现代汉语分析系统UCAS的研究与实现

朱靖波　王宝库　姚天顺
东北大学计算机科学研究所
辽宁.沈阳 110006

【摘要】本文介绍了一个汉语分析系统UCAS。主要介绍汉语分析系统UCAS的知识集及其机内组织形式、基于优化图学习的分析机制、知识表达和存储管理。最后本文简单的讨论了系统的二次开发可行性以及系统的展望。

关键字：ＳＯＣ算法，汉语分析器，扩展PS&U规则

Design and Research of Chinese Analysis System UCAS

Zhu Jingbo, Wang Baoku & Yao Tianshun
Dept. of Computer Science and Engineering
Northeastern University
Shenyang, 110006
P.R.China

ABSRACT: This paper introduces a Chinese analysis system UCAS, and main discussed knowledge set, and parsing mechanism based on SOC algorithm, knowledge expression andstorage management. At last will discuss improvement and forecast of UCAS system.
KEY WORDS: SOC algorithm, Chinese parser, extended PS&U language

一、前言

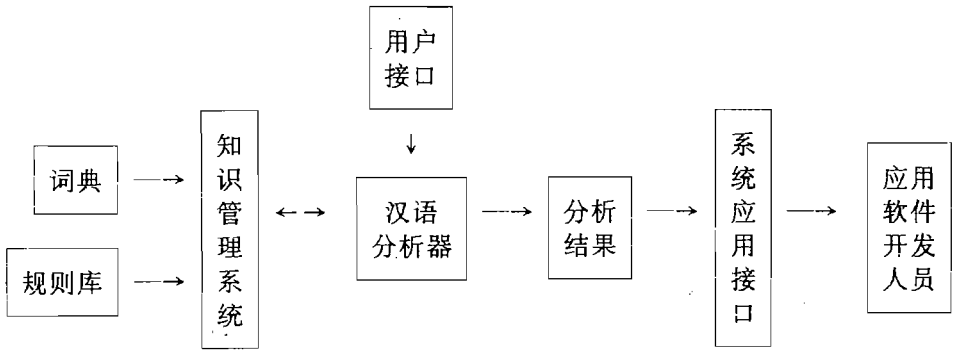
在中文信息处理的许多领域中，汉语分析被认为是至关重要的‘瓶颈’问题，计算机要实现对汉语的理解，则必须建立一个符合计算语言学要求的汉语知识库和分析器。本文介绍的汉语分析系统UCAS的知识库由汉语词典和汉语分析规则库组成，分析器则由预分析器、控制器和规则解释器组成。由于系统的知识库模型是建立在复杂特征集结构基础上的，使得分析器在工作过程中能够方便地进行复杂特征集的“建立”及“合一”运算。

为了建立一个十分高效的汉语分析机制，我们从优化解析序列图的构造和搜索的角度提出一种十分有效的基于上下文无关文法的自然语言分析算法SOC算法，其中引入基于复杂特征集的消歧机制，自学习机制和优化的图搜索策略，以实现对汉语句子的有效分析。

系统的各子部分都曾有专文详细论述过，词典和句法规则体系曾在文[2,3,5]中论述过，分析机制和知识表示曾在文[1,4,5]中论述过。本文将全面论述汉语分析系统UCAS，最后讨论一下系统的展望。

二、系统体系结构

UCAS系统目前已经在中文Windows 3.1环境下开发成功，主要包括三个部分：一是知识库管理模块，二是分析器模块，三是系统应用接口模块。系统的体系结构如下：



三、知识库模型

§3.1 词典

为了适应计算机的处理，系统将词典分为源词典和目标词典。源词典包括句法词典、语义词典和框架词典，它以数据库的形式存放。而目标词典指系统分析使用的词典。目标词典信息来源于源词典，两者的结构不一样，因此需进行转换。词典通用转换机制参见文[1]。一个汉语机器词典的构造好坏对汉语分析系统的效率和质量影响很大。为了提高系统的分析效率，分析词典（目标词典）采用统一的目标词典结构，该词典共包括五类信息，其结构如下：

