹配的过程恰巧和增字最大匹配法相反,匹配过程中不断减字,一旦匹配成功,这轮匹配就算结束,结果就是该匹配成功的字符串,如匹配不成功则减去一个字,然后重新进行匹配。

通常认为:如果一个字段存在不同的切分形式,则称该字段为歧义字段。歧义字段又分为交集型歧义字段和组合型歧义字段。如果 ABC 中 AB 和 BC 都是词,则称 ABC 为交集型歧义字段。可以用有向图来表示交集型歧义字段,如下所示:

上图中有向弧线 1-3、2-4、3-5、4-6 表示在这个交集歧义字段中存在着四个词语,四条有向弧的相交点表示了交集歧义存在。三个交叉点表明了存在三个交集歧义。

如果 AB 中 A、B、AB 都是词则称 AB 为组合型歧义字段。而在这两种歧义字段中交集型歧义字段占歧义字段总数的 90 %左右。所以要消除歧义,首先应消除交集型歧义字段的错误切分。

由于单用最大匹配法无法满足汉语自动分词的需要,无论是前向最大匹配法还是逆向最大匹配法都会造成相当的错误切分,如[组合成分子]无论是前向还是后向都无法作出正确的切分。而如果采用一些分词规则来对最大匹配法作一些补充,则效果较好。即采用正向增字最大匹配法来作粗切分,分出歧义字段,然后把可能存在歧义切分的字段交由分词规则来处理。如[大学生活很丰富],粗切分得到[大学生活/ / 很/ / 丰富],而其中[大学生活]是歧义字段,存在几种切分可能[大学/ / 生活][大学生/ / 活][大/ / 学生/ / 活],如果能识别出[大学生活]是一歧义字段,就有可能通过一些规则来对其作出正确的切分。

3　多交集型歧义字段的切分算法

歧义切分校正主要是针对交集型歧义字段和组合型歧义字段而言的。对于交集型歧义字段采用一些优先规则来处理。而对于组合型歧义字段主要采用特殊处理的方法来解决。

对于多交集型歧义字段的切分算法则相对比较复杂,如[组合成分子时],由于存在着 5 个交集型歧义,因此不能套用简单的规则来解决这以切分问题。在这以问题上,我们遵循了以下原则:

1) 坚持把歧义字段切分成一个个词。

2) 分析所有的切分可能,选择最优化的一种。

为了实现这样的设想,在上面所介绍的改进的最大匹配法的基础上,作进一步的分析,以求能得到最佳的切分效果。

令 C = C1C2. . . Cm 是待切分的字符串,W = W1W2. . . Wn 是切分的结果。P(W/ C)是给定字符串的条件下所产生的输出词串的 W 的概率。W1, W2,. . . Wn 是所有可能的切分方案。那么如果使用最小错误概率决策,基于统计的分词模型就是寻找这样的词串 ,使得 W 满足:

$$P(W/C) = MAX(P(W1/C), P(W2/C), \ldots P(Wn/C)) \quad (1)$$

即估计概率为最大的词串。我们称 为评价函数。一般的基于统计的分词模型的评价函数都是基于贝叶斯公式。

根据贝叶斯公式,我们可以有如下结果:

$$P(W/C) = P(W) * P(C/W) / P(C) \quad (2)$$

在(2)式中,对于 C 的多种切分,C 是给定的,P(C)是一个常数,在计算中不起作用,P(C/ W)是在给定词串的情况下字串出现的概率,可以假设为 P(C/ W) = 1。由式(2)我们可以得出以下结论:用 P(W)来代替 P(W/ C) ,也就是求 P(W)的概率最大值。一般来说,最直接的估计就是利用 N 元语言模型,用 N 阶马尔可夫模型来表示,即:

$$P(W) = \prod_{i=1}^{n-1} P(W_i/W_1 \ldots W_{i-1}) \times \prod_{i=n}^{N} P(W_i/W_{i-n+1} \ldots W_{i-1})$$

但由于词和词之间的关系难以用简单的数学模型来表示,所以在这个问题上我们忽略了词与词之间的松紧关系,认为他们之间的关系式相互独立的,那么 P(W)可简单表示如下:

$$P(W) \approx \prod_{1 <= i <= n} P(W_i)$$

而 $P(W_i)$ 可以用下面的公式来计算:$P(W_i) = f(W_i) / N$

其中 $f(W_i)$ 为 W_i 出现的次数,N 为语料中短语的总数。在本系统中,$f(W_i)$ 和 N 来源于《现代汉语常用词词频词典》,其中包含了 2500 万的语料抽样。

基于上述的统计模型,我们对一些常见的交集歧义进行了分词,结果如下:

