维吾尔语的词性标注校对初探

牛洪梅　吐尔根.伊不拉音

（新疆大学信息科学与工程学院计算机系）

摘　要：近些年来语料库语言学的发展较为迅速，语料库的建设成为一项重要的工作。在对语料加工的过程中，保证词性标注的一致性也成为建设高质量语料库的重要问题。目前国内外对汉语语料库词性标注结果的校对,还停留在人工校对上,对词性标注结果不一致现象还没有进行系统的研究。对于词性标注方法不足很成熟的维吾尔语语料库来说，词性校对方面的研究工作更是甚少。本文首先概要介绍了一种维吾尔语的标注方法，并受一些文献的启发，根据维吾尔语的特点对其进行词性标注自动校对的研究，并分析其适用于维吾尔语词性校对的可行性。进而提高维语词性标注的正确率。

关键字：语料库；词性标注;自动校对

Basic research on proofreading for part of speech tagging on Uighur corpus

NIU Hongmei, Tuergen Yibulayin

(School of Information Technology and Engineering, Xinjiang University)

Abstract: In recent years, the development of corpus linguistics is more rapidly, building corpus has became an important task. In the processing of corpus ， it has been a primary problem to assure the consistence of part- of- speech tagging in building the high quantity corpus。The current, proofreading of part-of-speech is still stuck in the artificial proofreading and we have not do some detailed research on the inconsistence of part-of-speech tagging 。For the corpus of Uighur whose the part －of-speech tagging is still young ,we do research on it even less。At first the paper outlines a method of Uighur word tagging, and inspired by some of the documents, according to the characteristics of Uighur word we do some research about the auto-proofreading of part-of-speech and analyze the feasibility of its use for Uighur word, thus enhancing the correctness ratio of the part of speech tagging on Uighur word.

Keywords: corpus; part-of-speech tagging; the auto-proofreading

1. 引言

从国内外语料库建设来看：一个计算机语料库的功能主要和下面三种因素密切相关，即语料库的规模、语料的分布和语料的加工深度。因为库容量的大小直接影响到统计结果的可靠性，语料分布的考虑则关系到统计结果的适用范围，而加工深度则决定了该语料库能为自然语言处理提供什么样的知识。

在对语料库的加工处理过程中，词类标注是一项很重要的工作。它的任务就是给语料库中的每个词赋一个合适的词类标记。近年来国内外对词性标注的研究有很多,大多是采用基于规则和基于统计的方法。清华大学和山西大学对基于统计的汉语语料库自动标注方法进行了一些研究和探索，提出了一套用于汉语语料库标注的词类标

作者简介：牛洪梅，新疆大学信息科学与工程学院计算机系 2004 届在读研究生，Email：niuhongmei2004@163.com

记集，标注正确率也达到了95%左右。相比之下，国内对维语语料库的词类标注研究则起步较晚，但是随着维吾尔文信息处理技术的发展和维吾尔语研究的成果为开展维吾尔语语料库加工创造了条件。新疆大学从2002年起开始建设现代维吾尔语语料库系统，计划包括5个部分：语料库、电子语法信息词典、规则库、统计信息库和检索统计软件包。其中语料库部分又分成生语料库（经初步整理的原始语料）和加工语料库（经过标注和校对的语料）。目前已有生语料800万词。新疆师范大学也建立了200万词的维吾尔语语料库。针对维吾尔语言的特性提出的多种标注方法也有很高的正确率，使维吾尔语料深加工工作进一步的发展。

目前国内外对汉语语料库词性标注结果的校对，还停留在人工校对上，对词性标注结果不一致现象还没有进行系统的研究。对于词性标注方法不是很成熟的维吾尔语语料库来说，词性校对方面的研究工作更是甚少。本文受一些文献提出的对汉语语料库进行词性自动校对方法的启发，在考虑维吾尔语和汉语相同点和不同点及维吾尔语的特殊结构的基础上，结合一种维吾尔语词性标注原则，将此方法进行阐述，并分析其适用于维吾尔语词性校对的可行性。

2. 对词性标注方法的分析

维吾尔语是粘着性语言、其形态变化比较丰富，兼类词和同形词较多，而且使用比例较高。标注时，除了正确的统计词干形式的同形词以外，还得解决附加成分与非附加成分同形的还原。这是计算机自动切分和标注的难点。

根据维吾尔语各类词和构词、构形附加成分的结构特征，词干与附加成分的结合规律，归纳了一部规则。这部规则主要由形态变形结构规则、分布特征结构规则等组成。形态变形结构规则是一组基于维吾尔语各类构词，构形附加成分与词干结合规律的规则集。这些规则主要由词干、词干变化、词干附加成分的匹配、附加成分和附加成分的匹配等条件作为识别当前词性的依据。分布特征结构规包括标记搭配规则、词与标点符号搭配规则、词与词的搭配规则、固定搭配规则。主要方法和步骤如下：

