第36卷 第10期 计 算 机 工 程 2010年5月
Vol.36 No.10 Computer Engineering May 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)10—0231—02 文献标识码：A 中图分类号：TP391.1

基于上下文的二阶隐马尔可夫模型

刘洁彬[1]，宋茂强[1]，赵 方[1]，杨志宇[2]
(1. 北京邮电大学软件学院，北京 100876；2. 北京航天航空大学软件学院，北京 100083)

摘 要：为体现上下文信息对当前词汇词性的影响，在传统隐马尔可夫模型的基础上提出一种基于上下文的二阶隐马尔可夫模型，并应用于中文词性标注中。针对改进后的统计模型中由于训练数据过少而出现的数据稀疏问题，给出基于指数线性插值改进平滑算法，对参数进行有效平滑。实验表明，基于上下文的二阶隐马尔可夫模型比传统的隐马尔可夫模型具有更高的词性标注正确率和消歧率。

关键词：词性标注；二阶隐马尔可夫模型；参数平滑；Viterbi 算法

Second-order Hidden Markov Model Based on Context

LIU Jie-bin[1], SONG Mao-qiang[1], ZHAO Fang[1], YANG Zhi-yu[2]
(1. College of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876;
2. College of Software, Beihang University, Beijing 100083)

【Abstract】 To better represent the influence of the context to the part of speech of the current word, this paper proposes a second-order hidden Markov model based on the traditional hidden Markov model and applies it to part-of-speech tagging in Chinese. In the improved statistical model, sparse data problem occurs due to the shortage of training data. To solve this problem, an improved smoothing algorithm based on index linear interpolation is proposed, which provides effective smoothing. Experiments show that the second-order Hidden Markov Model(HMM) based on the context has higher correct rate and disambiguation rate of part-of-speech tagging than the traditional hidden Markov model.

【Key words】 part-of-speech tagging; second-order Hidden Markov Model(HMM); parameter smoothing; Viterbi algorithm

1 概述

现有的词性标注方法主要分为基于规则的方法和基于统计的方法[1]。基于规则的方法具有方法简单、易于实现等特点，但由于其灵活性和自适应能力均比较差，在实际应用中很难被广泛地应用。基于统计的方法恰能弥补这一缺点，它能够在统计的基础上找到最大可能性的词性标注，以达到自主学习的目的，因此，具有较好的灵活性和自适应性。目前，应用于基于统计的词性标注模型有很多，例如，隐马尔可夫模型、最大熵模型、条件随机场等模型，其中，隐马尔可夫模型是统计模型中应用效果较好的模型之一。

虽然隐马尔可夫模型及其衍生模型均有很好的词性标注效果，但现有模型的共同特点是当前词的词性仅根据上文的信息判定[2]，即仅考虑了上文对当前词的影响，未考虑下文对当前词的影响。然而中文自然语言是上下文依赖关系很强的一种语言，针对这一特性本文提出一种基于上下文的二阶隐马尔可夫模型，并将改进后的模型应用于汉语的词性标注。

2 隐马尔可夫模型及其应用

2.1 隐马尔可夫模型

隐马尔可夫模型(Hidden Markov Model, HMM)[3]是Rabiner等人在20世纪80年代提出的一种特殊Markov模型，并通过最大期望算法(expectation maximization)训练模型参数。HMM具有很强适应性使得特别适合于处理随机序列数据，在语音识别、手写字符识别、图像处理等诸多领域得到了广泛地应用。隐马尔可夫模型由以下几个部分组成：

状态集合 $S=\{S_1,S_2,\cdots,S_n\}$
输出字母表 $K=\{K_1,K_2,\cdots,K_m\}$
起始概率向量 $\boldsymbol{\pi}$
状态转移概率矩阵：
$\boldsymbol{A}=\{a_{ij}\},1\le i\le N,1\le j\le N$
符号发生概率矩阵：
$\boldsymbol{B}=\{b_{ik}\},1\le i\le N,1\le k\le M$
状态序列 $X=X_1,X_2,\cdots,X_t$
输出序列 $O=O_1,O_2,\cdots,O_t$

其中，S，$\boldsymbol{A}$，X，$\boldsymbol{\pi}$ 基本与Markov模型中的概念一致，HMM增加输出序列 O、输出字母表 K 和一个符号发生概率矩阵 $\boldsymbol{B}$。实际上，HMM是一个二重Markov随机过程，它包括了状态转移的随机过程和观察值输出的随机过程，其中状态转移随机过程是隐式的，它通过观察序列随机过程表现出来。HMM的意义在于可以通过观察到的表层事件反映深层事件。为了使模型不过于复杂，实际应用中一般使用简单的一阶HMM，即假设状态的转移概率只与前一个状态有关，但这样也限定了解决状态转移关系复杂问题的准确性。