组合成分子:/ / 组合/ / 成/ / 分子/ /

组合成分子时:/ / 组合/ / 成/ / 分子/ / 时/ /

大学生活:/ / 大学/ / 生活/ /

大学生活动:/ / 大学生/ / 活动/ /

学生活动:/ / 学生/ / 活动/ /

[注:]'/ / '表示分割符。

这种基于统计模型的估计方法的优点是:在语料样足够大情况下,它能够比较精确第体现 P(Wi)。但也有其缺点:首先对大规模语料库进行统计本身就是一件比较困难的事情;其次如果在计算结果中,包含了几个切分结果的值相近,就不能简单的取概率最大的一种切分结果,否则就会产生错误。在本系统中我们采取了这样一种方法,即设定一个阀值,只要切分结果的概率值大于这个域值,就把这个切分结果作为一个待选结果。然后从这几个待选结果

中用其他方法来进行筛选。比方说可以辅之以最少成词规则,自然成词规则等等。这些规则在其他的一些文章中已有详尽的描述,本文就不再赘述。

4 多音字的处理

多音字指同一个字在不同的词里由几种不同的读音。

多音字按意义是否相同可以分为多义多音字和同义多音字两类。多义多音字一个字具有两种或两种以上读音,不同的读音表示不同的意义。例如"难"字有两种读音,nan2 和 nan4。读 nan2 是"不容易、不可能、不好办"的意思,读 nan4 是"灾难、困苦"的意思。读音不同,意义也不一样,所表达的语法功能也不一样。nan2 是形容词,nan4 是名词。同义多音字一个字虽然也具有两种或两种以上的读音,但并没有明显的意义差别。例如"壳"在"蛋壳"中读 ke2,在"地壳"中读 qiqo4,读音不同,意义却没有什么不同。

多音字的产生主要是由于语音分化,异音通假,词性转化,字型的归并所造成的。

据《词海》中的 16339 个汉字的统计结果,其中多音字有 2641 个。多音字在现代汉语中占有相当的比例,但其中大部分多音字为古字,已经不再具有生命力。本文语转换系统中处理了大约 440 个常用多音字,大约涉及到一万多个词。由此可见,如果不对多音词进行处理、识别,那么在合成语音时会造成非常明显的错误,将直接影响到文语转换语音合成的自然度。

一般来说,多音字的读音可以从辩词性、意义和用法三个方面来加以辨别。但由于汉语语法和语义分析的困难性、复杂性,我们在文语转换系统采用了穷举法和一些简单的规则来解决多音字的异读问题。

穷举法

一字多音的产生,一般来说是用以区别这个字所代表的词汇意义,或者是表示不同的语法作用。据统计资料,大量的多音字异读现象和这个字所处的词汇有一一对应关系,也就是说处于特定词汇中的多音字的读音可以认为是固定的。例如:[壳]有两种读音,ke2 和 qiao4,而[壳]只有在和[地]组成[地壳]一词时才读成 qiao4。推而广之,如果一个多音字其中的一种发音只有在某种特定的词汇中才会出现,或含有这种发音的词可以用穷举的方法全部列出来,这样即可增加分词的精度,又解决了大部分的多音字的发音问题。

简单规则

a) 量词规则:比如[重],在作量词时读成 chong2,由于作量词时,前面有明显数字可以辨别,像这样的多音字还有[宿]、[服]、[发]、[行]等等。

b) 简单的上下文语法:比如[个子长高了]和[裤子太长了]中的[长]字读音不同,而且[长]字和其前后也不成词。但可以通过简单的上下文分析来解决这个问题,[太]是副词,副词可以修饰形容词,[长]在形容词时读成 chang2,[高]时形容词,可以在动词后面作补语,可以确定[长高]的[长]为动词,读成 zhang3。

5 汉语姓名的自动识别

汉语分词系统中,姓名识别一直是一个较难处理的部分,不仅因为中文姓名具有任意性、多样性,而且,中文姓名中的姓和名在很多情况下又可以作为句子的其他成分参与句子的构造和活动。中文姓名在文章中的出现频率虽然不高,但绝不可以忽略,否则在分词系统中将会出现不可预料的错误。而且由于姓名中的多音现象存在,如果不进行姓名识别会直接导致语音合成系统的发音错误。例如:[金利来有限公司董事会主席曾宪梓先生],如果不能识别出[曾宪梓]是一个中文姓名,那么,发音时[曾]字就会发成[ceng2],这样合成出来的语音听起来就十分别扭。

