基于隐马尔科夫模型的中文分词研究

魏晓宁
（南通大学 计算机科学与技术学院，江苏 南通 226019）

摘要：一直以来，汉语自动分词是公认的汉语信息处理瓶颈。反思现有汉语自动分词技术，发现均有隐含两大假设：语言是规律的、词具有确定边界。这与语言的复杂性、组合性、动态性、模糊性特征不符。本文采用一种基于隐马尔科夫模型（HMM）的算法，通过CHMM（层叠形马尔科夫模型）进行分词，再做分层，既增加了分词的准确性，又保证了分词的效率。

关键词：自动分词；隐马尔科夫模型（HMM）；N-最短路径粗切分；统计模型
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-3044(2007)21-40885-02

HMM-Based Of Study On Chinese Language Classifying Words

Wei Xiao-ning
(College of Computer Science&Technology Nantong University,Nantong 226019,China)

Abstract: All along, Chinese language automatic classifying words is universally acknowledged bottlenecks during processing Chinese language .There stand two concealing supposes .By introspecting existing current Chinese language automatic classifying technology .For languages have the character of regularity and words have their own determining frontier ,which don't accord with their complication ,compose ,tendency and indistinct. The paper provided a HMM-based arithmetic ,via CHMM to classify the words and then to divide layers once more. This way can assure the precise and efficiency of classifying the words.

Key words: Automation participle;Hidden Markov Model(HMM);Most fault route segments N-roughly;Count a model

汉语自动智能分词是中文信息处理的基础与关键。随着中外文机器翻译研究的深入和自然语言理解，电子词典等中文词语处理技术应用的扩展，对汉语自动分词软件的要求越来越高。近年来我国已经开发了多种现代书面汉语自动分词软件，国内众多研究机构已经在计算机汉语文本自动分词方面进行了大量的研究，并取得了很多成就。虽然这方面的研究和应用正在不断深入，但到目前为止还没有评价此类软件的标准模型和方法。

1 中文分词方法

汉语自动分词不同于英文中的分词，汉语文本是大字符集上的连续字串，以字为单位，句子中所有的字连起来才能描述一个意思。中文句子和段落可以通过明显的分界符来简单划界，而句中词与词之间并没有明显的界限标志，因此在分词时尤为困难。

针对于中文语句的这一特性，在处理分词时就必须要考虑几个方面的问题。词语切分、未定义词识别、词性标注。常用的分词方法有：1.基于字符串匹配的分词方法；2.基于统计的分词方法；3.基于规则和基于统计相结合。

2 基于语料库的统计语言学方法

近年来，基于语料库分析的自然语言处理方法受到了越来越多的计算语言学家的重视和应用。在规则方法即理性主义方法屡受挫折的事实面前，语料库语言学的发展促使计算语言学家们越来越重视数理统计在语言学中的应用。

传统语言学给我们积累了丰富的语言实例，但对于语言规律的把握，人类至今还没有找到最好的方法。但是，数理统计方法已经发展的比较成熟，值得信赖。语料库是经过处理的大量领域文本的集合，通过对语料库中的文本进行统计分析，可以获取该类文本的某些整体特征或规律。如果能够充分地利用这些统计现象、规律，就可以构造基于语料库的统计学信息抽取算法。

统计的分析方法多种多样，近期研究的热点主要集中于由随机过程发展而来的理论和方法。其中最重要的是应用隐马尔科夫模型（HMM）进行自然语言处理的方法。

3 隐马尔科夫模型（HMM）简介

3.1 马尔科夫（Markov）过程的定义

一般地，考虑只取有限个（或可数个）值的随机过程$\{X_n|n=1,2,\cdots\}$：若$X_n=i$，就说过程在n时刻处于状态，假设每当过程处于状态i，则过程在下一时刻处于状态j的概率P_{ij}为一定值，即$\forall n\geq 1$有：

$$P_{ij}=P(X_{n+1}=j|X_n=i,X_{n-1}=i_{n-1},\cdots,\cdots,X_1=i_1)=P(X_{n+1}=j|X_n=i)$$

这样的随机过程称为Markov链（给定过去的状态$X_1,\cdots,X_{n-1}$和现在的状态X_n，将来的状态X_{n+1}的条件分布独立于过去的状态，只依赖于现在的状态——这就是Markov性）。

一个马尔科夫模型（MM）M就是一个Markov链加上一个转移概率矩阵。显然，它可被视为一个随机有限状态自动机，其每个状态都代表一个可观察的事件，之间的转换都对应一定的概率。

3.2 隐马尔科夫模型（HMM）的概念

对于马尔科夫模型而言，每个状态都是决定性地对应于一个可观察的物理事件，所以其状态的输出是有规律的。然而，这种模型限制条件过于严格，在许多实际问题中无法应用。于是人们将这种模型加以推广，提出了隐马尔科夫模

收稿日期：2007-09-12
作者简介：魏晓宁（1977-），女，江苏省南通市人，讲师，硕士研究生，研究方向：中文信息处理。

型(HMM)。隐马尔科夫过程是一种双重随机过程。即：观察事件是依存于状态的概率函数，这是在HMM中的一个基本随机过程，另一个随机过程为状态转移随机过程，但这一过程是隐藏着的，不能直接观察到，而只有通过生成观察序列的另外一个概率过程才能间接地观察到。

