基于二阶隐马尔可夫模型的文本信息抽取

周顺先[1,2],林亚平[1],王耀南[2],易叶青[1]
(1.湖南大学计算机与通信学院,湖南长沙 410082;2.湖南大学电气与信息工程学院,湖南长沙 410082)

摘 要: 隐马尔可夫模型是文本信息抽取的重要方法之一.在一阶隐马尔可夫模型中,假设状态转移概率和观察值输出概率仅依赖于模型当前的状态,一定程度降低了信息抽取的精确度.而二阶隐马尔可夫模型合理地考虑了概率和模型历史状态的关联性,对错误信息有更强的识别能力.提出了基于二阶隐马尔可夫模型的文本信息抽取算法;分析了二阶隐马尔可夫模型在文本信息抽取中的有效性;仿真实验表明,新的算法比基于一阶隐马尔可夫模型的算法具有更高的抽取精确度.

关键词: 文本信息抽取;一阶隐马尔可夫模型;二阶隐马尔可夫模型;精确度
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2007) 11-2226-06

Text Information Extraction Based on the Second-Order Hidden Markov Model

ZHOU Shun-xian[1,2], LIN Ya-ping[1], WANG Yao-nan[2], YI Ye-qing[1]
(1. *College of Computer and Communication, Hunan University, Changsha, Hunan* 410082, *China*;
2. *College of Electrical and Information Engineering, Hunan University, Changsha, Hunan* 410082, *China*)

Abstract: Hidden Markov model is one of important approaches for text information extraction. In the first-order hidden Markov model, there is the hypothesis that the transition probability of state and the output probability of observation are only dependent on the current state of the model, which debases the precision of information extraction comparatively. The relationship between the probability and the model's historical states is considered reasonably in the second-order hidden Markov model which has stronger performance of recognition for incorrect information. An algorithm of text information extraction based on the second-order hidden Markov model is proposed. The validity of the second-order hidden Markov model in information extraction is analyzed. Simulation Experiments show that the new algorithm has higher precision than the algorithm based on the first-order hidden Markov model.

Key words: text information extraction; the first-order hidden Markov model; the second-order hidden Markov model; precision

1 引言

从 20 世纪 60 年代以来,文本信息抽取理论的研究得到了不断发展,成为自然语言处理领域的一个重要研究分支.隐马尔可夫模型(Hidden Markov Model,HMM)是文本信息抽取的一种重要方法,得到了广泛关注和研究,抽取性能从不同的方面得到了改善.文[1]应用 HMM 抽取计算机科研论文的头部信息;文[2]使用"收缩(shrinkage)"技术改进 HMM 概率的估计;文[3]使用随机优化技术动态选择最适合的 HMM 结构;文[4]结合文本分块的方法改善 HMM;文[5]利用主动学习技术减少训练 HMM 时所需的标记文本.但是,简单的一阶 HMM 没有考虑文本上下文特征等信息对抽取性能的作用以及状态转移概率和观察值输出概率与模型历史状态的关联性;最大熵模型可以利用文本上下文特征等信息改善 HMM,提高抽取性能[6];而在最大熵模型中存在的标记偏置(label bias)问题则可以通过条件随机场(Conditional Random Fields,CRFs)模型来解决.CRFs 模型是一种概率图模型,具有表达长距离依赖性和交叠性特征的能力,能利用文本上下文特征信息、领域知识、或邻域系统提高英文命名实体识别、中文词性标注、中文机构名识别及文本信息抽取的性能,也能解决标记偏置问题[7~10].而二阶 HMM 则从时间上考虑模型某一时刻的状态转移概率和观察值输出概率与模型历史状态的关联性,而这种关联性能提高文本信息抽取的正确性.文[11]研究和推导了二阶 HMM 的主要学习算法.文[12]将二阶 HMM 应用到语音识别中,发现比一阶 HMM 具有更好的识别性能.因此,本文提出将二阶 HMM 应用

收稿日期:2007-02-05;修回日期:2007-06-05
基金项目:国家 863 高技术研究发展计划(No.2006AA01Z227);湖南省重点自然科学基金(No.06JJ20049)

到论文头部信息抽取中,并分析了这种方法的有效性,仿真实验结果也表明,二阶 HMM 比一阶 HMM 具有更高的抽取精确度.

