# 现代维吾尔语同形词词性自动标注探析

赛麦提·麦麦提明
( 新疆大学 人文学院 , 乌鲁木齐 830046)

摘要: 本文在介绍词性自动标注系统原理的基础上, 初步探讨了对维吾尔文语料进行自动处理、统计分析过程中排除由词语的兼类和同形而引起的歧义的三种方法, 即词语结构分析法、搭配词统计法和分布特点规则法等。

关键词: 维吾尔语; 同形词; 词性; 自动标注
中图分类号: H215　　文献标识码: A　　文章编号: 1001- 0823( 2006) 03- 0035- 04

维吾尔语词性自动标注是维吾尔语词法、句法分析自动化的重要突破口, 在维吾尔语自然语言的理解、处理和机器翻译系统开发中具有重要意义。词性自动标注能有效解决同形词( 组) 引起的歧义问题。同形词词性自动标注的方法, 不仅适用于维吾尔语自然语言的理解、维汉( 汉维) 机器翻译, 还可用于词频统计、文体分析等语言学领域。

维吾尔语是粘着语, 其形态变化比较丰富。名词有数、格、人称等语法范畴; 动词有数、人称、时态、语态的变化; 形容词有级的范畴。形态变化一方面提供了一些深层语法信息, 为词法分析、词性标注带来极大的方便, 另一方面也增加了自动标注的复杂性。维吾尔语中兼类词和同形词数量较多, 且使用频率较高, 这也为词性自动标注带来了一定的困难。因此, 研究词性自动标注必须首先解决同形词识别问题。本文重点探讨维吾尔语词性自动标注系统的基本原则和同形词词性自动标注的方法。

## 1.　维吾尔语词性自动标注基础

词性标注的作用就是采取适当的方法, 根据上下文语境和词语的形态标志, 消除句子中词的语法兼类, 给句子中的每个词一个合适的词类标记。对维吾尔语来说, 词性自动标注是词频统计、机器翻译、语法分析等工作的基础, 也是维吾尔语文本自动处理的难点。

自然语言的词性标注都有其各自的特点, 英语与汉语不同, 维吾尔语也不同于英语和汉语。因为汉语是孤立语, 汉语词类自动标注系统一般采用词典标注、规则标注、统计标注等手法, 其中统计标注法占的比重较大。维吾尔语是粘着语, 其形态变化非常丰富, 适合词典标注法和规则标注法相结合的方法进行词性自动标注。另一方面, 与其它语言一样, 维吾尔语有相当多的兼类词, 同形词的现象也很普遍, 这些因素加大了词性自动标注工作的难度。

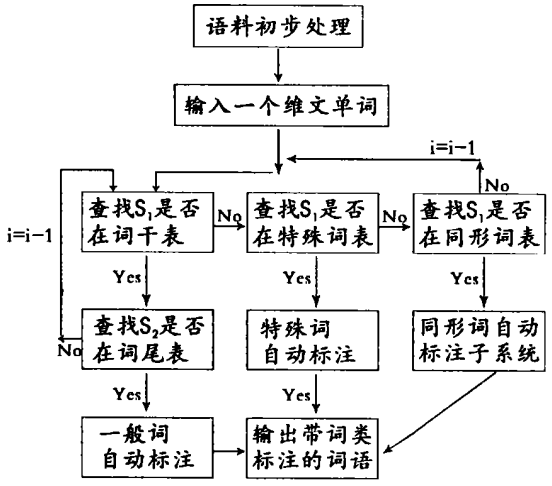
对维文语料的词性标注, 我们采取了分两步走的方法。第一步, 先对非兼类词和非同形词进行词性标注。这些词的词性标注主要靠词典和基于词干与附加成分搭配的构形规律。第二步, 对各类同形词( 兼类词和其它同形词) 进行词性标注。因为, 具有词性标注的上下文和同形词的形态结构特点会给确认同形词的词类提供更多的信息。

### 1.1　总体设想

我们首先在 Windows/xp 环境中用编程语言 Visual basic. net 开发了一个维吾尔语词性自动标注系统。该系统在各类词库和规则库的基础上通过程序从语料中提取词的切分、自动标注等。以下是该系统总的工作流程图:

( $S_1$ 为词干, $S_2$ 为词尾, i 为字符串的长度)

