知识组织与知识管理

基于双向最大匹配和 HMM 的分词消歧模型*

麦范金[1] 王 挺[2]
[1](桂林工学院现代教育技术中心 桂林 541004)
[2](桂林工学院电子与计算机系 桂林 541004)

【摘要】提出一种消减分词切分歧义的模型。利用正向和逆向最大匹配方法对中文文本信息进行分词,基于隐马尔科夫模型对两次最大匹配的分词结果进行对比消歧,得到较为精确的结果。整个过程分为歧义发现、歧义抽取、歧义消除 3 个过程。测试结果显示,该模型能有效地降低分词歧义引起的错误切分率。

【关键词】分词 最大匹配 隐马尔科夫模型 歧义消减
【分类号】TP391.1

Sense Disambiguation of Chinese Segmentation Based on Bi-direction Matching Method and HMM

Mai Fanjin[1] Wang Ting[2]
[1](Modern Education Technology Center, Guilin University of Technology, Guilin 541004, China)
[2](Department of Electronic and Computer Science, Guilin University of Technology, Guilin 541004, China)

【Abstract】This paper puts forward a model which can eliminate sense ambiguity of Chinese segmentation. This model segments word based on MM and RMM at first. Then it compares the segmentation results with each other, and output a more accurate result for the segmentation. The process can be divided into three parts: discovery, extraction and disambiguation. The test result shows that this model is able to reduce the error rate of segmentation, which is caused by the ambiguity of word segmentation.

【Keywords】Word segmentation Maximum matching method HMM Sense disambiguation

1 引 言

词是自然语言中最小的有意义的构成单位。中文自然语言处理的首要基本工作就是要进行中文分词。现在比较流行的分词方法主要有三大类,一类是基于规则的方法;另一类是基于统计的方法;还有一类是综合方法。这些方法大大推动了汉语分词研究的发展,但在实际应用中,一些因素和问题依旧影响着分词的精度,其中最严重的是词语切分歧义和未登录词识别这两大问题[1,2]。本文主要针对如何减少词语切分歧义、提高分词正确率,提出一种分词系统模型。该模型首先基于正向和逆向最大匹配方法对中文文本信息进行分词,然后基于隐马尔

收稿日期:2008-04-25
收修改稿日期:2008-06-12
*本文系广西教育厅科研项目"基于语意理解的垃圾邮件处理模型研究"(项目编号:桂教科研 2006[26]号)的研究成果之一。

科夫模型对两次最大匹配的分词结果进行对比消歧，最后得到较为精确的结果。

2 最大匹配法与切分歧义

正向最大匹配法（Maximum Matching Method，MM 法）和逆向最大匹配法（Reverse Maximum Matching Method，RMM 法）是两种常用的基本分词方法。它们具有原理简单、时间复杂度低、易于实现等优点，但是由于切分歧义的存在，往往容易导致切分错误[3]。根据梁南元的统计，正向最大匹配法错误切分率为 1/169，逆向最大匹配法的错误切分率为 1/245[4]。

这些切分歧义主要分为两种：一种是交叉歧义，另一种是组合歧义[1]。前者如“他的确切目的”，按正向最大匹配法切分为“他 | 的确 | 切 | 目的”，这显然是错误的。后者如“美国会制裁伊拉克”若按正向最大匹配法切分结果为“美国 | 会 | 制裁 | 伊拉克”；而按逆向最大匹配法的切分结果为“美 | 国会 | 制裁 | 伊拉克”。可见，在没有切分歧义的情况下，正向和逆向最大匹配法能得到同样的结果。然而，一旦文本中含有歧义切分字符串，那么，这两种方法将得到不同的结果，甚至是错误的结果。因此，单独采用机械分词方法效果不佳，有必要探索其他方法来消除切分歧义。

