基于功能词缀串的维吾尔语词性标注方法*

王海波[1,3]，祖漪清[2]，力提甫·托乎提[3]

（1.中国社科院民族学与人类学研究所，北京 100081；2. 安徽科大讯飞科技股份有限公司，安徽 合肥 230088；
3. 中央民族大学维吾尔语言文学系，北京 100081）

摘要：维吾尔语作为一种典型的黏着语，通过丰富的功能词缀来表达各种语法和语气。本研究探讨了“词干词性标注方法”与“词缀词性标注方法”在维吾尔语自然语言处理中的优缺点，以力提甫·托乎提教授的维吾尔语生成语法理论为指导，本研究提出以常用词缀串作为基本单位的分析方法，在大规模语料库中统计了常用词缀串的数量、频次和覆盖度，以此来判断该方法在自然语言处理中的可行性，并且对词缀串的词性标注进行了相应的语法定义，在实际语料中进行了小规模的基于词缀串的词性标注实验。本文提出的基于词缀串的词性标注方法不仅适用于维吾尔语，也适用于有着大量相似词缀的突厥语族其他语言。

关键词：维吾尔语;词缀串;词性标注

中图分类号：TP391　　**文献标识码：**A

The Uyghur POS-Tagging method based on functional suffix strings

Abstract: As a typical agglutinative language, Uyghur has rich suffixes to express syntax and mood. Firstly, this research contrasts two kinds of POS-Tagging methods in Uyghur language processing: one is based on the stem，the other is based on the suffix. Further, this research proposes a new method to segmentation the words of Uyghur based on the theory of Prof. Litip Tohti: using the suffix string as an unit. And statistics the sum, the frequency, and the cover degree of common suffix strings in a big corpus, aims to judge the feasibility of suffix string POS-Tagging method. Further, this research defines the regulation of suffix string POS-Tagging and label some corpus based on the regulation. This method we propose in this paper is not only useful to Uyghur, but also to other Turkic languages which have much similar suffixes.

Key words: Uyghur; suffix strings; POS-Tagging

引言

维吾尔语作为一种典型的黏着语，有两种词缀：构词词缀和构形词缀。其中，构词词缀是派生词的实现方式，例如名词“iš”（事，活儿）后接构词词缀“-či”派生出 “iš-či”(工人)[1]；构形词缀是语法范畴（格、数、人称、时、体、态、语气、否定等）的实现方式，在短语和句子两个层次上都起作用，例如名词“öj”（房子）后接位格词缀“-dä”构成“öj-dä”（在房子里），动词“bar”（去）后接第一人称过去时后缀“-dim”构成“bar-dim”(我去了)。

有些构词词缀派生能力非常强，语义也逐渐虚化，比如“-liq/-lik/-luq/-lük”，本文将这类派生能力强的构词词缀与所有的构形词缀一起，合称为“功能词缀”。除了词缀形式之外，维吾尔语的功能词类还有“功能独立词”形式，比如后置词（dair“关于”）、连词（yaki “或者”）、叹词（wah“哇”）、代词[2]（bu“这”）等。

在传统的维吾尔语词性标注中，词缀属于词法的一部分，词性标注实际上标注的是词干词性，本文称之为“词干词性标注方法”。

力提甫·托乎提教授的“维吾尔语生成句法理论”[1]，不赞成词干词性标注方法，他建议将“功能词缀”作为标注的核心部分，以词缀的类作为整词的类，比如 “öj-dä”（在房子里）标注为“位格短语”，“öy-din”（从房子）标注为“从格短语”，“öy-gä”（向房子） 标注为“向格短语”，本文称之为“词缀词性标注方法”。而在“词干词性标注方法”中，以上三个词标注的则都是 öy 的词性“名词”。

“词干词性标注方法”和“词缀词性标注方法”，到底哪一种更适用于维吾尔语的自然语言处理？这是非常值得研究和分析的。我们首先从理论上进行讨论：

1）从汉语和维吾尔语的语法差异上，目前常用的“概率统计模型”对不同语系语言的适用度问题。

* **收稿日期：2013.06.12**　　**定稿日期：2013.07.15**

作者简介：王海波（1979—），女，助理研究员，语音学与计算语言学；祖漪清（1956—），女，高级研究员，多语种语音合成；力提甫·托乎提（1953—），男，教授，维吾尔语句法研究。

