

维语自然语言理解及处理研究的应用成果

吾守尔·斯拉木

新疆大学计算机系

毛居德·祖农

新疆大学电子系

摘要: 本文主要介绍了维语自然语言理解及处理研究领域的一些重要应用成果, 主要包括维语语音识别、维语语法识别及维语句法分析的自动处理、维语文/语转换等方面的应用。论述了中心语驱动文法的维语句法分析器结构、维语词法句法分析实验系统的实现方法及语音合成技术等。

关键词: 词法句法分析 韵律调整 逆自动机模型

Achievements on Applications of the Research in Comprehending and Processing Uyghur Natural Language

Wushour. Islam

Department of computer science xinjiang University

Maojuda.Zunun

Department of Electronic Xinjiang University

Abstract: In this paper. We introduce some main achievements on applications in the research feild of comprehending and processing Uyghur atrualanguage. These applications include: Distinguish in Uyghur Pronunciation. Distinguish in Uyghur Grammar Auto-processing in analysing Uyghur Syntax. Conversion between Spoken Uyghur and Writen Uyghur and son on. And we also state the structures of analysing Uygur Syntax Machine. which works by means that central words drive rammar. mean while offer the method to realize an experimental system of analysing Uyghur morphology and syntax. and phinic compounding the technology based on Uyghurp honetics and linguistics.

1. 引言

几年来我们重点做了维吾尔语自然语言理解及处理方面的基础研究, 该内容的研究在维语句法的分析, 维语快速分音节算法的实现, 维语语法识别逆自动机模型, 维语文/语转换系统的研究, 维语印刷体文字识别研究等方面都获得了重要的应用。并对维语语音识别系统, 机器翻译, 图书馆信息电子化管理(图书资料自动检索, 自动分类)等智能应用系统的研究提供了非重要的理论基础及智能化技术和方法。本文重点介绍维语自然语言理解研究的几个重要应用成果。

2. 维语句法分析器

文法描述和分析：根据现代维语的特点定义句法范畴的属性，构造分析维语句子的规则集是进行句法分析的前提。对于自然语言描述和分析依靠句法，语义等知识将输入的句子“分解”，从而决定输入句子的句法结构。其结构能够说明该句子的词与词，词组与词组之间的关系。自然语言分析的目的就是确定自然语言句法的合法性，并建立句法结构。我们采用基于合一的文法来描述和分析维语句子，因为这种文法突出了自然语言句子的核心成分，以核心成分作为句子中心来传递匹配各成分之间的信息。采用合一算法，保证了信息的传递，是区别成分匹配合法性的有效算法。

句法分析器的算法：根据句法分析器传送来的信息，将句子的每一个单词依序放在 ZH1, ZH2, ZH3, … ZHn 单元之中，然后根据维语完全语法树进行句子分析，分析步骤如下：

IF(是否句子谓语)

{ZHn 单元里的所有信息存放在 DDn 单元里，并且进行下一次的分析工作}

else {IF(是否短语) {是则显示这个合法短语}
{不是则结束短语分析}}

IF(是否句子主语)

{ZHn-i 单元中的所有信息存放在 DDn-i 单元里，并且进下一次的分析}

else {是无主句，将句子的谓语放入分析模块里进行分析}

IF(句子谓语分析模块里的 ZHn-1 和 DDn 单元的词是否匹配)

{ZHn-1 单元和 DDn 中的所有信息存放在 DDn-1 单元里，且进行下一次的分析}

else {这个句子是非法的}

{[谓词] [定语] [宾语]}这是个错误的句子，因为维语中定语只能定义宾语。……

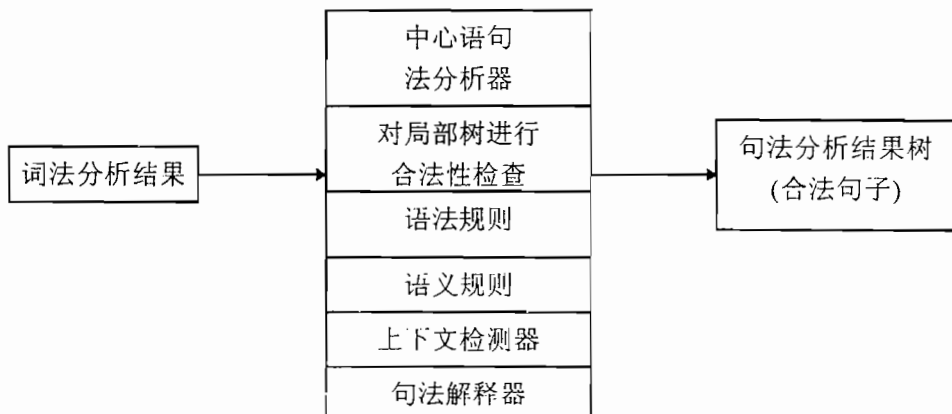
IF(谓语分析模块里的 ZHn-i-1 和 DDn-i-2 单元的词是否匹配)

