文章编号：1003-0077（2011）00-0000-00

基于感知器算法的维吾尔语词性标注研究

帕提古力依马木[1]，买合木提买买提[1]，卡哈尔江阿比的热西提[1]，

（1.新疆大学 信息科学与工程学院，新疆 乌鲁木齐市 830046）

摘要：维吾尔语自动标注是维吾尔语信息处理后续维吾尔语句法分析，语义分析及篇章分析中必不可少的基础工作。词性是词的重要的语法信息，假如一个词的词性无法确定或一个词给予错误的词性，对后续句法分析造成直接的影响句法分析等一些工作。本文中使用感知器训练算法和 viterbi 算法对维吾尔语进行词性标注并在词性标注时利用词的上下文信息作为特征。实验结果表明，用该方法对维吾尔语词性标注有良好的效果。

关键词：词性标注；感知器算法；维吾尔语词性标注

中图分类号：TP391　　**文献标识码：**A

Uyghur speech tagging based on perceptron algorithm research

，　，，

(1.Information Science and Engineering Technology Institute, Xinjiang University, Urumqi 830046, China)

Abstract:Uyghur automatic tagging follow Uyghur statement analysis, semantic analysis and discourse analysis become an indispensable foundation work of uyghur language information processing. Speech is an important gra mmar information of the word, if the speech of word unable to determine or a term given to the wrong speech, it w il directly make bad influnce to the some work, such as syntactic analysis. In this paper, perceptron training algorithm and viterbi algorithm are used for uyghur language to conduct speech tagging and use the conte xt information of the word as a feature when tagging the speech. Experiment results show that, this method have g ood results to the uyghur language speech tagging.

Key words: The speech tagging. Perceptron algorithm; Uyghur speech tagging

1 引言

目前基于统计的词性标注方法得到了广泛的应用并取得了良好的效果。在基于统计方法的词性标注中，对兼类词和未登录词的处理是需要解决的问题。对兼类词和未登录，可以根据词的上下信息来确定词在句子中的词性。

维吾尔语属粘着语，其形态变化比较丰富。名词有数、格、人称等语法范畴；动词有数、人称、时态、语态的变化；形容词有级的范畴。形态变化一方面提供了一些深层语法信息，为词法分析、词性标注带来极大的方便，另一方面也增加了自动标注的复杂性。维吾尔语同其它语言一样也存在词性歧义现象（词兼类现象）。在维吾尔语中兼类词数量较多，且使用频率较高，这给维吾尔语词性标注带来了很大的困难。兼类现象是词性标注中的一个不可回避的重点和难点，词性是词的重要的语法信息，假如一个词的词性无法确定，对后续句法分析造成直接的影响句法分析就无法进行。如果一个词赋予错误的词性，将导致严重的句法分析错误，所以，维吾尔语词性标注在自然语言处理中有至关重要的意义。本文中使用感知器算法进行维吾尔语的词性标注。目前基于感知器算法的模型在各个领域都表现出很好的性能，本文主要利用感知器算法的优点，在进行词性标注时利用词的上下文信息作为特征，在维吾尔语词性标注中取得了好的效果。

2 相关工作

目前词性标注方法可分为三类：基于规则的词性标注方法、基于转换的错误驱动词性标

注方法以及基于统计的词性标注方法。

1）基于规则的词性标注方法

基于规则的词性标注方法首先由语言学家制定相应的规则，在规则中使用大量的上下文信息来对词性进行判断。词性标注的性能与规则制定者的语言学知识具有很大的关系。其次要构造一套对语言的各方面特性都覆盖的规则是一个艰难很耗时的工作，而且随着规则数量的增加，各规则之间往往会产生冲突。最具有代表性的基于规则的词性标注系统是 1971 年开发的 TAGGIT 标注系统[1]。对于维吾尔语来说，基于规则的词性标注有吐尔根等开发的基于词典的维吾尔语词性标注[2]。

2）基于转换的错误驱动词性标注方法

为了克服手工制定规则带来的问题，1995 年 Eric Brill 提出了基于转移的错误驱动的词性标注方法[3]。该方法最初用于英语的词性标注，基本处理步骤是:首先为每个句子赋以初始词性序列，然后通过将这些句子与正确词性标注的句子相比较，自动学习一些结构转换规则，最后将这些规则作用于新的被赋以同样初始词性序列的句子上，就可以得到正确的词性标注。重复以上的过程直到不再获取新的转换规则，这样就可以构建一个词性标注规则集[4]。该方法的优势在于能有效地利用语言的词和语法的规则和一定的上下文信息。实验结果显示，此方法可以用较小的训练集达到较高的分析准确度。

