·人工智能及识别技术· 文章编号:1000-3428(2017)07-0211-06 文献标志码:A 中图分类号:TP18

基于二阶HMM的中医诊断古文词性标注

刘 博,杜建强,聂 斌,刘 蕾,张 鑫,郝竹林

(江西中医药大学 计算机学院,南昌 330004)

摘 要: 针对传统隐马尔可夫模型(HMM)的词性标注存在捕获上下文信息有限的问题,提出一种改进的二阶隐马尔可夫模型。该模型考虑上下文联系,精确标注中医诊断文本。对训练过程中出现数组下溢的问题,采用生词处理及增加比例因子的方法对其加以修正。实验结果表明,改进后的二阶HMM比传统HMM模型具有更高的词性标注正确率。

关键词: 中医诊断古文;词性标注;上下文联系;比例因子;二阶隐马尔可夫模型;生词处理

中文引用格式: 刘 博,杜建强,聂 斌,等. 基于二阶HMM的中医诊断古文词性标注[J]. 计算机工程,2017,43(7):211-216.

英文引用格式: Liu Bo,Du Jianqiang,Nie Bin,et al. Part-of-speech Tagging of Traditional Chinese Medicine Diagnosis Ancient Prose Based on Second-order HMM[J]. Computer Engineering,2017,43(7):211-216.

Part-of-speech Tagging of Traditional Chinese Medicine Diagnosis Ancient Prose Based on Second-order HMM

LIU Bo,DU Jianqiang,NIE Bin,LIU Lei,ZHANG Xin,HAO Zhulin

(School of Computer Science,Jiangxi University of Traditional Chinese Medicine,Nanchang 330004,China)

【Abstract】 Aiming at the deficiency of traditional Hidden Markov Model(HMM) in solving part-of-speech tagging,this paper proposes an improved second-order HMM. This model can better connect with contextual information,making part-of-speech tagging of the diagnosis of Traditional Chinese Medicine(TCM) accurate. Method of taking scale factor is proposed to solve the array of underflow and the new word processing method is given in the course of training. Experimental results show that,compared with traditional model,the improved second-order HMM has higher accuracy in part-of-speech tagging.

【Key words】 Traditional Chinese Medicine(TCM) diagnosis ancient prose; part-of-speech tagging; context relations; scale factor; second-order Hidden Markov Model(HMM); new word processing

DOI: 10.3969/j.issn.1000-3428.2017.07.035

0 概述

中医诊断古文是中华民族几千年来防病治病宝贵经验的结晶,不仅传承了中医药学理论知识,而且大量的经验方直接指导临床[1]。将这些传统资源转变为可用的语料库进行多层次的分词和检索,对中医诊断古文进行标注是一个极其重要的步骤。

目前采用的词性标注方法主要有基于隐马尔可夫模型(Hidden Markov Model,HMM)的标注方法[2-3]、基于规则的标注方法[4]、基于最大熵模型的标注方法[5]、基于条件随机场的标注方法[6-8]等。而基于HMM是基于统计的方法,算法健壮性强,是目前应用最广泛的词性标注方法。传统的HMM的标注方法是基于以下假设[9]:1)状态转移的假设。时刻$t+1$的状态转移只取决于时刻t的状态,而与t时刻以前的状态无关。2)输出值的假设。在时刻t输出观测值的概率,只取决于时刻t的状态,而与t之前的状态无关。这样假设所导致的结果是仅仅只考虑了上文对当前词的影响,而忽略了下文对当前词的影响。中医古文里面词汇上下文依赖关系决定了

基金项目:国家自然科学基金(61363042,61562045);江西省高校科技落地计划项目(LD12038);江西省研究生创新基金(YC2015-S350)。

作者简介:刘 博(1990—),男,硕士研究生,主研方向为文本挖掘、医药数据挖掘;杜建强(通信作者),教授、博士;聂 斌,讲师、硕士;刘 蕾、张 鑫,硕士研究生;郝竹林,助教。

