现代维吾尔语词干词类标注标记集验证性研究

艾孜尔古丽[1]，米尔夏提[2]，玉素甫·艾白都拉[1]

（1. 新疆师范大学计算机科学与技术学院，新疆维吾尔自治区 乌鲁木齐 830054；2. 新疆大学信息科学与工程学院，新疆维吾尔自治区 乌鲁木齐 830046）

摘要：本文以维吾尔语小学语文教材语料为验证对象，使用从语法语义相结合角度制定的《现代维吾尔语词干词类标注标记集》，对维吾尔语小学语文教材词干进行了词性标注，验证该标记集规范的可行性、适应性和可靠性。首先介绍小学语文教材电子语料库；其次讨论《信息处理用现代维吾尔语词干词类标注标记集》的基本情况和多策略现代维吾尔语词干标注标系统模型设计与算法；最后分析现代维吾尔语词性标注标记集验证结果，并验证《信息处理用现代维吾尔语词干词类标注标记集》科学性，补充和改正部分词类的语义分类及标注代码，提出了规范的扩充建议。

关键词：现代维吾尔语词干；词类标注；标记集；验证

Confirmatory study of Modern Uyghur Word Stem POS Tag set

Azragul[1], Mirxat[2], Yusup Abaydula[1]

(1 School of Computer Science & Technology Xinjiang Normal University, Urumqi, Xinjiang, 830054, China; 2 School of Information Science and Engineering Xinjiang University, Urumqi, Xinjiang, 830054, China)

Abstract: Based on the Uyghur word stem POS tagging of the Uygur language textbooks which are in use in primary schools as the verification object, we validate the feasibility, adaptability and reliability of Modern Uyghur Word Stem POS Tag set which is made from the perspective of grammatical semantic combination. This article first describes electronic corpus of primary school Uyghur language textbooks; Secondly, it discusses the basic situation of ″The part-of-speech and tagging set standards of modern Uyghur word stem information processing″, and the design and algorithm of multi-strategy modern Uyghur Words Stem tagging system model; Finally, we analyses the experimental results, validate the scientificity of Modern Uyghur Word Stem POS Tag set, supplement and correct part of The semantic classification and code and recommends a substantial expansion of the standard.

Keywords: Modern Uyghur Words Stem; POS tagging; tag set; confirmatory study

1 引言

从维吾尔语信息处理目前的情况看，研制出一套面向信息处理的、具有较强通用性的现代维语词语标注规范是当务之急。无论是维吾尔语的自动切分、识别、校对、词类标注、还是编制通用的现代维语语法信息词典、维汉机器翻译，其最基本的前提就是需要有一个能够全面反映维吾尔语的基本情况，且易于形式化和机器处理的词语词类规范体系。只有词类标注规范，同一个系统内部才能运行，不同系统才能相互兼容。制定面向信息处理的现代维语词词类规范，不仅会对各种应用系统提供一个完整、实用的词类标注规范体系，而且还会对政府有关部门制定现代维吾尔语各类规范与标准提供一个较为可靠的依据。

* **收稿日期：** **定稿日期：**

基金项目：教育部人文社会科学一般项目（项目编号：14YJC740001）；新疆维吾尔自治区自然科学基金（项目编号：2014211A045）；新疆维吾尔自治区哲学社会科学研究规划基金项目（项目编号：14CYY093）；新疆维吾尔自治区高校科研计划青年教师科研启动基金（项目编号：20140706213103147）；国家自然科学基金重点项目（项目编号：61132009）；国家自然科学基金项目（项目编号：61262066）；国家社科基金重点

信息处理用词类标注标记集在语言研究工作中，特别是自然语言信息处理研究早已成为一个最重要的研究课题。汉、英语信息处理研究工作开始的较早，通过长期探讨已经取得了引人瞩目的成果。现代维吾尔语是黏着性语言，语法、语义、语用三位一体构词结构，三个范畴融合在一起的复杂语言。目前的现代维吾尔语词性标注研究主要是按词语的语法功能进行词性标注，没有考虑语义方面的特征。这种词性标注方法无法满足短语分析、句法分析、机器翻译等现代维吾尔语信息处理方面的需求。只有语法语义特征相结合，才能正确的表达词语的有关信息。新疆大学从语法角度来考虑词性标注标记集，有关专家在理论方面提出了《信息处理用现代维吾尔语词干词类标注标记集（以下简称：标记集)》，但是目前为止，还没有用具体材料来验证该标记集的可行性、适应性和可靠性。