[1] 后缀和词干之间采用“-”进行切分。

[2] 代词是实义词还是功能词是有争议的，我们倾向于将代词作为功能词处理。

从语法理论上讲：汉语主要通过语序和虚词来表达句法，其中语序在句法中的作用显著。常用的“概率统计模型”是通过统计词性之间的前后同现概率来进行词性预测以及边界位置预测，这种统计方法可以有效地利用汉语语法的语序特征，从而取得较好的处理结果。而维吾尔语起主要语法作用的是词缀，语序虽然也在句法中起作用，但作用远远小于汉语等语言。如下例所示：

U meni zal - γa kir-güz-mi-di.
他 把我 礼堂-向格 进-使动-否定-过去时 （“他没有让我进礼堂。”）

将该句的词调换位置，得到如下所示的其他句子：

U zalγa meni kirgüzmidi. Meni u zalγa kirgüzmidi. Kirgüzmidi meni u zalγa. U kirgüzmidi meni zalγa. ……

解释：这些句子都是合语法的，并且除了语义焦点不同之外，基本语义也是相同的，这是因为每个词都携带有自己的语法标志[3]。

采用“词干词性标注方法”和“词缀词性标注方法”，分别对上例的标注结果如下：

词干词性标注方法：U/代词 meni/代词 zalγa/名词 kir-güz-mi-di/动词.[4]
词缀词性标注方法：U/代词 meni/代词宾格 zalγa/向格 kir-güz-mi-di/使动.[5]

如上例所示，采用词干词性标注方法，不能反映维吾尔语言的语法本质，在根据词性的前后同现概率来预测词性和韵律边界时，将会因为无法体现维吾尔语的主要句法信息而得不到一致性统计结果。而采取词缀词性标注方法，可以充分利用维吾尔语这种显式语法标记语言的特点，使得词性同现概率更加体现维吾尔语的句法规律。

2）从维吾尔语语料处理和标注的复杂度来看，在不考虑未登录词的前提下，对语料库采取“词干词性标注”方法，只要词干匹配到，就可以根据词干词典中的词性定义，对语料库中的整词[6]进行词性标注；而采用“词缀词性标注方法”则需要首先解决词干与词缀之间的自动切分问题，然后要进行词缀串词性规则定义，最后才进行词性的自动标注工作，实现起来比词干词性标注更复杂、涉及的语言学问题更多。这可能也是为什么很多维吾尔语自然语言处理系统采用的是“词干词性标注方法”的原因之一。

3) 从数据处理效果上看，词缀词性标注方法，不仅能够体现维吾尔语的句法实质，并且能解决“词干词性标注”的诸多瓶颈问题：

a. “未登录词”问题

维吾尔语的“未登录词”比汉语等语言的处理难度更大，维吾尔语从汉语、英语等语言中借用的新词数量日益庞大，无法在词典中全部登录，这些新词以音译为主，有些可能违反维吾尔语的音系规则，但是依然严格遵循维吾尔语的词缀语法规则。在这种情况下，采取词缀词性标注方法，有两个优点：

一是只要匹配上词缀，就可以进行整词标注，而不需考虑词干是否是未登录词。

二是可以通过词缀预测词干词性，预测未登录词的词性。举例如下：

“未登录词干-da/-ta” 解释：在位格词缀-da/-ta 前的未登录词词干很有可能是个地点名词。
“未登录词干-lar/-lär” 解释：在复数词缀-lar/-lär 前的未登录词干很有可能是个表示人的名词。

---

b. 韵律边界预测问题

如何预测韵律边界，是语音合成系统必须考虑的问题。在维吾尔语的韵律边界预测上，采取“词干词性标注”和“词缀词性标注”的优缺点，如下例所示：

Dölit-imiz-niŋ härqaysi jay-lir-i-da // nopus-niŋ jayliš-iš ähwa-i oxšaš-ma-ydu.
国家-从属-领属格 各 地-复数-从属-位格 户口 领属 分布-名词化 情况-从属 相同-否定-非过去时
“我国各地的人口分布情况不一样。”

---

[3] 主格的语法标志是空（Ø）。
[4] 该例标注的是一级词性，根据实际需要，可能标注二级词性或者三级词性（比如代词-人称代词）。
[5] 具体标注规则详见文章后面部分。
[6] 整词是指语料库中以空格切分的“词干-词缀”这种词。