收稿日期:2016-05-24 修回日期:2016-08-03 E-mail:252142817@qq.com

对于该词汇的理解和运用。文献[10]应用一阶HMM于中医诊断古文;一阶HMM虽然在一定程度上能标注中医诊断词汇,但捕获上下文的信息有限。文献[11]研究和推导了二阶HMM的主要学习方法。本文提出一种改进的二阶隐马尔可夫模型,并将改进后的模型应用于中医古文的词性标注。

1 中医诊断词性标记集

1.1 中医诊断古文特点

中医诊断文本中“证”是对病变中机体整体反应状态的阶段性病理本质概括。“辨证”是根据中医理论,对证候进行分析,认识其本质——证素,并作出证名诊断的思维过程。证素,即辨证的要素,是辨证的核心与关键。而证素[12]又可分为病位要素和病性要素。病位要素包括心、肺、脾等;病性要素包括风、寒、暑等。通过分析大量中医文本会发现其是由病位和病性相互组合形成的短文本,其文本特征有如下特点:

1)每个症状描述句子较短,大多情况可分为一个症状部位对应一个症状表现,如“口干”的症状部位为“口”,症状表现为“干”。同时还存在一个症状部位对应多个症状表现:如“小便短黄”和多个症状部位对应一个症状表现,如“脘腹坠胀”。

2)部分症状不存在症状部位,只有症状表现存在。如“盗汗”“自汗”。

3)部分症状的症状部位对应的症状表现是在相应的时机才发生,如“午后颧红”,“颧红”发生的时机为“午后”,部分症状不存在症状部位,只有症状表现和症状时机存在,如“遇寒则痛”,症状表现为“痛”,症状时机为“遇寒”,“则”是连词。

4)部分症状的症状部位对应的症状表现还对应着相应的症状表现描述,如“颧红升火”,症状部位为“颧”,症状表现为“红”,“升火”是对症状表现“红”的描述,部分症状不存在症状部位,只有症状表现和相应症状表现描述存在,如“痛如针刺”,“针刺”是对症状表现“痛”的描述。

5)部分症状存在症状部位,症状表现和相应症状表现值,如“口吐涎沫”,症状部位为“口”,症状表现为“吐”,症状表现值为“涎沫”。

1.2 词性标记集

在对中医诊断古文进行词性标注时,需要选取相应的词性标注集。传统的词性标注集包含名词、动词、方位词、数词、形容词、代词等。但如果对中医诊断文本仍采用传统的词性标注集,会完全影响到病性和病位的分布,并且传统的大量词性标注在中医诊断文本中并不能发挥其原有的作用。所以,本文采用文献[13]提出的基于键值对模型的中医诊断词汇标注集,把病位标注为键,把病性标注为值。在对词汇标注集进行存储时也会简便。基于中医诊断古文文本特征的键值对模型的标记集如表1所示。

表1 基于键值对模型的中医诊断词性标记集

标记符号	标记名称	标记解释
K	键	症状部位
V	值	症状表现
ZN	中医名词	中医专业诊断名词
T	时机	病发时机
A	属性	症状部位属性
E	附加描述	症状修饰
W	标点符号	标点符号
UL	无用信息	对诊断无用信息

2 二阶隐马尔可夫模型

词性标注问题可以描述为在给定词汇序列 $O=\{O_1,O_2,\cdots,O_m\}$ 的条件下,找到词性序列 $S=\{S_1,S_2,\cdots,S_n\}$,使 $P(S|O)$ 最大:

$$P(S|O)=\frac{P(S)\,P(O|S)}{P(O)} \tag{1}$$

其中,由于分母 $P(O)$ 是个归一化常数,可以忽略不计。

基于HMM前面提到的转移假设和输出值发射假设,得到的实际公式为:

$$P(S|O)=P(s_1)\,P(o_1|s_1)\times\prod_{i=2}^{n}P(s_i|s_{i-1},s_{i+1})\,P(o_i|s_i,s_{i+1},o_{i+1}) \tag{2}$$

