基于规则的哈萨克语句法分析算法研究

牛 娜,古丽拉·阿东别克
(新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046)

摘 要:哈萨克语的理解一般分为以下步骤:原文输入、词语切分及词语属性特征标注、语法及句法分析、语义及语用和语境分析、生成目标形式表示、句群及篇章理解等。句子分析上接篇章理解,下联词汇分析,起着承上启下的作用。由于哈萨克语句法分析结果的准确度将对后续机器翻译的研究产生影响,在掌握哈萨克语词法分析技术的基础上,结合现代哈萨克语句法结构特点,首先介绍了厄尔利算法、GLR算法和线图算法三种基于规则的句法分析算法。通过实验对比发现,线图分析算法在哈萨克语简单句的分析中具有运算速度快和占用空间小的综合优势。针对传统线图分析算法冗余边较多造成分析准确率不高的现象引入规则库优化的改进线图算法,实验结果表明,改进后的线图算法使得准确率提高了4.19%,运行时间缩短了20倍。

关键词:哈萨克语;句法分析;线图分析算法;规则库;句法树
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2015)09-0043-05
doi:10.3969/j.issn.1673-629X.2015.09.009

Research on Parsing Algorithm Analysis of Kazakh Based on Rule

NIU Na, GULIA · Altenbek
(College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

Abstract: The understanding of the Kazakh is generally divided into the following steps, the original input words, word segmentation and attribute features labeling, grammar and syntax analysis, semantics and pragmatics, and context analysis, generating target form, sentence group and text understanding, etc. Sentence analysis discourses text understanding, allying lexical analysis, playing the essential role. Because the Kazakh syntactic analysis result accuracy influences the followed machine translation, based on mastering Kazakh lexical analysis technology, combined with the characteristics of modern Kazakh syntactic structure, first introduce the three rule-based parsing algorithms including Earley algorithm, GLR algorithm and chart analysis algorithm. The chart analysis algorithm has fast speed and small footprint of the comprehensive advantages in simple Kazakh sentences analysis found by experimental comparison. The rule base optimization chart analysis algorithm is introduced to aim at the problem of low accuracy caused by more side redundancy, experimental results show that the algorithm makes the accuracy improved 4.19%, the running time shortens 20 times.

Key words: Kazakh; syntactic analysis; chart analysis algorithm; rule base; syntax tree

0 引 言

自然语言处理是人工智能的重要学科,研究领域不仅仅局限于汉语英语的应用,如今利用计算机处理少数民族语言成为了顺应社会发展的重要课题之一。语言的研究层次大致可以分为分词、词性标注、句法分析、语义分析、篇章分析和信息抽取等[1]。由此可见,句法分析处于关键地位,其结果的准确度对后续深入研究具有很大的影响。

句法分析目的是识别出句子包含的所有句法成分以及这些成分之间的关系,最终的分析结果以句法树的形式展示出来。常用的句法分析方法是规则的方法和统计的方法。基于规则的方法是以语言学理论为基础,从构词法、短语构造法、句子构造法等最本质的特征出发[2]。文中在掌握哈萨克语词法分析的基础上,为了减少冗余边,提高准确率,引入规则库优化的改进线图分析算法,最终设计出句法分析模型。通过实验

收稿日期:2014-11-14 **修回日期**:2015-02-16 **网络出版时间**:2015-08-26
基金项目:国家自然科学基金资助项目(61063025,61363062)
作者简介:牛 娜(1989-),女,硕士,研究方向为自然语言信息处理;古丽拉·阿东别克,教授,博士生导师,研究方向为自然语言信息处理、人工智能等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150826.1603.070.html

测试表明,该算法在大大缩短运行时间的基础上使得准确率提高了 4.19%。

哈萨克族是新疆的主要少数民族之一,哈萨克语属阿尔泰语系突厥语族,在形态结构上属黏着语类型[3]。哈萨克语结构特点[4]:

