中 文 信 息 学 报
第 20卷 第 3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 3

文章编号: 1003 - 0077(2006)03 - 0043 - 06

面向自然语言信息处理的维吾尔语名词形态分析研究*

阿依克孜 ·卡德尔[1],开沙尔 ·卡德尔[2],吐尔根 ·依布拉音[2]

(1. 新疆大学 人文学院,新疆 乌鲁木齐 830046; 2. 新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046)

摘要:名词是人类语言中的基本词类之一。维吾尔语是一种形态变化很复杂的语言,其中名词是一种形态变化复杂的词类。因此名词的形态分析研究无论在语法研究还是在语言信息处理中都非常重要。本文对维吾尔语名词的形态变化(名词的数、人称、格等语法范畴)进行了形式化的描述和分析。指出了维吾尔语名词的基本形态参数,总结出参数的组配规律并统计了其类型,探索了维吾尔语名词的削尾方法。这些工作将为维吾尔语名词形态处理提供有效的方法和新的思路。

关键词:人工智能;自然语言处理;维吾尔语信息处理;名词;形态

中图分类号: TP391 **文献标识码**: A

Morphological Analysis of Uighur Noun for Natural Language Information Processing

AYKIZ · KADIR[1], KAYSAR · KADIR[2], TURGUN · IBRAHIM[2]

(1. College of Liberal Arts, Xinjiang University, Urumqi, Xinjiang 830046, China;
2. Information Science & Engineering College, Xinjiang University, Urumqi, Xinjiang 830046, China)

Abstract: Noun is one of the basic word classes in human languages. As Uighur language is a highly inflectional language, morphological analysis of Uighur noun, a highly inflectional word class, is very important for study of Uighur grammar and Uighur language information processing. This paper concerns the formalized morphological description and analysis of Uighur noun (number, person and case etc). It points out the essential morphological parameters of Uighur noun, sums up the rule of its composition, statistical type and gives a method for paring suffixes. This approach provides an effective way for noun analysis in Uighur language information processing.

Key words: artificial intelligence; natural language processing; Uighur language information processing; noun; morphology.

1 引言

维吾尔语信息处理的研究工作始于 80年代初。最初涉及的主要是文字处理方面的工作。新疆大学一直致力于维、汉文多语种操作系统的开发。从 1984年到 1994年研发了维、哈、柯、汉、英 DOS2.0到 DOS6.2等一系列维、哈、柯、汉、英 DOS多语种处理平台。从 1994年到 2000年前后研究开发成功 Window3.x多文种和纯维文两种处理平台、Windows95多文种平台"民文视窗"、Windows9x/2000系列多文种处理平台等[1]。从 2003年起 863课题组在开发民文

* **收稿日期**: 2005 - 08 - 21 **定稿日期**: 2005 - 11 - 24
基金项目:国家自然科学基金资助项目 (60263004)
作者简介:阿依克孜 ·卡德尔 (1974—),女,语言学博士生,主要研究方向为对比语言学、语言信息处理.

Linux系统和民文全文检索系统。维吾尔文的处理已达到可用程度。除此之外新疆的许多单位、公司及电脑爱好者都在开发其他一些维、汉、英 Windows多文种平台及应用软件。

随着信息技术的发展和互联网络的普及,近年来现代维吾尔语料库建设、机器翻译,语音识别、自动校对、智能检索等等工作也得到了重视。有关研究部门已开始进行大型语料库建设及研制机器翻译系统[2]工作等。

语言信息处理的不断发展要求借助更多的语言学知识,对于维吾尔语来说更加如此。因为维吾尔语是一种形态发达的粘着语,各词类一般都具有一定的语法范畴,主要表现在形态变化上。也就是说,每一个词在生语料中以不同的形态出现,这主要表现为在词干上按照一定的循序缀接不同的词尾。因此在维吾尔语信息处理中,形态分析问题,尤其是名词、动词等基本实词的形态分析是一个很重要的问题。基于这一问题,我们在此首先分析维吾尔语名词语法范畴(这里指的是形态范畴)及其形态变化规律。

2 维吾尔语名词形态分析的必要性

在所有的语言中,名词是最基本的实词之一。再说维吾尔语中名词的形态变化非常丰富,有必要进行深入研究。

2.1 从第一语言习得的角度看

第一语言习得的研究表明,在儿童早期的语言习得过程中,首先习得的基本单位是一些未经形态变化的单词,然后慢慢再习得语言能力中的语法规则。因此一个两岁大的小孩会说:“apakɛldi(妈妈来了。)”,“mɛn øj ketimɛn(我要回家。)”这样的形态变化不完整的句子。实际上,上述例句中的“apa(妈妈)”应该发生形态变化,词干上附加人称词尾“m”,表示说话者的妈妈。“øj(家)”的词干上应该附加向格词尾“gɛ”,表示行为的趋向。这说明了人的语言知识中,尤其是形态丰富的维吾尔语语言知识中,名词的形态问题也占据相当重要的地位。