其中,$P(s_i|s_{i-1})$ 称为词性转移参数;$P(o_i|s_i)$ 称为词汇发射参数。

式(2)中的词性转移参数 $P(s_i|s_{i-1})$ 只是考虑了前一状态对当前状态的影响,忽略了下文对当前状态的影响。如:短文本“脉虚数”中“虚”是连接上文“脉”和下文“数”的关系,而一阶HMM只能表示“脉”“虚”的联系。“口吐涎沫”中一阶HMM只能标注“口”到“吐”之间的联系,而“涎沫”则会被单独标注出去。词汇发射参数 $P(o_i|s_i)$ 规定了当前词汇只依赖于当前的词性,当词汇具有多个词性时,就不能通过下文来判断其拥有哪个词性。如“午后颧红”和“里急后重”中“后”在前者中词性为时机,而在后者中词性为症状部位。

这些问题都会直接影响标注的准确率。为了解决这些问题,本文提出一种改进的基于上下文的二阶隐马尔可夫模型。

2.1 二阶HMM模型假设

考虑到中医诊断领域中症状部位与症状表现、症状时机与症状属性描述之间上下文联系,本文对

HMM 假设限制放宽为:

1) 状态转移的假设: 时刻 t 的状态转移不仅与时刻 $t-1$ 有关,还与时刻 $t+1$ 有关。

2) 输出值的假设: 在时刻 t 输出观测值的概率,不仅取决于时刻 t 的状态,还与时刻 $t+1$ 所处的状态和观测值有关。

用公式可描述为:

$$P(S|O)=P(s_1)P(o_1|s_1)\times\prod_{i=2}^{n}P(s_i|s_{i-1},s_{i+1})P(o_i|s_i,s_{i+1},o_{i+1}) \tag{3}$$

2.2　二阶 HMM 模型参数

二阶 HMM 所对应的参数可在文献[14]的基础上,根据本文提出的输出值的假设对观测概率矩阵和句末词汇概率做出相应的修改,参数描述如下:

1) π 为初始状态概率,其中 π_i 为词性 s_i 作为句首出现的概率。

2) N 为标记集中词性的个数。

3) M 为词汇集中词汇的个数。

4) 词性状态转移概率矩阵:

$$\boldsymbol{A}=(a_{ijk})_{N\times N\times N}$$

其中:

$$a_{ijk}=P(q_t=S_j|q_{t-1}=S_i,q_{t+1}=S_k)\quad 1\leqslant i,j,k\leqslant N \tag{4}$$

5) 观测值概率矩阵:

$$\boldsymbol{B}=(b_{ij}(O_t))_{N\times N\times M}$$

其中:

$$b_{ioj}(O_t)=P(o_t=v_k|q_t=S_i,q_{t+1}=S_j,o_{t+1}=v_{k+1})\ 1\leqslant j\leqslant N,1\leqslant l\leqslant M \tag{5}$$

其中,$b_{ioj}(O_t)$ 表示在 q_t 的状态 S_i,且在 q_{t+1} 在状态 S_j 和 o_{t+1} 输出值为 v_{k+1} 的条件下,o_t 输出为 v_k 的概率。

6) 句末词汇概率:

$$C=\{c_i(w_k)\}$$

其中:

$$c_i(w_k)=P(o_T=v_k|q_t=S_i,q_{t-1}=S_j,o_{t-1}=v_{k-1}) \tag{6}$$

其中,$c_i(w_k)$ 表示当句子末尾词的词性为 S_i; 末尾词前一个词词性为 S_j; o_{t-1} 输出值为 v_{k-1} 的条件下; o_T 输出为 v_k 的概率。

