文章编号:1007-2985(2020)02-0015-04

基于 MEM 和 HMM 的中文词性标注方法*

周　潭,莫礼平,胡美琪,李航程
(吉首大学信息科学与工程学院,湖南 吉首 416000)

摘　要:为了进一步提高中文语料库中语料的词性标注效率,在分析最大熵模型(MEM)和隐马尔科夫模型(HMM)所涉及理论、算法及其在中文词性标注技术中的应用的基础上,进行了基于 MEM 和 HMM 的中文词性标注实验.实验结果显示,基于 MEM 和 HMM 的中文词性标注算法都获得了一致性很好且覆盖率较高的标注效果,中文词性标注的准确率、召回率和 F_1 这 3 个指标均达到 92%以上;MEM 的标注效果总体上比 HMM 的稍佳.

关键词:最大熵模型;隐马尔科夫模型;中文词性标注
中图分类号:TP391.1　　**文献标志码**:A　　**DOI**:10.13438/j.cnki.jdzk.2020.02.004

词性标注(Par-of-Speech Tagging,POS Tagging)就是赋予每个词语一个正确候选词性的过程,它是自然语言信息处理研究的重要内容.2015 年,梁喜涛等[1]对现有词性标注方法进行了分析整理,将传统的词性标注方法归纳为 3 类:(1)基于规则的方法.该类方法简单,易于实现,但构造规则是一项非常艰难的任务.(2)基于统计的方法.该类方法客观性强,准确性较高,但需要处理兼类词和未登录词的问题.基于最大熵模型(Maximum Entropy Model,MEM)和隐马尔科夫模型(Hidden Markov Model,HMM)的词性标注方法是统计类方法的典型代表,因其能够获得一致性很好且覆盖率较高的标注结果而被广泛关注[2].(3)基于规则和统计的方法.该类方法有效地利用了规则类方法和统计类方法的优势,但标注效果依赖于建立的规则或人工的选取特征,且与任务领域的资源有很大的相关性,一旦领域变化,标注效果就会受较大影响.因此,笔者将对基于 MEM 和 HMM 的中文词性标注方法进行理论分析和对比实验.

1　基于 MEM 的中文词性标注方法

1.1 MEM 理论

在热力学中,熵是大量微观粒子的位置和速度的分布概率的函数,用"热熵"表示分子状态混乱程度.1948 年,Shannon[3]借鉴热力学的概念提出"信息熵"的概念.为了描述信源的不确定度,Shannon 将信息中排除了冗余后的平均信息量称为"信息熵",并给出了计算信息熵的数学表达式.通常,一种信息源的不确定性越大,其信息熵就越高;反之,其信息熵就越低.1957 年,Jaynes[4]提出了基于概率统计的最大熵方法.最大熵方法通过将各种不同来源的信息知识聚集在一个框架下面,用以解决一些复杂的问题.1992 年,Della 等[5]首次将最大熵方法应用于自然语言处理.经过近 30 年的发展,基于 MEM 的自然语言处理技术取得了令人瞩目的成果.

* 收稿日期:2019-07-26
基金项目:国家自然科学基金资助项目(61462029);湖南省自然科学基金资助项目(2019JJ40234);吉首大学本科生科研项目(JDX1809);湖南省大学生研究性学习和创新性实验计划项目(湘教通〔2018〕255 号);吉首大学生研究性学习和创新性实验计划项目(JDCX2018012)
通信作者:莫礼平(1972—),女,湖南益阳人,吉首大学信息科学与工程学院高级实验师,主要从事自然语言处理、Petri 网理论及其应用研究.

最大熵方法的本质就是从满足约束的模型中选择熵值最大的.利用 MEM 需要解决特征选择和模型选择这 2 个基本问题:特征选择就是选择一个能表达这个随机过程的统计特征的集合;模型选择就是参数估计或模型估计,为入选的特征集合估计权重.假设现有 n 个特征,约束的集合定义为

$$C \equiv \{p \in P \mid E_p(f_i) = E_{\tilde{p}}(f_i), i \in \{1,2,\cdots,n\}\}. \tag{1}$$

最大熵方法就是求解满足约束(1)的模型.这样模型可能不只 1 个,所以需要找到一个最均匀分布的概率模型.概率模型的均匀性可以用如下的条件熵来衡量:

$$H(p) \equiv -\sum_{x,y} \tilde{p}(x)p(y \mid x)\log p(y \mid x), \tag{2}$$

其中 $p(y \mid x)$ 表示在 x 出现的情况下 y 出现的概率.满足(2)式的唯一的最均匀分布模型可表示为 $p^* = \arg\max_p H(p)$.