2.2 从第二语言学习的角度看

名词的形态对于成年的操维吾尔语者的问题主要限于书面语范围,但对于第二语言学习者来说,问题就会普遍得多、严重得多。对操母语者来说十分简单的形态变化,经常使第二语言学习者,尤其是那些母语是缺乏形态变化的分析性语言的第二语言学习者错误百出。第二语言学习者往往根据母语的语法规则,通过词对词翻译的方式造出目标语中的组合形式,而这些组合形式又往往在目标语中是不符合语法规则的。比如:

汉语的:“明天乌鲁木齐晴,有时多云”,在维吾尔语中应该用“ɛtɛ yrymtʃidɛ hawa otʃuq, bɛzidɛ bulutluq bolidu”表达。其中“yrymtʃidɛ”是名词“yrymtʃi(乌鲁木齐)”的时位格形式,表示处所,在这里就不能用名词原形“yrymtʃi”。

可见对于第二语言学习者来说,在掌握形态丰富的维吾尔语时,必须重视名词的形态。

2.3 从自然语言处理的角度看

2.3.1 自然语言理解

自然语言理解是用结构语法和语义分析对句子逐词加以解析[3]。它是知识信息中的核心问题。由于维吾尔语是典型的粘着语,名词的形态由不同的词尾的不同的组配构成,所以名词的形态分析是维吾尔语言理解中的难点之一。如果总结出维吾尔语名词的形态规则,并统计出名词词尾的组配类型,就有助于理解同一名词的不同形式并把它们还原为原形。比如,可以弄清“kitabtin(从书上), kitabni(把书), kitablar(一些书)”都是“kitab(书)”一词的不同形态,只不过在附加意义上有所区别。

2.3.2 语言生成

自然语言生成是将“意义”(深层结构)通过计算机主动生成所需要的某种特定语言[3]。计算机自然语言生成与计算机自然语言理解一样,都需要语言学知识[4]。因此总结出维吾尔语名词的形态规则,并统计出名词词尾的组配类型之后,按照词尾的组配及其缀接规则,可以生成同一名词的不同形式,使语句准确地输出。比如,按照规则附加词尾,“øj(房子、家)”一词一般可以生成 120种可使用的形式(见以下维吾尔语名词形态参数组配表)。

2.3.3 机器翻译

名词的形态分析,对机器翻译是非常重要的。因为机译系统的核心是分析、转换和生成。只有仔细分析维吾尔语各词类的形态规则,才能解决目标语和原语言之间的形态转换问题,其中名词是首先要处理好的词类之一。因此,名词的形态分析对汉维机器翻译就显得非常重要。

3 维吾尔语名词的语法范畴

维吾尔语的名词有数、人称、格等语法范畴。由于这里的研究是面向计算机的,所以形式化的描述当然是必要的[5]。下面我们将对维吾尔语名词的数、人称、格等语法范畴尽可能详细地进行形式化的描述,为此设置了以下形态参数:

N (noun) =名词; n (number) = (复)数词尾; p (person) =人称词尾; c (case) =格词尾

3.1 数范畴

维吾尔语名词的数范畴是通过名词的形态变化表示事物的数量的语法范畴[6],表达的概念包括单数和复数两类。名词的原形就是名词的单数形式,其复数形式是在名词词干上附加构形词尾,词尾的形态参数如下:

n =复数词尾　　lar/lɛr

比如:

	单数	复数
①	kitab(书)	kitab + lar → kitablar(表示一本以上的书)
②	adɛm(人)	adɛm + lɛr → adɛmlɛr(表示一个以上的人)

名词的复数词尾是第一个附加在词干上的词尾。

3.2 人称范畴

维吾尔语名词的人称范畴是通过名词的形态变化表示名词所指事物的领属关系的语法范畴。表达的概念包括第一人称单数、复数,第二人称单数、复数和第三人称(形式上不分单复数)[5];词尾的基本形态参数如下:

p_1 =第一人称单数　　m/im/um/ym (p_{11}, p_{12}, p_{13}, p_{14})

p_2 =第一人称复数　　miz/imiz (p_{21}, p_{22})

p_3 =第二人称单数　　ŋ/jŋ/uŋ/yŋ/iŋiz (p_{31}, p_{32}, p_{33}, p_{34}, p_{35})

p_4 =第二人称复数　　ŋlar/iŋlar/uŋlar/yŋlar (p_{41}, p_{42}, p_{43}, p_{44})

p_5 =第三人称单、复数　　i

比如: ① kitab + im → kitabim(我的书)　② kitab + imiz → kitabimiz(我们的书)