(1)哈萨克语句按"主—宾—谓"的顺序排列,因为主语和宾语的词性特点,在分析时容易产生歧义,这给研究带来一定的难度,得到的分析树可能会出错。

(2)哈萨克语单词都是由词干和词缀组成。

(3)哈萨克语句法结构相当严谨。主要表现在句法结构边界明晰、名词具有双重身份标记、词类与句法成分对应相对严整、句法结构的转换机制系统严密等方面[5]。

(4)哈萨克语的书写形式是字母,词和词之间是分开的。哈萨克文是从右向左横写,每个字母没有大写小写之分,字母是由笔画和标点符号组成,书写不当都会影响字母的形体,而造成无法拼读或改变词义的结果[6-11]。

1 基于规则的句法分析算法

1.1 Earley 算法

算法核心是按照句子中各元素的顺序,相应地做出分析表序列。分析表中各个项目是通过 Earley 算法中的三种运算,即预测运算、完成运算和扫描运算[12-13]得到的。

(1)预测(Predictor)运算:预测可以生成新的状态,把带有点的非终结符进行扩展;

(2)完成(Completer)运算:把已分析完成的字符串加入到点规则所指向的字符后面,并后移点,把新的状态加入线图;

(3)扫描(Scanner)运算:把需要匹配的词进行规则匹配,如匹配则后移点,更新状态加入线图。

Earley 算法的描述:

S 是文法初始符,S' 是新增非终结符,表示空集。

给定输入字符串:$X = x_1, x_2, \cdots, x_n (x_i \in \mathrm{VT}, i = 1, 2, \cdots, n)$;

① $S = \{[S' \rightarrow \cdot S]$(初始项目)$\}$;

②for $i = 0$ to n

do

顺序处理每一个字符串 $s \in S$,对相应字符串依次使用如下可行的操作,直至不产生新的项目;

预测;

完成;

扫描;

if $S+1$ = NULL 退出循环;

end for

③if $i = n$ and$[S' \rightarrow S \cdot, 0] \in S + n$,生成句法 S

else 跳出

1.2 GLR 算法

GLR 算法伪代码[14-18]:

给定上下文无关文法<VN,VT,P,S>,分析表

把初始化的字符串 $T = T_1 T_2 \cdots T_n$ 放入缓冲区并准备入栈

图栈 CS = null,共享树 T = null,Action 表 A = 图栈中所有栈的项目;

Step1:初始化。把图栈的所有栈顶按先进后出的方式存入 A 中,分析指针指向带分析的系统终结符位置,清除终止标识符。

Step2:从 A 中取出一状态,查 Action 表中以状态为行,以待分析字符串 T_i 为列的格子动作,设为 X;

if X = 移进,将当前状态及当前符号入栈,分析指针下移;

if X = 规约,检查规则文法。若匹配,则组成当前非终结符的句法结构树,并压入符号栈;弹出状态栈各中间状态;查 Goto 表,将新状态压入状态栈。同时,将中心词指针指向相应的中心词。若条件不满足,则终止;

if X = 终止:置终止标志;

if X = 接收:将符号栈顶句法树弹出,Return;

if X = 出错:恢复初始状态,Return

Step3:重复以上步骤,直到 A = Null。

1.3 Chart 算法

Chart 算法[19]的数据结构主要有:

(1)活性边:如果文法右部的句法未被完全匹配,则称这条规则为活性边。

(2)非活性边:如果文法右部的句法被完全匹配,则称这条规则为非活性边。

(3)Agenda:用来存储非活性边或词的词性、词的左间隔点和右间隔点。

(4)Activearc:用来记录活性边的集合。

算法流程图如图 1 所示。

2 改进的线图分析算法

2.1 传统线图算法的不足

线图分析算法虽然避免了回溯的过程,但是在实际句法分析过程中,存在这种情况:首先确定了 β 的句法成分,根据点规则形成活动边 $A \rightarrow \beta A$,继续搜索匹配,由于规则库中存在大量 $A \rightarrow ij$ 规则,就会形成许多条第一个儿子为 β 的活动边。这样句法分析过程中规则匹配循环次数会进行很多次,大多数的无用活动边也由此产生。由于存在大量的活动边带来了更多可能的结构需要搜索,基于线图的句法分析器的速度就不