2　二阶 HMM 的定义

HMM 是指一个不可观察的马尔可夫链(称为状态过程)以及与每一个状态相关联的观察值输出的随机过程.一个 HMM 包含一个可观察层和一个隐藏层,可观察层是待识别的观察序列,用观察值输出概率描述;隐藏层是一个马尔可夫过程,用状态转移概率描述.应用 HMM 模型,主要解决评估问题、学习问题和解码问题[13].

在一阶 HMM(记为 $HMM^{(1)}$)中,计算状态转移概率时,假设状态序列中的每一个状态只与前一个状态有关;计算观察值的输出概率时,假设任意时刻观察输出概率只依赖于系统当前时刻所处的状态.因此,一个 $HMM^{(1)}$可以看成一个五元组$\{S, O, A, B, \pi\}$,其定义参见文[13].

二阶 HMM(记为 $HMM^{(2)}$)与 $HMM^{(1)}$的区别在于两个重要假设:

假设 1　在 $HMM^{(2)}$中,隐藏的状态序列是一个二阶 Markov 链,即在 $t+1$ 时刻的状态 q_{t+1}的转移概率不仅依赖于 t 时刻的状态 q_t,同时依赖于 $t-1$ 时刻的状态 q_{t-1}.

假设 2　在 t 时刻释放观察值 V_k 的输出概率,不仅依赖于系统当前所处的状态 S_j,同时依赖于系统前一时刻所处的状态 S_i.

因此,$HMM^{(2)}$定义为一个七元组$\{S, O, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{B}_1, \boldsymbol{B}_2, \pi\}$:

(1) S:$HMM^{(2)}$中,Markov 链的状态集.记为 $S=\{S_1, S_2, \cdots, S_i, S_j, \cdots, S_N\}$,$t$ 时刻,系统所处的状态为 q_t,$q_t \in \{S_1, S_2, \cdots, S_i, S_j, \cdots, S_N\}$.

(2) O:模型输出的观察集.记为 $O=\{O_1, O_2, \cdots, O_M\}$,$t$ 时刻观察到的观察值为 O_t,$O_t \in \{O_1, O_2, \cdots, O_M\}$.

(3) $\boldsymbol{A}_1, \boldsymbol{A}_2$:状态转移概率矩阵.$\boldsymbol{A}_1=(a_{ij})_{N\times N}$,$\boldsymbol{A}_2=(a_{ijk})_{N\times N\times N}$,其中:

$$a_{ij}=P(q_2=S_j \mid q_1=S_i), 1\leqslant i,j\leqslant N \tag{1}$$

$$a_{ijk}=P(q_{t+1}=S_k \mid q_t=S_j, q_{t-1}=S_i), 1\leqslant i,j,k\leqslant N \tag{2}$$

(4) $\boldsymbol{B}_1, \boldsymbol{B}_2$:观察值释放概率矩阵.$\boldsymbol{B}_1=(b_j(O_t))_{N\times M}$,$\boldsymbol{B}_2=(b_{ij}(O_t))_{N\times N\times M}$,其中:

$$b_j(O_t)=P(O_t=O_l \mid q_t=S_j), 1\leqslant j\leqslant N, 1\leqslant l\leqslant M \tag{3}$$

$$b_{ij}(O_t)=P(O_t=O_l \mid q_t=S_j, q_{t-1}=S_i), \quad 1\leqslant i,j\leqslant N, 1\leqslant l\leqslant M \tag{4}$$