2.3　二阶 HMM 模型参数估计

由于本文所采用的中医诊断文本只有部分经过专家人工标注,因此采用 Baum-Welch 算法训练模型得到参数。在用 Baum-Welch 算法对训练样本进行训练的过程中发现,经过几次迭代后算法中的参数前向变量和后向变量都迅速趋近于 0,出现数组下溢导致程序中断,分析其原因是计算过程中所有的计算参数都是小于 1,其中必定会有 $<<1$ 的量,经过几次迭代后就会出现数组下溢,为了解决前向变量和后向变量的下溢,本文采用增加与时间 t 相关的比例因子 θ_t 的方法对其加以修正。

1) 计算前向变量 $a_t(i,j)$:

(1) 初始化

$$\alpha_2(i,j)=\pi_i b_i(O_1)a_{ij}b_{ij}(O_2),\theta_2=\sum_{i=1}^{N}\alpha_2(i,j),\alpha_2^*(i,j)=\frac{a_2(i,j)}{\theta_2},1\leqslant i,j\leqslant N \tag{7}$$

(2) 递归

$$a_{t+1}(i,k)=[\sum_{i=1}^{N}\alpha_t^*(i)a_{ijk}]b_{jk}(O_{t+1}),\theta_{t+1}=\sum_{i=1}^{N}a_{t+1}(j,k)$$

$$\alpha_{t+1}^*(j,k)=\frac{\alpha_{t+1}(j,k)}{\theta_{t+1}},1\leqslant i,j\leqslant N,2\leqslant t\leqslant T-1 \tag{8}$$

2) 计算后向变量 $\beta_t(i,j)$:

(1) 初始化

$$\beta_T(i,j)=1,\beta_T^*(i,j)=\frac{1}{\theta_T},1\leqslant i\leqslant N \tag{9}$$

(2) 递归

$$\beta_t(j,k)=\sum_{i=1}^{N}aijkb_{jk}(O_{t+1})\beta_{t+1}^*(i,j),\beta_t^*(j,k)=\frac{\beta_t(j,k)}{\theta_t},1\leqslant i,j\leqslant N,t\in[T-1,1] \tag{10}$$

其中,参数 π_i,a_{ijk},b_{ioj} 按照文献[15]推导的重估公式进行计算,本文给出 w 的重估公式:

$$\bar{w}=\frac{\sum_{j=1}^{N}\sum_{k=1}^{N}\xi_T(i,j,k)\cdot\delta_{o_T,v_l}}{\sum_{j=1}^{N}\sum_{k=1}^{N}\xi_T(i,j,k)},1\leqslant l\leqslant M \tag{11}$$

2.4　未登录词处理

未登录词仍然是困扰基于 HMM 统计词性标注方法的最主要问题之一。文献[16]中给出了传统 HMM 中生词处理的一种方法,但其只能运用在传统的隐马尔可夫模型上,为了使它更好地适用于改进后的二阶隐马尔可夫模型,需要在该文所提出的方法基础上把上下文信息考虑在内。假设有输入的句子 $W=o_1,o_2,\cdots,o_{j-1}x_jo_{j+1}$,其中,$W$ 表示整个句子; o_i 表示单个的词; x_j 表示生词。基于本文的假设,x_j 的词性 s_j 由前一词 o_{j-1} 的词性和后一词 o_{j+1} 的词性共同决定,如图 1 所示。

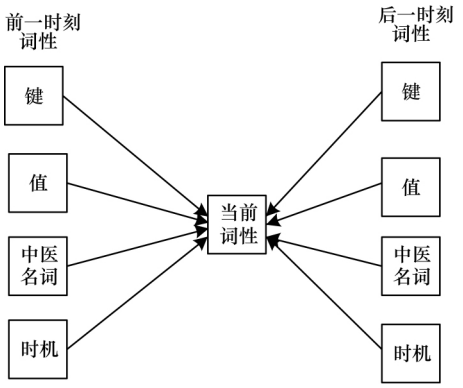

图 1 当前词性决定图

$$P(s_j|x_j)=\sum_{m=1}^{M}P(s_m,s_j,o_j|o_{j-1})\,p(s_j|s_m)\times\sum_{n=1}^{M}P(s_{j+1},s_{j+2},o_{j+2}|o_{j+1})\,p(s_j|s_n)\quad(12)$$

