现代维吾尔语句子成分分析技术研究

努尔艾合买提·艾买提[1]，艾孜尔古丽[1]， 玉素甫·艾拜都拉[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：句子成分分析是自然语言处理研究中的重点和难点。首先陈述现代维吾尔语短语和句子之间的关系、语类间的相互关系及现代维吾尔语单句成分划分问题；其次讨论现代维吾尔语语料库的预处理、短语标记集、句子成分划分基本思路、句子成分分析算法等研究内容；探索现代维吾尔语谓语的识别算法设计、其他句子成分的识别、自动界定预测算法等关键技术；最后分析实验结果，利用规则和统计相结合的方法实现现代维吾尔语句子成分分析，取得相对较高的准确率。

关键词：现代维吾尔语；短语结构；句子成分；分析技术

The Research of Modern Uyghur Language Sentence Constituents Analysis Technology

Nurahmat[1], Azragul[1], Yusup Abaydula[1]

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

Abstract: The sentence component analysis is emphasis and difficulty in natural language processing research. Firstly, it states relationship between phrases and sentences, the relationship between language class, division problems of modern Uyghur single Sentence component; Secondly we discussed these Research Contents which are pretreatment of modern Uyghur language corpus, phrase tag set, the basic idea of dividing the sentence elements and Sentence component analysis algorithm, and explores key techniques like design of modern Uyghur predicate elements recognition, identify other sentence components, automatic definition of prediction algorithms; Finally, the experimental results are analyzed and discussed. The method used to realize modern Uyghur language sentence constituents analysis that is using rules combined with statistic, and have a relatively high accuracy.

Keywords: Modern Uyghur language; phrase structure; sentence component; analysis technology

1 引言

句子成分分析是自然语言处理研究中的重点和难点。现代维吾尔语句子的形式复杂多样，简单的维吾尔语句子可以只有一个词或一个短语，复杂的维吾尔语句子可以接近一个段落。基于浅层句子成分分析的思想,本文介于线性词序列和完整句法树表示之间浅层句子成分分析。若只靠语法知识是不能满足要求的，语法语义相结合的深层句子成分分析是一个难题。

现代维吾尔语单句成分划分在句子中，词与词之间有一定的组合关系，句子成分由词或词组充当。现代维吾尔语里一般的句子成分有六种：即主语、谓语、宾语、定语、状语和补语。

---

* **收稿日期：** **定稿日期：**

基金项目：国家自然科学基金项目（项目编号：61262066）；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）；国家社科基金重点（项目编号：14AZD11）。

维吾尔语句子成分主要格式为（主语）（宾语）（谓语），一般谓语都在句子的后面。例如：

①．写人格式：

“谁”＋“干什么”

吐尔逊‖写字　　تۇرسۇن خەت يازدى

主语	谓语	宾语
吐尔逊	写	字
تۇرسۇن	يازدى	خەت

②．写物格式：

我读了书。　مەن كىتابنى ئوقۇدۇم

主语	谓语	宾语
我	读了	书
مەن	ئوقۇدۇم	كىتابنى

经过复杂句子的分析发现，现代维吾尔语句子成分划分为三个中心概念。以上三大成分具有一定的特征。还有其他成分，如：定语、状语、补语也具有一定的特征。这维吾尔语句子成分分析带来一定的方便。

2 现代维吾尔语句子成分的识别与实现

2.1 现代维吾尔语短语和句子之间的关系

为了描述句子成分分析的结果，对句子成分制定成分标注标记集。主语表示为 1，谓语表示为 2，定语表示为 3，宾语表示为 4，状语表示为 5，补语表示为 6。例如：

مەن [قىزىل تاشلىق كىتاب]/NP300805B1KP0 سېتىۋالدىم.