[作者简介] 赛麦提·麦麦提明( 1980— ), 男, 维吾尔族, 新疆大学人文学院语言学及应用语言学硕士研究生, 研究方向: 现代维吾尔语。




1.2 数据库的建立

我们在数据库管理系统 Access 的帮助下建立了本研究所需要的各类数据库。库中有词干表、特殊词表、同形词表和词尾表等。其中词干表中存放着《维吾尔语详解词典》中的 34554 个词条及其词性、弱化、脱落、词源、使用领域、页码等属性。特殊词表里有一些不规则词、专用名词等计算机不能自动标注的词语及其词根、词缀、词性等属性。在同形词表中存放着维吾尔语的所有同形词的词干形式、兼类词及其语法结合规则等信息。在词尾表存放着我们收集的所有基本词尾和它们可能组合的各种形式及其语法属性。

1.3 词干和词尾切分算法

词干和词尾的切分实际上是个搜索过程。在人工智能中一般将搜索过程分为正向搜索和逆向搜索,即从开始状态出发的正向搜索和从目标状态出发的逆向搜索。我们首先决定为本系统提出上述两种可以解决词干和词尾切分问题的方法——逆向搜索和正向搜索。下面以 باشلىقىنى 一词为例来比较两种算法的性能。

(1) 正向搜索

A. 从词首 i 个字符开始在词干表中进行搜索(i 的初值为 2)。

B. 若无匹配词和 i<N (N 是词的总长度)转下一步;若有匹配词则转到 D。

C. i=i+1,若 i<N 则转到 A;否则转下一步。

D. 若查找成功,则将剩下的字符串在词尾表中进行匹配。若成功,则返回并转到下一步操作,否则将结果返回。

以上算法是从字符串前部开始正向进行搜索,所以称为正向搜索。用该算法我们得出的词干是 باش,词尾是 لىقىنى。正确的切分应该为:词干是 باشلىق,词尾是 نى。因为该方法中对词尾表的要求很高,出错的可能性比较大,所以我们决定选用第二种方法。

(2) 逆向搜索(最大匹配法)

初值:字符串变量 $S_1$=单词(词干),字符串变量 $S_2$=词尾,N=词的总长度,i=N

A. 若词(词干)的长度是 N,从词首 i 个字符开始在词干表中进行搜索;$S_1$= Left($S_1$,i)。

B. 若无匹配词和 i>2,则转下一步;若有匹配词,则转到 D。

C. i=i−1 转到 A。

D. 若查找成功,则在词尾表中对剩下的字符串进行匹配。若成功,则转到下一步操作,否则将结果返回。$S_2$= Right($S_1$,N−i)。

以上算法是从字符串(词语)后部开始逆向进行的搜索,所以称为逆向搜索。由该算法我们得出词干是 باشلىق,词尾是 نى 的结论。所以该算法对词尾表的要求比正向搜索低,出错的可能性也比较小。

2. 维吾尔语同形词的分类

一般来说,同形词是指形式(拼写)相同,意义不同的一组词。例如 ئوت(火),和 ئوت(草)。但是在语言自动处理时,同形词所指的范围比一般所说的同形词的范围更大,因为对计算机来说,只要形式相同就算同形词。本文将在这个基础上对同形词进行分析。即将词干形式相同的词语、形态变化相同的词语、兼类词都看作同形词来处理,以便更好地解决维吾尔语同形词词性自动标注问题。比如,以下词语都将被看作同形词。

A. 以词干形式出现,属于不同词类的同形词:

这类词语是指词典上作为不同词条,属于不同词类,而且在文本中以词干形式出现的词语(下面画线的词)。如:جۇش كۆردۈم (我做梦了,名词)和تېز جۇش (快下来,动词)中的 جۇش; ئات مىلتىق (开枪,动词)和 ئات مىنىپ كەلدىم (我骑马来了,名词)中的 ئات 等。这类同形词一般出现得不多,特别是作为动词常以第二人称单数的形式出现。

B. 以词干形式出现,属于同一词类的同形词:

这类词语是指在词典上作为不同词条,属于

同一词类,而且在文本中以词干形式出现的词语。如:ئوت كەتتى(着火了,名词)和ئوت يېدى(吃草了,名词)中的ئوت;ئۇ مەندىن بەش ياش چوڭ(他比我大五岁,名词)和ئۇ ياش ئاققۇزدى(他流泪了,名词)中的ياش等。虽然这类同形词的数量不多而且出现频率低,但它们没有明显的区分性语法标志。因此,识别这类同形词是很困难的。