2.2 隐马尔可夫模型在词性标注中的应用

本文给出一个隐马尔可夫模型应用于中文词性标注的示例。首先建立隐马尔可夫模型，假设词汇集为 W 为HMM中输出字母表 K；词性集 T 为HMM中状态集合 S；词汇序列 O 为HMM中输出序列 O；词性序列 S 为HMM中状态序列 X；π 为初始状态概率分布，π_i 表示初始状态为词性 t_i 的概

基金项目：国家“863”计划基金资助项目“高精度高鲁棒性室内定位关键技术装置研究”(2007AA12Z321)
作者简介：刘洁彬(1986－)，女，硕士研究生，主研方向：自然语言处理，通信软件；宋茂强，教授；赵 方，博士研究生；杨志宇，硕士研究生
收稿日期：2009-11-20 **E-mail：**liujiebin1986@126.com

率；参数 $\boldsymbol{A}$ 为词性状态转移矩阵，即 HMM 中状态转移概率矩阵，其中 a_{ij} 为从词性 t_i 转移到词性 t_j 的概率；参数 $\boldsymbol{B}$ 为词汇发射概率矩阵，即 HMM 中符号发生概率矩阵，其中 $b_j(w_t)$ 为当词性标注为 t_j 情况下输出词汇 w_t 的概率。

词性标注问题可以描述为给定词汇序列 $O=o_1,o_2,\cdots,o_m$ 条件下，找到词性序列 $S=s_1, s_2,\cdots,s_m$，使 $P(S\mid O)$ 最大。HMM 中 Viterbi 算法能够计算出全局最优的词性标注序列。

2.3 Viterbi 算法

Viterbi 算法如下：

初始化：

$$\delta_1(i)=\pi_i b_i(o_1) \qquad 1\le i\le N \tag{1}$$

$$j_1(i)=0 \qquad 1\le i\le N \tag{2}$$

递归：

$$\delta_m(j)=\max_{1\le i\le N}[\delta_{m-1}(i)a_{ij}]b_j(o_m),2\le m\le M,1\le j\le N \tag{3}$$

$$j_m(j)=\arg\max_{1\le i\le N}[\delta_{m-1}(i)a_{ij}],2\le m\le M,1\le j\le N \tag{4}$$

终结：

$$p^*=\max_{1\le i\le N}[\delta_M(i)] \tag{5}$$

$$q^*_M=\arg\max_{1\le i\le N}[\delta_M(i)] \tag{6}$$

从 Viterbi 算法的计算过程中可以看出传统隐马尔可夫模型在计算词性标注最佳序列时，仅使用 a_{ij} 这个参数，即 w_t 的词性标注概率仅依赖与 w_{t-1} 的词性。因此，传统隐马尔可夫模型具有考虑上下文信息少的缺点，使其准确性不能满足现代分词准确率的要求。

3 基于上下文的二阶隐马尔可夫模型

传统隐马尔可夫模型计算词性标注最佳序列时，仅考虑上文词性对当前词词性的影响。针对这一不足，本文提出一种基于上下文的二阶隐马尔可夫模型。

3.1 二阶隐马尔可夫模型的建立

传统隐马尔可夫模型中参数的定义仅和上文有关，要做到与上下文有关首先就要重新定义参数。本文对基于上下文的二阶隐马尔可夫模型 $\theta=(N,M,\pi,A,B,C)$ 中各参数重新定义如下：

(1)N 为词性标注系统中所用到标记集词性的个数；

(2)M 为词汇集词汇的个数；

(3)π_i 为词性 t_i 作为句首出现的概率；

(4)词性概率转移矩阵 $\boldsymbol{A}=\{a_{ijk}\}$，其中，

$$a_{ijk}=P(s_m=t_j\mid s_{m-1}=t_i,s_{m+1}=t_k)$$

$$3\le m\le M-1,\ 1\le i,j,k\le N \tag{7}$$

状态 t_j 的词性转移概率不仅和上文 c_{m-1} 的词性 t_i 有关系，并且和下文 c_{m+1} 的词性 t_k 也有关系，这样就能同时获取上下文的信息；

