基于形态分析的现代维吾尔语名词词干识别研究

艾孜尔古丽[1]，阿力木·木拉提[1]，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 乌鲁木齐市 830054）

摘要：现代维吾尔语名词词干识别是自然语言处理领域的重要基础性研究，主要目的是从句子中提取名词词干，提高名词识别效率。首先陈述形态分析概念和通过这些形态特征可以准确地识别其词性的意义。其次讨论维吾尔语的词类划分标准、名词的形态特征分析，总结词缀歧义及消解规则。本文提出研究总体思路，设计现代维吾尔语新词中名词识别算法，其中包括特征选择及参数估计、词内部特征、前后依存词特征等。最后将初中、高中物理维吾尔语教材作为验证对象，对名词词干进行统计与分析。

关键词：现代维吾尔语；形态分析；名词词干识别

Based on the morphological analysis of the modern Uyghur noun stems recognition research

Azragul[1], Alim Murat[1], Yusup Abaydula[1]

(1 School of Computer Science & Technology Xinjiang Normal University, Urumqi, Xinjiang, 830054, China)

Abstract: Modern Uyghur noun stems recognition is an important basic research in the field of natural language processing. The main purpose of the study is extracted noun stems from the sentence, thereby improving the efficiency of nouns recognition. Firstly, statement morphological analysis concepts and the meaning of through these characteristics can accurately identify the parts of speech. Secondly discuss the part of speech division standard, morphological analysis of nouns, summarizing the affix and ambiguity resolution rules in Uyghur language. Then put forward the overall train of thought, the design of modern Uyghur language new nouns recognition algorithm, include feature selection and parameter estimation and internal characteristics, the characteristics of the dependent word before and after the word. The teaching material of middle school, high school physics Uyghur as test object, statistics and analysis of the noun stems.

Key words: The modern Uyghur language; Morphological analysis; Noun stems recognition

1 引言

现代维吾尔语名词词干识别是自然语言处理领域的重要基础性研究，主要目的是从句子中提取名词词干。现代维吾尔语名词具有丰富的句法和语义信息，识别的结果可以广泛应用于维吾尔语名词短语分析、词性标注、命名实体识别、句法分析、机器翻译等领域。现代维吾尔语名词词干识别还应用在平行语料的对齐上，以提高词对齐的效果；由于维吾尔语歧义切分问题导致句法分析的任务变得十分复杂，维吾尔语名词词干的识别能部分消解这些歧义；随着新事物的不断涌现，新词术语的识别、统计分析、翻译也是亟待解决的重要问题。

维吾尔语在形态结构上属于粘着语类型，作为粘着语类型的语言，词的词汇变化和各种

* **收稿日期：** **定稿日期：**

基金项目：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）；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（项目编号：14AZD11）。

语法变化都是通过在实词词干上缀接各种附加成分的方式来表现的。维吾尔语词形态的多变性是维吾尔语的最突出的特点之一。 “形态是在语言中词与词组合时形式要发生变化，同一个词与不同的词组合就有不同的变化。这些不同的变化形成一个聚合，叫做词形变化，或者叫做形态。”[2]维吾尔语名词类词语的特点就集中体现在形态的变化上，容易产生歧义。而影响短语分词、机器翻译质量的核心问题是机器处理歧义的能力。本文结合现代维吾尔语语言学、形态学、计算语言学、计量语言学等学科，通过研究维吾尔语名词的形态特征，并根据名词特定的构词规则和格式，很大程度上提高了名词的识别率和机器翻译的准确性。比如

4	-	“承诺书”
5	- - - -	“筛子”、 “耳机”

表 2-3 表示人、物的维吾尔语名词词缀样例

序号	词缀	举例
1		“操劳的人”、 “酒鬼”
2	-	“水壶”， “暖瓶” “”
3		“知道（名）” “熟人”
4		“叛徒”、 “俘虏”
5		“同事”、 “亲戚”、 “同义”

表 2-4 表示地域的维吾尔语名词词缀样例

序号	词缀	举例
1		饭馆 ， 书店 “厕所”
2		花园 ， 草坪 “葡萄园”
3		会所 ， “景区”
4		“阿富汗” “”
5		“日本”、 “土耳其”

表 2-5 不能构成名词的维吾尔语名词词缀样例

	9		7		5		3		1
	10		8		6		4		2

2.3 词缀歧义及消解规则

同一词缀接在词干上也会产生不同的词类。对于这个类型的问题，本文专门列出具有歧义的词缀及其消歧规则，已提出 7 种，共 19 个词缀歧义及消解规则，有效地提高了维吾尔语名词的识别率。样例如表 2-6 所示。

表 2-6 词缀歧义及消解规则样例

序号	词缀	词性	举例	歧义的规则
1		①构成名词	“骗子”、 “吃醋”、 “艺术家”	
		②构成动词	“将要来”、 “将要去”	若该词缀前有 ，该词语被视为动词。
2		①构成名词	“小羊”、 “小湖”、 “钱包”	
		②构成形容词	“爱打扮的”、 “爱偷懒的”	若含有该词缀的词语前一个词是带有加重级（ ）的形容词，那这个词语被判断形容词。