5.1 中文姓名识别的准备工作

·姓的分类

我们对同济大学的将近两万人的学生库进行了统计分析,考虑到样本的数量并不是太大,得到的词频并不是十分正确,所以仅对姓作了粗分,分成常用姓和非常用姓。

·名的分类

同样,从同济大学的学生库中选取了使用度较高的一部分词作为常用名,而把剩下的所有单字作为非常用名。

·简单上下文

称谓:称谓经常和姓名一起出现,可以作为姓名的边界。如:张三先生、李四教授、校长王二等等。

指介动词:指介动词经常出现在姓名的后面,可以用来帮助判断姓名的右边界。如:张三指出:……;李四说道:……等等。

特征字库:有相当一部分字[是、吧、说、叫、嚷、……]作为姓名最后一个字的概率非常小,而又经常出现在姓名的后面,同样我们可以利用它来帮助我们判断姓名的右边界。

两个特殊的字表:

a) 非常用姓氏兼最常用单字[百、次、从、大、道、发、国、过、和、红、还、回……]等等。

b) 非常用名字用字兼最常用单字[和、有、中、上、不、个、为、也、就……]等等。

5.2 中文姓名识别的过程

·姓氏的确认

如果常用姓氏的前面是前称谓语或该姓氏的后

二、三词中包含后称谓语,那么确定该姓氏为[确定姓氏]。

如果常用姓氏的前面无前称谓语或该姓氏的后二、三词中也无后称谓语,那么暂且称该姓氏为[可能姓氏]。

如果非常用姓氏的前面是前称谓词或该姓氏的后二、三词中包含后称谓,那么暂且称该姓氏为[可能姓氏]。

·姓名右边界的确定

如果该姓氏为[确定姓氏],则:

如果姓名右边第一字为常用名,标注该字为单名。

如果姓氏右边第二字为常用名,标注该姓氏右边第一字和第二字为双名。

如果姓氏右边无常用名,但右边存在标点,特征字或指介动词,同样可以定出姓名的右边界。

如果该姓氏为[可能姓氏],则:

如果姓名右边第一字为常用名,标注该字为单名,该姓氏为[确定姓氏]。

如果姓氏右边无常用名,但右边存在标点或称谓,可以确定该姓氏为[确定姓氏]。

5.3　中文姓名识别的结果

我们随即抽取了八篇文章来进行测试,测试结果如下:

文章	姓名(个)	分错	漏分	误识
1	25	1	0	1
2	17	0	1	2
3	3	0	0	0
4	20	1	0	1
5	21	2	0	3
6	12	0	2	0
7	26	1	0	2
8	4	0	0	12
总计	128	5	3	21

[注]:分错的姓名指姓名的右边界定错,如[李应发]切成[李应/ / 发]。

[注]:漏分的姓名指该姓名被误认为不是姓名。

[注]:误识的姓名指不是姓名的字被误识为姓名,如[万达俱乐部]中的[万达]被识别成姓名,[20多名干警]中的[干警]被识别成姓名。

召回率 = 文本中的中文姓名被识别的比例。

正确率 = 辩识为中文姓名中真正为中文姓名的比例。

由此我们可以得出中文姓名识别的召回率为(125/ 128) * 100 % = 97. 7 %,

正确率为(128 - 3) / (128 + 21 - 3) * 100 % = 86. 2 %。

由于本系统是一个文语转换系统,对姓名的识别要求并不象机器翻译等领域内那样严格,像某些被误识的姓名,如[万达]、[干警],恰恰可以满足文语转换系统中的分词要求。比如[万达]其实是一个机构的名称,虽然不是严格意义上的任命,但从文语转换系统的角度来说,并无大的妨碍。

6　结束语

本系统在 Intel Pentium MMX 166 上实现,分词速度在 1000 字/ 秒以上,分词精度在 99 %以上,完全可以满足文语转换系统对分词速度和精度的要求。如果对词库进一步完善,还可以提高分词精度。

笔者认为,自动分词系统是中文信息处理中的一个主要组成部分,是中文自然语言理解、文献检索、机器翻译即语音合成系统中最基本的一部分,分词的目的就是为了替后续工作作好准备。而汉语自动分词是一个极其复杂的过程,如果要达到令人满意的效果,必须对要分词的文章,句子做完善的词法、语法、语义分析,这不仅涉及到自然语言处理,还与计算机的人工智能和模式识别的发展有密切关系。而目前的分词系统一般都是基于形式分词和语法分词的基础之上,虽不能和人脑相比,但 99 %以上的分词精度应该可以满足一些对分词精度要求不高的系统。

参考文献

[1]　王永成,等. 中文词的自动处理[J]. 中文信息学报, 1989,4(4).

[2]　刘挺,等. 关于歧义字段的思考与实验[J]. 中文信息学报,1987,12(2).

[3]　王永成,等. 汉语的自动分词[J]. 上海交大学报, 1979,23(2).

[4]　姚天顺,等. 基于规则的汉语自动分词系统[J]. 中文信息学报,1989,4(1).

[5]　梁南元. 书面汉语自动分词系统[J]. 中文信息学报, 1987,2(2).

[6]　孙茂松,等. 中文姓名的自动识别[J]. 中文信息学报, 1994,9(2).

[7]　张俊盛,等. 多语料库作法之中文姓名辩识[J]. 中文信息学报,1989,4(2)

[8]　梁南元. 汉语计算机自动分词知识[J]. 中文信息学报,1989,4(4)

[9]　刘挺,等. 关于歧义字段切分的思考与实验[J]. 中文信息学报,1996,12(2)