1.2 MEM 在中文词性标注中的应用

最大熵方法应用于中文词性标注,需要根据上下文信息确定约束条件,从而建立 MEM.基于 MEM 的中文词性标注方法的重点是根据中文的特殊性进行特征选取.当某一现象出现多次时,就认为该现象不是偶然的,而是表现了数据某一方面的特征.因为人工选取特征耗时耗力,所以一般是由机器自动在训练数据中寻找这种特征.特征的选取一般分为 2 步[6]:第 1 步,利用特征模板从语料中获取候选的特征;第 2 步,从候选特征集中选取特征.在国内的词性标注研究中,大多采用基于词的上下文特征.但汉语不同于英文,汉语的每个字一般都有其自身的意义,而英文的单个字母没有具体意义,因此在对汉语进行特征选择时考虑字的编码信息,会有助于有提高词性标注的准确率[7].

现以对文本"把这次演讲安排一下"中的"下"进行词性标注为例,说明如何将 MEM 应用于中文词性标注中.首先,将文本标注为"把 $/q-p-v-n$ 这 $/t$ 次 $/p$ 演讲 $/v-n$ 安排 $/v-n$ 一 $/m-c$ 下 $/f-q-v$",其中每个词后的字母代表该词所可能具有的词性.由该标注序列可知,"下"在此句子中可能有 f,q,v 这 3 种词性.用 t_1,t_2,t_3 来表示这 3 种词性,即 $t_1 = f, t_2 = q, t_3 = v$,则根据"下"的 3 种词性得到第 1 个约束条件:

$$P(t_1) + P(t_2) + P(t_3) = 1. \tag{3}$$

基于约束(3),即可找到词"下"的词性标注的合适模型.但满足约束(3)的模型可以有无限个,例如,$M_1 = \{P(t_1) = 0.5, P(t_2) = 0, P(t_3) = 0.5\}$,$M_2 = \{P(t_1) = 1, P(t_2) = 0, P(t_3) = 0\}$.模型 M_1 和 M_2 都只做了粗略假设,没有任何的经验判断.假设当前词语的词性只有 3 种候选,那么最直观的合适模型就是 $M_3 = \{P(t_1) = 1/3, P(t_2) = 1/3, P(t_3) = 1/3\}$.在模型 M_3 中,3 种可能词性出现的概率相同,是均匀模型.同时注意到,在训练样例中 90% 的"一下"中的"下"的词性为 t_2.据此可得第 2 个约束条件:$P(t_2) = 0.9$.此时,还有许多的概率分布都能同时满足上述 2 个约束条件.在没有其他约束条件下,合理的选择仍然是概率分布最均匀的模型.即在满足上述 2 个约束的同时,尽可能平均分配它的概率分布:$P(t_1) = 0.05, P(t_3) = 0.05, P(t_2) = 0.9$.

2 基于 HMM 的中文词性标注方法

2.1 HMM 理论

基于统计的方法是最常使用的一类词性标注算法.对于给定的输入词串,基于统计的方法先确定其所有可能的词性串,再对它们打分,选择得分最高的词性串作为最佳的输出结果.在所有基于统计的方法中,基于 HMM 的词性标注算法最常见[8].目前,HMM 已应用于各种语言的词性标注并取得极高的标注准确率,基于 HMM 的中文词性标注方法研究也受到人们的重视.HMM 是在离散马尔科夫过程的基础上改进的.它包含 2 个随机过程,一个是已知的观察序列,另一个是隐含的状态转移序列.状态转移序列是不可观测的,需要通过观察序列来推断[9].