C. 带词尾的词和以词干形式出现的词相互为同形词:

一个词以词干形式出现,另一个词以词尾形式出现(大部分为动词)互为同形词。如:ئۇ بازاردىن ئالما ئالدى(他在街上买了苹果,名词)和باشقىلارنىڭ نەرسىسىنى ئالما(别拿别人的东西,动词)中的ئالما。

D. 带词尾的词和带词尾的词互为同形词:

原本互为同形的一对词带词尾以后仍互为同形;或者原本不是同形的一对词在带了词尾后反而互为同形词。如:قورساق ئاچتى(他肚子饿了)和ئۇ ئىشىكنى ئاچتى(他打开了门)中的ئاچتى。

E. 以词干形式出现的兼类词

兼类词的词性在运用中才能体现出来,同一个词在不同的句子中可能属于不同词类,因此词性标注必须确定其词类。如:ماشىنىنى تېز ھەيدە(快开车,副词)和تېز سۈرئەتتە ھەل قىلدى(快速解决了,形容词)中的تېز等。

F. 带词尾的兼类词

带词尾的兼类词也可能产生同形歧义现象。如:بازاردا ئادەم بەك كۆپ(街上人太多,形容词)和كۆپىنىڭ كۆزى كۆپ(群众的眼睛是雪亮的,名词)。词性标注也必须处理这类词。

## 3. 同形词自动标注过程

词性自动标注系统(请看上述自动标注系统工作流程图)如果在同形词表里发现了某个词(词干),就会把它看作同形词,而且对它进行词性自动标注。该系统将会对以上的六种同形词进行分析和自动标注。同形词词性自动标注方法主要有:

### 3.1 词语结构分析法

词语结构分析法是根据词语的形态结构分析词语语法特点的一种方法。此方法同样作为一种有效方法在同形词词性自动标注中广泛使用。词语结构分析法适合于带词尾的同形词(兼类词),因为"词干+词尾"格式中"词尾"是词语的语法标志,也是自动标注的基础。通过此方法可解决C、F类同形词的标注问题。词语切分(切开词语的词干和词尾部分)时往往会出现一些同形歧义现象。比如:كۆپىنىڭ كۆزى كۆپ(群众的眼睛是雪亮的)中的كۆپىنىڭ将会被切分为كۆپ+نىڭ。其中كۆپ一般可做名词、形容词和副词,会成为同形歧义现象。但是نىڭ是名词的词尾,所以自动标注系统将会在词尾库中找到نىڭ以及其它信息,根据这些信息自动标注为كۆپىنىڭ(n) كۆزى كۆپ。

### 3.2 搭配词统计法

搭配词统计法是一种常用的研究方法,即从加工的语料库中将关键词(同形词)的所有搭配词提取出来,然后用统计手段测量各搭配词与关键词共现的程度,以确定它们之间在多大程度上存在着相互期待和相互吸引,从而概括、描述和建立反映关键词的典型搭配情况的词语搭配信息库,在此基础上将该词在文本中的搭配与库中的搭配进行比较,以确定同形词的词性。

一个搭配是一个任意的、可重复出现的词对。设a、b两个词,(a,b)表示这两个词同现,并且出现顺序为a在前b在后,如这两个词对经常一起出现,则称(a,b)为一个搭配,并且称b为a的右搭配,a为b的左搭配。如:ئىنقىلابىي قەھرىمان(革命英雄)这个搭配中ئىنقىلابىي是قەھرىمان的左搭配,قەھرىمان是ئىنقىلابىي的右搭配(计算机默认的顺序)。算法:

```
If 右搭配 Then
    Else If 下一个词的左搭配
    Else
    不搭配标注
End If
```

这种方法适合于以词干形式出现的、不能根据词的内部结构确定词性(词项)的同形词。通过搭配词方法可解决B、D类同形词的问题。其中D类同形词虽然有形态变化,但是它们的词干和词尾相同,不能直接确定词性。如果将它们还原到词干形式仍然会出现A类同形词的情况。文本中的搭配为确定此类同形词词性提供有关信息,但前提是建立好标准的词语搭配库。比如:文本有个句子بۇ يىل چوڭ ئوت ئاپىتى يوق بەرمىدى(今年没发生大火灾),计算机遇到这个句子没办法确定ئوت一词是指"火"还是"草"。但是在同形词库中找到该