3 隐马尔科夫模型与 N－Gram 模型

马尔科夫模型（Markov Model，MM）是一个 5 元组 $M=\{\Omega,\Sigma,P,A,\Pi\}$，其中 $\Omega=\{s_1,s_2,\cdots,s_n\}$ 是所有状态的集合；$\Sigma=\{\sigma_1,\sigma_2,\cdots,\sigma_n\}$ 是所有观察序列的集合；$P=\{p_{ij}\}_{n\times n}$，p_{ij} 是概率转移矩阵，表示马尔科夫链中状态从 s_i 到 s_j 的概率；$A=\{\alpha_{ij}\}_{n\times m}$，其中 α_{ij} 是发射概率矩阵，表示在状态 s_i 观察到 σ_j 的概率；$\Pi=\{\pi_1,\pi_2,\cdots,\pi_n\}$ 是初始概率向量[1]。在马尔科夫模型中，如果第 i 个状态出现的概率与它前面 i－1 个状态有关，那么这个马尔科夫模型就被称之为 i－1 阶马尔科夫模型。在语言模型中，某个词出现的概率依赖于它前面出现的词[5]，即第 n 个词出现的概率只与它之前的 n－1 个单词有关，这就是一个语言的 n－1 阶马尔科夫模型。在语言学中，它被称为 N－Gram 模型，也叫 N 元文法模型[5]。当马尔科夫模型的观察序列已知而实际状态未知时，就被称之为隐马尔科夫模型（Hidden Markov Model，HMM）[1]。

4 切分歧义消解策略

4.1 消歧模型

分词歧义消解一般是与分词过程结合在一起的，可以通过分析上下文语境来解决[1]。实验证明，基于统计的方法进行分词的精确度要高于基于规则的分词方法[2]，但其速度较慢[5]。为了兼顾分词精度与速度，可以采用将基于规则的方法和基于统计的方法相结合的办法来进行。首先通过正向和逆向最大匹配法对输入文本进行分词。然后对两种分词结果进行字符串对比，抽取其中含有切分歧义的部分，将没有切分歧义的字符串直接过滤。如果在此过程中发现未登录词，则将该词存放入分词词典。然后，通过分析上下文语境，统计语料信息，将抽取出来的含有切分歧义的字符串视为某些随机过程的状态集合，通过考察观察状态，预测可能形成此观察状态的实际状态，即预测可能组成该歧义字符串的实际词语。将在此过程中发现的未登录词存放入分词词典，给出分词最终预测结果，同时将此结果更新进语料库，供未来的统计使用。整个歧义消解过程可分为 3 个步骤，即歧义发现、歧义抽取和歧义消除。图 1 是分词消歧的模型示意图。