3）基于统计的词性标注方法

基于统计的方法是目前应用最广泛的词性标注方法。基于统计的词性标注方法将词性标注看作是一个序列标注问题，为每一个词语赋予一个正确的候选词性。基于统计的词性标注有基于隐马尔科夫模型的词性标注方法，基于最大熵的词性标注方法，基于支持向量机的词性标注方法，基于条件随机场的词性标注方法等。基于统计的维吾尔语词性标注方面文献[5]提出基于N元模型的维吾尔词性自动标注方法,使用N元语法模型和动态规划的方法进行维吾尔语的的词性标注在测试中把训练语料库和测试语料库的比例设置为19:1，并分析了二元,三元模型对维吾尔语词性标注的效率. 训练和测试语料库的规模差距较大,该测试基本上接近于封闭测试。根据[6]的错误分析，模型性能下降的主要原因是未登录较多。实际上，大多数未登录词在训练库里已有词干形式,只是因为词干附加了词缀发生形态导致模型与训练库进行的匹配失败。文献[7]中提出基于条件随机场的词性标注方法,有效地利用了所有可用的信息并选择不同的模板进行试验。最后选用模板C建立基于条件随机场的维吾尔语词性标记标注模型。文献[7]还提出基于混合策略的维吾尔语词性标注并取得了良好的结果。常见的基于统计的方法还有神经元网络、决策树、线性分离网络标注模型等等。

3 训练算法和特征选择

本文中主要用感知器算法进行训练并根据维吾尔语的特点选择特征。以下详细地介绍感知器算法和选择的特征。

3.1 感知器算法

目前基于统计的方法是词性标注，文本识别等方面的主流方法。在基于统计的方法中，

问题描述为统一的序列标注问题，即给定一个观测序列X = $(x1, x2, \cdots, xn)$，需要求解最优

的标记序列Y = (y1, y2, ⋯, yn)。 其中一类方法从概率的角度来估计 X 和 Y 的概率分布，

这类方法常用的统计模型有最熵模型(Maximum Entropy Model，ME)[8]，隐马尔科夫模型（Hidden Markov Model，HMM）[9]、以及条件随机场模型（Condition Random Fields

Model，CRF）[10]等。在另一类序列标注算法中，定义观测序列X = (x1, x2, ⋯, xn)对应状

态序列Y = (y, y2, ⋯yn)的分数（score）为公式（1）所示：

$$score(X,Y) = \quad (1)$$

其中(X,Y)为特征函数,为第 k 个特征对应的权重。
当特征函数取特定值时，则该模板被实例化，得到具体的特征。。特征值一般可以定义为下面的一个二值函数形式：

$$f(x,y)=\begin{cases}1 & \text{如果x和y满足一定的条件}\\ 0 & \text{否则}\end{cases}$$

给定观测序列X = (x1, x2, ⋯, xn)，最好的状态序列 Y 为 score 最大的状态序列，即如公

式（2）所示：

$$Y = argmax_{Y'}\ score(X,Y') \qquad (2)$$

当通过训练得到每个特征对应的权重后，我们可以使用动态规划算法快速的得到 score 最大的状态序列。

在线算法[11]是一种常用的训练算法，在在线算法中，每次仅仅使用一个实例对参数进行更新，而不像梯度下降之类的批处理训练算法，每次更新参数都需要用到所有的训练语料，导致对资源的巨大消耗。感知器算法[12]是一种典型的在线算法。感知器算法每次使用一个训练实例对模型参数进行更新，在更新参数时每次将需要更新的参数重加 1 或者减 1。感知器算法的代码如下所示：

Input: Training examples ()
Initialiazation:←0
Define: F(x)= (y).
Algorithm: **For t =1 .. T,i=1 ..n**
F()

ifthen←+← (,)

Output: **Parameters**

为了防止模型对数据的过拟合，常对参数进行平均化操作，即Average Perceptron 算法。把这个算法也可以如图1这样表达:

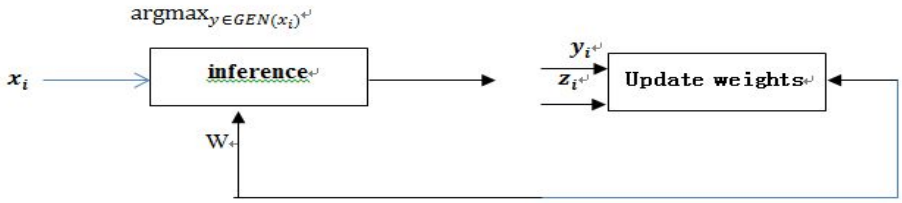

图1 算法的表达形式

该算法使用一组训练的例子估计参数向量 . 对于每一个例子,它发现在得分最高的候选人,候选人使用当前参数值。如果得分最高的候选人不是正确的,它更新参数向量的不同的特征向量. 这种方式的参数更新增加了正确的候选人的参数值并减少了竞争者的参数值。该算法具有收敛性,经过有限次迭代(算法中的t表示迭代数)后,可收敛到正确的权值(参数向量)。算法中的是训练的句子，它也可以描述为=。是标注序列，它抽象为=,其中n是训练例子的个数，i是一个句子的长度。利用函数GEN（X）列举输入X的结果候选人。用映像每个训练例子(x, y) X × Y的特征向量_ (x, y). 对于一个输入字符序列x,我们的目标是找到一个输出F(x), F(x)满足如公式(3):

F(x) = argmax_ (x, y) .(3)

这里的 (x, y). 是(x, y) 和参数向量 的内积。

用简单的例子描述感知器算法在维吾尔语词性标注的过程:

gold-standard: NB MI AO NB VN (x,y)

Yiraqtin bir qara at këliwatatti

Current output : NB MI AO VN VN (X,Z)

Yiraktin bir qara at këliwatatti

假设有以下的特征:

词性 ti-1ti;词/词性组合Wi/ti 根据上面讲的算法，对参数进行更新。

weights ++: (AO, NB)(NB→at)

weights --: (AO, VN)(VN→at)

通过感知器训练算法得到每个特征对应的权重后,我们可以使用动态规划算法快速的得到最优的状态序列。

3.2 特征选择

特征选择是指针对特定的任务，为模型选取特征集合。词性的正确判断依赖于可靠的特征信息。维吾尔语词性自动标注模型的关键是利用对词性歧义消除的特征构建特征模块,尽量减少冲突的特征。根据维吾尔语的语言知识，维吾尔语词的结构，形态等特征信息与词性的关系并维吾尔语的语法特点，本文中主要使用以下基本特征如表1所示：

表1 使用的基本特征表

	特征属性	意义		特征属性	意义

1	w	当前词	7	W1 W2	当前词的后两个词
2	w_1	当前词的前一个词	8	t	当前词的词性
3	w_2	当前词前第二个词	9	t_1	当前词的前第一个词的词词性
4	W1	当前词的后一词	10	t_2	当前词的前第二个个词的词性
5	W2	当前词的后第二个词	11	t_1 t_2	当前词的后两个词
6	w_1 w_2	当前词的前两个词的词性	12	词和词性组合的	

4 解码算法

本文中使用Viterbi算法快速的得到最优的状态序列。Viterbi算法是基于动态规划（Dynamic Programming）的思想，找“正确”的状态序列--词性。具体的就是先解决最基本的子问题，然后再寻找整个问题即最优解。对已知词序列 …，词性标记序列 …… ，寻找该词序列上可能性最优的词性序列 …

Viterbi 算法有三个步骤：（1）初始化；（2）推导；（3）终止和读取路径（最优解）。下面给出标准的viterbi算法：

定义一个局部概率(i)，它是表示的是时刻 t 到达状态Ci

的所有序列概率中最大的概率。再定义一个反向指针(i)，它用来表示的是时刻 t 到达最佳状态的路径。

（1）初始化： t=1，表示所处状的初始概率

（i）=() 1 i N

(i)=0 1i N

(2)推导阶段

递归计算通向词的词性标记的最佳路径

（i）=max[(i)] () 2 m M 1 i j N

(i)=argmax[] 2 m M 1 i j N

(3)终止和读取路径（最优解）

终止，即到达最后一个词时的最佳词性标注

P==argmax

从最后一个词开始，回退求取每个词的最佳状态序列：

=() m=M-1 M-2, ……1

这样可以得到最优词性标序列 …

5 实验与分析

维吾尔语中有名词，形容词，数词，代词，副词，量词，连词，语气词，叹词，后置词，动词等12个词类。新疆大学多语种信息技术实验室自然语言处理组对维吾尔语规则进行深入研究，结合实际文本制定了现代维吾尔语词性标注集（共计137，,1个一级标注，71个二级标注，51个三级标注），该标注集主要用于新疆大学多语种信息技术实验室将要研究的维吾尔语词法分析器，句法分析器，机器翻译等领域。本文章用二级标注,实验主要使用新疆大学自然语言处理实验室构建的维吾尔语语料库，此语料库已进行人工标注，它作为统计数据