由于维吾尔文小学语文教材通用性强，教材中使用的大多数词语在日常生活中经常使用，语言最为规范，代表性强，故本文以小学维吾尔文语文教材（普通版）为研究对象。首先，使用语法语义特点相结合的标记集，对小学维吾尔语词类进行词性标注，建立标准的小学维吾尔文语文教材词性标注标记熟语料库。其次，在构建熟语料库过程中，不断验证、补充和完善现有的标记集，考察分析对教材中各种词类标注标记集的满意度情况。最后，分析本研究中所得到的标记集结果，根据标记集的满意度概况，验证和补充标记集代码，并且对该标记集提出扩充建议。

维吾尔语词的分类问题是在维吾尔语言研究中存在的重要问题之一 。至今还没有得到一致意见，对语言教学和维吾尔语信息化带来了一定的困难。本文以小学维吾尔文语文教材语料库作为对象对该规范进行验证性的研究，进一步证明该规范的可行性、可靠性和适应性。本研究为开展语法分析、语义分析、语音识别、机器翻译、信息过滤、舆情分析、内容理解等诸多领域的研究提供统一标准，为开发智能软件通用性和兼容性提供标准支持服务。

2 关于信息处理用现代维吾尔语词干词类标注标记集

利用语法语义相结合的方法，《信息处理用现代维吾尔语词干词类标注标记规范草案》提出了词语、词干词性标注的标记集规范。该标记集规范了 18 种词类作为语法框架，制定了词干标注代码。在 18 种框架基础上，制定了《信息处理用现代维吾尔语词干类标注标记规范》。根据现代维吾尔语语法语义结合的方法和维吾尔语词干语义特点，将词干分为一级、二级、三级、四级等多个层次代码，每类词干根据具体词类情况而确定，总共 1349 种。其中：

名词（N）代码长度 7 个字符组成，语法语义相结合角度分四个层次，其中首位为字母 N，其余 6 位是数字。例如：

图　　现代维吾尔语名词分类结构图

动词（V）代码长度由 7 个字符组成，根据动词的语法语义特点，动词分为三个层次，关系更加系统化。静态关系（用 V100 符号表示）、心理活动（用 V200 符号表示）、动态行为（用 V300 符号表示）等分三个子类。其中动态行为根据行为语义分为变化（用 V300301 符号表示）、气象（用 V300302 符号表示）、身体活动（用 300V303 符号表示）等 13 个子类。每一类动词具有直接陈式（11 种）、间接式（8 种）、转述式（11 种）、陈述式（10 种）等四种实态，总共 931 种。如图 2-2 所示。

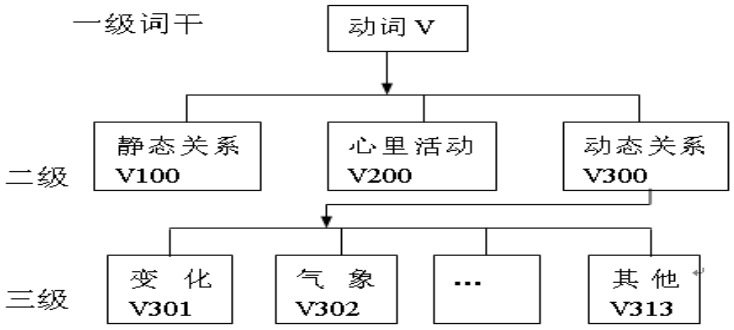

图　　现代维吾尔语动词分类结构图

形容词（A）分为程度和非程度两种，代码长度为 6 个字符，其中首位为字母 A，其余 5 位由数字表示，总共 80 种。程度形容词（A1）共分为 56 类，其中一级共 1 类、二级共 1 类、3 级共 7 类、4 级共 47 类，例如：A10102(浓度：