2(سېتىۋالدىم)　4[4(كىتاب) 3(قىزىل تاشلىق)]　1(مەن)

（我）1　（买了）2　　(红皮的)3　　(书)4

以上例子中有一个最大的短语是红皮的书，它的词类编码为 NP300805B1KP0。

维吾尔语句子成分的数量是能确定短语的多与少。句子成分分析是先进行词汇、短语分析。

维吾尔语的短语类是两个或两个以上的单词语类构成的更长的单位，由于每一个短语里只有一个核心词（head），其他成分都起补足语（complement）或附加成分(adjunct)的作用，该短语类就以核心词类名来命名。例如一个名词短语（NP）的核心词是名词，一个动词短语（VP）的核心词是动词，一个形容词短语（AP）的核心词是形容词等。由于居上关系和属下关系考虑，一个短语类里包含更大的语类。这些语类中短语语类居上，单词语类属下。整个句子层面考察 IP 居最上，S 与 I 直接属下于 IP。而 S 直接居于 KP 与 VP 之上，KP 与 VP 直接属下于 S。但是，KP 与　VP 有自己的内部结构，可以居于其他成分之上。

2.2 语类间的相互关系

语类的居上和属下关系其实就是句法层次关系。语类间的居上和属下关系可以是直接的，也可以是间接的。语类间还有前后关系。维吾尔语言中有些结构可以用语类之间的姊妹关系来下定义。在 NP 里，领属名词(Possessor)带领属格 نىڭ 出现在前；从属名词(Possessee)带从属人称词尾(即 Pos)，出现在后。从属人称词尾必须与领属者的人称和数相一致。如：مېنىڭ كىتاۋىم “我的书”。句中 مېنىڭ“我的”表示第一人称单数，表示领属格。这些特征中一致关系明显地体现出来。因此，服从领属从属名词短语的一致性规则。如：领属名词和从属名词必须在人称和数量上的一致。

3 句子成分分析算法

3.1 维吾尔语语料库的预处理

一般需经过分词处理、语法语义标注、句法分析等不同阶段进行处理，才能一个“生”语料库逐步加工成为一个包含各种语言知识的“熟”的语料库，作为下一步自然语言信息处理中非常有用的知识库，图 3.1 表示维吾尔语语料库进行多级加工过程。

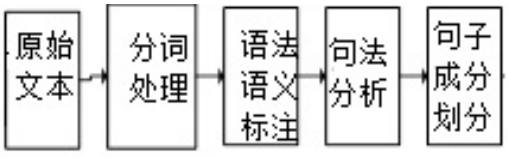


图 3.1 多级加工过程

3.2 维吾尔语句法标记集

为进行维吾尔语语料库句子成分的划分，词一级标注的语料需要短语标注。本文使用新疆师范大学提供的“面向信息处理的现代维语短语一级词性标记集”标准。详见表 3.1

表 3.1《面向信息处理的现代维语短语一级词性标记集》样例

格短语	KP	كىلىشلىك سۆز بىرىكمىسى
名词短语	NP	ئىسىم سۆز بىرىكمىسى
动词短语	VP	پېئىل سۆز بىرىكمىسى
形容词短语	AP	سۈپەت سۆز بىرىكمىسى
副词短语短语	DP	رەۋىش سۆز بىرىكمىسى
数词短语	MP	سۆز بىرىكمىسىسان
量词短语	QP	سۆز مىقتار بىرىكمىسى
代词短语	RP	ئالماش سۆز بىرىكمىسى
摹拟词短语	ZP	تەقلىدىي سۆز بىرىكمىسى
叹词短语	EP	ئىملىق سۆز بىرىكمىسى
引语	QT	كۆچۈرۈلمە سۆز بىرىكمىسى

3.3 句子成分划分基本思路

根据维吾尔语特点，打字员打字过程中有些问题很容易忽略，很容易出现字母录入错误和词语录入错误。

（1）字母录入错误：维吾尔语 8 个元音字母双元字符组成，打字员录入时先打入（ى），然后录入（ۇ）后完成录入完成（ۈ）元音字母的录入任务。但是具体语料中统计发现（ىو）、（ىۇ）现象。针对解决这种问题，利用维吾尔语的元音字母构成规则，自动完成调成任务。