三、基于形态分析的现代维吾尔语名词识别方法

3.1 识别总体思路

现代维吾尔语名词识别主要包括维吾尔语词汇统计、词性标注（基于词典、统计）、名词识别等关键技术与方法，图 3-1 所示。

图 3-1 名词识别流程图

3.2 现代维吾尔语普通新词中名词识别算法研究

本文提出一种融合现代维吾尔语形态变形特征的最大熵名词识别模型。根据上述总结的维吾尔语构词特点，定义上下文特征模板，提取特征集，再通过人工设置规则筛选模板．然后，训练最大熵概率模型参数．经实验结果表明，融入多个语言形态特征的最大熵模型能获得较好的性能。

最大熵原理的主要思想描述为：将已知事实作为制约条件，求得可使熵最大化的概率分布作为正确的概率分布．该模型的形式是

$$p_\lambda(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i f_i(x,y)\right) \quad (1)$$

$$Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right) \quad (2)$$

其中，$Z_\lambda(x)$为归一化函数；$f_i(x,y) \in (0,1)$为特征函数；λ_i是特征函数的权重，代表每个特征函数的重要性，每个λ_i对应于一个特征函数．

3.2.1 特征选择

（1）特征选择依据

使用最大熵模型对维吾尔语名词进行识别，是根据当前词的上下文特征，确定它的信息。本文的模型特征选择依据维吾尔语名词本身的构词特点。

（２）特征模板定义

根据维吾尔语构词特点和统计结果，本文共设计了词内部特征、前后依存词特征。

3.2.2 词内部特征

词内部特征表现的是一个词的内部变化，包括词干信息和词缀信息．维吾尔语词是通过在一个词干之后连接不同的词缀（构词词尾）构成，词缀信息表现词性等语法意义，本文设计了以下 2 个类型的词内部信息特征模板。

（１）词干信息

因为构形词尾并不影响整个词的词类信息，对于维吾尔语词干、词根上连接构形词尾构成的词，只需考虑其词干或词根的标注信息。比如，“

$$f_j(x_i, y_i) = \begin{cases} 1, \text{如果} stem(w_i) = \text{"ئۇيغ"} \ \&t_i = N \\ 0, \text{否则} \end{cases}$$

w,t	Current word
stem(w),t	The stem of current word
suffix(w),t	The suffix of current word
stem(w),suffix(w),t	The stem and suffix of current word

表 3-1 词内部信息特征模板

（2）词缀信息

尽管维吾尔文的构词和构形都是以词根、词干上连接不同词尾来完成，形成各类词，但是词尾信息是有限的，根据“维吾尔文语法语义信息词典”收录为准维吾尔文词缀中过滤的100余种名词词缀。设计例如“

由表 4-1 所示，中学物理教材中名词在整个教材词汇的平均比例为 46.37%，本教材作为实验语料合理、可行。

教材		识别出的名次数	识别出的正确名词数	准确率（%）
初中物理		5974	5413	91
高中	必修	4183	3764	90
	选修	8764	7887	90

本实验将一些带领属性人称的代词、缀接一些词缀的动词命令式等也被识别成名词。还有一些又不带附加成分的，又不在名词词根库中的名词容易被忽略，需要丰富名词词根库。

五、总结

本文介绍了现代维吾尔语名词词干识别方面的一些研究工作，重点维吾尔语名词的形态分析和在最大熵模型特征的选择。本文根据维吾尔语的特点，选取词内部词干和词缀、词前后信息等形态信息作为特征，构建了名词识别系统。实验结果表明，利用维吾尔语形态特征和最大熵模型，有效地利用上下文信息，得到了较好的识别率，尤其是对普通新词的名词识别有显著的效果。

参考文献

[1] 艾孜尔古丽等 《现代维吾尔文网络媒体用词研究》，计算机应用与软件，2012.2。
[2] 艾孜尔古丽等，基于网站用词调查的现代维吾尔语词干提取和应用，计算机应用与软件，2012.3。
[3] 艾孜尔古丽等，现代维吾尔语语言资源监测中数据分析技术研究，计算机应用与软件，2013.4。
[4] 艾孜尔古丽等，九年义务教育维吾尔语文新课标普通班教材用词研究，2012 第四届全国少数民族青年自然语言信息处理学术研讨会,8 月 6-7 号。
[5] 艾孜尔古丽等，现代维吾尔语语言监测中词频与词种分析技术研究，第六届青年计算语言学国际会议(YCCL-2012),11 月 17-18 号。
[6] 2011 中国语言生活状况报告,《维吾尔语文小学、初中语文教材用词用语调查》，中国语言生活绿皮书,商务印刷馆 2011，作者之一（拼音排序）。
[7] 2012 中国语言生活状况报告,《维吾尔语高中语文教材用词调查》，中国语言生活绿皮书,商务印刷馆 2012，作者之一（拼音排序）。
[9] 玉素甫.艾白都拉等,《维语中心语驱动文法句法分析器中的上下文相关处理》,《计算机应用与软件》[J]，1999/6，P45-48 页。
[10] 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》《计算机应用与软件[J]》 2002/4，P59-62 页。
[11] 玉素甫.艾白都拉等,“现代维吾尔语语料库的词类标注研究” [J]，民族语文，2005/4，P63-66 页。

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；

阿力木·木拉提（1988-），男，博士研究生，自然语言处理、机器翻译。

Email:xjalmjan@gmail.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：