（1） 先对带有构形附加成分的一部分词通过筛选法分组进行词类标注。这里所说的筛选法的查找步骤是利用字符匹配法，数据库的附加成分字段和词干字段，从语料中循环比较逐词查找词干和附加成分相对应的词，如果查找成功，自动标注。不成功的则参与下一轮查找。

（2）运用数据库对部分单独出现的非兼类词、带有附加成分的兼类词和同形词进行分组词类标注。自动标注时利用数据库的统配符解决了词干变化、词干末音节弱化等现象。

通过上述的步骤，初步解决了大部分词的标注。但仍剩下部分单独出现的兼类词、同形词。这是自动标注的难点。

（3）对单独出现的兼类词和同形词通过分布特征结构规则来进行标注。分布特征结构规则包括(1)标记搭配规则；(2)词与标点符号搭配规则；(3)词与词的搭配规则；(4)固定搭配规则。

以上概要的介绍了一种维吾尔语标注方法，并在小规模维吾尔语料库进行了试验，完成了词性标注，在此项试验中还获得了另外一个结果，像英语，汉语一样，维吾尔语中也存在着大量的兼类词现象，这给语料的自动词类标注带来了很大的困难。因此如何排除词类歧义，是语料库建设中自动词类标注研究的关键问题。这里受山西大学提出的汉语语料库词性标注自动校对方法的启发，根据汉语和维语的特性求同存异，将其应用于维语的词性标注的自动校对工作。

3. 词性标注自动校对

3.1 向量模型

这里兼类词词性标注是否正确，主要是根据其特定的上下文语境来判断的，所以我们以每个兼类词及其上下文语境所形成词性标记序列作为研究对象。

以下我们建立含有兼类词的词性标记序列表用来描述兼类词的上下文语境：

表 1 词性标记序列表

词	前三词	前两词	前一词	兼类词	后一词	后两词	后三词
词性标注	词性一	词性二	词性三	词性四	词性五	词性六	词性七

这里我们选了离兼类词前后三个词作为词性标记序列，这些词由于离兼类词的距离不同,对兼类词的词性影响程度也不同,称之为位置属性。建立下表描述各个词的位置属性值:

表 2 词性序列位置属性值表

词	前三词	前两词	前一词	兼类词	后一词	后两词	后三词
位置属值	1/22	1/11	2/11	4/11	2/11	1/11	1/22

用向量 X={(1/ 22) ,(1/ 11) ,(2/ 11) ,(4/ 11) ,(2/ 11) ,(1/ 11) ,(1/ 22)}表示。

此外兼类词词性标记序列前、后词的词性和词性标记的位置,对确定兼类词的词性影响程度不同,称之为词性属性。用一个7 行m列的二维矩阵来描述。其中:行表示兼类词词性标记序列前、后三个词及兼类词本身；列表示语料库所采用的词性标记集的标记。

例句：去年的一场疾病剥夺了他生活的自理能力

ئۆتكەن يىلدىكى بىر مەيدان ئېغىر كېسەل ئۇنى ئۆزىنىڭ تۇرمۇشىدىن خەۋەر ئالىدىغان

ئىقتىدارىدىن مەھرۇم قىلدى.

ئۆتكەن/去 يىلدىكى/年 بىر/一 مەيدان/场 ئېغىر/疾 كېسەل/病 ئۇنى/他

تۇرمۇشىدىن/生活 خەۋەر/ئالىدىغان/自理 ئىقتىدارىدىن/能力 مەھرۇم/剥夺

قىلدى/了

词性标记序列是: (v n r n q m n)

设:词性标记集为:{n v a d u p r m q c w I f s t b z e o l j h k g y}“病”的词性属性矩阵:

Y = 0 1 0 0 0 □
1 0 0 0 0 □
0 0 0 0 0 □
1 0 0 0 0 □
0 0 0 0 0 □
0 0 0 0 0 □
1 0 0 0 0 □

(当词性标记序列不完整时,即某个位置没有词性,则该位置所对应的行都标为0)

位置属性向量与词性属性矩阵的乘积定义为词性标记序列向量。即: Vec = X×Y

例句的词性标记序列向量如下:

Vec = (1/ 22 ,1/ 11 ,2/ 11 ,4/ 11 ,2/ 11 ,1/ 11 ,1/ 22) ×Y =(11/22 ,1/11 ,0 ,0 ,0 ,0 ,0 ……)