语法语义标注是目前语法语义分析的一种主要实现方式,它采用“谓语动词—角色”的结构形式,标注句法成分为给定谓语动词的语义角色,每个语义角色被赋予一定的语义含义。例如“[早晨（动作发生的时间）] [我（施事）]向[单位（受事）][已去（核心动词）]。”由于各种机器学习方法都已经比较成熟,仅依靠单纯机器学习算法的改进,在性能上很难提高水平。所以,丰富有效的特征提高了语法语义标注的效率。

3 现代维吾尔语词干词类标注系统

3.1 现代维吾尔语词干词类标注标系统模型设计

本文利用现代维吾尔语词干词类标注标记集作为基础，维吾尔语小学语文教材验证对象，利用现有的“现代维吾尔语多策略词干标注系统”，验证了《信息处理用现代维吾尔语词干词类标注标记》。该系统总体模块图 1 所示。

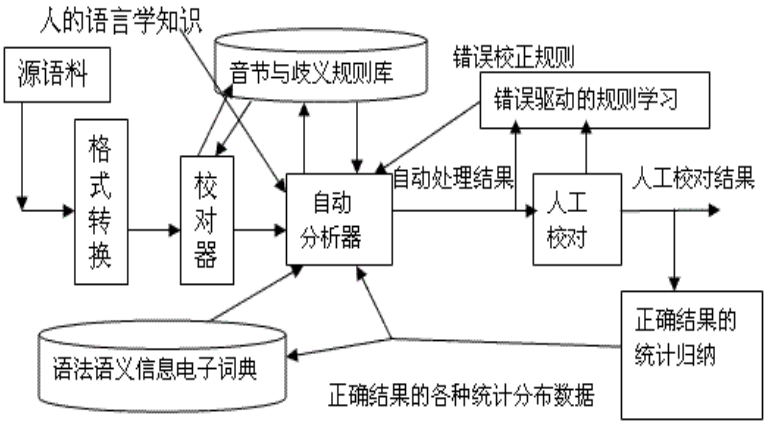

图 1 现代维吾尔语词干标注系统模块图

3.2 现代维吾尔语词干标注标算法

图 1 所示的自动分析器由词干提取、词干词性标注、统计分析等多个模块组成，其中词干标注模块利用隐马尔可夫模型（Hidden Markov Model，HMM）进行词干词性标注。

本研究中词干是词性标注的对象，因此设 N 是词干集(词干-Nomenclature 的首字母)，T 是词干词性标记集库（标记-Tag 的首字母)。S 是待标注词干串（待-Stay 的首字母)，S=S1S2S3 …Si…Sm（Si∈N)。C 是 S 对应的词性标记串，C=C1C2C3 …Ci…Cm（Ci∈T)。那么，词性标记串 C 的出现频率 P（C）可由以下公式来计算：

$$p(c)=p(c1)p(c2|c1)\ p(c3|\ c1\ c2)\cdots p(ci|\ c1\ c2\ \cdots ci-1)\cdots p(cm|c1\ c2\ \cdots cm-1)$$

$$=p(c1)\prod_{i=2}^{m}\ p(ci|\ c1\ c2\ \cdots ci-1)。 \qquad （公式\ 1）$$

对于一元模型，将以上公式可简化为：

$$p(c)=p(c1)\ p(c2|c1)\cdots p(cm|cm-1)=\prod_{i=1}^{m}\ p(ci|\ ci-1) \qquad （公式\ 2）$$

其中找出概率 p（cici-1)的公式是：

$$p(ci|ci-1)=\frac{count(c_ic_{i-1})}{\sum_{c_i} count(c_ic_{i-1})}。 \qquad （公式\ 3）$$