(5)词汇概率矩阵 $\boldsymbol{B}=\{b_{ij}(w_k)\}$，其中，

$$b_{ij}(w_k)=P\{o_m=w_k\mid s_m=t_i,s_{m+1}=t_j\}$$

$$1\le m,k\le M,1\le ij\le N \tag{8}$$

其中，$b_{ij}(w_k)$ 表示在 s_m 的状态 t_i，且 S_{m+1} 在状态 t_j 条件下，O_m 输出为 w_k 的概率；

(6)句末词汇概率 $C=\{c_i(w_k)\}$，其中，

$$c_i(w_k)=P\{o_M=w_k\mid s_M=t_i\},\ 1\le i\le N,1\le k\le M \tag{9}$$

其中，$c_i(w_k)$ 表示当句子末尾词的词性为 t_i 条件下，O_M 输出为 w_k 的概率。

3.2 参数平滑

最初改进后二阶隐马尔可夫模型中 a_{ijk} 和 $b_{ij}(w_k)$ 参数值采用极大似然估计进行计算：

$$a_{ijk}=N(t_i,t_j,t_k)/N(t_i,t_k) \tag{10}$$

$$b_{ij}(w_k)=N(w_k,t_i,t_j)/N(t_i,t_j) \tag{11}$$

采用该方法获取数据时，对语料中未出现的情况均将其值设为 0，这样 0 值会一直传递下去，使得用极大似然估计方法所得数据矩阵为稀疏矩阵。为了解决这一问题，必须把少量概率分配到未出现情况当中，称之为数据平滑。

常用的数据平滑模型分为 2 类：回退模型和线性插值模型。回退模型的代表是 Katz Smoothing 算法，是一种基于 Good-Turing 估计的退化平滑模型。基本思想是：从那些已知中按一定比例扣除一些频率，分配给未知事件；线性插值模型的代表算法为 Jelinek-Mercer Smoothing。其基本思想是：如果没有足够数据准确估计高阶 n 元模型，可以使用更低阶的 n-1 元模型信息对高阶模型进行插值，因为低阶模型受数据稀疏问题的影响相对较小。

文献[4]提出了一种基于插值模型基础上 Bi-gram 指数线性插值模型，并且证明出此平滑模型优于传统的线性插值模型。该技术有效地解决了数据稀疏问题，且需要训练的参数较少，花费时间和空间复杂度低，平滑效果却很好。由于文献[4]中插值模型的应用环境与本文的应用环境有些不同，因此，在对改进二阶隐马尔可夫模型进行数据参数平滑时本文将文献[4]中的插值模型做了进一步改进，得到改进后二阶隐马尔可夫模型进行数据参数平滑后的公式如下：

(1)词性概率转移矩阵

$$a_{ijk}=\lambda_0\frac{N(t_i,t_j,t_k)}{N(t_i,,t_k)}+\lambda_1\frac{N(t_i,t_j)}{N(t_i)}+\lambda_2\frac{N(t_j,t_k)}{N(t_k)}$$

$$(\lambda_0+\lambda_0+\lambda_0=1) \tag{12}$$

(2)词汇概率矩阵平滑

$$b_{ij}(w_k)=(1-\lambda)\frac{N(w_k,t_i,t_j)}{N(t_i,t_j)}+\lambda\frac{N(w_k,t_i)}{N(t_i)}$$

$$(\lambda=e^{-N(w_k,t_i,t_j)}) \tag{13}$$

3.3 改进的 Viterbi 算法

改进后的二阶隐马尔可夫模型对参数的定义进行了相应调整，因此传统 Viterbi 算法不能很好地应用于改进后的模型。针对修改后的模型 Viterbi 算法进行如下改进：

初始化：

$$\delta_1(i,j)=\pi_i b_{ij}(o_1),1\le i,j\le N \tag{14}$$

$$j_1(i,j)=0,1\le i,j\le N \tag{15}$$

递归：

$$\delta_m(j,k)=\max_{1\le i\le N}[\delta_{m-1}(i,j)a_{ijk}]b_{jk}(o_m)$$

$$2\le m\le M-1,1\le j,k\le N \tag{16}$$

$$j_m(j,k)=\operatorname{argmax}_{1\le i\le N}[\delta_{m-1}(i,j)a_{ijk}]$$

$$2\le m\le M-1,1\le j,k\le N \tag{17}$$

终结：

$$p^*=\max_{1\le j,k\le N}[\delta_{M-1}(j,k)]c_k(o_M) \tag{18}$$

$$q^*_M=\operatorname{argmax}_{1\le j,k\le N}[\delta_{M-1}(j,k)]c_k(o_M) \tag{19}$$

4 实验结果

针对上述基于上下文二阶隐马尔可夫模型在中文词性标注理论的分析，本文对词性标注的性能进行实验分析。实验采用的语料是从 1998 年人民日报中随机抽取的。该语料是 26 个词类组成的小标注集，分别对 20 万、25 万、30 万词的语料进行训练。语料涉及政治、经济、文艺、体育、报告文学等题材。从训练集中随机抽取 5 万语料作为封闭测试集，

(下转第 235 页)