向量模型的算法既考虑词性标记序列的位置属性,也考虑词性属性。对每个含兼类词的词性标记序列进行向量化表示,然后求出任何两个向量之间的相似度。

$$S_{i,j} = (x_i, y_i)' V^{-1}(x_i, y_i) \qquad (一)$$

（其中：xi和yi 是两个任意的词性标记序列向量）

$$V=\frac{1}{m-1}\sum_{i=1}^{m}(xi-\bar{x})(xi-\bar{x})'$$

3.2 聚类和阈值计算

聚类是把某些对象按其相似性加以分组的一种数据划分。它是通过较为少数的聚类簇去表现大量的数据，每个聚类簇都有自己的特征。这里采用的是基于重心的聚类方法。词性标记序

列向量集合中任一向量x_i与重心向量x_j间的距离d_{ij} 满足 ：$k-1\ \Sigma d_{ij}\leqslant H$

称集合对于H 组成一类。

其中：k 为集合中元素个数，H 为阈值。

H 值是通过训练范例求得的，具体步骤如下：

(一)：随机选取一些含有兼类词的句子，进行人工校正，分别计算含有同一兼类词且其词性相同的所有词性标记序列的向量的平均值VA，这个平均值就是该词的这个词性类的重心向量，根据公式(1) 计算所有该词性类的词性标记序列向量与VA 的马氏距离Distance (Vi ，VA) ，其中Vi为该词性类的词性标记序列向量集合中的任一向量。

(二)：计算(一)中求出的每一词性类的马氏距离的平均值:Average (Distance(Vi,VA))该平均值就是该词性类的阈值H。

3.3. 自动校对模型

我们要解决的就是把词性标注不正确的词性归类，针对这个问题，本文提出了按离词性类别重心最近的原则归类。具体步骤如下：

(一)：把词性标注不正确的词性序列中的兼类词的词性分别换成词表中该词的其他词性，生成新的词性标记向量，求它在这个词性类中与重心向量的距离。用Distance (V(i) ，VA(i))表示，其中i为兼类词的某一词性类。

(二)：找出距离最小的那一类的兼类词词性作为校对词性。即找出Min{ Distance(V(i) ， VA(i))} ，并求出距离最小时对应的I值。

如果某词(有n个词性) 包含第i- 1个词性的词性序列向量，但它不能归为第i- 1 类，即出现了词性标注不正确。按照自动校对模型我们把该词的词性分别换成其他的n－1个词性，分别计算形成的词性序列向量与各自的词性类的重心的向量距离Distance (V(i) ， VA(i)) ，找出Min{ Distance (V(i) ，VA(i))} 为第I 类，我们就认为该词在这个序列中的校对词性应为第I个词性，即把第I 类词性作为该词在这个序列中的校对词性。

4. 小结

本项实验是在维吾尔语小规模语料库的基础上进行初步的探索，尽管这个对维吾尔语的自动校对只是个初步的研究，但是介于它对提高汉语词性标注的准确率起了很大的作用，在一定程度上解决了机器自动标注中容易出现而且较难解决的一致性问题，也大大的提高了人工校对的效率和质量，我们相信结合维吾尔语的特点，继续对自动校对进行深入的实验研究，一定也会大大提高维语词性标注的质量。任何词语所包含的信息可以说是一个多维空间，不同属性之间会产生相互的影响。在进行词性标注正确性检查时，如果只考虑词性标注和词性与词性之间的关系是不全面的，词语的其他属性也可以用来辅助判断词性。我们将继续对维语词性标注及其正确性检查进行研究，希望能借鉴比较成熟的汉语标注工作成果，充分利用数学知识，语言知识，为建设合理，有用的维吾尔语料库提供更好的方法。

参考文献：

[1] 张虎,郑家恒,刘江,语料库词性标注一致性检查方法研究，《中文信息学报》,Vol118 No15.2005

[2] 齐璇,王挺,陈火旺. 义类自动标注方法的研究[J] . 中文信息学报,2001

[3] 汉语语料库词性标注规范，清华大学计算机系智能技术与系统国家重点实验室技术资料，1998.10

[4] 海米提 现代维吾尔语语法（词汇学）北京：民族出版社。1987

[5] 俞士汶主编，现代汉语语料库加工——词语切分与词性标注规范与手册，北京大学计算语言学研究所，1999 年

[6] 俞士汶、朱学锋、段慧明，大规模现代汉语标注语料库的加工规范，《中文信息学报》，2000 年6 期

[7] 米吉提.阿不力米提 在多文种环境下的维吾尔语文字对系统的开发研究。系统工程理论与实践 2003。23（5）