Män yurt-um-γa qayt-mi-γili // ikki yil bol-di.
我 家乡-从属-向格 回 否定 副词化 两 年 成为-过去时 （“我没回家乡，已经有两年了。”）

在上面两个例子中，如果我们采取词缀词性标注，在边界前的词“jayliri-da”标注成“位格”，“qayt-mi-γili”标注成“副词化”，在这种情况下进行概率统计时，就能得到“位格词缀后、副词化词缀后出现停顿的可能性较大”这种符合维吾尔语语言学的统计结果，如果采取传统的词干词性标注，其概率统计结果可能是模糊甚至是错误的。

综上所述，“词缀词性标注方法”理论上要比“词干词性标注方法”更适用于概率统计模型的维吾尔语自然语言处理和语音合成工作。并且这种基于词缀的词性标注方法，也适用于其他黏着性语言，尤其是有着大量相似词缀的突厥语系其他语言。

1. 词缀词性标注处理的前提条件

吐尔根·依不拉音等（2006）[2]，哈里旦木·阿布都克里木等（2010）[3]，阿里甫·库尔班等（2010）[4]，努尔比娅·塔依尔等（2011）[5]，尼加提·纳吉米等（2012）[6]，麦热哈巴·艾力等[7]都对维吾尔语词干与词缀切分以及词性标注进行了专题研究，不过词缀切分方法基本都是切分到单个词缀，整词词性标注也基本以词干词性为准。

在碰到词缀切分有歧义、词缀有音变、词干是未登录词等情况时，单个词缀的切分方法的正确率会受影响，而且当词干附加不只一个词缀时，按照单个词缀的切分方法无法实现对整词的基于词缀的词性标注。基于此，在维吾尔语的词缀切分中，本文提出以常用词缀串作为一个单元的词缀切分方法，不对其进行单个词缀切分，这种“词缀串短语”式的切分方法，既可以提高词干与词缀之间的切分正确率又可以实现我们的词缀词性标注方法。

以“词缀串”作为一个单元进行切分和标注，首先碰到的问题是：在真实语料中，到底常用词缀串的数量是否是可控的？这种处理方法的可行性如何？

我们知道，维吾尔语的词缀是用来表达不同语法和语气的，比如：同一个动词后可以附加“时、体、态、人称、语气、否定”等多类词缀，每类词缀又都有多个子类，每个子类又有多个语音变体，所以从理论上讲，可以产生的词缀串的数量是非常巨大的。但是在实际语料中的词缀串数量会远远小于理论值。这是因为：首先，一个动词后不可能同时添加所有类型的词缀，词缀串里同时有四五个词缀就算多的了；其次，不同词缀之间是有约束关系的，比如语气词缀“-sa/sä”一般不会与时态语缀同时出现；最后，同一体裁的语料，其语法表达方式是相对收敛的（尤其是新闻语料），也就是说常用词缀串的数量也是相对收敛的。

我们用 3.5 万的词干词典对 27 万不重复的整词语料库（以新闻语料为主）进行了“词干－词缀串”的粗略匹配切分，统计了词缀串的数量、出现频率以及语料覆盖度。对匹配上的语料进行抽查发现：词缀串频率越高，切分出错的可能性越小，词缀串频率越低，切分出错的可能性就越大。基于抽查结果，认为词缀串频率为 1 的情况切分错误的可能性非常大，目前都忽略不计，共得到了 14145 个不重复的词缀串，其中频率>=3 的有 7754 个，共覆盖匹配语料的 88％。也就是说，在新闻体裁的语料中，常用词缀串的数量是可控的，我们的词缀词性标注方法是可以实现的。

高频词缀串的例子如下：

-i-ni	1756(frequency)	-lir-i	1654(frequency)
-lir-i-ni	1242(frequency)	-i-niŋ	1180(frequency)
-lar-ni	1118(frequency)	-lar-niŋ	1070 (frequency)

