基于历史模型的蒙古文自动词性标注研究

赵建东，高光来，飞龙

（内蒙古大学计算机学院，呼和浩特 010021）

E-mail:djzshadow@126.com，csggl@imu.edu.cn，csfeilong@imu.edu.cn

摘要：蒙古文自动词性标注方面的研究工作较少，制约了对蒙古文的机器翻译、语法分析及语义分析等领域的深入研究。针对于此，提出了加入 *lookahead* 学习机制的基于历史模型的蒙古文自动词性标注方法。实验表明，加入 *lookahead* 学习机制的基于历史模型的蒙古文自动词性标注方法对蒙古文的未登录词、集内词、总体词自动词性标注的准确率分别达到了 71.2766%、99.1482%、95.3010%，说明此方法可以较好的进行蒙古文的自动词性标注。

关键词：历史模型；*lookahead*；蒙古文；自动词性标注

中图分类号：TP391　　**文献标识码：**A

Research on History-based Mongolian Automatic POS Tagging

ZHAO Jiandong, GAO Guanglai, BAO Feilong

(College of Computer Science, Inner Mongolia University, Hohhot 010021, China)

E-mail:djzshadow@126.com, csggl@imu.edu.cn, csfeilong@imu.edu.cn

Abstract: The researches on Mongolian machine translation, syntax analysis and semantic analysis are restricted because of few researches on Mongolian automatic Part-Of-Speech (POS) tagging. In view of this, we proposed a history-based Mongolian automatic POS tagging method which incorporating a *lookahead* mechanism into the decision making process. Experiment results showed that the POS tagging accuracy of Mongolian unknown words, known words and all words are 71.2766%, 99.1482% and 95.3010%, respectively, which demonstrate that our method is appropriate for Mongolian automatic POS tagging.

Key words: History-models; learning with *lookahead*; Mongolian; automatic POS tagging

1　引言

词性标注是自然语言处理领域的基础性研究课题之一，该项研究对于机器翻译、语法分析、语义分析等领域的研究有着重要的意义。在汉语、英语等方面已经有了许多自动词性标注方面的研究，采用的方法主要有基于转换[1]、隐马尔科夫模型[2]、支持向量机[3]、最大熵模型[4]等方法。

蒙古文作为少数民族语言，词性标注的研究开始较晚，研究工作也比较少。目前为止，胡冠龙，张建，李淼做了改进的基于转换方法的拉丁蒙文词性标注[5]；艳红，王斯日古楞用基于隐马尔科夫模型(HMM)的方法对蒙古语自动词性标注进行了研究[6]；张贯红，斯•劳格劳，乌达巴拉做了最大熵蒙古文词性标注的研究[7]等。这些工作对于研究蒙古文的自动标注具有重要的作用，但是每种方法都有一些不足。基于转换的蒙古文词性标注方法，是一种基于规则的方法，这种方法通常采用手工来编制复杂的词性标注规则系统，可以充分利用人的语言知识，但是带有很强的主观性，容易产生规则冲突，对语言学专家的依赖性强，存在知识获取的瓶颈问题，并且加工、调试规则费时费力。基于 HMM 的蒙古文词性标注方法是基于统计的方法，由于人工标注训练语料库的易得性和统计模型的健壮性，使它成为主流的词性标注方法，但是这种方法无法使用复杂特征，对未登录词词性标注的准确率较低，而蒙古文中外来词汇量很大，这需要寻求其他方法来进行改善。最大熵模型能够较好的包容各种约

收稿日期：2013-6-1　　**定稿日期：**

基金项目：国家自然科学基金项目(No. 61263037)；内蒙古自然科学基金重大项目(No. 2011ZD11)

作者简介：赵建东(1988-)，男，在读硕士，主要研究方向为语音合成、语音识别；高光来(1964-)，男，教授。主要研究方向为文字识别、语音识别、图像识别、计算智能和数据挖掘；飞龙（1985-），男，博士，研究方向主要为蒙古文信息处理、语音识别、语音合成、语音检索。