能满足实际需求。

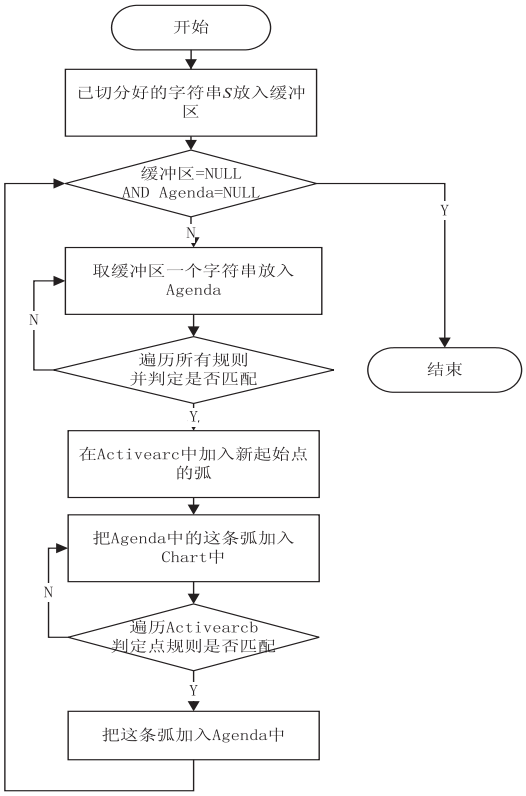

图 1　Chart 算法流程图

2.2　改进算法的描述

为了减少冗余边的产生,节省出较多时间和空间,达到提高句法分析的准确率的目的。文中提出了规则库优化方法,即消除文法左递归。在线图分析算法运行前首先将这种存在左递归的规则进行压缩存放,在规则库中生成 $A \rightarrow \beta ij$ 的规则,消除 $A \rightarrow \beta A$ 的规则,这样在算法匹配过程中会减少循环次数,无用的活动边的产生数目大大降低,算法时间效率明显提高。

改进算法的伪代码描述如下:

Input:

RuleNum:规则库规则数

RuleList:带有左递归规则

RuleBase:规则库

Output:NewRuleBase:新规则库

Begin:

Step1:从第一条规则文法开始遍历,用规则 i 的左标识匹配规则 j 的左标识,for $i \leftarrow 0$ to RuleNum

Step2:判断规则 i 的左部与规则 j 的左部是否相同并且规则 i 为左递归规则

RuleList←This. Rule←Rule[i]

This. Rule. left. tag = Rule [j]. left. tag&&Both (Rule [j]. right. tag = ε)

find location(This. Rule)

Step3:将两条规则文法合并为新的规则并加入规则库

create new Rule(This. Rule,Rule[j])

RuleBase. Add(NewRule)

RuleBase. Remove(RuleList)

Step4:直到所有规则被遍历完

End

2.3　哈萨克语句法分析器的设计

哈萨克语句法分析的最终目标是满足大规模真实文本的高效和准确兼顾的分析。针对现有的基于规则的自然语言句法分析器的研究,大致总结出建立准确率较高的哈萨克语句法分析器须考虑如下几个问题:

(1)基于规则的句法分析算法的选择。针对上述实验分析比较,选择线图分析算法作为句法分析算法,根据词与词之间的搭配关系(即句法规则),把输入的词序列利用算法思想最终规约为一棵句法分析树。

(2)需要建立知识库,该知识库应该包括规则库和电子词典两个。建设哈萨克语规则库主要通过哈萨克语语言专家人工总结,这样构造的规则库缺点是有很大的不完整性。文中实验采用常用的哈萨克语的典型句型作为研究对象,总结出哈萨克语句法产生规则。在常用句型的基础上总结出了 70 条句法规则和 5 种常用短语类型进行测试。所用词性标记集如表 1 所示。