(5) π:模型的初始状态概率.$\pi=(\pi_1, \pi_2, \cdots, \pi_N)$,其中:

$$\pi_i=P(q_1=S_i), 1\leqslant i\leqslant N \tag{5}$$

3　基于 $HMM^{(2)}$的论文头部信息抽取

将 $HMM^{(2)}$而应用到文本信息抽取中,需解决模型的学习问题和解码问题:首先从训练文本中学习 $HMM^{(2)}$,对于已完全标记训练文本,采用 $HMM^{(2)}$的 ML(Maximum Likelihood)算法(记为 $ML^{(2)}$);对于部分标记的训练文本,采用 $HMM^{(2)}$的 BW(Baum-Welch)算法(记为 $BW^{(2)}$);然后采用 $HMM^{(2)}$的 Viterbi 算法(记为 Viterbi$^{(2)}$)将待抽取的文本序列标记为最大概率的状态标签序列.本文中,考虑从完全标记的训练文本中学习 $HMM^{(2)}$,因此,只讨论 $ML^{(2)}$算法中模型参数的计算公式和 Viterbi$^{(2)}$算法.

3.1　$ML^{(2)}$算法中模型参数计算公式

$ML^{(2)}$算法以统计的方法从完全标记的训练文本中得出 $HMM^{(2)}$的初始状态概率、状态转移概率和观察值释放概率等模型参数.根据 $HMM^{(2)}$的定义,假设整个已标记的训练文本中,序列总个数为 N,则在 $ML^{(1)}$算法的基础上,可推导 $ML^{(2)}$算法的主要计算公式如下:

(1)初始状态概率计算公式:

$$\pi_i = Init(i) \Big/ \sum_{j=1}^{N} Init(j), \quad 1\leqslant i\leqslant N_s \tag{6}$$

其中:$Init(i)$为整个已标记的训练文本中,以 S_i 作为开始状态的序列的个数;$\sum_{j=1}^{N} Init(j)$为整个已标记的训练文本中,以所有状态作为开始状态的序列的个数之和.

(2)状态转移概率计算公式:

$$a_{ij} = C_{ij} \Big/ \sum_{k=1}^{N} C_{ik}, \quad 1\leqslant i,j\leqslant N \tag{7}$$

$$a_{ijk} = C_{ijk} \Big/ \sum_{u=1}^{N} C_{iju}, \quad 1\leqslant i,j,k\leqslant N \tag{8}$$

其中:C_{ij}表示从状态 S_i 到状态 S_j 的转换次数;$\sum_{k=1}^{N} C_{ik}$表示从状态 S_i 到所有状态的转换次数之和;C_{ijk}表示 $t-1$ 时刻状态为 S_i,t 时刻状态为 S_j,转换到 $t+1$ 时刻 S_k 的次数;$\sum_{u=1}^{N} C_{iju}$表示 $t-1$ 时刻状态为 S_i,t 时刻状态为 S_j,转换到所有状态的次数之和.

(3)观察值释放概率计算公式:

$$b_j(O_k) = E_j(O_k) \Big/ \sum_{i=1}^{M} E_j(O_i), \quad 1\leqslant j\leqslant N \tag{9}$$

$$b_{ij}(O_k) = E_{ij}(O_k) \Big/ \sum_{i=1}^{M} E_{ij}(O_u), \quad 1\leqslant i,j,k\leqslant N \tag{10}$$

其中：$E_j(O_K)$表示在状态 S_j 时，释放观察值 O_k 的次数；$\sum_{i=1}^{M} E_j(O_i)$表示在状态 S_j 时，释放所有观察值的次数之和；$E_{ij}(O_k)$表示 $t-1$ 时刻状态为 S_i，t 时刻状态为 S_j，释放观察值 O_k 的次数；$\sum_{u=1}^{M} E_{ij}(O_u)$表示 $t-1$ 时刻状态为 S_i，t 时刻状态为 S_j，释放所有观察值的次数之和.

3.2 Viterbi$^{(2)}$算法

Viterbi$^{(2)}$算法[11]解决在给定条件下求最佳状态序列的解码问题，即给定一个观察值序列 $O=(O_1,O_2,\cdots,O_T)$和一个 HMM$^{(2)}$ $\lambda=(\pi,\boldsymbol{A}_1,\boldsymbol{A}_2,\boldsymbol{B}_1,\boldsymbol{B}_2)$，求使得 $P(Q|O,\lambda)$最大的状态序列 Q，最佳状态序列记为 $Q^*=(q_1^*,q_2^*,\cdots,q_T^*)$.