束信息及与自然语言模型相适应等优点，在蒙古文词性标注研究中取得了比较好的效果。但算法收敛的速度较慢，所以导致它的计算代价较大，时空开销大，而且数据稀疏问题比较严重。针对这些问题本文用加入 *lookahead* 学习机制的基于历史模型的方法研究了蒙古文自动词性标注，这种方法具有 HMM 模型和最大熵模型的优点。通过与基于最大熵模型的方法进行对比，实验结果表明，加入 *lookahead* 学习机制的基于历史模型的蒙古文自动词性标注方法优于基于最大熵模型的蒙古文词性标注方法。这对于进一步研究蒙古文自动词性标注具有重要的意义。

本文的结构安排如下，第二部分详细介绍基于历史模型的词性标注的方法，第三部分为蒙古文自动词性标注实验，最后一部分为结论。

2 基于历史模型的词性标注方法

2.1 基于历史模型的发展

基于历史模型的方法[8]已经被广泛应用于词性标注等一系列的自然语言处理任务中。它的思想是将复杂的结构预测问题分解成一系列的分类问题，并把过去的决策和部分完成的结构信息作为特征，然后用基于机器学习的分类器来做每一个决策。基于历史模型的方法有很多优点，但是，由于标注偏置问题在词性标注等方面的准确率经常不如全局优化的方法。近年来基于全局优化模型的方法越来越流行，但同时也暴露出了计算复杂度高的缺点。

研究[9]发现，在决策过程中采用 *lookahead* 学习机制可以显著的提高基于历史模型方法的准确率。这里的 *lookahead* 和句法分析中的“lookahead”的含义不同，句法分析中的“lookahead”是根据观察正确的词来选择正确的解析操作。这里的 *lookahead* 是指选择最佳操作的过程，在这个过程中要考虑未来操作的不同序列，并评价由这些序列构成的结构。换句话说，就是在未来的操作空间执行搜索时采用了 *lookahead* 机制。

2.2 “移进-归纳”分析器

为了介绍基于历史模型中的 *lookahead*，先简单介绍一个基于历史模型的依存分析方法的例子—“移进-归纳”分析器。“移进—归纳”算法主要包含堆栈和队列两种数据结构。堆栈用来储存中间的解析结果，队列用来储存读取的单词。移进和归纳两种操作方式在这两种数据结构中构成一个接一个的关系。通过移进和归纳操作，“移进-归纳”分析器从单词序列中读取单词并且构建一个依存关系树，同时把依存关系储存在堆栈中。我们在一个状态时，通常不知道是应该选择移进还是归纳。传统的方法是用多级分类器去决定下一个操作，分类器要根据堆栈和队列中词的表面形态、词性等信息去选择最合理的操作。

在 *lookahead* 策略中，是根据未来的状态去做出决定。由于未来的状态能够提供更多的附加信息或者有用的消歧信息，所以从理论上讲用这种方法可以提高准确率。但是，这样也会带来一些问题，随着 *lookahead* 的深度的增加，*lookahead* 方法的计算代价会增加。不过这种代价被证明是有价值的。

2.3 *Lookahead* 学习机制

我们用状态表示部分完成的分析，相当于基于历史的确定性分析中在每个决策点收集的历史信息，状态还可以根据允许的操作来进行转移。在词性标注中，状态就是单词，我们需要将词性的标签分配给当前的目标单词，允许的可能的操作就是词性集合中的所有词性标签。

2.3.1 搜索算法

在 *lookahead* 学习机制中需要一个搜索算法来搜索每个可能的操作，图 1 为搜索算法的伪代码。该算法采用深度优先策略执行搜索，在状态空间中找到得分最高的状态，搜索复杂度由深度 *d* 决定。这个搜索过程是执行一个递归函数，递归函数以剩余的搜索深度和当前的状态作为它的输入，最终以得分最高的状态及其得分作为返回值。假设一个线性得分模型，每个状态 S 的得分就是当前的权值向量 ω 和代表状态的特征向量 $\Phi(S)$ 的点积。得分从搜索