表 1　哈萨克语词性标记集

词性	说明	举例
名词	n 表示人或事物的名称	تعلەك “愿望”
动词	v 表示人或事物的动作	عندع “扫”
形容词	adj 表示人或事物的特征	بايلاولع “捆着的”
数词	num 表示人或事物的数量	ٷكەدەن-ٷشتەن “三三两两”
副词	adv 修饰动词或形容词	ايرەقشا “特别地”
代词	pron 用来代替名词、数次或形容词等	ول “他”
摹拟词	ono 表示人或事物的声音	ارسعلدا “汪汪”
感叹词	int 表示呼应或情感	ويصعرعماي “天爷呀”

哈萨克语常用短语类型有:

· 名词短语(Noun Phrase,NP);

· 动词短语(Verb Phrase,VP);

· 形容词短语(ADJective Phrase,ADJP);

· 副词短语(ADVerb Phrase,ADVP);

· 数词短语(NUMeral Phrases,NUMP)。

(3)句法分析器输入的是经过分词、词性标注和兼类处理的词序列,输出的是一棵句法分析树,因此前期的词法分析模块必须具有较高的准确率,否则直接影响后期句法分析模块的正确率。

句法分析器设计过程中使用的关键技术包括:

(1)词法分析模块中的分词模块,文中采用的是在最大词长匹配算法的基础上输入一个简单哈萨克语句子,利用该算法将句子分解成一个个独立的词序列。

(2)词法分析模块中的词性标注,采用相对频率训练的统计方法,从已标注好的语料库中训练得到概率参数,进而对给定的句子进行词性标注的过程。

(3)以改进的线图分析算法(即引入消除左递归文法)作为核心算法分析哈萨克语句子,最终得到对应句子的句法分析树。

句法分析器的功能模块图如图 2 所示。

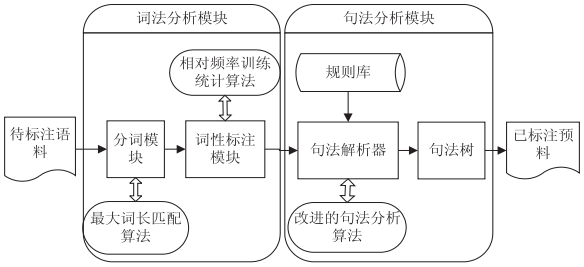

图 2 哈萨克语句法分析原型系统功能模块图

3 实验结果与分析

3.1 Earley、GLR 和 Chart 算法对比分析

衡量算法的两个重要指标为时间复杂度和空间复杂度,因此本实验在从新疆日报(哈萨克语版)电子版(20 万多词)中选取 16 条已经预处理的句子进行测试时,将空间复杂度和时间复杂度作为选取研究算法的评价标准。总词长为 105,平均句长为 6.6 个词的语料,针对以上三种算法进行句法分析得到的实验结果如表 2 所示。

表 2 三种常用句法分析算法对比表

	GLR	Earley	Chart
句子数	16	16	16
平均句长	6.6	6.6	6.6
词数	105	105	105
时间/ms	6 703	230 098	234 624
时间复杂度	$O(kn)$	$O(n^3)$	$O(n^3)$

线图分析算法的空间复杂度较低,因为它无需构造类似 LR 那样的静态分析表的数据结构,但是分析过程中时间效率不高。而 GLR 算法恰巧和它不同,其需要花费一定空间存储分析表,但时间复杂度是优于 Chart 算法的。由上表可看出,GLR 算法运行时间较其他两种快很多,但 GLR 算法使用图栈和共享节点包的方式来节省存储空间,这样换取了少量时间。但 GLR 算法仍需要从共享节点包回溯找出所有的句法分析树,所以要额外花费一定数量级的时间。由于哈萨克语句法分析处于初级阶段,句子长度较短,语料规模不大,不必花费较大的空间换取时间。而 Earley 和 Chart 相比较都是运用了点规则的方法,二者相似,但后者更简单、可视化更好,可以对生成的句法树一目了然。综上所述选取 Chart 算法进行下面的实验。