为了理解 HMM,先看一个实例:缸和球的实验.设有 N 个缸,M 种不同颜色的球,每一个缸都装有很多不同颜色的球,球的颜色由一组概率分布描述.首先,根据某种随机过程选择 N 个缸中的某个缸,记为 Z_1,再根据这个缸中球的颜色概率分布,随机选择一个球,记该球的颜色为 O_1,并将球放回缸中;然后,根据缸的状态

转移概率分布,随机选择下一个缸,记为 Z_2,再根据该缸中球的颜色的概率分布,随机选择一个球,记该球的颜色为 O_2,并将球放回缸中……如此循环,一共进行 T 次实验,得到缸的选取序列 $Z=(Z_1,Z_2,\cdots,Z_T)$ 和球的颜色序列 $O=(O_1,O_2,\cdots,O_T)$.称可以直接观察到的球的颜色序列为观察序列,称在后台进行的缸的选取序列为隐藏状态序列.通常,HMM 可用一个五元组 $\lambda=(N,M,\boldsymbol{A},\boldsymbol{B},\boldsymbol{\pi})$ 来表示[9]:(1)N 表示模型中隐含状态的数目.用 T 表示状态的集合,$T=\{T_1,T_2,\cdots,T_N\}$,t 时刻的状态为 T_j,$1\leqslant j\leqslant N$.(2)M 表示模型中观察值的数目.用 o 表示观察值的集合,$o=\{o_1,o_2,\cdots,o_M\}$,t 时刻的观察值为 o_k,$1\leqslant k\leqslant M$.(3)$\boldsymbol{A}$ 表示状态转移概率矩阵.$\boldsymbol{A}=(a_{ij})$,其中 $a_{ij}=P(q_t=T_j\mid q_{t-1}=T_i)$,$1\leqslant i\leqslant N$,$1\leqslant j\leqslant N$,表示状态从 T_i 转移到状态 T_j 的概率.(4)$\boldsymbol{B}$ 表示符号的发射概率矩阵,它描述了 HMM 模型中每个状态下出现各个观察值的概率.$\boldsymbol{B}=(b_{jk})$,其中 $b_{jk}=P(x_t=o_k\mid q_t=T_j)$,$1\leqslant j\leqslant N$,$1\leqslant k\leqslant M$,表示在 t 时刻、状态 T_j 时观察值为 o_k 的概率.(5)$\boldsymbol{\pi}$ 表示初始状态概率向量.$\boldsymbol{\pi}=(\pi_j)$,其中 $\pi_j=P(q_1=T_j)$,$1\leqslant j\leqslant N$,表示在初始时刻($t=1$)、状态为 T_j 时的概率.

为了便于计算五元组中的 3 个概率矩阵,在实际应用中常做如下假设:(1) 马尔可夫假设.在状态转移过程中,当前的状态只与前一时刻的状态有关,而与之前的其他状态无关.该假设用数学表达式可表示为 $P(q_t\mid q_{t-1},\cdots,q_1,\lambda)=P(q_t\mid q_{t-1},\lambda)$.(2) 输出独立性假设.即当前时刻输出的观察值的概率只与当前时刻的状态有关.该假设用数学表达式可表示为 $P(x_1,\cdots,x_T\mid q_1,\cdots,q_T,\lambda)=\prod_{t=1}^{T}P(x_t\mid q_t)$.(3) 不动性假说.即假设 2 个状态的转移与具体的时间无关.

2.2 HMM 在中文词性标注中的应用

HMM 可以用来解决 3 个基本问题:第 1 个问题是评估问题,即根据给定的 HMM 求解一个观察序列的概率,可用向前算法求解此类问题;第 2 个问题是解码问题,即求解生成一个观察序列的最优隐藏状态序列,可用 Viterbi 算法求解此类问题;第 3 个问题是学习问题,即已知观察序列 O,求解 HMM 的参数,可用向前向后算法求解此类问题.

词性标注问题实际上就是解码问题.将 HMM 应用于词性标注,那么在五元组 $\lambda=(N,M,\boldsymbol{A},\boldsymbol{B},\boldsymbol{\pi})$ 中:N 为词性的数目;M 为词汇的数目;$\boldsymbol{A}$ 为词性状态转移概率矩阵,a_{ij} 表示词性从 T_i 转移到 T_j 的概率;$\boldsymbol{B}$ 为词汇的发射概率矩阵,b_{jk} 表示词性标注为 T_j 的情况下输出词汇 o_k 的概率;$\boldsymbol{\pi}$ 为初始状态概率分布,π_j 表示初始状态词性为 T_j 的概率[10].HMM 五元组中的参数 N 和 M 易求,故只要计算出 $\boldsymbol{A}$,$\boldsymbol{B}$,$\boldsymbol{\pi}$ 这 3 个参数值,就可利用 Viterbi 算法来找出最优的词性序列.