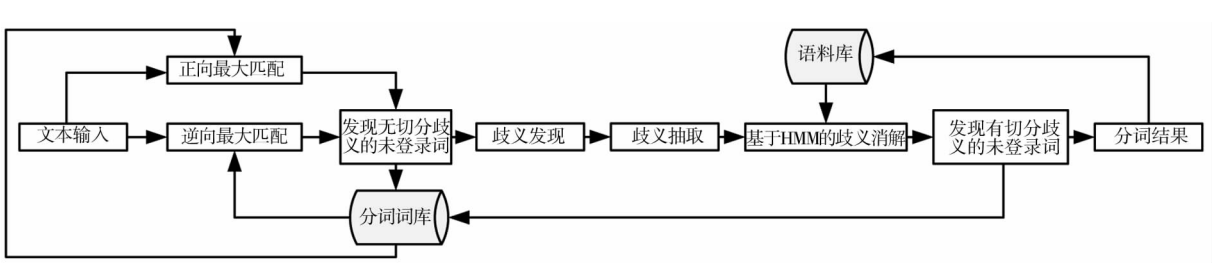

图 1 分词消歧模型示意图

4.2　基于 HMM 消除切分歧义

N－Gram 模型是一种考虑上下文语境的统计语言模型[1]。已知含有切分歧义的字符串及其 N－Gram 模型，也就已知了一个观察序列和这个序列的 HMM。利用这个 HMM，能够计算出组成这个歧义切分字符串的可能的真实词，也就是计算 HMM 中最有可能产生这个观察序列的状态序列，这是 HMM 的解码问题[6]。一个直观的方法是穷举法，即算出所有可能状态序列的概率，然后选择可能性最大的状态序列。但这样做的计算复杂度太高。另一种想法是保留每一次计算后的最优值，而把其他可能情况都省略。下一轮计算仅在上一轮最优的基础上展开。因此，在整个计算过程中，并不把所有的情况都进行穷举，计算速度大大提高，而且能够保证得到某种意义下最有可能产生的词组序列。其算法的伪码如例 1 所示：

例 1：

```
Begin initialize path←{ },AmbiguousString,SubString←{ }
  While( AmbiguousString. Length >0)
    {
      //只考虑以当前 HMM 第一个状态开始的匹配序列
        Substring←以 AmbiguousString 中的第一个字为基准,
          取出所有可能的匹配符串
      For each SubString
        {
          //提供当前情况下所有的概率,为判断歧义作参考
        计算当前每一种可能情况的概率 P(SubString)
        }
        //选择概率最大的 SubString 添加到 Path
        将 argmax(P(SubString)) 添加 Path
        //准备考察除去最大概率的 SubString 后的 Ambiguous-
          String,从 HMM 序列首部开始,除去所有的匹配状态
           AmbiguousString. Remove (0, argmax ( P ( String )).
             Length)
    }
    Returm Path
End
```

4.3　概率计算与数据平滑

要计算可能状态序列的概率，就需要对语料库进行统计。因此必须首先进行语料训练。假设 c(w) 表示词 w 在训练文本中出现的总次数，N 为训练文本中的总词数。采用最大似然估计，则词 w 出现的概率为：

$$P_{ML}(w)=\frac{c(w)}{N} \tag{1}$$

在 N－Gram 模型中，

$$P_{ML}(w_i|w_1,w_2,\cdots,w_{i-1})=\frac{c(w_1,w_2,\cdots,w_i)}{c(w_1,w_2,\cdots,w_{i-1})} \tag{2}$$

可将 $w_1,w_2,\cdots,w_i$ 简写为 w_1^i，则式(2)可简写为：

$$P_{ML}(w_i|w_1^{i-1})=\frac{c(w_1^i)}{c(w_1^{i-1})} \tag{3}$$

但是，对分词系统进行的语料训练具有局限性，如语料数量的局限性、语料采集年代的局限性等，在进行正式文本测试时，很可能会因为这些局限性，导致一些本应出现的语料没有出现，因此得到不可靠的估计结果，即零概率问题；或者一些本应相对大量多次出现的语料仅出现很少的几次，从而使得所得到的估计不可靠，即数据稀疏(Data Sparseness)问题。数据稀疏和零概率问题几乎永远存在，这对于基于统计的自然语言处理来说是非常严重的。可以采用统计数据平滑技术(Data Smoothing)来避免数据稀疏和零概率问题导致的统计失误[7]。现在比较常见的平滑方法有加法平滑(Additive Smoothing)[8]、Good－Turing 平滑[9]、线性插值平滑(Linear Interpolation Smoothing)[10]、Katz 平滑[11]、Kneser－Ney 平滑[12]和 Witten－Bell 平滑[13]等。

4.4　未登录词的处理

在分词消歧模型内未登录词的处理分为两次进行。第一次在对照两种最大匹配分词结果后，提取不含切分歧义的字符串中的未登录词。其具体办法是：如果在规则匹配方法分词过程中，直到最后一个单字依然无法找到匹配项，则将这个单字放入暂存器。如果下一轮匹配还有类似的情况发生，则把两次所放入暂存器的字符串组合起来，继续放在暂存器中，直到有一轮匹配能够成功完成为止。这时，先输出暂存器中的组合字符串，再输出新匹配完的字符串。暂存器中先输出的组合字符串就是新识别的未登录词。这种方法比较简单，且能够识别未登录词。但是如果有两个或多个未登录词紧挨在一起，系统将会认为这仅是一个未登录词。为了尽可能地避免这种情况的发生，需要对整个系统进行大量的语料训练。通过大规模的语料训练，降低两个或多个未登录词同时出现的概率，然后通过概率统计的方法确定未登录词。