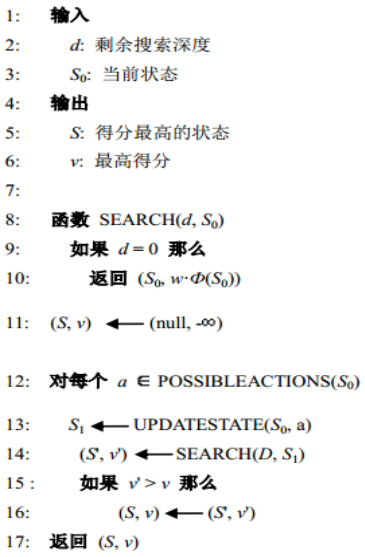


图 1 搜索算法

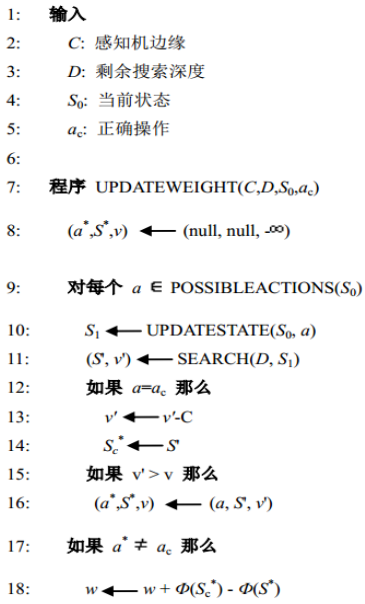


图 2 感知机权值更新算法

树的每个叶子节点开始计算并将其备份到根节点（在词性标注等实际应用中，如果每次都从叶子节点开始计算，这样效率会太低，所以通常是状态每次被一个操作更新时就直接计算得分的增量，以此来得到每个状态的得分）。采用这种搜索算法进行标注的时间复杂度是 $O(nm^{D+1})$，其中 n 是处理句子需要的操作次数，m 是每个状态可能的平均操作数。D 是搜索的深度。由于这种搜索的方式是基于历史的，所以并不需要局部的特征，即我们可以根据任意的特征进行决策。可以说这种学习的机制权衡了全局最优参数学习和特征的灵活选择性。

2.3.2 训练一个边缘感知机

为了优化 *lookahead* 搜索算法中的权值，我们采用了边缘感知机的学习算法[10]。边缘感知机和支持向量机类似，与没有采用边缘的感知机相比，它能够产生更加精确的模型。图 2 给出了我们采用的学习算法的伪代码。这个学习算法与标准的边缘感知机的训练算法相似，即当感知机出现错误时，我们用两个不同的特征向量来更新权值，一个特征向量对应正确的操作；另一个特征向量对应得分最高的操作，当边缘不足够大的时候，也可以对应为得分次最高的操作。需要注意的是次最优操作向量可以用正确操作的得分减去边缘距离这样一个简单的技巧而自动选择，即图 2 中的第 13 行。这种权值更新算法与边缘感知机的标准算法相比，唯一不同的是，这种算法用到了根据 *lookahead* 搜索算法得到的状态以及状态的得分（图 2 中第 11 行），这些状态及其得分从搜索树的叶子节点就开始被备份。研究证明如果训练数据是线性可分的，则这种训练算法是收敛的，并且边界最后至少能达到真正边界的一半[9]。同许多采用感知器的研究[11]一样，文中采用的训练算法在整个训练迭代结束时，也将权值向量进行了平均。

3 蒙古文自动词性标注实验

3．1 语料库的构建

蒙古文是一种拼音文字，它的拼写规则是以词为单位竖写，词与词之间用空格分开，采取从上到下的书序，从左到右的行序。蒙古文字母在词中变化有很多，在一个蒙古文单词中，蒙古文字母在上、中、下位置不同而导致的写法也不同，并且蒙古文字母中形同音不同的现象比较普遍。鉴于蒙古文中元音与辅音的形式变化多样问题，对蒙古文进行处理时，通常采用拉丁转写的方法。这样有助于蒙古文的校正、统计和研究。

构建语料库时，我们选取了 10990 句蒙古文句子，先根据转换规则将蒙古文转到拉丁转写，再进行词性标注。采用的词性主要有：语气词、名词、动词、连词、副词、形容词、

数词、量词、后置词、构成附加成分、模拟词、情态词、感叹词、时位词、时间词、不确定词、复合词、代词、标点。实验方法是随机选取其中的 9900 句作为训练语料，剩余的 1090 句作为测试语料，详细的实验数据如表 1 所示。