定义 $\delta_t(i,j)$为 t 时刻沿路径 $q_1,q_2,\cdots,q_t(q_{t-1}=S_i,q_t=S_j)$，释放观察值序列 $O_1,O_2,\cdots,O_t$ 的最大概率，则：

$$\delta_t(i,j)=\max_{q_1,\cdots,q_{t-2}} P(q_1,\cdots,q_{t-1},q_t=S_j,O_1O_2\cdots O_t|\lambda),\quad 1\leqslant i,j\leqslant N,2\leqslant t\leqslant T \tag{11}$$

通过推导(略)，可得 $t+1$ 时刻：

$$\delta_{t+1}(j,k)=\max_{1\leqslant i\leqslant N}[\delta_t(i,j)a_{ijk}]b_{ij}(O_{t+1}),\quad 1\leqslant j,k\leqslant N,2\leqslant t\leqslant T-1 \tag{12}$$

寻找最佳状态序列的 Viterbi$^{(2)}$算法过程为：

(1)初始化

$$\delta_2(i,j)=\pi_i a_{ij}b_i(O_i)b_{ij}(O_2),1\leqslant i,j\leqslant N \tag{13}$$

$$\Psi_2(i,j)=0,\quad 1\leqslant i,j\leqslant N \tag{14}$$

(2)递归

Where $(1\leqslant i,j\leqslant N,2\leqslant t\leqslant T-1)$

$$\delta_{t+1}(i,j)=\max_{1\leqslant i\leqslant N}[\delta_t(i,j)a_{ijk}]b_{ij}(O_{t+1}) \tag{15}$$

$$\Psi_{t+1}(j,k)=\arg\max_{1\leqslant i\leqslant N}[\delta_t(i,j)a_{ijk}]$$

(3)终结

$$P^*=\max_{1\leqslant i,j\leqslant N}[\delta_T(i,j)] \tag{16}$$

$$q_{T-1}^*,q_T^*=\arg\max_{1\leqslant i,j\leqslant N}[\delta_T(i,j)] \tag{17}$$

(4)求取最佳状态序列

$$q_{t-1}^*=\Psi_{t+1}(q_t^*,q_{t+1}^*),t=T-1,T-2,\cdots,2 \tag{18}$$

算法中，$\Psi_{t+1}(j,k)$为记录节点的数组.

3.3 基于 HMM$^{(2)}$的论文头部信息抽取算法

应用 HMM$^{(2)}$进行文本信息抽取时，考虑论文头部信息中某类信息(比如文章标题、作者等)均包含多个单词，如果对同类信息进行分组，并以分组为单位进行抽取，显然比以单个单词为单位进行抽取的效率要高.因此，本文将文本分组的思想[4]应用到信息抽取中，首先对训练文本和待抽取文本进行分组预处理.基于 HMM$^{(2)}$的文本信息抽取算法的基本步骤可以归纳如下：

第一步：HMM$^{(2)}$的训练

(1)初始化 HMM$^{(2)}$：确定模型的状态数，初始化模型的参数.

(2)训练文本分组预处理：扫描训练文本，依据排版格式，分隔符等信息将标记好的论文头部文本序列转换为由分组构成的序列，每一分组都用 HTML 语言进行状态标记.

在分组预处理时，要保证分组足够小，使得每一个分组内所有单词只属于一个状态，但连续的几个分组内的单词可以属于同一个状态.

Web Mining: Information and Pattern Discovery
on the World Wide Web
R. Cooley, B. Mobasher, J. Srivastva
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455, USA

例如，对上面的论文头部信息，以换行和标点符号为分组依据，可满足分组的要求，这样，可以分为 10 个分组，以每个分组为一个观察值，再以分组为单位进行抽取.