3.2 传统 Chart 与改进算法的对比分析

实验所选句子和规则库同上文实验,句法分析的评价指标如下:

正确率:得到分析准确句法树的句子数与实验语料总句数的比值。

CPU 时间:分析输入句子,算法花费的所有 CPU 时间。

活动边数目:活动边数目的多少描述了占据空间存储的大小,如果活动边数目降低,利用空间换取时间的概念体现出算法时间效率的提高。

实验过程中所有程序的开发及实验测试均是在 Windows 7 Professional 操作系统平台上,程序的编写使用 VS 2010(Microsoft Visual studio 2010)开发环境,C# 编程语言实现。结果如表 3 所示。

表 3 传统 Chart 与文中算法对比

	Chart(传统)	改进后 Chart
规则数目	28	53
句子数	16	16
平均句长	6.6	6.6
活动边数目	3 424	2 480
时间/ms	234 624	11 488
正确率/%	68.52	73.71

结果表明,改进线图算法在处理自然语言时具有明显的优势,其贡献在于减少了活动边生成数目,压缩了分析过程中的句法树,从而达到:

(1)活动边数目减少,相比传统 Chart 算法,减少了 944 个;

(2)算法运行时间大大减少,相比传统的 Chart 算法,改进后的算法运行时间缩短了 20 倍;

(3)句法分析准确率提高。改进算法将准确率提高了 4.19%。

3.3 改进算法的句法分析器的实现

基于改进线图算法的哈萨克语句法分析系统的处理过程为:首先输入哈萨克语句子,通过词法分析模块进行分词并标注出相应词性,然后根据这些词性序列查找规则库系统得到此句子所匹配的文法规则,最后采用改进的 Chart 算法进行句法分析,输出句法树。

根据句法分析原型系统的工作原理分析一条哈萨克语句子:

ول شىگىن جامىلىپ سىرىتقا شىقتى

经过分词模块得到如下结果:

<word pos="pron" stem="ول" affix=" " var= "1" > ول
</word>他、她、它
<word pos="n" stem="شىگىن" affix="گىن"var="0">شىگىن
</word>外套
<word pos="v"stem="جامىلىپ"affix="ىپ" var="0">جامىلىپ
</word>披着
<word pos="adv" stem="سىرت" affix="قا" var="0">سىرتقا
</word>外面
<word pos ="v" stem=" شىق" affix="تى" var="0">شىقتى
</word>出去
<punction>.</punction>

符号 pos 表示单词的词性,stem 表示单词词干,affix 表示单词词缀;对于词类的标记用符号 var 来表示(var 值为 0 时,表示这个词的词性唯一;var 值为 1 时,表示这个词为兼类词;var 值为 2 时,表示这个词为未登录词)。

规则库:S→pron VP,VP→VP VP,VP→n v ,VP→adv v ;