表 1 实验数据

语料	句子数	词数	未登录词个数
训练语料	9900	62890	-
测试语料	1090	6810	940

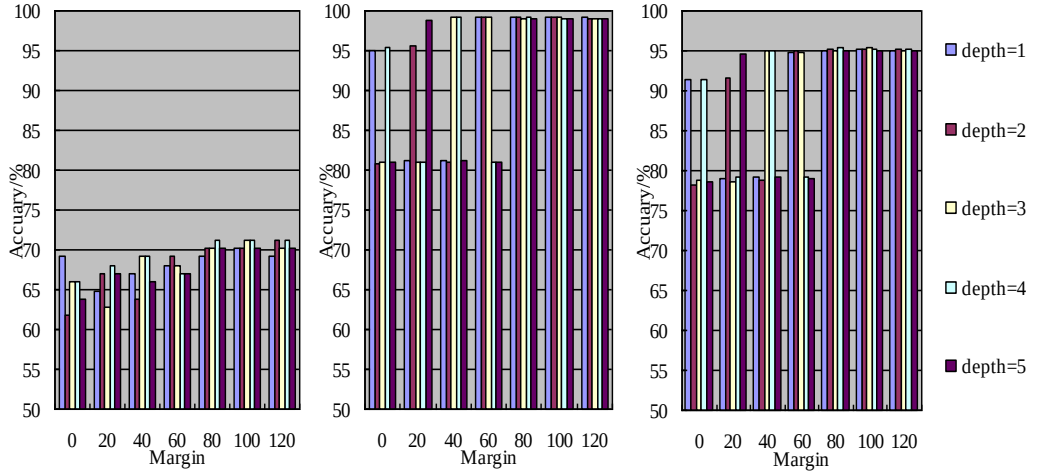
3．2 实验结果

我们用加入 *lookahead* 学习机制的基于历史模型的方法进行了蒙古文自动词性标注实验，工具为 lapos[12]。为了评价蒙古文自动标注系统性能，主要采用词性标注正确率进行评价，即

$$P_T = \frac{N_T}{N_{ALL}} \times 100\%$$

其中：N_{ALL}表示用来测试的语料库中词的总数量；N_T 是测试时标注正确的词的数量。

为了研究这种方法中搜索深度和边缘对蒙古文自动词性标注准确率的影响，我们进行了多组交叉实验，实验中深度分别设定为 1、2、3、4、5，边缘分别设定为 0、20、40、60、80、100、120。实验的结果如图 3 所示。



(1)未登录词　　（2）集内词　　（3）所有词

图 3 蒙古文自动词性标注准确率实验结果

分析图 3，我们可以得到几个信息：（1）当感知机边缘小于等于 60 的时候，准确率随深度变化而变化的幅度较大，说明边缘小时模型不稳定，所以为使模型稳定需要增大边缘；（2）随着边缘的增大，模型趋于稳定。但是当大于等于 80 的时候，再增大边缘对于准确率的影响不大；（3）在边缘适宜的时候，深度对于集内词和总体词词性标注的准确率的影响并不是很大，相比对于未登录词的影响要稍大些。从本文第二部分我们知道增大深度会增加搜索时间的复杂度，综合这些信息，深度为 3、边缘为 100 时候本文的模型表现最佳，此时未登录词、集内词、总体词词性自动标注的准确率分别达到了 71.2766%、99.1482%、95.3010%。

为了将本文提出的加入 *lookahead* 学习机制的基于历史的蒙古文自动词性标注方法与基于最大熵模型的蒙古文自动词性标注方法进行比较，我们用相同的实验数据做了一些基于最大熵模型的蒙古文自动词性标注实验，采用工具是 maxent[13]。实验结果如表 2 所示。

表 2 基于最大熵模型蒙古文词性自动标注结果

迭代次数	未登录词准确率/%	集内词准确率/%	所有词准确率/%
20	67.0213	93.5264	89.8678
60	70.2128	93.5264	90.3084
100	68.0851	93.3560	89.8678
140	68.0851	93.3560	89.8678
180	68.0851	93.0153	89.5742
220	68.0851	93.0153	89.5742

从表 2 中可以看出，当迭代次数为 60 的时候，最大熵词性标注方法的准确率达到最好的效果，此时未登录词、集内词、总体词词性自动标注的准确率分别达到了 70.2128%、93.5264%、90.3084%。图 4 中给出了这两种方法在表现最佳的情况下的准确率对比结果。从中可以看出加入 *lookahead* 学习机制的基于历史模型的蒙古文自动词性标注的准确率要高于基于最大熵模型的蒙古文自动词性标注的准确率。