{ZHn-i-1 和 DDn-i-2 中的所有信息存放在 DDn-i-1 单元里并结束谓语部分的分析}

else{这个句子是非法的}

根据维语完全语法树的定义维语句子的主语分析模块的分析方法与维语句子的谓语分析模块的分析方法完全一样。

中心语句法分析器的结构与算法：



对从词法分析器传来的信息，中心语句法分析器用局部树进行合法性条件检查作为分析程序的算法。分析程序主要包括四大步。

第一步：查询句子中的谓语。

第二步：查询句子中的主语，并使用合法性条件检查程序检查是否合法。

第三步：对句子的连贯成分进行合法性条件检查。先检查局部树父，子范畴属性值与语法规则主范畴和子范畴对应范畴属性值是否一致；然后对句部树依次进行规则匹配；依据语法规则中所指定的需要进行合一的范畴属性检查局部树中对应范畴属性间是否满足合一条件。不满足，则这棵树是非法的，分析结束。否则这棵树是合法的，则被加入到句法分析树的相应位置上，成为句法树的一部分。

第四步：句子的主语部分和谓语部分连接在一起形成一个完整的句子。

3. 维语快速分音节算法

设组成词的第 i 个字母为 $C(i)$ ，该词中的字母数为 N 。输入是顺序的字符流，输出是插入分割符的字符倒续流，分割符的数目就是音节数，抽取音节时从后往前即可。

第一步：对组成词的 N 个字母生成相应的编码序列 $X(i)$ ；

$$x(i)=\begin{cases} 1 & \text{当 } c(i) \text{ 是元音时} \\ 0 & \text{当 } c(i) \text{ 是辅音时 (其它)} \end{cases}$$

其中 $0 \leq i < N$ ，第一个字母 $c(0)$ 对应的码字是 $x(0)$ ，最后一个字母 $c(N-1)$ 的码字是 $x(N-1)$ 。初始化：设置角标变量 $i=N-i$ ，元音计数器 $v=0$ 。

第二步：计算 $v=v+x(i-1)$ 。

第三步：如果 $i=0$ 则执行第四步，否则输出字符 $c(i)$ 。如果 $v=0$ ，结束算法退出，否则输出音节分割符并结束算法退出。

第四步：计算 $v=v+x(i-1)$ ，并输出字符 $c(i)$ 。

第五步：如果 $v=2$ ，则输出分割符，并置 $v=0$ 。

第六步：计算 $i=i-1$ ，并从第三步开始继续执行。

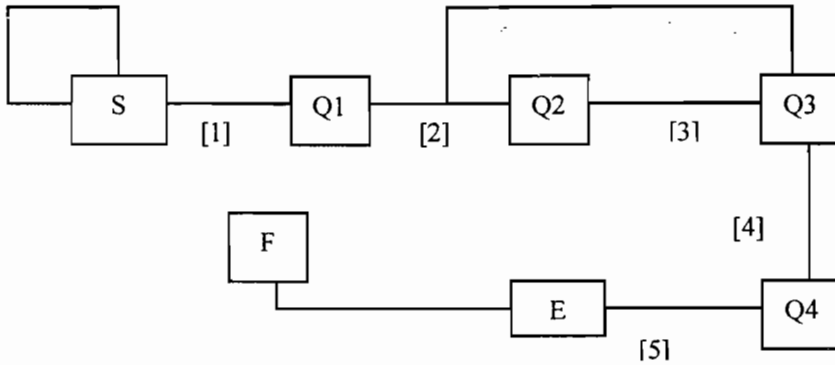
以上算法，如果生成编码采用查表法，则整个算法中，只进行两次比较，一次是测试角标溢出比较，一次是测试元音数目比较，其执行速度相当快。

4. 维语语法识别逆自动机模型

逆自动机模型： $G=(S, Q, P, E, F)$ 其中 Q 为所有状态集合， Σ 为元素集合， $\Sigma=W \cup C$ ， S 为起始状态集合， F 为结束状态集合， P 为状态转移集合， E 为拒绝状态集合。当某一字符串被逆自动机识别通过时，和自动机的情况相反，表明该字符串不符合维语语法规则。相应的，如果被逆自动机识别拒绝时，说明符合维语语法规则。这样克服了系统语法库的不完备性带来的一系列问题。自动机模型识别范围很有限，有些符合维语语法规则的语句会被语法自动机拒绝，不能通过语法校验层。由于语法库的不完备性，语法校验层反而会大大降低识