词条	语法信息	语义信息	搭配信息	其　它

例如词语：

＜通知＞：

词条：　拼音=tongzhi/同音=2/同音调=2/同字词=2/信息出处=8/音节数＝2/字数=2

语法信息：

（词法属性）体谓=谓/虚实=实/后动量词=动/兼类=N/在正在=在/间接可无=可

（句法属性）动结=结/兼语句=兼/双宾=双/体谓准=体|谓/有的宾语=有/小句宾=句/前体后句=句/直宾可无=可

语义信息：

（语义类别）使令/222145

（语义约束）目的=[事件]/施事=人类/受事=人类/材料=-/成果=-/程度=-/当事=-/对象=-/方向=-/分事=-/基准=-/客事=-/类事=-/领事=-/履事=-/属事=-/指事=-/终点=-

搭配信息：着了过=了|过/格标1=把/受事=受/了=A/着=C
其它：　　~完

§3.2 句法规则体系

由于汉语的独特特点，使得在分析汉语时会有好多不同于西文分析的问题出现。几年的汉语分析研究使人们越来越意识到要彻底分析汉语句子，简单的语法信息是不够的，还须在同时加入语义的信息。加入的方法就是对于上下文无关文法描述的短语结构规则，先用句法属性进行约束。若仍有歧义(一般来说这是肯定的)，就再用语义属性进行约束。所有约束均被满足后结合成的新结构的句法和语义属性，通过“属性传递”来得到。这种描述自然语言的句法功能，语义关系等对短语结构的约束的一套形式化语言就是规则描述语言，简称扩展PS&U语言[2][3]。

句法体系不包括词性判别和词义选择规则，只描述短语和句子的结构规则以及句法和语义上的约束。目前短语结构规则的产生式已达480左右条。一个扩展PS&U单元对应一个短语文法产生式文件，用以存放其约束规则。

扩展PS&U的规则文件的基本形式：

<短语文法产生式>:<规则1>.<规则2>. ...<规则N>

每个规则的基本形式又是：

属性约束1，约束2，...，约束n；属性传递1，传递2，...，传递n

各约束式是“且”的关系，只有均约束成功才能执行传递动作。约束(传递)包括属性值约束(传递)、结点的约束(传递)等。它们都是通过CFS的路径来描述的。

下面给出一些规则示例 (Ctr表语法树路径，Net表示语义网路径)：

Vp <-- Vm Np:	
1: <Vm Ctr Hed Syn 体谓准>= 体&~准,	相当于Vm.Hed.体谓准=体&~准，即要求Vm的Hed子结点可带体词性宾语而不可带准谓词宾语
<Np Ctr Hed Syn 词类>=~V, <Vm Ctr Hed Evl 受事 语义限制> =<Np Ctr Hed Sem核心语义>,	Np的Hed子结点的词类不能是动词 要求Np.Hed能做Vm.Hed的受事
<Vm Net Ker 受事>= Null; <Vp Ctr>:=<Vm Ctr>,	Vm.Ker没有受事子结点 上面约束式全部通过就做传递动作．此处表示将Vm.Hed作Vp1的Hed子结点
<Vp Ctr Obj>:=<Np Ctr>, <Vp Net Ker>:=<Vm Net Ker>, <Vp Net Ker 受事>:=<Np Net Ker>.	将Np.Hed作Vp的Obj子结点 将Vm.ker作Vp的Ker后继结点 将Np.Ker作Vp.ker的受事后继结点

四、基于优化图学习的分析机制

系统采用基于优化图学习的自然语言分析算法——SOC算法[4][5]。下面简单介绍一下系统的分析机制。

例子：

句子1：决定(1:V)战争(2:N)胜负(3:N)的(4:U)是(5:V)人(6:N)。(7:W)

其中V、N、U、W分别表示动词、名词、助词、标点符号，数字表示词结点号。

为了描述方便，我们引进四元组结构来描述汉语的分析过程，定义如下：

G=(BNO, ENO, PHRSYM, CURRSYM)