分词消歧模型第二次处理未登录词是在消歧完成之后，对有切分歧义的字符串中的未登录词进行处理，将其存放入分词词典。

5　系统实验分析

5.1　实验平台建设

根据以上论述，笔者采用 ASP. NET 构建了一个基

于双向最大匹配和隐马尔科夫模型的中文分词消歧模型。其中的 N－Gram 模型采用二元文法模型,即 Bigram 模型。在 Bigram 模型中,式(3)可写成:

$$P_{ML}(w_i|w_{i-1}) = \frac{c(w_{i-1},w_i)}{c(w_{i-1})} \quad (4)$$

其中的数据平滑方法采用 Katz 平滑技术。该技术是 Katz 根据 Good－Turing 平滑提出的一种回退式数据平滑(Backing－off Smoothing)算法。其基本思想是:设定 k 为 Katz 系数,对出现次数介于 0 和 k 之间的 N 元词组,将其按折扣率 d_r 扣除一些频度,分配给出现次数为 0 的 N 元词组;如果 N 元词组出现的次数大于 k,则采用最大似然估计。根据 Katz 的建议,k 值一般取 5[11]。在 Bigram 模型中 Katz 平滑的公式是:

$$P_{Katz}(w_i|w_{i-1}) = \begin{cases} P_{ML}(w_i|(w_{i-1}) & c(w_{i-1},w_i) > k \\ d_r \cdot P_{ML}(w_i|w_{i-1}) & 0 < c(w_{i-1},w_i) \le k \\ \alpha(w_{i-1}) \cdot P_{ML}(w_i) & c(w_{i-1},w_i) = 0 \end{cases} \quad (5)$$

其中,d_r 是折扣率,$\alpha(w_{i-1})$ 与 d_r 是保证模型参数的归一化约束条件。

$$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}} \quad (6)$$

$$\alpha(w_{i-1}) = \frac{1 - \sum_{c(w_{i-1},w_i)>k} P_{ML}(w_i|w_{i-1} - \sum_{0<c(w_{i-1},w_i)\le k} d_r \cdot P_{ML}(w_i|w_{i-1})}{\sum_{c(w_{i-1},w_i)=0} P_{ML}(w_i)}$$
$$= \frac{1 - \sum_{c(w_{i-1},w_i)>k} P_{ML}(w_i|w_{i-1} - \sum_{0<c(w_{i-1},w_i)\le k} d_r \cdot P_{ML}(w_i|w_{i-1})}{1 - \sum_{c(w_{i-1},w_i)>0} P_{ML}(w_i)} \quad (7)$$

r^* 是对于 N－Gram 模型中出现 r 次的 N 元词组,根据 Good－Turing 估计[9],该 N 元词组的出现次数:

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \quad (8)$$

5.2 语料训练

在进行语料训练之前,首先建设一个分词词典,包括 128 000 个中文词汇,不包括英文词汇和标点。为了使后续的分词效果更好,该词典涵盖了文献《普通话三千常用词表》[14]。

所选择的训练用语料包含两部分:一是改革开放至今,由人民教育出版社出版的小学语文课本教材中的课文 304 篇,共计 183 349 字,其中包含标点符号。这些课文不包括古诗词、现代诗歌、戏剧、文言文和早期白话文;二是选择 2007 年 10 月 1 日到 10 日的《人民日报》,共计 459 059 字,其中包含标点符号。两项总计 642 408 字,其中包含标点符号。

5.3 实验结果

选择 2007 年 10 月中旬《人民日报》上的两部分文本和人民教育出版社出版的初中语文课本中《纪念白求恩》、《皇帝的新装》两篇课文,总共 23 274 字,对所提出的分词消歧模型进行测试。