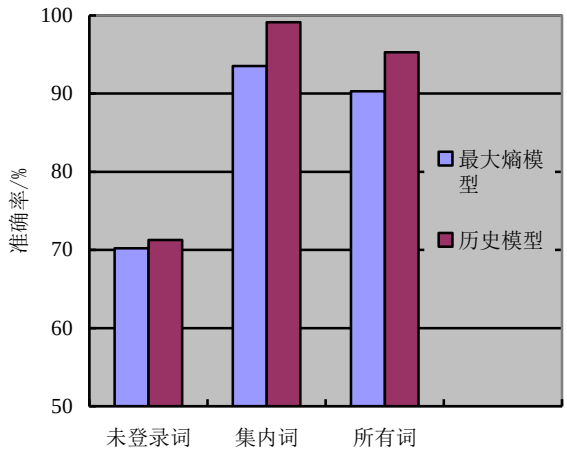


图 4 本文方法与基于最大熵模型词性标注方法准确率对比

4 结论

针对当前蒙古文词性自动标注研究工作较少，制约了对蒙古文的机器翻译、语法分析及语义分析等领域的深入研究这个问题，本文采用加入 *lookahead* 学习机制的基于历史模型的方法对蒙古文自动词性标注进行了研究。实验结果表明在深度为 3，边缘为 100 时的未登录词、集内词、总体词词性自动标注的准确率分别达到了 71.2766%、99.1482%、95.3010%。在相同的训练集和测试集下与基于最大熵模型的方法相比准确率的提高比较显著，说明本文的方法有一定的优势。虽然取得了一定的成绩，但是我们训练的语料库还是比较小，且包含的领域比较单一。未来我们将搜集更多的蒙古文语料库进行词性标注，以此来提高并验证模型的健壮性。另一个问题是我们与其他自动词性标注的方法对比实验还较少，需要做进一步的研究工作。

参考文献：

[1] Brill E. Transformation based error driven learning and natural language processing: A case study in part of speech tagging[J]. Computational Linguistics, 1995, 21(4): 543-565

[2] Brants T. TnT: A statistical part of speech tagger[C]. Proc of the 6th Conf on Applied Natural Language Processing. Stroudsburg, Association for Computational Linguistics, 2000: 224-231

[3] Gimenez J, Marquez I. Fast and accurate part of speech tagging: The SVM approach[C]. Proc of the 4th Int Conf on Recent Advances in Natural Language Processing. 2003:158-165

[4] Ratnaparkhi A. A maximum entropy model of part of speech tagging[C]. Proc EMNLP. Computational Linguistics, Cambridge, MIT Press, 1996: 133-141

[5] 胡冠龙，张建，李淼. 改进的基于转换方法的拉丁蒙文词性标注[J]. 计算机应用, 2007, 27(4): 963-965

[6] 艳红，王斯日古楞. 基于HMM的蒙古文自动词性标注研究[J]. 内蒙古师范大学学报（自然科学版）, 2010, 39(2): 206-209

[7] 张贯红，斯•劳格劳，乌达巴拉. 融合形态特征的最大熵蒙古文词性标注模型. 计算机研究与发展, 2011, 48(12): 2385-2390

[8] Andrew McCallum, Dayne Freitag, and Fernando Pereira. Maximum entropy markov models for information extraction and segmentation[C]. In Proceedings of ICML. San Francisco, Morgan Kaufmann Publishers, 2000: 591–598

[9] Yoshimasa Tsuruoka, Yusuke Miyao, Jun'ichi Kazama. Learning with Lookahead: Can History-Base Models Rival Globally Optimized Models?[C]. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Stroudsburg, Association for Computational Linguistics, 2011:238-246

[10] W Krauth and M Mezard. Learning algorithms with optimal stability in neural networks[J]. Journal of Phisics A, 1987, 20(11): L745–L752

[11] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms[C]. In Proceedings of EMNLP. Stroudsburg, Association for Computational Linguistics, 2002:1–8

[12] Yoshimasa Tsuruoka, Lookahead Part-Of-Speech Tagger [CP], 2012, http://www.logos.ic.i.u-tokyo.ac.jp/~tsuruoka/lapos/

[13] Zhang Le, Maximum Entropy Modeling Toolkit for Python and C++ [CP], 2011, http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

作者：赵建东

电话：13847142363

邮箱：djzshadow@126.com

地址：呼和浩特市大学西路 235 号内蒙古大学计算机学院

邮编：010021