（2）词语录入错误：A、“明天”（**نەتە**）是正确的，有时候 B、打“明天”（**نە تە**）是错误的。但是看出来很相似，实际上 B 时事错的。这个错误人发现不了，机器根据维吾尔词与词之间空格分开的分词原则，很容易误认为（**نە**）和（**تە**）两个词。这软件统计的词汇中大量出现非规范词汇。为了解决这个问题，根据维吾尔语本身特点和 11 条现代维吾尔语音节构成规则，基于词典与机器学习结合技术、人机互助方法解决此问题。

（3）语料需要解决和处理语料正字问题：解决这个问题，我们根据维吾尔语的音节规则和新疆维吾尔自治区语言文字工作委员会发布的维吾尔语正字正音词典为基础，新疆师范大学《网络信息安全与舆情分析》重点实验室开发的校对系统完成了语料校对任务。保证统计语料的规范性，并下一步处理模块提供标准词语打下基础。

（4）词语定义和汉族人名、外国人名、机构名、地名等问题：得到正确句子成分，本文中主要解决维吾尔语词语和汉族人名、外国人名、机构名、地名识别问题。

（5）研究词干提取问题：本次词干提取词干中利用基于词典和人机交互技术结合方法

提取词干，提取词干过程中现代维吾尔语词干词典维护来提取词干过程中出现的新词干的发现、机器词典中新词干的补充和机器学习功能增加等。

（6）提取的词干进行标注问题：根据现代维吾尔语的特点、现代维吾尔语词干标注标记集进行词干词类标注。

（7）短语标注和句子成分分类算法研究：根据现代维吾尔语分短语原则和现代维吾尔语短语类标注标记集，进行句子短语分类，设计句子成分分类研究。

3.4 句子成分分析算法研究

对句子进行分析句子成分。

. （买了）سېتىۋالدىم（苹果）ئالما（红）قىزىل（集市）بازاردىن（今天）بۈگۈن（我）مەن

语法标注后，成分分析结果图 3-2 所示。

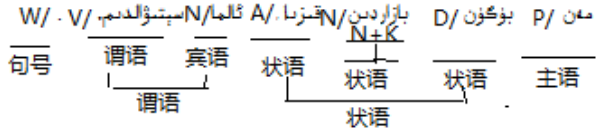


图3-2 句子语法标注成分分析结果

语法语义相结合标注后，成分分析结果图 3-3 所示。

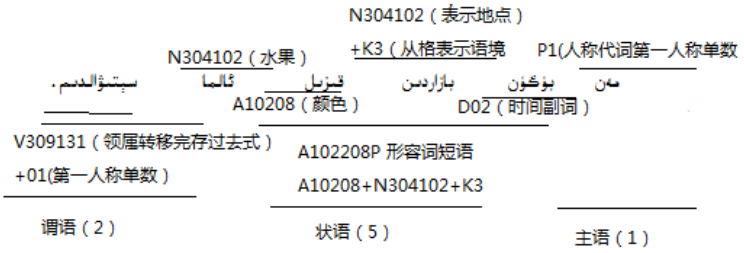


图3-3 句子语法语义标注成分分析结果

为了实现以上目标下面几个方面进行现代维吾尔语句子成分分析，在分析过程中首先识别谓语，然后逐步识别主语，宾语，定语，状语。

3.4.1 谓语的识别算法设计

经过粗筛选得到的集合 C 中的某个词可能是真正的谓语中心词，也可能不是，本文称这些词成为准谓语，记为 Pk，k∈N，Pk∈C。

可信度 R(Pk)：指句子中的准谓语 Pk 是句子真正谓语中心词的可能性的一种度量。输入的句子经过语片捆绑和基于规则的谓语粗筛选，得到一个谓语候选集 C。其中$\forall$Pk∈C 都有 Pk∈H∨Pk∈H(H 为谓语中心词集，$\overline{H}$为非谓语中心词集，H∪$\overline{H}$=C)。现在需要根据句子中 Pk 的上下文和 Pk 本身的词性，评价 Pk 的可信度 R(Pk)，评价依据在本文中称为特征。下面讨论相关特征的选择、特征权值的设定和调整、Pk 的可信度的计算。