其中,S 代表句子,pron 表示该词词性为代词,n 表示名词,adv 表示副词,v 表示动词,VP 表示该短语为动词短语。

通过改进线图分析算法得到例句的句法分析树,如图 3 所示。

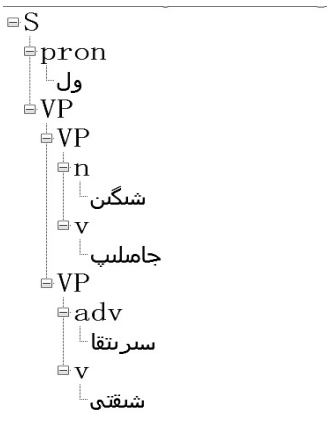

图 3　句法分析树

4　结束语

哈萨克语句法分析是哈萨克语信息处理学科中的一门重要课题,它的快速发展将带动基于句法的语义、情感的应用和发展。文中通过 Earley、GLR 和 Chart 三种算法的对比研究,选出 Chart 算法作为刚起步不久的哈萨克语句法分析研究的核心算法,针对其活动边冗余现象提出消除文法左递归的改进算法。对比实验结果表明,比起传统算法改进算法在时间方面,缩短了 20 倍;准确率方面,提高了 4.19%。结合改进算法和句法分析系统的功能分析,设计出哈萨克语句法分析器。该系统能够针对简单哈语句子分析得到完整分析树,不仅为哈语句法分析拉开新的篇章,也将为下一步完善哈萨克语句法分析系统奠定基础。

随着语料库的建设和统计模型的成熟,建立以规则和统计模型一体化的原型系统,进一步提高句法分析的效率和实用性将是哈萨克语句法研究很好的方向。

参考文献:

[1] Wright J H. LR parsing of probabilistic grammars with input uncertainty for speech recognition[J]. Computer Speech and Language,1990,4:297-323.

[2] Brian R. Probabilistic top-down parsing and language modeling[J]. Computational Linguistics,2001,27(2):1-28.

[3] 王　鹏,戴新宇,陈家骏,等. 基于规则的汉语句法分析方法研究[J]. 计算机工程与应用,2003,39(29):63-66.

[4] 吐尔根·依布拉音,袁保社. 新疆少数民族语言文字信息处理研究与应用[J]. 中文信息学报,2011,25(6):149-156.

[5] 王　花,古丽拉·阿东别克. 基于语料的哈萨克语词频统计研究[J]. 计算机工程,2010,36(24):59-61.

[6] 玛依来·哈帕尔,古丽拉·阿东别克. 哈萨克语文本分类系统的设计与实现[J]. 计算机工程,2011,37(5):196-198.

[7] 孙瑞娜,古丽拉·阿东别克. 哈萨克语基本名词短语自动识别研究与实现[J]. 中文信息学报,2010,24(6):114-119.

[8] 冯鲸华,古丽拉·阿东别克,吴守用,等. 基于位置概率模型的哈萨克语人名识别[J]. 计算机应用与软件,2010,27(12):21-23.

[9] 桑海岩. 哈萨克语固定词组自动抽取[D]. 新疆:新疆大学,2013.

[10] 张定京. 现代哈萨克语实用语法[M]. 北京:中央民族大学出版社,2004.

[11] 张定京. 哈萨克语语法结构特点概要(上)[J]. 语言与翻译,2010(2):3-8.

[12] 赵志国. 两个真歧义句的 Earley 算法演示[J]. 连云港师范高等专科学校学报,2013,30(4):48-55.

[13] Earley J. An efficient context-free parsing algorithm[J]. Communication of ACM,1970,13(2):94-102.

[14] 许　福,金茂忠,陈志泊,等. 面向软件逆向工程的 GLR 优化算法[J]. 计算机工程,2013,39(6):12-20.

[15] Zhou H P,Wang T,Chen H W. Using LR algorithm to analyze the grammar relations of Chinese[J]. Journal of Software,1999,10(9):967-973.

[16] Aycock J,Horspool N,Janousek J,et al. Even faster generalized LR parsing[J]. Acta Information,2001,37(9):633-651.

[17] Tomita M. An efficient augmented-context-free parsing algorithm[J]. Computational Linguistics,1987,13(1-2):31-46.

[18] 朱敬国. 基于 GLR 算法的维吾尔语句法分析研究[D]. 新疆:新疆大学,2011.

[19] 卢俊之. 基于语法功能匹配的句法分析算法[D]. 南京:南京师范大学,2013.

基于规则的哈萨克语句法分析算法研究

作者： 牛娜， 古丽拉·阿东别克， NIU Na， GULIA·Altenbek
作者单位： 新疆大学 信息科学与工程学院,新疆 乌鲁木齐,830046
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(9)