其中，M 表示词性种类个数；$\sum_{m=1}^{M}P(s_m,s_j,o_j|o_{j-1})\,p(s_j|s_m)$ 表示前一词性对当前词性的影响；$\sum_{n=1}^{M}P(s_{j+1},s_{j+2},o_{j+2}|o_{j+1})\,p(s_j|s_n)$ 表示后一词性对当前词性的影响。

生词的发射概率为：

$$P(x_j|s_j)=\frac{P(x_j)}{P(s_j)}\sum_{m=1}^{M}P(s_m,s_j,o_j|o_{j-1})\,p(s_j|s_m)\times\sum_{m=1}^{M}P(s_{j+1},s_{j+2},o_{j+2}|o_{j+1})\,p(s_j|s_n)\quad(13)$$

对上式中各个概率做最大似然估计，可以得到：

$$P(x_j|s_j)=\frac{C(x_j)}{C(s_j)}\sum_{m=1}^{M}P(s_m,s_j,o_j|o_{j-1})\,p(s_j|s_m)\times\sum_{m=1}^{M}P(s_{j+1},s_{j+2},o_{j+2}|o_{j+1})\,p(s_j|s_n)=\frac{1}{C(s_j)}\sum_{m=1}^{M}\left(\frac{C(s_ms_jo_jo_{j-1})}{C(o_{j-1})}\times\frac{C(s_js_m)}{C(s_m)}\right)\times\sum_{m=1}^{M}\left(\frac{C(s_{j+1}s_{j+2}o_{j+2}o_{j+1})}{C(o_{j+1})}\times\frac{C(s_js_n)}{C(s_n)}\right)\quad(14)$$

其中，$C(s_ms_jo_jo_{j-1})$ 表示当词 o_{j-1} 和词 o_j 表现为词性 s_j 和 s_m 时在训练语料中共现次数；$C(s_js_n)$ 表示词性为 s_js_n 共现的次数。

2.5 改进的 Viterbi 算法

由于 Viterbi 算法解码时，概率容易变的非常小，引入对数(lg)保证算法的稳定性。算法描述如下：

1) 初始化

$$\delta_t(i,j)=\lg\ [\pi_i b_{ioj}(O_1)],1\leqslant i,j\leqslant N$$

$$\varphi_2(i,j)=0,1\leqslant i,j\leqslant N\quad(15)$$

2) 递归

$$\delta_{t+1}(j,k)=\lg\ [\delta_t(i,j)\,a_{ijk}]+\lg\ [b_{ioj}(O_{t+1})]$$

$$2\leqslant t\leqslant T-1,1\leqslant j\leqslant N,1\leqslant k\leqslant N\quad(16)$$

$$\varphi_{t+1}(j,k)=\lg\ [\delta_t(i,j)\,a_{ijk}]\quad(17)$$

3) 终结

$$P^*=\max_{1\leqslant i\leqslant N,1\leqslant j\leqslant N}[\lg\,\delta_T(i,j)+\lg\,w(O_T)]\quad(18)$$

$$q_{T-1}^*,q_T^*=\arg\max_{1\leqslant i,j\leqslant N}[\lg\,\delta_T(i,j)+\lg\,w(O_T)]\quad(19)$$

4) 求取最佳状态序列

$$q_{t-1}^*=\varphi_{t+1}(q_t^*,q_{t+1}^*),t\in[T-1,2]\quad(20)$$

其中，$\varphi_{t+1}(j,k)$ 为记录节点的数组。

3 二阶 HMM 有效性分析

对中医诊断文本进行词性标注的基本步骤归纳如下：

步骤 1 对训练样本和测试样本进行文本分词等预处理，其中分词加载由专家自定义的中医专家定义中医诊断词汇词典。

步骤 2 训练模型，用改进的 Baum-Welch 算法计算二阶 HMM 的参数。计算出加入比例因子的前向变量和后向变量；用式(11)计算出模型的各个参数；输出训练好的模型。