别率。在语音识别中错误语句通过语法校验层的风险远小于正确的语句不能通过语法校验层的风险。我们试采用了维吾尔语语法的逆自动机识别模型，从而用语法校验层提高了识别率。

下面是逆自动机识别模型：其中[1], [2], [3], [4], [5]是词类。表示不带空格的转移， \rightarrow 表示带空格的转移。Q(q1, q2, q3, q4)。



5. 维吾尔文/语转换系统

文/语转换的关键是语音合成技术。我们采用的是波形存储法。这种技术用于语音合成时，选取音库中自然语言合成基元的波形，对这些波形进行编辑拼接后输出。所以我们首先以音节为单位进行录音，然后用语音编辑软件对语音进行截取，存储成格式为.WAV 的文件。对输入的文本流进行语言学处理，进行分词，分音节，将音皆内码与索引库中的内码相对应，在由音皆索引库按地址索引出该音节的发音参数在语音库中的位置，然后读取语音库进行语音合成。考虑到在实际中的应用，为了用最小的信息量合成出最接近人的自然声音，通过对 DSP, GSM, PCM, ADPCM 几种压缩编码方式的实验，我们采用了 ADPCM 压缩编码。使合成出的声音质量最为清晰自然。

在系统的实现过程中，根据维吾尔语自身的特点进行了特殊的韵律调整，总结了维吾尔语言学规则并应用于系统。这样就综合了维吾尔语言学，语音学及语音编码技术，使合成声音能够满足大多数场合的要求。基本反映了维吾尔的特殊韵律特性，达到了较好的自然度，提高了维吾尔语句的可理解性。

维吾尔语言学规则：在维吾尔语音合成系统中，语言学处理对改善可理解性起着关键性作用。我们研究了维吾尔语言的特点，总结了维吾尔语音处理规则并建立了维吾尔语音规则库，用于维吾尔语音合成系统中分段，分句，符号处理并控制其阅读形式。规则包括停顿规则和文本替换规则。停顿规则又分为第一类停顿（时间最长）：流程图中遇到两个或更多的空格，则判为标题或段结束。第二类停顿（时间稍短）：遇到标点符号，表格线符则判断为一句结束。第三类停顿（时间最短）：词与词之间的一个空格。文本替换规则就是对一些常用的外拼音符号用维吾尔语音来模拟表示，包括标点符号，数字符号，英文字母和其它一些特殊符号。

维吾尔语音合成的韵律规则：维吾尔的重音向其它突厥语一样，两个音节以上组成的

单词重音一般落在最后一音节上。而句子的语调根据句子成分有升调和降调等。所以较典型的韵律规则包括语调规则和元音的和谐规则。维吾尔的句子按其表达的意义和语气可分为陈述句，疑问句，祈使句，感叹句四类。其中陈述句的语调一般是平的，句尾用逗号或分号。疑问句的语调末尾是向上扬的，句尾用问号。祈使句和感叹句按谓语，感叹词，语气助词的不同而具有不同的语调。所以对陈述句我们通过判断句尾的标点符号使其合成出平声语调。由于维吾尔中韵律调节最重要的是语调调节。而维吾尔语调最大的特点是句子中语调的变化是由助词和感叹词决定的。所以在疑问句，祈使句，感叹句中，我们利用对助词和感叹词的控制来进行语调的调整。因此，做语音库时特别注意了助词和感叹词的音节参数提取。此外，在维吾尔语音体系和词法体系中，存在着突厥语所具有的粘着与和谐现象。这类规则比较多，又没有同性，所以实现起来非常困难。所以只对比较普遍使用的和谐规则做了归纳，使附加成分中的元音和词干中的最后一个元音相协调。在合成系统中引入上述规则，使合成语音的自然度显著提高。

6. 结束语

随着计算机应用在少数民族地区的不断推广，维吾尔自然语言研究的重要性就更加突出。上述研究成果已在各种维文系统中得到了广泛的应用。例如，维汉(汉维)机器翻译系统，各种维文中小学家庭电脑教师系统及各种多媒体教学环境系统等。

参 考 文 献

- [1] 现代维吾尔语法(维文), 哈米提·铁木尔著, 民族出版社, 1987.6
- [2] 汉语计算语言学, 吴蔚天等著, 电子工业出版社, 1994.7
- [3] 维吾尔语句描述和分析方法的探讨, 玉素甫·艾等, 中文信息, 1994.7
- [4] 计算机语音技术, 朱雄民, 1991年