文章编号:1001 - 9081(2009)07 - 2006 - 03

信息处理用维语词汇标注标记集的确定

玉素甫·艾白都拉,阿不都热依木·沙力,阿拉帕提古丽
(新疆师范大学 数理信息学院,乌鲁木齐 830054)
(yusup2002@ sohu. com)

摘 要:介绍了研究和制定信息处理用维语标注标记集的研究进展。讨论了研究与制订"维语标注标记集"必要性,说明了只靠语法角度提出的词汇一级"维语标注标记集"的有限性,论述国内外英语、汉语层次分析研究的最新成果,结合维语的本身特点研究维语语义分类体系迫切性,给出了"维语标注标记集"的工作思路、标记集制定的原则和语法语义词汇一级词性标记集的内容,比较了标记集两个版本的特点,展望了其发展。

关键词:维吾尔语;信息处理;标注标记集
中图分类号: TP391 **文献标志码**:A

Information processing with ascertaining Uyghur vocabulary and labeling marks set

YUSUP Abaydul, ABDIRYIM Sali, ARAPATGUL
(*College of Mathematics, Physics and Information Science, Xinjiang Normal University, Urumqi Xinjiang* 830054, *China*)

Abstract: This paper presented the research development of information processing with Uyghur set of tags marking progress. Firstly, this paper discussed that it is necessary to research and develop on "tag-dimensional language marked"; secondly, the survey stated that it is limited for "tag-dimensional language marked" depending on grammar terminology. And then the study introduced the latest results of English, Chinese-level analysis at home and abroad. According to the urgency of the classification system of Uighur semantic, the paper researched on "Uighur tagging tag collection" work ideas, formulation of the principles set tag and grammar semantic vocabulary of first-degree tag sets, and gave the future development prospects.

Key words: Uyghur Language; information processing; labeling marks set

0 引言

对维吾尔语(简称维语)语料库的多级加工处理,主要分为以下几个阶段:词类标注、短语结构标注、语义信息标注等。对于前两个阶段,我们已进行了一些研究和探索,提出了一种词类标注相融合的维语语料库多级加工方法,取得了较好的处理效果。目前的研究重点开始转向维语短语的自动划分和标注方法上,而这项工作的一个重要基础是确定合适的维语词汇一级标记集和短语标记集。

在维语中,词汇、短语具有特别重要的地位。它们的内部结构比较稳定,往往作为一个整体和句子中的其他成分发生作用,并且它的构造原则和句子的构造原则也基本一致。经过计算语言学角度多年研究发现,如果我们把每一个词汇类型确认,各类词组的结构和功能都足够详细地描述清楚,那么句子的结构实际上也就描述清楚了,因为句子不过是独立的词组而已。从这个意义上看,维语词汇、短语标注的研究具有很高的理论和实用价值。它的顺利完成,将为进一步进行词语依存关系的确定、以及维语理解和汉维机器翻译的研究打下良好的基础。

1 研究规范标记集的必要性

对现代维语语料标注处理的目的是给语料库中原始语料的每一个单词或词组、句子添加一定的特征。经过如此标注的语料库可以简称为"标注语料库"。这项工程对加工质量要求甚严。要建成高质量大规模的标注语料库,需要具备多方面的基础和条件,其中必须先行的一项工作就是制订完备的可供人机两用的标注规范。为了得到高精度的标注结果,必须制定明确可操作的标注规范。

为了得到高精度的加工处理结果,必须制订明确的可操作的加工处理规范标记集,同时实现人与计算机的合理分工与妥善配合。《现代维语语法》中的 12 个词类标记,不能满足计算机信息处理的需要,所以,从语法角度对"信息处理用维语词汇一级信息处理规范标记集"1.0 版本(简称:1.0 版本)实现了一类、二类、三类等的标记,例如:名词 1 个一类、5 个二类、16 个三类等推类(1/5/16/15/26/17/6),形容词(1/2/4),代词(1/7),数词(1/8),量词(1/2),副词(1/4),摹拟词(1/3),叹词(1/3),连词(1/2),后量词(1/2),语气词(1),谚语(1),成语(1),习用词(1),动词(1/4/74),标点符号(1/50),句子成分(1/6),形容词的级(1/5),动词语态(1/5),人称(1/4),数(1/3),动词是否式(1/3),词语来源(1/10),名词的格范畴(1/10)共划分了 24 个一类,138 个二类,94 个三类等。这在现代维语信息处理领域中,是首次制定的小标记集。这些人机两用的加工标准词一级的语料库加工处理起了很好的先导性作用。但是只靠词一级语料的加工处理,满足