步骤 3 结合训练好的二阶 HMM，加入生词处理的方法，利用改进的 Viterbi 算法求最佳状态序列。

步骤 4 输出标记好的词汇序列。

综上所述，基于二阶 HMM 的中医诊断文本词性标注流程如图 2 所示

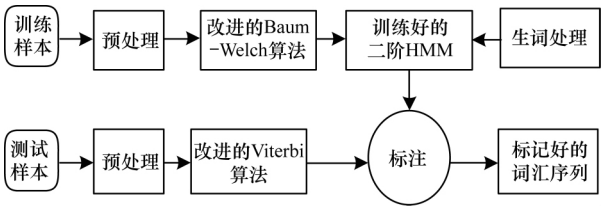

图 2 基于二阶 HMM 的中医诊断文本词性标注流程

其中，中医诊断文本词性标注在一阶 HMM 和二阶 HMM 标注时的示意图分别如图 3、图 4 所示。

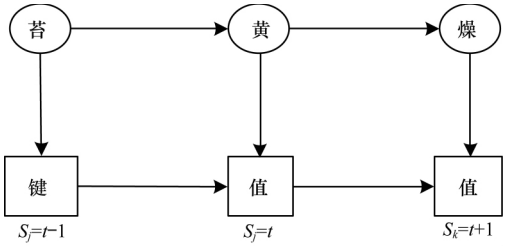

图 3 一阶 HMM 词性标注模型

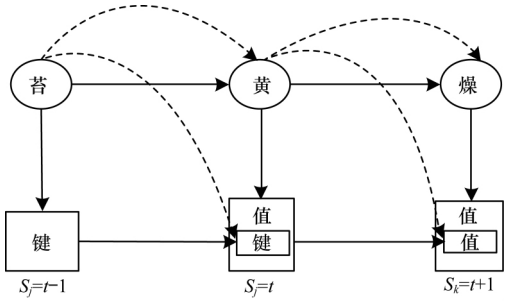

图 4　二阶 HMM 词性标注模型

当给中医诊断词汇“苔黄燥”标注时,在如图 3 所示的一阶 HMM 中,根据一阶 HMM 的假设,中医词汇“黄”的词性只是通过词汇“苔”的词性转移过来,而词汇“燥”的词性只是通过词汇“黄”的词性转移过来,这样会导致“苔”到“燥”的关系被割裂开来; 如图 4 所示的二阶 HMM,在“黄”的词性转移到“燥”的词性的过程中,“黄”本身还记载了“苔”传到“黄”时的信息(小方框内所标识的词性“键”) ,这样“黄”就在“苔”和“燥”之间起到承上启下的作用,“苔”和“燥”之间的关系就被成功地搭建起来。

词汇“午后颧红”和“里急后重”在用二阶 HMM 标注时,基于上文修改的输出值的假设词汇“后”在“午后颧红”中依赖于后一时刻的词汇“颧”以及其对应的词性(即图中所示的虚线) ,“后”会被成功地标注为词性时机的概率要远大于用一阶 HMM 标注时的概率,因为“颧“此刻的词性为键,“后”的标注增加了向后依赖的联系。同一个词汇在“午后颧红”中词性为时机,而在“里急后重”中词性为症状部位,依赖于后面词汇及其词性为其增加了限定条件,如果基于一阶 HMM 标注时词汇“后”在 2 个词汇中会被标注成一样的词性。中医古文的复杂性决定了一个词汇同时拥有多种词性的情况,用二阶 HMM 的限定条件标注时能使其被成功标注的准确率大幅增加。