③ kitab + iŋ → kitabiŋ(你的书)　④ kitab + iŋlar → kitabiŋlar(你们的书)

⑤ kitab + i → kitabi　(他的书 /他们的书)

在人称词尾单独缀接名词时,只需直接附加在名词词干上。如果与其它词尾同时出现,就要附加在复数词尾后,即附加在第二层上。比如:

⑥ kitab + im → kitabim

⑦ kitab + lar + im → kitablirim (这里的元音弱化规则下面再解释)

3.3 格范畴

维吾尔语名词的"格"指的是形态格。它是通过名词的形态变化表示名词与其它词之间的各种关系的语法范畴。维吾尔语中有主格(主要表示动作的主体)、属格(表示领属关系)、宾格(表示动作与客体的关系)、向格(主要表示行为动作的趋向)、从格(表示行为动作的起点或来源)、时位格(主要表示行为动作的发生,存在的时间或空间,也可以表示工具)、界限格(表示行为状态所持续的时间界限和空间界限)、范围格(表示人或事物存在的范围或领域)、相似格(表示事物和事物之间在性质、形状、特征等方面具有某些共性)、和量似格(表示事物和事物之间在量或程度上具有某种共性)十种[7],词尾的基本形态参数如下:

主格 0(无词尾)

c_1 =属格 niŋ

c_2 =宾格 ni

c_3 =向格 ʁa/qa/gɛ/kɛ(c_{31}, c_{32}, c_{33}, c_{34})

c_4 =从格 din/tin(c_{41}, c_{42})

c_5 =时位格 da/dɛ/ta/tɛ(c_{51}, c_{52}, c_{53}, c_{54})

c_6 =界限格 ʁitʃɛ/qitʃɛ/gitʃɛ/kitʃɛ(c_{61}, c_{62}, c_{63}, c_{64})

c_7 =范围格 diki/tiki(c_{71}, c_{72})

c_8 =相似格 dɛk/tɛk(c_{81}, c_{82})

c_9 =量似格 tʃɛ/tʃilik(c_{91}, c_{92})

比如:

① qiʃ kɛldi	(冬天到了。)	主格
② u akamniŋ oʁli	(他是我哥哥的儿子。)	属格
③ mɛn kitabni taptim.	(我把书找到了。)	宾格
④ akamʁa χɛt jazdim.	(我给哥哥写信了。)	向格
⑤ jurttin tuʁqanlar kɛldi	(从老家来了几个亲戚。)	从格
⑥ bir kyn øjdɛ olturdum.	(我整天都在家里。)	时位格
⑦ sahipχan mihmanlarni iʃikkitʃɛ uzitip tʃiqti	(主人一直把客人送到门口。)	界限格
⑧ bu tɛχsɛ biznin øjdiki tɛχsidin tʃirajliqkɛn.	(这盘子比我们家的盘子好看。)	范围格
⑨ bu qiznin mɛŋzi almidɛk qipqizil	(这女孩的面颊红得象苹果似的。)	相似格
⑩ akiŋiztʃilik birsi izdɛp kɛptikɛn.	(年龄同你哥哥差不多的一个人来找过你。)	量似格

维吾尔语名词的格词尾单独缀接名词时,直接附加在名词词干上,如果与其它词尾同时出现,就要附加在最外层。

比如:

qol + ni	→ qol + ni	(把手)	(单数)
qol + lar + ni	→ qollarni	(把手)	(复数)
qol + lum + ni	→ qolumni	(把我的手)	(单数)
qol + lar + im + ni	→ qollirimni	(把我的手)	(复数)

3.4 维吾尔语名词的形态规则

维吾尔语名词的形态变化会受一些特定的规律的限制。为便于语言信息处理工作,我们总结出了以下几种基本形态规则。

3.4.1 层次规则

维吾尔语词的结构也像其它语言单位的结构具有层次性。名词的词尾也按照一定的层次缀接词干。这可以用以下树性结构图表示:

其中最外层是格词尾,其次是人称词尾,最里层是复数词尾,参见图(a)。具体语言结构中有时某些层次上的词尾可能为零形式,参见图(b)。

3.4.2 形态音位变化规则

在词干与词尾的组合过程中,会发生一些形态音位变化。主要表现在以下几点:

1. 词干对词尾的元音和谐选择。

比如:qɛlɛm + im (我的笔), yzym + ym (我的葡萄)

("im"或"ym"的选择)

2. 复数词尾中的元音弱化。

比如:qol + lar + im + ni→qollirimni (我的手,复数)

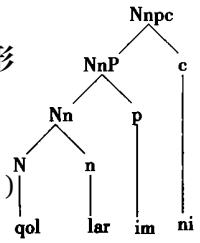
(其中"lar"变为"lir")

3. 词干中的元音弱化。

比如:bala + si→balisi([他]的孩子) (其中词干的"a"弱化为"i")

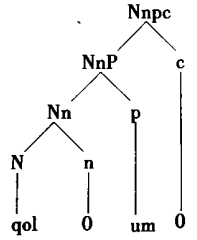
4. 词干与词尾之间增加音位。

比如:sija + im→sijajim (我的墨水) (词干与词尾之间增加了音位"j")



(把我的手,复数)

图(a)



(我的手)

图(b)

3.4.3 对词尾的语义选择规则

根据表达语义的要求,一种词尾类型中有一个词尾会被选择。比如:以格词尾为例,要表示行为动作的起点或来源选择 c_4 (从格),要表示动作与客体的关系,则选择 c_2 (宾格)。

3.5 维吾尔语名词形态参数组配表

根据具体的分析,一个维吾尔语名词按照以上规则一般可生成 120种形式。具体分析结果如下:

表 1 名词的形态参数及组配表

词尾的形态参数组合形态类型		数	人称					格									例子
		n	p					c									
			p_1	p_2	p_3	p_4	p_5	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	
1	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	øj(家,房)
2	Nn	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	øjlɛr
3	Np_1	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	øjym
4	Np_2	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	øjimiz
5	Np_3	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	øjyŋ
6	Np_4	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	øjyŋlar
7	Np_5	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	øji
8	Nnp_1	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	øjlirim
9	Nnp_2	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	øjlirimiz
10	Nnp_3	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	øjliriŋ
11	Nnp_4	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	øjliriŋlar
12	Nnp_5	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	øjliri
13	Nc_1	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	øjniŋ
14	Nc_2	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	øjni

续表

词尾的形态 / 参数 / 组合 / 形态类型		数	人称					格									例子
		n	p					c									
			p_1	p_2	p_3	p_4	p_5	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	
15	Nc_3	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	øjgɛ
16	Nc_4	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	øjdin
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
120	Nnp_5c_9	+	-	-	-	-	+	-	-	-	-	-	-	-	-	+	øjliritʃilik

注:该表只包括维吾尔语名词形态参数的标准组配形式,不包括不合规则的词尾重叠形式。比如,格词尾的重叠等。如果要考虑重叠形式,形成的形式会远远超过 120种。因篇幅所限,在此省略了表格的一部分。

4 名词词尾削尾方法

在维吾尔语信息处理中,按照以上所分析的形态范畴及其规则,可以选用以下两种削尾法来完成名词的形态分析、转换和生成。

4.1 基于具体统计的削尾方法

基于具体统计的削尾方法指的是,按照以上所分析的形态范畴及其规则,逐词统计出维吾尔语名词的所有形态形式,并把这些形式作为整体收入知识库中,与原形相对应。以便在具体语言信息处理中直接从知识库里查出其原形,从而完成名词的削尾任务。基于这种"大词库,小语法"的思想,其优点在于可以使语法得到简化,从而降低信息处理系统的复杂度[8]。这种方法虽然可以解决主要问题,但也有缺点,就是占据的空间大,并且对语言学家的语言研究来说,不能较详细地体现其语言本身的规律。

4.2 基于规则的削尾方法

为了节省存贮,机器词典一般仅存贮词干,其它形态变化均由削尾图示处理[9]。这种存贮和削尾方法的基本原则是,维吾尔语的名词存贮词干形式(仅存单数,无人称词尾的主格形式)查词时先到词典查询,如果查不到,则削尾后再查,循环到查到为止。相似的方法曾在维吾尔语词的切分中被运用过[10]。这种方法虽然可以节省存贮,思路也接近于语言本身的规律,但是有些问题还需要深入研究。比如说,在维吾尔语名词的词尾中,有些是成音节的,有些是不成音节的,而且形态音位变化规则也不能适用于每一个名词。因此对一些使用频率很高的名词和词尾,还可以考虑存贮它们的各种变形形式。