(3)训练模型：以分组为单位，应用 ML$^{(2)}$算法计算 HMM$^{(2)}$参数.

(a)用式(6)计算初始状态概率；

(b)用式(7)、式(8)计算状态转移概率；

(c)计算观察值释放概率：以分组为单位的观察值释放概率为分组内各单词的释放概率之和.假设以分组为单位的观察值序列为 $O=\{O_1,O_2,\cdots,O_M\}$，若第 k 组(即第 k 个观察值)长度为 L(包含 L 个单词)记为：$O_k=O_{k1}O_{k2}\cdots O_{kL}$，则 O_k 的释放概率为：

$$b_j(O_k)=\sum_{t=1}^{L} b_j(O_{kt}),1\leqslant t\leqslant L \tag{19}$$

$$b_{ij}(O_k)=\sum_{t=1}^{L} b_{ij}(O_{kt}),1\leqslant t\leqslant L \tag{20}$$

其中 $b_j(O_{kt})$，$b_{ij}(O_{kt})$分别用式(9)和式(10)计算.

(4)输出 HMM$^{(2)}$模型.

第二步：应用训练好的 HMM$^{(2)}$进行信息抽取

(5)对输入的论文头部文本进行分组预处理.

(6)结合训练部分输出的 HMM$^{(2)}$，利用 Viterbi$^{(2)}$算法求最佳状态序列.

(7)输出最佳状态标记序列.

综上所述，得到基于 HMM$^{(2)}$的文本信息抽取算法框架如图 1 所示.

4 HMM$^{(2)}$在信息抽取中的有效性分析

对于 3.3 小节中的论文头部信息，包含文章标题

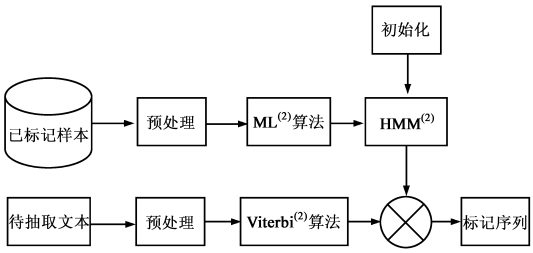


图 1　基于 HMM$^{(2)}$的文本信息抽取算法框图

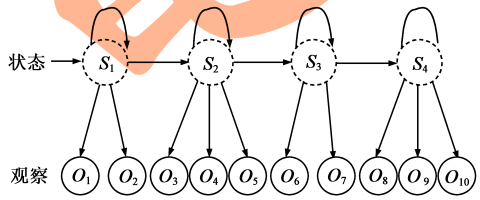
(title)、作者(author)、组织机构(organization)、地址(address)等 4 个状态,状态序列记为 $S=\{S_1,S_2,S_3,S_4\}$;通过分组预处理后,得到 10 个以分组为单位的观察值,观察值序列记为 $O=\{O_1,O_2,\cdots,O_{10}\}$.其对应 HMM 的状态序列、观察值序列在时间序列上的关系如表 1 所示.表中的列表示在某一时刻,模型转移的状态及在该状态下输出的观察值.

进一步可以得到针对上述论文头部信息构建的 HMM$^{(1)}$和 HMM$^{(2)}$示意图分别如图 2(a)和图 2(b)所示.

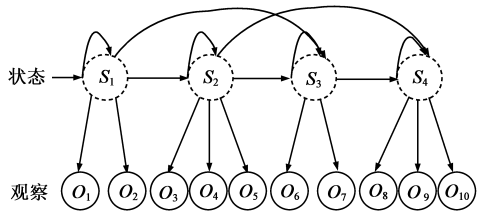
在图 2(a)所示的 HMM$^{(1)}$中,状态转移概率和观察值输出概率仅依赖于模型当前的状态,例如 $t+1$ 时刻状态 S_3 只由 t 时刻的状态 S_2 转移,观察值 O_6 的输出概率仅依赖于状态 S_3;在图 2(b)所示的 HMM$^{(2)}$中,$t+1$ 时刻状态 S_3 由 t 时刻的状态 S_2 和 $t-1$ 时刻的状态 S_1 混合转移,观察值 O_6 的输出概率依赖于状态 S_3 和前一时刻的状态 S_2.现以计算机科研论文头部信息抽取为背景,以观察值 O_6 为例,分析比较 HMM$^{(2)}$和 HMM$^{(1)}$抽取信息的正确性.