其中，BNO、ENO表示词结点位置，常用结点号表示。PHRSYM表示归约动作产生的结果，也称作短语类。CURRSYM表示执行归约动作时的当前符。

§4.1 解析序列

我们定义解析序列为：描述句子分析过程中不同词之间合并的顺序。这样句子1的某一个解析序列可以用四元组结构来描述如下：

1:(1,1,Vm,N)	7:(5,5,Vm,N)
2:(2,2,Np,N)	8:(6,6,Np,W)
3:(3,3,Np,U)	9:(5,6,Vp,W)
4:(2,3,Np,U)	10:(1,6,Sp,W)
5:(1,3,Vp,U)	11:(1,7,Ss,#)
6:(1,4,Dep,V)	12: OK

图 1 一个句子1的解析序列

§4.2 解析序列树

但由于汉语文法的LR分析表存在多入口问题，这必然导致一个句子存在二个或二个以上的解析序列。我们可以引入有向树结构来描述某句子的所有解析序列，称之为解析序列树。由于篇幅的关系，解析序列树的特性和描述方法参阅文[4]。

§4.3 解析序列图

对于基于解析序列树的分析算法，找全解的复杂度为N，N为解析序列数目。由于上下文无关文法的冗余性造成一个句子的解析序列数目随着长度将指数上增，这样造成分析器效率很低。为了提高分析效率，我们可以将解析序列树进行优化压缩，压缩的原则将BNO、ENO、PHRSYM、CURRSYM相同的结点合并，则形成解析序列图。经过统计分析，基于解析序列图的分析算法找全解的复杂度平均将下降为原来的十分之一。句子1的解析序列图部分可表示为(见下页)：