准谓语 Pk 是否在句子中充当谓语中心词是由 Pk 本身的词性及其上下文的特征等决定的。因此，本文选取了下文的几个特征。这样的选择减少了很大一部分的统计量，并且从对谓语识别的性能上来说，没有什么影响。在本文中，选取的特征有：

1)词类静态特征：考虑准谓语 Pk 的词性 T(Pk)对 Pk 充当谓语中心词的影响。表示为：

T(Pk)=t　　其中，T(Pk)为 Pk 的词性，t∈TagSet，TagSet 为汉语词语的词性集合等。

2)词类环境特征：准谓语 Pk 在句子中出现时上下文词类的特征，这里只取与 Pk 前后相连的两个词的词类。这两类特征分别表示为：

T(Pred(Pk))=t 和 T(Succ(Pk))=t　　其中,Pred(Pk)表示 Pk 的前一词,Succ(Pk)为 Pk 的后一词。

将上文选取的特征表示为 A1,A2,…,An,如设词类静态特征为 A1,谓语前一词的词性特征为 A2 等。对其中的某个特征 Ai,有 Ai=(ai1,ai2,…,aim),如词类静态特征 A1,准谓语的词性可能为 vgo、vgn、vh 等,此中的 aij 即为 T(Pk)=vgo,T(Pk)=vgn 等具体特征。

有了上述定义,我们设定 Ai(Pk)表示准谓语 Pk 出现时出现特征 Ai 这一事件,aij(Pk)表示特征 Ai 取某一值时的具体事件,$\overline{A_i}$(Pk)表示特征 Ai 不出现这一事件。则 aij(Pk)对 Pk 作谓语中心词的支持度 V(aij(Pk))定义如下:

$$V(a_{ij}(Pk))=\lg P(Pk\in H|a_{ij}(Pk))-\lg P(Pk\in H|,\overline{A_i}(Pk)) \quad (1)$$

其中,$$P(Pk\in H|a_{ij}(Pk))=\frac{P(Pk\in H\wedge a_{ij}(Pk))}{P(a_{ij}(Pk))} \quad (2)$$

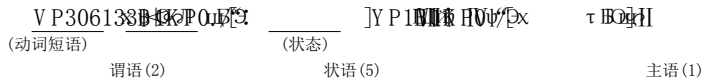
$$P(Pk\in H|\overline{A_i}(Pk))=\frac{P(Pk\in H\wedge\overline{A_i}(Pk))}{P(\overline{A_i}(Pk))} \quad (3)$$

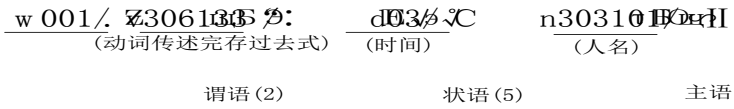
以上定义的论域是谓语候选集 C,HAC。在选定抽取的特征 A 后,支持度可以通过统计计算得到。其中　Ai(Pk)的情况如当准谓语 Pk 在句末时,无法考察 Pk 的后一词的词性特征,即 Succ(Pk)为空,特征 T(Succ(Pk))=t 不存在。[11]

3.4.2 其他句子成分的识别

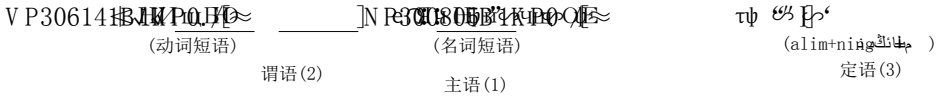
维吾尔语谓语成功识别后逐步实现宾语，定语，状语的实现。宾语和状语主要在谓语前面，如果词性和短语标记为 N,词尾为“ni”的词语或者短语在谓语前面，那这个词汇或者短语可看为宾语；比如：