表 1　论文头部信息对应的模型某时刻状态转移及观察值输出关系表

时刻	1	2	3	4	5	6	7	8	9	10
状态	S_1	S_1	S_2	S_2	S_2	S_3	S_3	S_4	S_4	S_4
观察	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}



(a) HMM$^{(1)}$意图



(b) HMM$^{(2)}$示意图

图 2

假设待抽取的计算机科研论文头部文本集中,观察值 O_6 可能存在以下两种单词组合(分别记为 O_6^1 和 O_6^2):O_6^1 = "Department of Computer Science and Engineering";O_6^2 = "Department of Chemistry and Chemical Engineering",其输出概率分别记为 P_6^1 和 P_6^2.对于科研论文,O_6^1 和 O_6^2 都可作为组织机构状态(S_3)的正确输出,但是,对于计算机科研论文,显然,O_6^2 不能作为组织机构状态的一个正确输出.下面分析比较应用 HMM$^{(1)}$和 HMM$^{(2)}$时,对这种可能存在的错误信息的识别能力.

应用 HMM$^{(1)}$时,由于只考虑了观察值输出对模型当前状态的依赖,所以,模型在 t 时刻,状态为 S_3 时,观察值 O_6^1 和 O_6^2 的输出概率分别为:$P_6^1=b_3(O_6^1)=P(O_t=O_6^1\mid q_t=S_3)$,$P_6^2=b_3(O_6^2)=P(O_t=O_6^2\mid q_t=S_3)$;此时,$O_6^1$ 和 O_6^2 作为状态 S_3 的输出概率几乎是相等的,即 $P_6^1=P_6^2$,所以,观察值序列"…Department of Chemistry and Chemical Engineering…"将作为正确的信息被抽取出来,这显然是一个错误的抽取结果.

应用 HMM$^{(2)}$时,由于考虑了观察值输出对模型当前状态和历史状态的依赖,所以,模型在 t 时刻,状态为 S_3,$t-1$ 时刻,状态为 S_2 时,观察值 O_6^1 和 O_6^2 的输出概率分别为:$P_6^1=b_{23}(O_6^1)=P(O_t=O_6^1\mid q_t=S_3,q_{t-1}=S_2)$,$P_6^2=b_{23}(O_6^2)=P(O_t=O_6^2\mid q_t=S_3,q_{t-1}=S_2)$;显然,由于 O_6^2 作为 S_2 的输出的可能性很小,$P_6^1>P_6^2$.所以,观察值序列"…Department of Chemistry and Chemical Engineering…"作为正确的信息被抽取出来的可能性很小,因而使得抽取结果中错误信息更少.

根据消息理解会议(MUC)对信息抽取精确度(P)的定义[14],假设用 ce 和 fe 分别表示对某个状态抽取出的正确信息个数和抽取出的错误信息个数,则精确度:

$$P=ce/(ce+fe) \tag{21}$$

根据上述分析可知,应用 HMM$^{(2)}$时,抽取结果中错误信息(fe)比应用 HMM$^{(1)}$时更少,因而抽取精确度更高.下面通过仿真实验进一步予以验证.

5　仿真实验与分析

仿真实验采用美国 CMU 大学 CORA 搜索引擎研制组提供的用 HTML 语言标记好的计算机科研论文头部数据集*.随机选用其中 100 篇作为待抽取数据,其他作为训练数据.用抽取精确度和总精确度作为抽取性能评价指标.用式(21)计算各状态抽取精确度,总精确度的计算定义为:

$$GP=\sum_{i=1}^{N} ce\Big/\sum_{i=1}^{N}(ce+fe) \tag{22}$$

* Data set for information extraction, http://www-2.cs.cmu.edu/~kseymore/ie.html

式中，N 为模型的状态数. 训练集从 100 篇开始不断增加到 800 篇，两种模型的抽取总精确度如图 3 所示；其中，以 500 篇作为训练集时各状态抽取的精确度比较如表 2 所示.