2. 基于词缀串的词性标注规则定义

对目前匹配到的几千高频词缀串进行了以下几方面的工作：

1） 根据“词缀词性标注规则”定义词缀串的类别；2）对词缀串的词干进行词性预测；

其标记内容如下所示：

词缀串	词缀串类别	词缀串前的词干词性预测
-i-ni	bg（宾格）	jc（静词）

参考力提甫·托乎提教授的理论[8]，本文对词缀串的类别定义如下所示：

[1]. 格:
[1.1] –niŋ 领属 ls [1.2]-GA[7] 向格 xg [1.3]-Din 从格 cg [1.4]-DA 位格 wg [1.5]-ni 宾格 bg
[1.6]-Diki 时位格 swg [1.7]-Gičä 界限格 jxg [1.8]-Däk 形似格 xsg [1.9]-čilik/ -čä 量似格 lsg
[2]. 静词化 jch
[2.1]名词化 mch [2.2]形容词化 xrch [2.3]副词化 fch
[3].语态 yt
[3.1]使动态 sd [3.2]被动态 bd [3.3]交互态 jh [3.4]反身态 fs
[4].从属成分 cs [5].复数 fs [6].示证范畴[8] sz [7]形容词级 ji [8]构词[9] gc [9].语气 yq

在实际语料中，词缀串的构成有以下几种形式：

（2.1）.词干是静词（名词、形容词、代词、数词等非动词的词性），其附加的词缀串是嵌套关系，以最外层的词缀作为该词缀串的词性。比如：

išči	dehqan	äskär-lir -	imiz	-	gä
工人	农民	士兵-复数-第一人称复数从属-向格	（“向我们的工农兵们……”）		

解释：-lir-imiz-gä 的词缀是 “-复数-第一人称复数从属-向格”，在这个嵌套关系中，其最外层词缀“向格”是该词缀串的词性。

（2.2）.词干是动词，并且附加静词化词缀，形成副词化、形容词化、名词化，在句子中实现类似“子句”功能。这种词缀串也是嵌套关系，以最外层词缀为词缀串词性。比如：

kör - idi	- γan
看-非过去时-形容词化	（“我们将要看的……”）

解释：其中最外层词缀是形容词化，以该词缀为词缀串词性。

（2.3）.在动词词干上附加时、体、态、人称等，或者在体助动词上附加时、体、态、人称等，这两种词缀串都不是简单的嵌套关系。[10] 比如：

bar-γuz - du	-	m.
去-使动-过去时-第一人称	（“我使某人去了…”）	

解释：“-γuz-du-m -使动态-过去时-第一人称”的几个词缀之间不是简单的嵌套关系。

oqu-p	bol-di.
读-副词化	完成体-第三人称过去时 （“他读完了”）

解释：oqu-p 虽然是副词化，依然是句子的主动词，只不过“时态、人称”等后缀移到了后面的体助动词上。体助动词上附加的这种词缀串也不是简单的嵌套关系。

当词缀串内部没有嵌套关系时如何标注词性？我们目前仅有一点粗略的想法：考虑到语态、语气这两类词缀对句子成分有限制作用，所以当词缀串中有语气、语态词缀时，以这两类词缀标注词缀串。如果不存在这两类的情况下，则直接标注词干词性，也就是标注动词或者体助动词。

实际词缀串标注词典中的标注实例如下所示：

---

[7] 力提甫·托乎提教授采用“形态音位”来代表具体的语音变体，比如-Din 代表 –din/-tin.

[8] 另一个术语是“传据”-kän/-ikän（“据说”）。

[9] 词缀串中虽然以构形词缀为主，但也有包含构词词缀的情况。

[10] 力提甫托乎提教授的最新理论中，对时态、人称、语气之间的关系进行了重新定义，也定义了相应的嵌套关系，该理论目前尚未在我们的词性标注中进行应用。

词缀	词缀的类	词缀前的词干词性
-ti-la[11]	yq（语气）	jc（静词）
-lär-ni[12]	bg(宾格)	jc（静词）
-sät-tuq[13]	sd（使动）	v（动词）
-ma-ydu[14]	v-（动词词干）	v（动词）
-ma-m-sän[15]	yq（语气）	v（动词）

在实际语料中有三类形式的词：词干-词缀（词干是实义词）、体助动词-词缀、独立词，其中独立词又有两类：实义独立词和功能独立词。按照本文 2 所示方法标注“词干-词缀”形式的词，其他的则按照下述词性类别进行词性标注（将根据实际标注需要，对一级词性类进行二级词性细化）：