总之,语言处理中,规则处理的优势在于能充分利用现有的语言学研究成果。统计处理的优势则在于可以获得很好的一致性和很好的覆盖率[11]。因此,最好把基于具体统计的削尾方法和基于规则的削尾方法结合起来运用,充分利用两种削尾方法的优势。

5 结语

综上所述,维吾尔语名词形态分析研究对语言信息处理是非常重要的。我们相信,以上对维吾尔语名词形态的形式化的描述和分析、形态规则归纳、形态参数组配表构建、削尾方法的制定将为维吾尔语及其它中国少数民族语言信息处理中的名词处理提供一定的理论依据,供大家在实际语言处理中借鉴。这将促进中文信息处理的发展。

(下转第 98页)

4 结论

本文提出了基于反馈学习的自适应方法。该方法考虑了话题追踪任务的两个特点:话题是动态发展的;追踪的故事是按照时间排序的。它能够无监督的根据反馈修正话题,并且能够防止由于错误的反馈导致的错误蔓延。本文采用 TDT4语料中的中文部分作为测试语料,实验结果显示基于反馈学习的自适应方法对话题追踪的性能影响较大。实验结果还表明打分归一化对话题追踪性能也有一定影响但影响不是很大。

参 考 文 献:

[1] James Allan. Topic Detection and Tracking: Event-based Information Organization [M]. USA: Kluwer Academic Publishers, 2002, 1 - 16.

[2] Thomas Galen Ault, Yiming Yang Information Filtering in TREC-9 and TDT-3: A Comparative Analysis [J]. Information Retrieval, 2002, (5): 159 - 187.

[3] V. R. Shanks, H. E. Williams TDT2001 Topic Tracking at RMIT University [A]. The Topic Detection and Tracking (TDT) Workshop [C]. 2001.

[4] 王会珍,朱靖波,陈文亮,等. 基于一元语法模型的中文话题追踪 [A]. 第二届全国计算语言学学生会议 [C]. 2004: 422 - 427.

[5] Aalbersberg, I J. Incremental Relevance Feedback [A]. In: proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C], 1992: 11 - 22.

[6] Tim Leek, Richard Schwartz, and Srinivasa Sista. Probabilistic approaches to topic detection and tracking. In James Allan, editor, Topic Detection and Tracking: Event-based Information Organization [M], USA: Kluwer Academic Publishers, 2002, 67 - 84.

[7] Linguistic Data Consortium. Creating the Annotated TDT - 4 Y2003 Evaluation Corpus [H], TDT 2003 Evaluation Workshop, NIST, 2003.

[8] The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation Plan [H], version 1. 0, http://www.nist.gov/speech/tests/tdt/tdt2002/evalplan.htm, 2004.

(上接第 48页)

参 考 文 献:

[1] 缪成. 基于红旗 Linux的维、汉、英多语种操作系统的设计与实现 [D]. 乌鲁木齐:新疆大学图书馆藏. 2004.

[2] 吐尔根·依布拉音. 艾尔肯·伊米尔. 阿布力米提·阿布都热依木. 基于翻译记忆库与基于规则的汉维-维汉机器辅助翻译系统 [A]. 北京:清华大学出版社. 2003: 405 - 411.

[3] 白锡嘉. 机器翻译与自然语言的理解 [J]. 中国科技翻译, 1996, (2): 31 - 34.

[4] 张晓龙. 姚天顺. 基于文本句法的文本生成模型 [J]. 中文信息学报, 1995, 9(1): 8 - 15.

[5] 俞士汶. 自然语言理解与语法研究 [A]. 马庆株编. 语法研究入门 [C]. 北京:商务印书馆. 1999: 240 - 251.

[6] 哈米提·铁木尔著. 现代维吾尔语语法 [M]. 北京:民族出版社. 1987: 47 - 48.

[7] 程适良. 现代维吾尔语语法 [M]. 乌鲁木齐:新疆人民出版社. 1996: 126.

[8] 孙宏林. 段慧明. 面向自然语言处理的现代汉语短语信息库 [J]. 术语标准化与信息技术, 1998, (2): 26 - 31.

[9] 刘涌泉. 乔毅. 应用语言学 [M]. 上海:上海外语教育出版社. 1991: 97.

[10] 古丽拉·阿东别克. 米吉提·阿布力米提. 维吾尔语词切分方法初探 [J]. 中文信息学报, 2004, 18(6): 61 - 65.

[11] 周强. 规则和统计相结合的汉语词类标注方法 [J]. 中文信息学报, 1995, 9(3): 1 - 10.