对于隐马尔科夫模型的应用，在语音识别领域已经取得了很好的成效，在信息抽取领域的应用也正在不断的尝试和推广中。

3.3 隐马尔科夫模型(HMM)的模型参数

(1)N：模型状态数。

(2)M：每个状态可能输出的观察符号的数目。

(3)T：观察符号序列的长度。

(4)$A=\{a_{ij}\}$ ：状态转移概率矩阵。

(5)$B=\{b_j(k)\}$ ：观察符号的概率分布集。

(6)$\pi=\{\pi_i\}$ ：初始状态概率分布。

一般地，由于当A、B确定后，M、N也随即确定，故通常将一个HMM描述为$\lambda(A,B,\pi)$。

3.4 隐马尔科夫模型的训练与优化问题

隐马尔科夫模型可描述为$\lambda(A,B,\pi)$，如何确定其中的A、B和π就是所谓的模型参数获取问题。

到目前为止，对于隐马尔科夫模型的参数选择和优化问题，还没有什么分析算法可以得到最优解。目前使用较广的处理方法是Baum-Welch估计算法(或称期望值修正法，即EM法)。该算法是一种迭代算法，初始时刻由用户给出各参数的经验估计值，通过不断迭代，使个参数逐渐趋向更为合理的较优值。算法可简单描述如下：

(1)初始化：$\gamma_1(i)=\bar{\pi}$，时间t=1时处于状态S_i的期望值：$\lambda=M(A_0,B_0,\bar{\pi})$

(2)迭代计算：令$\lambda_0=\lambda$

$$\bar{a}_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}=\frac{\sum_{t=1}^{T-1}a_t(i)a_{ij}b(o_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T-1}a_t(i)\beta_t(i)},$$

$$\bar{b}_j(k)=\frac{\sum_{t=1,o_t=v_k}^{T}\gamma_t(i)}{\sum_{t=1}^{T}\gamma_t(i)}=\frac{\sum_{t=1}^{T}a_t(i)\beta_t(i)\delta(o_t,v_k)}{\sum_{t=1}^{T}a_t(i)\beta_t(i)},\ \text{其中：}\ \delta(o_t,v_k)=\begin{cases}1, & if\ o_t=v_k\\ 0, & otherwise\end{cases}$$

$$\lambda=M(\bar{A},\bar{B},\bar{\pi})$$

(3)终止条件：$|\log P(O|\lambda)-\log P(O|\lambda_0)|<\varepsilon$，其中$\varepsilon$是预先设定的阈值。

本文应用Baum-Welch算法获取模型参数，但对算法做了适当的更改。最主要的修改是上述算法中的终止条件。与应用在语音识别中的隐马尔科夫模型不同，衡量模型质量时，并不是要求整个模型输出某一序列的总体概率最大就为最优，而是输出该序列时所经历的隐路径中最佳路径的概率最大为最优。所以，第三步应改为：

终止条件：$\left|\log P\left(\max_j Q^j|\lambda\right)-\log P\left(\max_j Q^j|\lambda_0\right)\right|<\varepsilon$，$\varepsilon$为阈值。

4 模块的主要功能及测试结果

本系统的主要设计思想是：先进行原子切分，然后在此基础上进行N-最短路径粗切分，找出前N个最符合的切分结果，生成二元分词表，再生成分词结果，接着进行词性标注并完成主要分词步骤。分词模块的主要功能其第一步是原子分词。所谓原子，是指该短句中不可分割的最小语素单位。但在进行原子切分之前，首先要进行断句处理。所谓断句，就是根据分隔符、回车换行符等语句的分隔标志，把源字符串分隔成多个稍微简单一点的短句，再进行分词处理，最后把各个分词结合起来，形成最终的分词结果。分成短句之后，即可进行原子分词。例如：索爱K-300型号的手机1元钱，则K-300、1都是一个原子，其它的每个汉字是一个原子。

按照这种方式，通过简单的汉字分割就形成了原子分词的结果，并对每个原子单位进行词性标注。npos=1表示是开始标记，npos=4表示结束标记，npos=0表示未识别词。经过原子分词之后，就可进行初次分词。经过原子分词后，源字符串成了一个个独立的最小语素单位。下面的初次切分，就是把原子之间所有可能的组合都先找出来。算法是用两个循环来实现，第一层遍历整个原子单位，第二层是找到一个原子时，不断把后面相邻的原子和该原子组合到一起，访问词典库看它能否构成一个有意义的词组。

系统在语料库评测中的测试结果：

表1 分词、词性标注精度表

语料领域	分词总数	分词正确率(%)	词性标注正确率(%)
IT	2348	97.01	86.77
财经	1524	96.40	87.47
法制	2668	98.44	85.26
理论	2225	98.12	87.29
教育	1765	97.80	86.25
总计	10530	97.58	87.32

从表1中可以看出，系统的中文分词精度在95%以上，这是一个比较理想的结果，基本上可以满足在分词精度方面的要求。

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。通过对中文分词技术的深入研究，高质量、多功能的分词系统的开发，必将促进中文信息处理系统的广泛应用，为用户提供更多的服务。

参考文献：

[1]张华平，刘群.基于N-最短路径的中文词语粗分模型[J].中文信息学报，2002,16(5):1-7.

[2]于江生.计算语言学中的概率统计方法[M].北京：北京大学计算语言学研究所，1999.

[3]于江生.隐Markov模型及其在自然语言处理中的应用[M].北京：北京大学计算语言学研究所.

[4]陈桂林，王永成.等.一种改进的快速分词算法[M].计算机研究与发展，2000.

[5]A Comparative Study On Chinese Text Categorization Methods[J].He Chew-Lim Tan 2000,24-25.