实义词：

[1].名词(N) ; [2].形容词 (A) ;[3].动词(V) [4].副词(ADV) ;
[5].数词(Num) [6].量词(Q) ; [7].代词(PRN) [8].模拟词(ONO)

功能独立词或功能词干词：

[1]后置词（hzc） [2]连词（lc） [3]叹词（tc） [4]体助动词(tzdc)[16] [5]语气助词（yqzc）

实际语料中的标注体例如下：

Bir/**num** čaγ-da/**wg** šimal/**n** šamil-i/**cs** quyaš/**n** bilän/**lc** bir/**num** yär-gä/**xg** kel-ip/**fch** kim-niŋ/**ls** biŋsi-si/**cs** köp/**adj** ikän-l-i-ki/**bd** toγri-liq/**gc** dätalaš/**n** qil-ip/**fch** qaptu/**tzdc**.

“某天北风和太阳一起争论到底是谁的本事强。”

3. 结论

本文基于词缀串的切分和词性标注方法，对维吾尔语的自然语言处理是一个较好的尝试，有助于解决目前维吾尔语传统的词干词性标注方法的诸多瓶颈问题。也适用于其他黏着性语言，尤其是有着大量相似词缀的突厥语系其他语言。

目前我们的工作主要是在理论论证和标注规则建立层面。词性标注的分类细化程度（包括词缀类的细化、独立词类的细化、词干词性标注与词缀词性标注的结合程度），受训练语料的规模和统计模型的处理速度影响。标注的类别越细，统计模型的精细度越高，越能反映具体的语法规律，但容易引起数据稀疏以及模型运行速度缓慢的问题，这需要在日后的实践中不断进行检验和修正。

4. 参考文献

[1] 力提甫·托乎提.《阿尔泰语言的句法结构-从短语结构到最简方案》[M]. 中央民族大学出版社, 北京, 2004.
[2] 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[A]. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论文集[C]. 2006: 148-152.

[11] -ti-la 中-ti 是-ta（向格）的元音弱化形式，-la 是语气词缀，以语气词缀。
[12] -lär-ni 中-lär 是复数词缀，-ni 是宾格词缀，是嵌套关系，以-ni（宾格）为词缀类。
[13] -sät-tuq 中-sät 是使动态词缀，-tuq 是第一人称过去时，不是嵌套关系，以-sät 为词缀类。
[14] -ma-ydu 中-ma 是否定词缀，-ydu 是非过去时（其中 y 是增音）。不是嵌套关系，又不包含语态和语气词缀，所以标注词干词性。可能是动词也可能是体助动词，但是体助动词的数量有限（大概 13 个），可以匹配与动词区分，所以在词缀串标注词典中只标注动词形式。
[15] -ma-m-sän 是-否定-疑问语气-第二人称后缀，不是嵌套关系，以疑问语气（yq）作为词缀串的类。
[16] 体助动词后要附加人称、时态等，是功能词干词。

[3] 哈里旦木·阿布都克里木. 吐尔根·依布拉音. 帕力旦·吐尔逊. 艾山·吾买尔. 阿布都热依木·热合曼. 阿布都克力木·阿不力孜. 基于短语结构语法的维吾尔语规则库建设[J]. 现代计算机(专业版), 2010(5):30-33.
[4] 阿里甫·库尔班. 吾买尔江·库尔班. 吐尔根·依不拉音. 面向信息处理的维吾尔语词语分类体系及标记研究(II)[J]. 新疆大学学报(自然科学版) Vol27，2010(1): 106-116.
[5] 努尔比娅·塔依尔. 地里木拉提·吐尔逊. 艾斯卡尔·肉孜. 面向韵律层边界自动划分的维吾尔语词性自动标注技术研究[J]. 计算机应用与软件 Vol28 2011(8):165-168.
[6] 尼加提·纳吉米. 买合木提·买买提. 吐尔根·依不拉音. 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, 48(25):137-140.
[7] 麦热哈巴·艾力. 姜文斌. 王志洋. 吐尔根·依布拉音. 刘群. 维吾尔语词法分析的有向图模型[J]. 软件学报, 2012(12): 3115-3129.
[8] 力提甫•托乎提.《现代维吾尔语参考语法》[M]. 社会科学出版社, 北京, 2013.