从图 3 可以看出，选取不同数量的已标记训练数据集时，基于 HMM$^{(2)}$ 的抽取总精确度均高于 HMM$^{(1)}$，随着训练数据集的增加，前者的抽取总精确度比后者更高，因为随着训练数据集的增加，HMM$^{(2)}$ 被训练得更加优化，其错误识别能力更强. 图中显示，当训练集从 500 篇增加到 600 篇时，两种模型的抽取总精确度分别有不同程度的下降，其原因可能是此时增加的 100 篇训练数据和测试数据集的匹配程度较差，影响了总的抽取精确度.

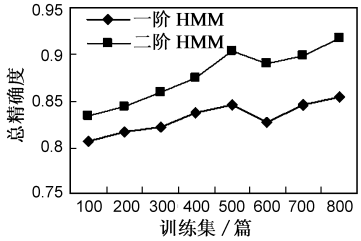


图 3 HMM$^{(1)}$和 HMM$^{(2)}$抽取总精确度比较

表 2 各状态抽取精确度比较

	HMM$^{(1)}$	HMM$^{(2)}$
States	Precision	Precision
Title	0.823237	0.836165
Author	0.853962	0.858543
Organization	0.835981	0.898625
Address	0.864763	0.953790
Email	0.888298	0.989547
Note	0.819986	0.906543
Web	0.902169	0.924587
Phone	0.916759	0.933909
Date	0.708491	0.755282
Abstract	0.905127	0.936794
Intro	0.856725	0.887543
Keyword	0.838184	0.867829
Degree	0.604369	0.725471
Pubnum	0.847641	0.934936
Page	0.988621	0.990217

从表 2 也可以看出，当训练集为 500 篇时，对各状态的抽取精确度，HMM$^{(2)}$ 高于 HMM$^{(1)}$. 因此，进一步验证了基于 HMM$^{(2)}$ 的抽取正确率优于 HMM$^{(1)}$. 同时，我们在实验数据中也发现，HMM$^{(2)}$ 对信息抽取的召回率没有提高(因此，有关召回率的数据在本文没有列出)，因为 HMM$^{(2)}$ 主要提高对错误信息的识别能力，这种排除错误信息的能力对信息抽取是非常重要的.

6 结论

HMM$^{(2)}$ 合理地考虑了概率和模型历史状态的关联，在自然语言处理上比 HMM$^{(1)}$ 具有更大的优势；应用到文本信息抽取中，其抽取结果更加符合客观事实，正确率更高. 本文从理论上分析了这种可能性，在仿真实验结果中也得到了进一步验证. 但是，当待抽取的文本更加复杂时，抽取模型也会更复杂，此时，模型参数估计值的准确性可能会下降，从而可能导致信息抽取精确度的下降. 因此，在进一步的研究中，我们将考虑在更加复杂的情况下，结合其他方法来改进 HMM$^{(2)}$，以确保抽取精确度不会下降，甚至得到进一步提高. 例如，当只能得到部分标记的训练文本时，将结合主动学习的方法训练 HMM$^{(2)}$；对于网上不同来源的、格式不尽相同的文本，将结合文本聚簇的方法改进 HMM$^{(2)}$；也将考虑文本上下文特征信息、领域知识等信息在信息抽取中的作用，结合最大熵模型或 CRFs 对 HMM$^{(2)}$ 进行建模，以进一步提高信息抽取精确度.

参考文献：

[1] Kristie Seymore, Andrew McCallum, Ronal Rosenfel. Learning hidden markov model structure for information extraction[A]. Procedings of the AAAI'99 Workshop on Machine Learning for Information Extraction[C]. Orlando, Florida: AAAI Press, 1999, 37 – 42.