词性和短语标记为 D,DP 的，词尾为“da,ta,de,te”的词语或者短语在谓语前面，那这个词汇或者短语可看为状语；比如：

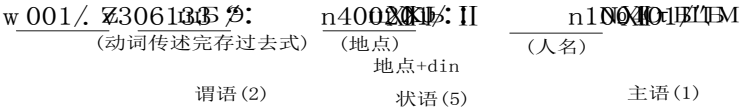




词性和短语标记为 A,AP 的，词尾为“ning”的名词词语或者短语在名词和名词短语前面，那这个词汇或者短语可看为定语；比如：



最后识别主语的问题，识别主语主要看谓语部分，通过谓语能确定主语部分，一般词性和短语标记为 N,Np ，R 的词汇或者短语可看为主语。主语主要在句子前面，中间。比如



识别主语不像识别谓语，宾语，状语，定语的一样，人工识别也有一定的难度，所以识别主语的时候我们偶尔依赖于人工处理，这样能保证句子成分识别的正确性。

给定一句经过正确切分和词性标注，短语标注后，如何利用其中的词语，词类，短语和句法特征信息，确定句子成分的边界位置，即哪个词语处于句子成分的左边界，哪个词语处于右边界，哪个词语处于句子成分中间的位置，是一个维吾尔语的界定研究所要解决的主要问题，如上面所示的例题，此问题的正确解决，对于进一步进行括号匹配和分析树生成进而完成维吾尔语句子成分自动划分，具有重大意义。

3.4.3 自动界定预测算法

想象一串句子成分界定情况，经过一系列奇怪的变换，以一串词语的形式输出，次那个人可以将句子成分界定预测问题抽象成如下的噪声信道模型[sh48]：

$B_1^n \rightarrow$ 噪声信道 $\rightarrow S_1^n$

其中$B_1^n = b_1, b_2, b_3 \ldots b_n$

为一串句子成分界定情况描述，b_i可去罢{0：句子成分中间位置，1：句子字成分左边界，2：句子成分边界}；$S_1^n = < W_1^n, T_1^n >$ 为观察到的原始句子，$W_1^n = w_1, w_2, w_3 \ldots w_n$为句子中的词语串，$T_1^n = t_1, t_2, t_3 \ldots t_n$为各词语的词类标记串，

这里的处理目标是：依据观察到的信道输出S_i^n,预测信源的句子成分界定分布情况B_i^n，对此模型，采用不同的计算方法，可以得到不同的句子成分界定预测结果。[14]

4. 研究现代维吾尔语句子成分分析系统

4.1 现代维吾尔语句子成分分析系统实现

句子成分分析的整体过程如图4-1：

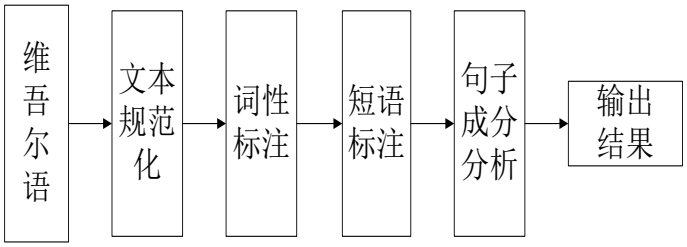


图4-1 句子成分自动分析过程

从图 4-1 看,整个句子成分识别过程分为文本规范化，词性标注，短语标注，句子成分分析四个主要部分。

1)文本规范化的主要原理是：有多个连续的空格时替换为一个空格，每一个维文单词和标点符号、标点符号和标点符号中间插入一个空格。按这规则规范化后的文本便于本系统得短语标注，句子成分分析模块儿的正常工作以及人的理解。

2)词性标注：句子分解成单词后，把单词和标注内码置换，如果遇到一些新的单词，这时系统靠操作员来处理新单词。

3)短语标注：短语分解过程为，首先把整个句子看成是一个短语，判断是不是一个正确的短语，是——构成一个短语结构体，否——从句子前部或后部剪取一个单词，剩下的部分看成一个短语进一步判断。短语分解过程中如果在当前的句子中各个短语的长度的和小于整个句子的长度减 3，对这句子进行人工分解短语，把人工分解后的新的短语生成一个新的短语规则添加到短语规则库里，以便下次遇到同样的短语时自动处理。