3　实验与结果

本实验采用 Python 语言编程实现基于 MEM 和 HMM 的中文词性标注算法,并在 Inter(R) Core(TM) i5-3470 CPU @3.20 GHz、4 G 内存、Win10 操作系统条件下进行实验.采用北京大学加工整理的《人民日报》1998 年 1 月份的新闻语料作为训练集和测试集.为了测试 2 个模型的实际标注效果,从训练的语料库中随机选取 1 000 行语料作为测试样本 1,随机选取 2 000 行语料作为测试样本 2.2 个模型的词性标注准确率、召回率和 F_1 这 3 个性能指标的比较见表 1.

表 1　2 个模型的中文词性标注的实验结果

Table 1　Experimental Results of Chinese Part-of-Speech Tagging Based on Two Models　　%

样本	模型	准确率	召回率	F_1
样本 1	MEM	92.52	93.12	92.82
	HMM	92.34	92.73	92.49
样本 2	MEM	93.03	93.48	93.25
	HMM	92.85	93.56	93.20

由表 1 可知,2 个模型的中文词性标注都获得了一致性很好且覆盖率较高的标注效果,准确率、召回率和 F_1 这 3 个指标都达到 92%以上.MEM 的标注效果总体上比 HMM 的稍佳,这与其灵活的特征机制有利于在词性标注的过程中更有效地利用上下文的信息有关.

4 结语

MEM和HMM是词性标注领域研究较多且应用较广的2个统计模型.基于MEM和HMM的中文词性标注方法具有更客观、适应性强和耗费资源少的优点,且可以通过训练更大规模的语料库来解决数据稀疏的问题.笔者分析了MEM和HMM所涉及理论、算法,并通过实验验证了2个模型用于中文词性标注的有效性,对于帮助人们更好地理解和掌握中文信息处理技术相关理论与方法具有一定的实用价值.接下来,笔者将利用MEM和HMM模型的优越性,尝试结合新型神经网络和智能优化算法对统计类中文词性标注算法进行改进.

参考文献:

[1] 梁喜涛,顾 磊.中文分词与词性标注研究[J].计算机技术与发展,2015,25(2):175-180.
[2] 魏 欧,吴 健,孙玉芳,等.基于统计的汉语词性标注方法的分析与改进[J].软件学报,2000,11(4):473-480.
[3] SHANNON C E. A Mathematical Theory of Communication[J]. Bell System Technical Journal, 1948, 27(3): 379-423.
[4] JAYNES EDWIN. Information Theory and Statistical Mechanics [J]. Physical Review, 1957, 106(4): 620-630.
[5] DELLA PIETRA S, DELLA PIETRA V, MERCER R L, et al. Adaptive Language Modeling Using Minimum Discriminant Estimation[C]// Proceedings of the Workshop on Speech and Natural Language. Association for Computational Linguistics, 1992: 103-106.
[6] 周雅倩.最大熵方法及其在自然语言处理中的应用[D].上海:复旦大学,2005:26.
[7] 赵法兴,赵 伟.平滑的最大熵模型在汉语词性自动标注中的应用[J].长春工业大学学报(自然科学版),2007,28(2):213-216.
[8] 唐 超.基于统计模型的汉语词性标注系统的改进方法研究[D].北京邮电大学,2009:4.
[9] 杨荣根,杨 忠.基于HMM中文词性标注研究[J].金陵科技学院学报,2017,33(1):20-23.
[10] 刘洁彬,宋茂强,赵 方,等.基于上下文的二阶隐马尔可夫模型[J].计算机工程,2010,36(10):231-232;235.

Chinese Part-of-Speech Tagging Method Based on Maximum Entropy Model and Hidden Markov Model

ZHOU Tan, MO Liping*, HU Meiqi, LI Hangcheng
(College of Information Science & Engineering, Jishou University, Jishou 416000, Hunan China)

Abstract: In order to further improve the efficiency of part-of-speech tagging in Chinese corpora, experiments of Chinese part-of-speech tagging methods based on the maximum entropy model (MEM) and the hidden Markov model (HMM) are designed according to the theoretical basis, algorithms, and application technology. The experimental results show that the Chinese part-of-speech tagging algorithms based on MEM and HMM have obtained a very consistent and high-coverage tagging result and the three indicators of tagging accuracy, recall rate and F_1 value have reached above 92%, with the effect of MEM better than that of HMM.

Key words: maximum entropy model; hidden Markov model; Chinese part-of-speech tagging

(责任编辑 向阳洁)