4　实验结果与分析

4.1　语料库

实验所用语料库是江西中医药大学中医研究所提供的中医诊断古籍文本,共 8 756 句症状描述句子,主要来源于中医证候诊断学的资料。该语料库是经该研究所名老中医专家长期总结并且归纳临床的数据,较全面地覆盖了中医古文常见证候。

在实验中采用的标注集是基于键值对模型的中医诊断标记集,随机选取中医诊断语料库中的 5 840 句作为训练集,剩余的 2 916 句作为测试集。

本文实验还对 1998 年人民日报的语料进行标注,对应的标注集采用传统的 26 个词类的标注集。

4.2　实验结果

为了评价中医诊断古文词性标注性能,定义如下评价函数:

$$P_{\mathrm{b}}=\frac{N_{\mathrm{tr}}}{N_{\mathrm{num}}}\times 100\% \tag{21}$$

其中,P_{b} 为对中医诊断古文词性标注的评价; N_{tr} 是标注结果中被正确标注的次数; N_{num} 是语料库中单词的总数。本文所改进的二阶 HMM 模型同样适用于人民日报的语料标注,只是所采用的标注集是传统的 26 个词类的标注集。

本文采用 HMM 模型和二阶 HMM 模型对数据集进行测试,来验证二阶 HMM 模型的准确性,实验结果见表 2 和表 3。

表 2　中医语料词性标注准确率实验结果

训练集大小	HMM 词性标注准确率/%	二阶 HMM 词性标注准确率/%
1 458	69.174 066	71.182 522
2 916	71.274 840	72.000 790
5 840	72.936 700	73.259 425

表 3　人民日报词性标注准确率实验结果

训练集大小/10^4	HMM 词性标注准确率/%	二阶 HMM 词性标注准确率/%
10	88.174 066	90.182 522
20	92.274 840	93.000 790
30	93.444 670	94.892 830

从表 2 可以看出,中医语料训练集的大小影响着标注结果。训练集越大,标注效果越好。随着训练集的逐渐增大,标注的准确率增大的趋势是逐渐减少的,训练集的大小与标注的准确率提高之间的关系是非线性的。从表 3 可以看出,本文所改进的二阶隐马尔可夫模型不仅适用于中医诊断古文,在传统的人民日报语料中也能取得较好的效果。而人民日报准确率稍微高于中医语料是因为跟其训练集大小有关。实验结果证明,本文改进的基于二阶 HMM 模型的标注结果准确率要高于传统的 HMM 模型的词性标注准确率,随着获取上下文信息的增多,其标注效果越好。

在计算中医文本的相似度时,不同的中医诊断文本是通过病位因素和病性因素相互搭配组成的文本,因此可以通过计算相同语境下病位因素和病性因素的互信息量来计算中医诊断文本的相似度。而中医文本中病位因素和病性因素的正确标注直接影响最后中医诊断文本相似度计算的结果。

5 结束语

本文针对传统隐马尔可夫模型在解决词性标注问题上的不足，提出改进的二阶隐马尔可夫模型。该模型能更多地联系上下文，使得在中医诊断文本的标注更为精确。在二阶 HMM 模型参数进行训练的过程中会出现数组下溢的的问题，本文引入比例因子加以修正。同时，还给出未登录词的解决方案。在相同的训练集和测试集下，二阶 HMM 模型在中医诊断古文的词性标注准确率明显提高。由于构建的中医诊断语料库规模有限，因此词性标记集和标注模型有待进一步研究。

参考文献

[1] 王 敏. 基于改进的隐马尔科夫模型汉语词性标注[D]. 太原: 山西大学, 2007.

[2] 王凤娥, 谭红叶. 基于最大熵的句内时间关系识别[J]. 计算机工程, 2012, 38(4): 241-243.

[3] 古丽拉·阿东别克, 侯呈风, 古丽拉·阿东别克. 改进的 HMM 应用于哈萨克语词性标注[J]. 计算机工程与应用, 2010, 46(36): 147-149.