[2] Dayne Frietag, Andrew McCallum. Information extraction with HMMs and shrinkage[A]. Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction[C]. Orlando: AAAI Press, 1999. 31 – 36.

[3] Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization[A]. Proceedings of the Eighteenth Conference on Artificial Intelligence[C]. Edmonton: AAAI Press, 2002. 584 – 589.

[4] 刘云中，林亚平，陈治平，等. 基于隐马尔可夫模型的文本信息抽取[J]. 系统仿真学报，2004，16(3)：507 – 510.
Liu YZ, Lin YP, Chen ZP. Text information extraction based on hidden Markov model[J]. Journal of System Simulation, 2004, 16(3): 507 – 510. (in Chinese)

[5] Scheffer T, Decomain C, Wrobel S. Active hidden Markov models for information extraction[A]. Proceedings of the Fourth International Symposium on Intelligent Data Analysis[C]. Berlin: Springer, 2001. 309 – 318.

[6] 林亚平，刘云中，周顺先，等. 基于最大熵的隐马尔可夫模型文本信息抽取[J]. 电子学报，2005，33(2)：236 – 240.
Lin YP, Liu YZ, Zhou SX, et al. Using hidden Markov model for text information extraction based on maximum entropy[J]. Acta Electronica Sinica, 2005, 33(2): 236 – 240. (in Chinese)

[7] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[A]. Proceedings of the 18th ICML[C]. San Francisco: Morgan Kaufmann, 2001, 282 – 289.

[8] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and Web-en-

hanced lexicons[A]. Proceedings of the 7th CoNLL[C]. Edmonton, Canada: Morgan Kaufmann, 2003. 188 – 191.

[9] 洪铭材,张阔,唐杰,李涓子.基于条件随机场(CRFs)的中文词性标注方法[J].计算机科学,2006,33(10):148 – 151.
Hong MC, Zhang K, Tan J, Li JZ. A Chinese part-of-speech tagging approach using conditional random fields[J]. Journal of Computer Science, 2006, 33(10): 148 – 151. (in Chinese)

[10] 周俊生,戴新宇,尹存燕,陈家骏.基于层叠条件随机场模型的中文机构名自动识别[J].电子学报,2006,34(5):804 – 809.
Zhou JS, Dai XY, Yin CY, Chen JJ. Automatic recognition of Chinese organization name based on cascaded conditional random fields[J]. Acta Electronica Sinica, 2006, 34(5): 804 – 809. (in Chinese)

[11] 史笑兴,王太君,何振亚.二阶隐马尔可夫模型的学习算法及其与一阶隐马尔可夫模型的关系[J].应用科学学报,2001,19(1):29 – 32.
Shi XX, Wang TJ, He ZY. The learning algorithm of the second order HMM and its relationship with the first order HMM [J]. Journal of Applied Sciences, 2001, 19(1): 29 – 32. (in Chinese)

[12] Shahin I. Using second-order hidden Markov model to improve speaker identification recognition performance under neutral condition[A]. Proceedings of the 10th IEEE ICECS [C]. United Arab Emirates: Sharjah, 2003, 124 – 127.

[13] Lawrence E. Rabiner. A tutorial on hidden markov models and selected application in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257 – 286.

[14] Douthat A. The Message Understanding Conference Scoring Software User's Manual[EB/OL]. http://www-nlpir.nist.gov/related_projects/muc/muc_sw/muc_sw_manual.html, 1998-04-28.

作者简介:

周顺先　男,1968 年生于湖南新邵.湖南大学计算机与通信学院博士研究生.研究方向为机器学习.E-mail:13007312816@hn165.com

林亚平　男,1955 年生于湖南邵阳,湖南大学计算机与通信学院教授,博士生导师.主要研究方向为计算机通信网络、机器学习.

王耀南　男,1957 年生于湖南长沙,湖南大学电气与信息工程学院教授,博士生导师.主要研究方向为智能信息处理.

易叶青　男,1976 年生于湖南邵阳,湖南大学计算机与通信学院博士研究生.主要研究方向为机器学习、计算机通信网络.