4)句子成分分析：通过短语标注，词性标注对维吾尔语文本进行句法分析，根据得到的数据确定边界预算。

4.2 实验数据分析

本文所用的统计和测试语料主要选小学语文教材为验证对象。选取了 120 篇文章共 9000 个句子,约 65,000 个词。这些句子中的大部分包括 1～3 个分句,本系统对每个分句进行句子成分识别。整个测试分两步进行:

1)第一种测试利用基于规则的句子成分识别器进行句子成分识别和边界预算,并经过人工校对。然后对带标记的语料集进行特征学习得到带权特征,进行测试,记录其准确率,其结果 60%～75%左右。

2)第二种测试利用统计的方法进行分析。从结果中看出,最开始的时候,由于训练集和测试集的特征的差别比较大,其准确率 69%左右,通过特征的不断学习,开始的时候准确率的波动较大,随后逐渐趋向稳定,接近 70%。

句子 测试	句子数量	词汇数量	句子成分数量	正确划分的句子成	准确率
第一种测试	9000	65000	39675	28564	71.99%
第二种测试	9000	65000	39675	27231	68.64%
合并以后	9000	65000	39675	29753	74.99%

从这个表可看出,如果本系统把两种方法合并起来使用,那本系统的识别正确率约在75%左右,高于规则和统计的方法近6个百分点。因此,当训练集特征与测试集的比较接近或训练集足够大时,可以取得相对较高的准确率。

参考文献

[1] 高士杰，维吾尔语语法　中央民族大学出版社　1998.2

[2] 程适良现代维吾尔语语法　，阿不都热西提　新疆人民出版社　1996.9

[3] 玉素甫，阿不都热依木.沙力，阿依木古丽。论现代维吾尔语词性标注系统“，新疆师范大学，2006/1

[4]、玉素甫，潘伟民，热孜万。笔式维吾尔文识别中的文字切分研究，民族语言文字信息技术研究，西苑出版社，2007 年 2 月。

[5]、玉素甫，阿不都热依木.沙力，木沙江。面向现代维吾尔语处理的信息库构造方法，民族语言文字信息技术研究，西苑出版社，2007 年 2 月。

[6]、玉素甫.艾拜都拉潘伟民阿布都热依木.沙力　。论面向信息处理现代维吾尔语资源库构建，2007 国家语言资源与应用语言学高峰论坛，2007 年。

[7]、玉素甫.艾拜都拉，　潘伟民，力提甫.托乎提，　面向信息处理的维吾尔语短语结构规则与标注集研究，中文计算技术与语言问题研究，电子工业出版社出版，2007

[8]、阿不都热依木.沙力，　玉素甫等《从语义角度探讨动词的分类》,《语言与翻译》，2005/1。

[9] Yusup Abaidula，Rezwangul，　Abdiryim Sali，“The Research and Development of Computer Aided Contemporary Uighur Language Tagging System ”，Journal of Chinese Language and Computing 15 (4):(203-210)　2005。

[10] 玉素甫.艾拜都拉等,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。1999 年　第 16 卷　第 6 期。　国内核心刊物。

[11]龚小谨,罗振声,骆卫华汉语句子谓语中心词的自动识别(中文自动校对技术)中文信息学报 2003,17(2)

[12]、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。2002 年　第 19 卷　第 4 期。国内核心刊物。第一作者。

[13]艾孜尔古丽，齐向卫，玉素甫·艾拜都拉　基于网站用词调查的现代维吾尔语词干提取和应用研究（计算机应用与软件）2012 年 3 月

[14] 周强　汉语语料库的短语自动划分和标注研究（博士论文）P18-P19

作者简介：

努尔艾合买提（1984——），男，硕士研究生，主要研究领域为计算语言学、自然语言处理；

Email:526193972@qq.com

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;

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：