

维语词法分析器的探讨

玉素甫·艾白都拉
新疆哈密电视大学

吾守尔·斯拉木
新疆大学计算机科学系

赛依提·阿不都拉
新疆哈密三道岭电视大学

The Discussion for Lex Analysis of Uighur

Yushup. Abaidula Ushur. Slam Sayit. Abudula
Hami TV University Department of Computer Xinjiang University Hami SanDaoLing TV University

维语词法分析器是维语句法分析器的重要接口、是句法分析器中不可缺少重要组成部分。研究维语词法分析器是研究维语句法分析器重要突破口。本文介绍维语词法、词类的词法变化以及词尾模型、单词词法变化的语法手段、词法分析器模块结构、分析算法以及与相关机器词典组织等主要内容。初步机器上实现结果表明,用于维语自然语言的理解、维汉(汉维)机器翻译、词频统计、文本分析等语言学领域、计算机辅助(CAI)教学是正确和有效。

词的分类是一个基础工作,每个词都有自己不同于其他词的语法,语义和语用,或者说具有某种特殊语法功能,语法性质和语义功能、语义性质。按照词的语法功能的共性可划分为词类(Word class)。词类可以看成是一个层次系统。根据各个词之间的不同作用分为子类(SUB-CLASS)。词类划分子类及子类细分的研究十分重要的。在单词这一级中,维语是一种形态变化十分丰富的语言,单词的词性变化和词尾变化非常丰富。这个特点,一方面增加了词法分析的复杂性;另一方面却也提供了一些深层语法信息;为句法分析、词类细分带来极大的方便。通过研究语法形式,语法范寿,词尾变化规律和词与词组组合规律,建立了维语词的词尾模型。Suffix model{* [[数(S,P)],* [格(case suffixes)],* [人称(Person)],* [时态(verb tense)],* [级(voice)]]}。其中括号[]前面的*符号表示括号[]内的内容可有可无。在维语词中,存在的兼类现象采用一个树型语义分类体系来进行语义处理。词法分析的基本方法就是分析单词可能的词尾,并得到它的属性。因此建立一个词根词典和三级词尾词库组成的词库集进行词法分析。词根词典Table来表示,词尾词库Attr_x1、Attr_x2和Attr_x3来表示其中的x1、x2、x3都是变量,它们的值自动的词类范畴来决定。词的识别过程本质上是一个基于词典查找过程。对于一个给定的词,只要它已收在词典中经过若干次查找最终查到该词并确定其词性。词法分析以后的成熟词存放在特征结构(Head-drive Phars Structure Grammar)中。维语词法分析器如下几步完成词法分析任务:1、切分输入(短语、句子或短文)分成一个个短语或句子,准备其切分一个词处理。2、切分输入(短语或句子)分成一个个词。3、识别过程中确定该词的词类。4、根据该词的词类获得语法属性(如:数、格、人称等等)的所有语法信息传给后续的分析器。