笔者采用两个指标对该分词模型的分词效果进行评价,一是公认的最终结果:分词准确率;二是梁南元所说的错误切分率[4]。测试结果如表 1 所示,其中,准确率采用如下公式计算:

$$准确率 = \frac{正确的切分数}{所有的切分数} \times 100\% \quad (9)$$

错误切分率采用如下公式计算:

$$错误切分率 = \frac{错误切分数}{总字数} \quad (10)$$

表 1 测试结果

文本	字数	总切分数	错误切分数	准确率	错误切分率
《人民日报》第 1 部分	10 307	5 794	14	99.76%	1/736
《人民日报》第 2 部分	9 017	5 054	17	99.66%	1/530
《纪念白求恩》	986	595	6	98.99%	1/164
《皇帝的新装》	2 964	2 038	7	99.66%	1/423
合计	23 274	13 481	44	99.67%	1/529

对比正向最大匹配法 1/169 的错误切分率和逆向最大匹配法 1/245 的错误切分率,可以认为基于双向最大匹配和隐马尔科夫模型的中文分词消歧模型具有较好的分词和消歧能力,能够有效地同时消除交叉歧义和组合歧义,具有一定的实用性。

6 结 语

有关中文分词方法的研究已经进行了许多年,但是,至今一直没有出现完全令人满意的解决办法。为了兼顾分词的速度与精度,本文提出一种规则方法和机械方法相结合的分词模型:利用正向和逆向最大匹配方法对中文文本信息进行分词;基于隐马尔科夫模型对两次最大匹配的分词结果进行对比消歧,同时识别未登录词;得到较为精确的分词结果。测试结果显示,该模型能够有效地提高分词准确率,降低错误切分率,较一些基本方法更具实用价值。

参考文献:

[1] 王晓龙,关毅. 计算机自然语言处理[M]. 北京:清华大学出版社,2005.

[2] 黄昌宁,赵海. 中文分词十年回顾[J]. *中文信息学报*,2007,21(3):8 - 19.

[3] 刘颖. 计算语言学[M]. 北京:清华大学出版社,2002.

[4] 梁南元. 书面汉语自动分词系统——CDWS[J]. *中文信息学报*,1987(2):44 - 52.

[5] 王小捷,常宝宝. 自然语言处理技术基础[M]. 北京:北京邮电大学出版社,2002.

[6] Duda R O, Hart P E, Stork D G. Pattern Classification[M]. 2nd Edition. York:Wiley New, 2001.

[7] Jurafsky D, Martin J H. Speech and Language Processing:An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition[M]. USA:Prentice Hall, 2000.

[8] Jeffrey H. Theory of Probability[M]. Oxford:Oxford University Press, 1948.

[9] Good I J. The Population Frequencies of Species and the Estimation of Population Parameters[J]. *Biometrika* , 1953, 40(3 - 4):237 - 264.

[10] Jelinek F, Mercer R L. Interpolated Estimation of Markov Source Parameters from Sparse Data[C]. In:*Gelsema E. S. and Kanal L. N.* (*eds.*) *Pattern Recognition in Practice*, North Holland, Amsterdam, 1980:381 - 397.

[11] Katz S M. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer[J]. *IEEE Transactions on Acoustics*, *Speech*, *and Signal Processing*, 1987, 35(3):400 - 401.

[12] Kneser R, Ney H. Improved Backing - off for M - Gram Language Modeling[C]. *In*:*Proceedings of the IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 1995(1):181 - 184.

[13] Witten I H, Bell T C. The Zero - frequency Problem:Estimating the Probabilities of Novel Events in Adaptive Text Compression[J]. *IEEE Transactions on Information Theory*, 1991, 37(4):1085 - 1094.

[14] 郑林曦. 普通话三千常用词表[M]. 北京:语文出版社,1987.

(作者 E - mail:328dickwong1981@163.com)