的来源。本实验中使用90%的语料库训练模型，10%的语料库用于测试.为了更好地评价维吾尔文词性自动标注的结构，采用计算正确率。表达式如下：
词性自动标注正确率=（标注结果正确词数/语料的总词数）100%
做其它方法（n元，crf，基于混合策略方法等）的学者们用各种训练测试比例做试验。本实验90%训练，10%测试的比列做的。所以为了与其他方法比较，本文中主要选择了90%训练，10%测试的实验结果做比较。
实验结果如表2所示：

表2 维吾尔语词性自动标注算法比较结果

标注方法	语料形式	词性标注正确率
基于 N 元的维吾尔语词性标注	封闭式测试	97.1%
	开放测试	90.2%
基于 crf 的维吾尔语词性标注	封闭式测试	92.26%
	开放测试	91.32%
基于感知器算法的维吾尔语词性标注	封闭式测试	95.4%
	开放测试	94.7%

从上面的例子和表可以看出，感知器算法对维吾尔语词性标注尤其是对兼类现象标注有更大的贡献。

6 总结

本文中使用基于感知器算法的序列标注方法进行词性标注。本方法的优点是可以充分利用多个任意的特征并具有在线算法的优点，每次使用一个训练实例对模型参数进行更新，在更新参数时每次将需要更新的参数重加1或者减1。这优点对维吾尔文词性标注尤其是标注中处理词性歧义（兼类现象）有很大的贡献。目前根据维吾尔语的特点，选择考虑词的上下文信息的特征，从而维吾尔语词性标注方法能够取得很好的标注效果。虽然标注效果好，但还是需要加其它的特征并用别的训练测试比列进行实验。因此今后进一步扩充语料库规模，同时加入更多的特征信息进行研究。

参考文献

[1]吐尔根·依不拉音，阿里甫·库尔班.基于词典的现代维吾尔语词性自动标注系统的研究[A].中文输入技术发展历程及输入方案汇编(论文集)[C]，2006-11.

[2] Màrquez, Lluís, LluisPadro, and Horacio Rodriguez. A machine learning approach to POS tagging. Machine Learning 39.1 (2000): 59-91.

[3] Brill, Eric. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational linguistics21.4 (1995): 543-565.

[4] 周明，吴进，黄昌宁. 用于词性标注的一种快速学习算法对Brill 的基于变换算法的一项改进[J].计算机学报,1998 (4) :357-366

[5] 买合木提·买买提，吐尔根·依布拉音.基于n - gram 的维吾尔语词性标注研究[C].第二届中国少数民族青年自然语言处理学术研讨会.2008 年10 月，中国安徽合肥,pp:185 - 189.

[6]艾斯卡尔·亚克甫,肖克来提,玉素甫·艾白都拉.维吾尔语词频统计子系统的体系结构[J].新疆师范大学学报 (自然科学版)2006 ,25(2),PP:16-20

[7] 艾山·吾买尔 维吾尔语词法句法分析关键技术的研究[D].博士论文，2010 年，新疆大学
[8]Ratnaparkhi A. A maximum entropy model for part-of-speech tagging[C]. Proceedings of the conference on empirical methods in natural language processing. 1996, 1: 133-142.
[9] Dobrushin R L. Central limit theorem for nonstationary Markov chains. I[J]. Theory of Probability & Its Applications, 1956, 1(1): 65-80.
[10]Lafferty, John, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML 18(2001):45-54.
[11]Manshadi V H, Gharan S O, Saberi A. Online stochastic matching: Online actions based on offline statistics [J]. Mathematics of Operations Research, 2012, 37(4): 559-573.
[12]Freund Y, Schapire R E. Large margin classification using the perceptron algorithm [J]. Machine learning, 1999, 37(3): 277-296.

作者简介：

作者一

帕提古力依马木（1989——），女，研究生，主要研究领域为自然语言理。Email:patigul0908@163.com ；

作者二

买合木提买买提（1980年——），男，博士生，主要研究领域为自然语言处理。 Email: 76472080@qq.com;

作者三

卡哈尔江阿比的热西提（1984年——），男，讲师，主要研究领域为自然语言处理.。Email: kaharjan@xju.edu.cn 。

作者四

吐尔根依布拉音(1958一)，男，教授，博士生导师，CCF会员，主要研究领域为自然语言处理，软件工程. Email:turgun@xju.edu.cn（通讯作者）