收稿日期:2008 - 11 - 07。 **基金项目**:国家自然科学基金资助项目(60463005);教育部少数民族语言文字规范标准建设与信息化项目(MZ115-80);新疆维吾尔自治科技厅高新计划项目(200723114);新疆师范大学鼓励启动基金资助项目。

作者简介:玉素甫·艾白都拉(1958 -),男(维吾尔族),新疆乌鲁木齐人,教授,主要研究方向:自然语言处理、模式识别; 阿不都热依木·沙力(1963 -),男(维吾尔族),新疆乌鲁木齐人,副教授,主要研究方向:计算语言学; 阿拉帕提古丽(1980 -),女(维吾尔族),新疆乌鲁木齐人,讲师,主要研究方向:信息处理。

不了真正的自然语言处理需求。从第一语言习得、第二语言习得及自然语言处理三方面来考虑,词语搭配或词汇化短语是人类语言知识库中的一个有机部分,它是词典和语法规则不能代替的。所以,对现代维语短语语料标注加工处理工作非常重要。在短语一级维语语料加工处理中发现,从语法角度出发制定的 1.0 版本的标注标记集还不能满足维语短语语料加工处理的需要。

根据维语本身的特点和维语短语语料加工处理的需要,必须研究并制定语法语义相结合的"信息处理用维语词汇一级信息处理规范标记集"2.0 版本(简称:2.0 版本)。在研究 2.0版本之前必须研究对维语的语义分类体系。

2　维语语义分类体系

国内外对维语语义分类体系的研究没有成熟的成果,但在汉语方面已经有了不少成果。但由于各家分类体系的目的及应用范围不同,对同一事物可能有不同的定义与归类。如"动物"在一个语义体系中分为:"兽类,鸟类,鱼类,虫类,爬行类",而在另一个体系中分为"脊椎动物,腔肠动物,软体动物"。但这些分类体系都是基于自然科学或常识而独立于语法的。在实际语言分析中,如何将这些语义知识与语法知识有机地结合起来是一件很困难的事情。本文从计算语言学的角度,以维语短语语料标注处理为目的进行语义分类研究,在分类的深度与广度上下了功夫。策略上应用语义知识应着重解决那些仅靠语法规则难以解决的问题。借鉴英语、汉语语义分类经验与思想,根据维语语义特点与分类目的,在现有维语词汇语法分类基础上,利用语法语义相结合方法进行,并且只对名词、动词、形容词等实词进行语义分类描述,而那些带有明显标志,通常用句法形式就可以表示的语义关系,如各类虚词,则不作为重点语义分类研究的对象。经过 3 年应用检验与研究发现,这种分类法是很有前途和实用价值的。

1)名词上下位关系更加系统化:首先,将具体事物、抽象事物与过程、时间、空间并列为 5 大类;然后再逐层细分:具体事物分为生物和非生物 2 类,生物里再把人与动物、植物、微生物相并列,非生物中则进一步区分开人工物、自然物、排泄物和外形。然后,根据机器词典中的内容,补充了一些低层小类。

2)把机器词典中的动词分类借鉴过来,但根据维语的实际作了相应改造。

3)形容词的分类更加细化,现在的 5 大类 29 小类,与名词的分类互相照应,从而可以更细致地刻画形名搭配关系。维语在每类词的基本分类下,只有大大小小,彼此具有上下位关系的同义词集合(synset),而不再设立低层的语义类名称。因此,我们对维语语义类主要限于名词、动词的基本语义类,然后,根据维语句子分析的需要适当地补充一些 synset 作为小类,如"Artifact(人工物)"下面的"创作物,药物,设施,工具"等,而不可能也没有必要把该语义类的直接下位概念全部都照搬过来。总的来说,调整后的新语义分类更趋合理,名词、动词、形容词的分类相对较细,数词、副词的分类较粗,只要能揭示出与名词性成分,动词性组合成分的不同组合类型即可。目前我们已实际完成了 6.6 万词语的语义类划分与属性描述。

根据以上原理,数词、副词等其他词类也进行了语义分类。以上是描述维语实词语义知识。除此之外,描述维语实词语义知识时特别强调在短语一级描述语言成分的语义信息。理由是在从词向短语的组合过程中,一个成分的语义搭配能力可能发生变化,描述这种变化,有助于计算机得到正确的分析结果。

3　维语词汇一级标注标记集的确定

3.1　现阶段的工作及技术思路

较为复杂的自然语言处理系统离不开词语分类、词性标注和语义解释,而这些环节必须有个统一的规范,只有规范,在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维语词语规范,不仅会为各种应用系统提供一个完整、实用的规范体系,而且还会为政府有关部门制定维语各类规范与标准提供一个较为可靠的依据。

所以我们现在正在做一些基础性工作,旨在研制面向信息处理的维语词语规范体系。其具体内容包括:

1)运用计算语言学、工程语言学的理论与语法语义相结合的方法进行面向信息处理的维语词语分类,制定维语词语标记集。这个分类与传统语法中的词类划分有明显差别:

①覆盖面广,不仅包括词,还包括标点符号、数字、字母等各类字符以及成语、惯用语等语言单位;②通用性强,因为它面向不同应用系统,所以并不基于特定的语法规则体系;③利用语法语义结合方法面向信息处理的维语进行语法语义分类体系;④操作性强,它不仅仅是理论上的归纳,还要落实到每一个词语、字符上,所以,有很强的可操作性。

2)通过对维语词语进行统计、分析,确定各类词的语法、语义属性,设计出易于机器处理的各种属性字段及取值规范。

引用文献[13]中的一段话:加强语言文字规范标准建设,特别应该注意下面几个问题。

①为保证规范标准的科学性、可行性,规范标准制定一定要以科学研究为基础。

②规范标准的制定要有系统性。相关规范标准的制定要保持连续性,同一规范标准要根据实际变化,及时扩充、升级。研究比较成熟的规范标准先制定,社会急需的先制定,基础性规范标准先制定。

③注意制定与维护相结合,对已发布的规范标准要根据应用需要,及时进行整合、修订。

④规范标准发布前要反复征求各方面意见,经由权威审定机构审定,向社会发布的每项规范标准都要发挥应有的社会效益。我们所开展的工作也就是遵循这一原则来进行的。

3.2　现代维语词语词性标记集制定的原则

制订"面向信息处理的现代维语词语词性标记集"的内容原则如下:

1)词语标注规范尽可能同已有的中国新疆维吾尔自治区《现代维语正字正音细解词典》保持一致。由于现代维语词语切分的目的是为词性标注服务的,所以建立了一部《现代维语机器电子词典》。此词典包括词根词典和词尾词典,可作为标注的依据。

2)词性标注使用小标记集。除了使用《现代维语语法》中的 12 个词类标记外,还在大部分词类中分了其子类。原来的语法角度考虑词类分类方法改为语法语义相结合,增加颗粒度,提高机器描述能力和处理能力。

3)与已有资源的配合。尽管使用的是小标记集,但标注语料库同《现代维语机器电子词典》是紧密联系的。在自然语言处理应用系统中,以文本中的词语及词性为入口,可以快速、准确地检索到词典中词语的丰富的语法属性信息。这就是说,经过切分、标注的语料库同《现代维语机器电子词典》相结合,可以形成一个超文本的语言知识库。

4)对专有名词(人名、地名、团体机构名等)进行了标注。并用方括号标出短语型专有名称。

5)规范既适应语言信息处理与语料库语言学研究的需要,又能为传统的语言学研究和语言研究提供充足的素材;既

适合计算机自动处理,又便于人工校对。

3.3 现代维语语法语义词性标记集

根据以上原则和以前研究成果 Version 1.0 基础上,利用语法语义相结合的方法研究制定"面向信息处理的现代维语语法语义词汇一级词性标记集"Version 2.0。其内容如下:

1)名词(1个一类,6个二类,25个三类,169个四类),编号:N。例如:

دېھقان(农民),代码编号:N100107,其中 N 表示名词、100 表示人有关的名词、107 表示行业。

بۆلۈم باشلىقى(科长),代码编号:N100205,其中 N 表示名词、100 表示人有关的名词、205 表示管理者。

كالا(牛),代码编号:N300101,其中 N 表示名词、100 表示动物有关的名词、101 表示食物动物。

2)动词(1个一类,12个二类,223个三类),编号: V。

كۆپەيدى(增加),V301111,其中 V 表示动词、301 表示事物的变化、111 表示动词直接一般过去式。

ياخشى كۆرگەدەك(喜欢),V200144,其中 V 表示动词、200 表示心理活动、144 表示动词间接完存现在式。

يۈگۈر(跑),V306101,其中 V 表示动词、306 表示位移、101 表示祈使式。

特殊说明:在动词编号中 V 表示动词,前三位(100)表示语义,后三位(112)表示语法。篇幅原因其他类不再一一介绍。

3.4 两个版本的比较

1)1.0 版本从维语语法角度出发对维语文本词汇一级进行标注处理。但是短语层面处理维语文本时,它完全不能满足需要。

2)2.0 版本的研究目标是从语法语义相结合的角度出发对维语文本词汇一级进行更详细标注处理。不仅满足了维语文本词汇一级进行标注处理,而且在短语层面处理维语文本时,能基本满足需要。