在公式 3 中，count（cici-1）表示词性标记对 cici-1 在训练语料中出现的次数。

以上模型为基础及利用算法设计与实现“现代维吾尔语多策略词干标注系统”。

4 现代维吾尔语词性标注标记集验证性分析

4.1 名词的验证性分析

从验证结果来看，全套教材中的所有 28133 词中，有 12926 种名词服从标记集规范，有

一部分名词没有放在该规范的语义分类框架中。例如：

05	代词	P	9	0	100	820	0	100
06	数词	M	9	0	100	223	0	100
07	量词	Q	2	0	100	407	0	100
08	模拟词	O	4	0	100	210	0	100
09	感叹词	E	4	0	100	125	0	100
10	连词	C	20	0	100	110	0	100
11	后置词	T	3	0	100	124	0	100
12	语气词	Y	8	0	100	135	0	100
13	缩略词	B	4	0	100	145	0	100
14	标点符号	W	65	0	100	65	0	100
15	合计	14	1345	17	99.5	28133	318	99.72

表 1 可见，在现代维吾尔语词干词类标注标记集中对该套教材中标注满足率 99.5%，其中名词、动词、形容词满足率 96.4%、99.14%、97.5%，其他词类标记集满足率 100%；在教材中无法标注的词干数 318 种，标注满足率 99.72 %，其中名词、动词、形容词满足率 98.89%、98.45%、98.8%，其他词类标记集满足率 100%；扩展四级词类 17 种，其中名词、动词、形容词分别 7、8、2 种，其他词类标记集满足率 100%。以上分析可以结果说明：

（1）“信息处理用现代维吾尔语词干词类标注标记集”的顶层设计符合现代维吾尔语特点和计算机信息处理要求。第一层语法角度考虑、共 14 种，第二层开始语法框架基础上语义角度进行第三、第四层。

（2）设计每一层时，考虑了适应性和扩建性，提出比较科学的层次。

（3）基本满足构建通用信息处理用现代维吾尔语词干词类标注熟语料库，提供科学、有效、可行和规范，并作为新疆维吾尔语词干标注标记集标准打下良好的基础，提供科学依据，进一步优化后，可以推荐地方标准，下一步提升国家标准层面提供科学依据。

本结论维吾尔语小学语文教材为对象，进行的实验性验证结果分析基础上提出来的。本课题组下一步初中、高中维吾尔语文教材进行实验性研究，进一步优化和补充“信息处理用现代维吾尔语词干词类标注标记集”，提高标记集的可行性和完整性，提高社会服务能力。

参考文献

[1]玉素甫.艾白都拉,张海军,艾孜尔古丽.信息处理用现代维吾尔语词干词类标记集研究[J].信息技术与标准化，2011.6
[2]玉素甫.艾白都拉,阿布都热依木.沙力.现代维语语料库的词类标注研究[J]民族语言，2005.4
[3]阿尔斯兰.阿不都拉.台赫尔.现代维吾尔语（3 本）[M].新疆人民出版社，2010.12
[4]玉素甫.艾白都拉.信息处理用现代维吾尔语词类标注标记集规范草案 [C].新疆师范大学，2011.6
[5]尼加提.纳吉米，买合木提.买买提，吐尔根.依布拉音.基于 N 元模型的维吾尔语词性标注实验研究[J].计算机工程与应用，2012.4
[6]牛洪梅，加米拉.吾守尔,吐尔根.依布拉音.现代维吾尔语的词性标注校对技术研究[J].伊犁师范学院（自然科学版），2007.3
[7]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，2009.2
[8]吐尔根.依不拉音，阿里甫.库尔班，阿不都热依木.基于词典的现代维吾尔语词性自动标注系统的研究[J].新疆大学,2011.6
[9]陈鹏.基于语料库的维吾尔语词干提取和词性标注[D].新疆大学.2006.2

作者简介：艾孜尔古丽（1987——），女，讲师，主要研究领域为计算语言学、自然语言处理。Email:Azragul2010@126.com；

米尔夏提·力提甫（1959-）男，副教授，自然语言处理、机器翻译。

2629200592@qq.com

通讯作者：玉素甫·艾白都拉（1958——），男，教授，主要研究领域为计算语言学、自然语言处理。Email:ysp2002@126.com。

第一作者：

第二作者：

通讯作者：