[4] 袁里驰. 基于改进的隐马尔科夫模型的词性标注方法[J]. 中南大学学报(自然科学版), 2012, 43(8): 3053-3057.

[5] 姜尚仆, 陈群秀. 基于规则和统计的日语分词和词性标注的研究[J]. 中文信息学报, 2010, 24(1): 117-122.

[6] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]//Proceddings of ICML'01. San Fransico, USA: Morgan Kaufmann, 2001: 282-289.

[7] McCallum A, Li W. EarlyResults for Named Entity Recognition with Conditional Randomfields, Feature Induction and Web-enhanced Lexicons [C]// Proceddings of CoNLL'03. Edmonton, Canada: Morgan Kaufmann, 2003: 188-191.

[8] Rabiner L E. Atutorial on Hidden Markov Models and Selected Application in Speech Recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[9] 王国龙, 杜建强, 郝竹林, 等. 中医诊断古文的词性标注与特征重组[J]. 计算机工程与设计, 2015, 36(3): 835-841.

[10] 韩 普, 姜 杰. HMM 在自然语言处理领域中的应用研究[J]. 计算机技术与发展, 2010, 20(2): 245-248.

[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.

[12] 唐亚平, 姜瑞雪, 樊欣荣. 证素及证素辨证的研究状况[J]. 时珍国医药, 2008, 19(10): 27-29.

[13] 周顺先, 林亚平, 王耀南, 等. 基于二阶隐马尔可夫模型的文本信息抽取[J]. 电子学报, 2007, 35(11): 2226-2231.

[14] 杜世平, 陈 涛. 与观测信息相关的二阶隐马尔可夫模型的参数估计[J]. 西南师范大学学报(自然科学版), 2006, 31(3): 24-27.

[15] 方 浩, 许鸿文, 蔡益宇. 一种基于语义关系改进的隐马尔可夫模型研究[J]. 通信技术, 2008, 41(5): 157-159.

[16] 张孝飞, 陈肇雄, 黄河燕. 词性标注中生词处理算法研究[J]. 中文信息学报, 2003, 17(5): 157-159.

编辑 索书志

(上接第 210 页)

[6] 王 朋, 陈树中. 基于混合模型 HMM/RBF 的数字语音识别[J]. 计算机工程, 2002, 28(12): 136-138.

[7] Li Ma, Crawford M M, Yang Xiaoquan, et al. Local-manifold-learning-based Graph Construction for Semisupervised Hyperspectral Image Classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(5): 2832-2844.

[8] Alain G, Bengio Y. What Regularized Auto-encoders Learn from the Datagenerating Distribution[J]. The Journal of Machine Learning Research, 2014, 15(1): 3563-3593.

[9] Gisbrecht A, Hammer B. Data Visualization by Nonlinear Dimensionality Reduction [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2015, 5(2): 51-73.

[10] 何 力, 张军平, 周志华. 基于放大因子和主延伸方向研究流形学习算法[J]. 计算机学报, 2005, 28(12): 2000-2009.

[11] Zhang Junping, Wang Qi, Zhou Zhihua. Quantitative Analysis of Nonlinear Embedding [J]. IEEE Transactions on Neural Networks, 2011, 22(12): 1987-1998.

[12] Jain A, Murty M, Flynn P. Data Clustering: A Review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[13] Borlard H, Kamp Y. Auto-association by Multilayer Perceptrons and Singular Value Decomposition [J]. Biological Cybernetics, 1988, 59: 291-294.

[14] Blake C L, Merz C J. UCI Repository of Machine Learning Databases [EB/OL]. (2010-11-21). http://www.ics.uci.edu/mlearn/MLRepository.html.

[15] Frey P W, Slate D J. Letter Recognition Using Hollandstyle Adaptive Classifiers [J]. Machine Learning, 1991, 6(2): 161-182.

编辑 刘 冰 陆燕菲