词就可以通过同形词自动标注子系统采用搭配词方法解决此问题。具体过程如下：

(1)在同形词表中找到后标注为同形词(同形词表有 1ئوت"草",n;2ئوت"火",n)。

(2)如果没有形态变化，词类也相同，则转下一步。

(3)在词语搭配库中搜索有关搭配(可以找到 ئورلـ ئوت1 ;ئوت2 ئاپەت; ئوت 1 يەـ;ئوت2 ئاپەت 等)。

(4)选择其中一个最相同的搭配(ئوت 2 ئاپەت)并标注为：

بۇ يىل چوڭ ئوت ئاپىتى يۈز بەرمىدى(n,2 ئوت)。

3.3 分布特点规则法

分布特点规则法是根据某类词语的语法规则确定同形词词性的一种方法。此方法的语言学依据是语言总是有规律的，词语、短语和句子都有内部结构和组合规律。其中各类词语的分布特征规则是对兼类词和不同词类的同形词进行自动词性标注的基础。通过此方法可解决 A、E 类同形词的标注问题。

词类是词语语法功能的分类，一个词的词类在具体的语境中才能表现出来，所以根据某个词类出现在什么位置、跟哪类词语组合等方面的规律可以确定某个兼类词(或同形词)的词性。本研究使用的词类符号有：名词(n)、动词(包括形动词、动名词、副动词)(v)、形容词(a)、副词(d)、代词(r)、数词(m)、量词(q)、模拟词(t)、后置词(k)、连词(c)、语气词(y)、叹词(i)、其它专用名词(j)等。其中实词在句中的词序大致如下：

(1) n+v(تاماق تەييارلاش)　(2)a+n(يېڭى بىنا)

(3)m+n(ئۈچ قەلەم)　(4)m+q(بەش كىلو)

(5)d+v(ئاستا ماڭ)　(6)a+v(ياخشى ئويىنماق)

(7)v+v(ئويناپ كەلمەك)　(8)d+a(بەك ئوبدان)

(9)r+n(بۇ ئادەم)

根据以上规律确定了兼类词词性标注的一些规则，通过这些规则可以初步解决标注问题。

(1) 如果一个名词 n 前有一个 x 词，且该词的变形形式为零(词干无构形附加成分)，那么 x 词前面自动记为 a。程序代码：If (x,n) then x=a。

(2) 如果一个动词 v 前有一个 x 词，且该词的变形形式为零(词干无构形附加成分)，那么 x 前面自动记为 d。程序代码：If (x,v) then x=d。

(3) 如果一个形容词 a 前有一个 x 词，且该词的变形形式为零(词干无构形附加成分)，那么 x 词前面自动记为 d。程序代码：If (x,a) then x=d

以上是同形词词性自动标注的几种方法，这些方法有时可以单独使用，但是很多情况下是混合使用的。可以根据实际情况选用其中一种或多种方法。

本文的研究结果表明，要提高词性自动标注的准确率，首先做好语法规则研究和一些实验性研究工作，在此基础上才能促进维吾尔语计算机处理和研究工作。

**参考文献：**

[1] 刘颖. 计算语言学[M]. 清华大学出版社，2002.

[2] 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，2001，(6).

[3] 玉素甫·艾白都拉. 维吾尔语句法分析器中的词义排歧问题研究[J]. 计算机应用与软件，2002，(4).

[4] 玉素甫·艾白都拉，吾守尔·斯拉木. 维吾尔语词法分析器成功[J]. 中文信息学报，1997，(4).

# On Automatic Tagging of Homographic Words in Modern Uygur

**Samat Mamatimin**

(*School of Humanities, Xinjiang University, Urumqi* 830046, *China*)

**Abstract**: This article introduces the basic principle of Uygur word-class-tagging system and on this base discusses disambiguation method of homographic words during Uygur corpus processing and statistical analysis. This method includes analysis of word structure, statistical approach of matched words, distributing rules.

**Key words**: Uygur; homographic words; part of speech; word-class-tagging