3)在 1.0 版本中,名词分 1 个一类,10 个二类,29 个三类、形容词分 1 个一类,2 个二类,4 个三类、动词分 1 个一类,4 个二类,74 个三类等,在 Version 2.0 中名词分 1 个一类,6 个二类,25 个三类,169 个四类,动词分类分 1 个一类,12 个二类,223 个三类。2.0 版本分类颗粒度小于 1.0 版本,对定义短语结构、短语边界判定和处理短语结构歧义更有帮助。

4)在研究 2.0 版本中考虑词汇、短语、句子和段落标注处理问题。

5)在研究 2.0 版本中发现维语除具有语义粘着性特点外还有语境粘着性特点。在维语中的格附加成分就表示语境等。例如:

موساجان قومۇلدىن ئۈرۈمچىگە قايتىپ كەلدى.(木沙江从哈密来到乌鲁木齐)。

其中دىن附加成分表示木沙江原来哈密,گە附加成分表示木沙江现在到了乌鲁木齐。这两个附加成分只表示木沙江存在的环境,所以,附加成分提供部分语境信息。

4 结语

本文提出了一个维吾尔信息处理用的基本规范标记集,是以400万词的维语语料作为研究对象基础上提出来的,并根据维语词根、词尾、短语(句法功能和结构)组成方面对不同的词语,特别是短语的性质进行了分析和探讨,以期为维语词语、短语划分和标注的自动处理和人工校对提供一个系统的处理标准。但这还是一个草案,还需各方面专家提出意见,更需具体系统运行、调试,在对大规模维语语料的标注实践中不断加以补充和完善。

参考文献:

[1] 力提甫·托乎提.维语及其他阿尔泰语言的生成句法研究[M].北京:民族出版社,2001.

[2] 力提甫·托乎提.从短语结构到最简方案:阿尔泰语言的句法结构[M].北京:民族出版社,2004.

[3] 玉素甫·艾白都拉,吾守尔.斯拉木.维语中心语驱动文法句法分析器中的上下文相关处理[J].计算机应用与软件,1999,16(6):22-25.

[4] 玉素甫·艾白都拉.维语句法分析器中的词义排歧问题的研究[J].计算机应用与软件,2002,19(4):59-62.

[5] YUSUP A, LUA K T. The development of Tagged Uyghur Corpus [EB/OL].[2008-10-28]. http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/12275/1/PACLIC17-228-234.pdf.

[6] 玉素甫.现代维语语料库的管理[C]// 中国人工智能学会第10届全国学术年会论文集.北京:北京邮电大学出版社,2005:1007—1010.

[9] 哈米提·铁木尔.现代维吾尔语语法[M].北京:民族出版社,1987.

[10] 俞士汶,朱学锋,王惠,等.现代汉语语法信息词典详解[M].北京:清华大学出版社,1998.

[11] 程适良.现代维吾尔语语法[M].乌鲁木齐:新疆人民出版社,1996.

[12] 孙茂松.信息处理用现代汉语分词词表的设计原则[C]// 计算语言学文集.北京:清华大学出版社,1999:193-198.

[13] 袁贵仁.以规范标准建设为核心,开创语言文字应用研究新局面[J].语言文字应用,2001(3):3-8.

(上接第2005页)

6 结语

本文建立了较为完善的现代藏文音节判断规则,并且对符合规则的编码序列能进一步确定各个编码间的位置关系,从而得到有利于排序的序列(空缺的位置用空格代替):"基本辅音 前加辅音 上加辅音 下加辅音 元音 后加辅音 又后加辅音";而将不满足判断规则的字母组合一律看作梵音藏文音节。从而完成了现代藏文音节和梵音藏文音节的区分。

参考文献:

[1] 孙怡荪.藏汉大词典[M].北京:民族出版社,1993.

[2] 安世兴:梵藏汉对照词典[M].北京:民族出版社,1991.

[3] 江荻,周季文.论藏文的序性及排序方法[J].中文信息学报,2000,14(1):56-64.

[4] 江荻,康才峻.书面藏语排序的数学模型及算法[J].计算机学报,2004,27(4):524-529.

[5] JIANG DI. The current status of sorting order of Tibetan dictionaries and standardization[C]// Proceedings of The 20 th Pacific Asia Conference on Language, Information and Computation. Beijing: Tsinghua University Press, 2006: 231-236.

[6] 林河水,程伟,曹晖,等.一种符合ISO14651语义的藏文排序实现方法[J].中文信息学报,2004,18(05):36-41.

[7] 洪锦玲,贾彦民,朱峰,等.藏文基本字符集的支持在OpenOffice.org中的实现方法[J].信息技术与标准化,2007(8):50-54.

[8] 格桑居冕,格桑央京.实用藏文文法教程[M].成都:四川民族出版社,2004.

[9] 新编藏文字典编写组.新编藏文字典[M].西宁:青海民族出版社,1989.