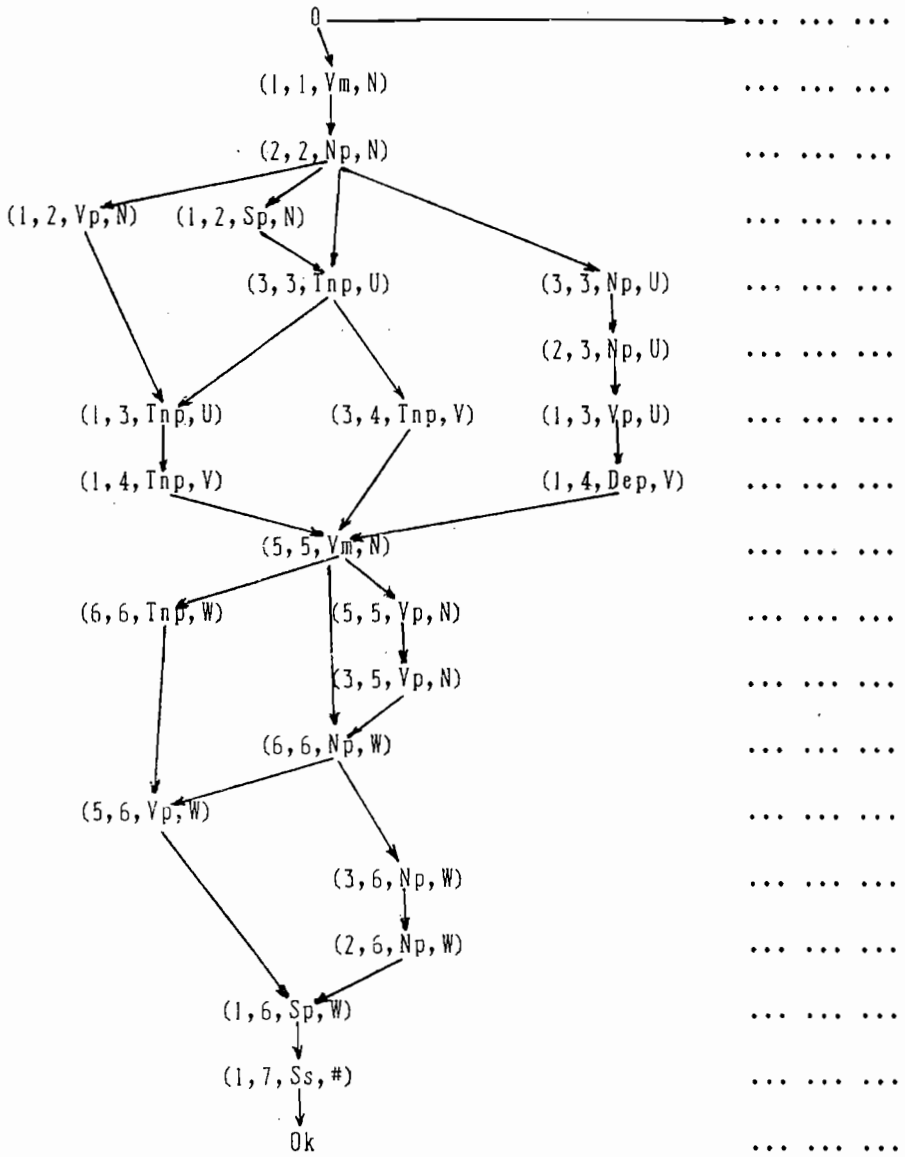


图 2 句子1的解析序列图的描述

§ 4.4 实现技术

归根结底，ＳＯＣ算法思想主要在于如何最快地寻找从0节点到OK节点的所有正确通路。为了解决由于汉语文法的非LR性和冗余性造成LR分析表的多入口问题和提高分析效率，基于上述结论，ＳＯＣ算法主要采用两种技术：优化构造解析序列图和优化搜索解析序列图。优化构造解析序列图采用动态构造解析序列图，允许在分析过程中随时对解析序列图进行

调整；优化搜索解析序列图采用宽度优先搜索驱动深度优先搜索策略搜索解析序列图，同时对每个规则动作调用相应的规则进行约束测试，一旦测试失败，从解析序列图中找到失败的断点，采用断点回溯技术[4]搜索其他解析序列；如果测试成功，则执行相应的拉树拉网动作，构造句法关系和语义关系。

五、分析结果的表示和存储

§5.1 分析结果的表示

系统采用具有复杂特征集的语法树、语义网综合在一起的数据结构，简称“综合网”SENT结构，作为分析和生成的信息载体，SENT的定义如下：

- SENT有且仅有一个主结点M；
- 如果任一结点T有儿子，则每一个儿子也是一个SENT；
- SENT的任一结点都有n个儿子（或后继）$n \geqslant 0$；
- SENT的非主结点有且仅有一个父亲和m个前趋 $m \geqslant 0$；
- SENT的每个结点都具有复杂特征集；
- 父亲（或前趋）结点和儿子（或后继）结点间存在关系结点。

从SENT结构可以看出，汉语分析过程实际上是线性链表到树和网的转换过程。如句子"我去过中国."的语法树和语义网的平面结构如下：

```
我去过中国(Sp,A1)
!
!-(Hed)->去过中国(Vp,A1)
!        !
!        !-(Hed)->去过(Vm,A1)
!        !        !
!        !        !-(Hed)->去(Vm,A1)
!        !        !        !
!        !        !        !-(Hed)->去(V,A1)
!        !        !
!        !        !-(A)->过(U)
!        !
!        !-(Obj)->中国(Snp)
!                 !
!                 !-(Hed)->中国(S)
!
!-(Sbj)->我(Np)
         !
         !-(Hed)->我(R)
```

如结点(Hed),(Obj),(Sbj)等是中间结点。

我去过中国（叙述性事件，从近处到远处．，Asp=曾经）

└(Ker)→去（自移/222141，从近处到远处．）

　　├(趋向)→中国（政区/14311）

　　└(施事)→我（人/121111）

。（，陈述句末尾停顿）

§5.2 分析结果的存储

系统将分析结果转换成属性－值系统描述，简称AV-描述。系统将分析结果采用AV-描述方式存储，希望能向访问机器词典一样去访问分析结果的静态和动态属性信息。

例如访问上例中句子的语义关系为“Ker”的结点，即可通过访问#10.Ker=#1实现。其中#10是词结点Sp的代码，#1是词结点“去”的代码，同时还可以通过#1.属性名的方式去访问“去”的其它静态属性和动态属性信息。

还有，系统为了具有二次开发能力，在系统应用开发接口提供一些访问分析结果的原子函数。由于系统是采用ＯＯＰ技术开发，用户和应用软件开发者可以通过继承原子函数来构造更加复杂、功能强大的应用开发接口。由于篇幅关系，应用接口开发具体实现技术将另文介绍。

六、结束语

整个系统采用软件工程的思想，采用开放性的设计方法，并设有标准接口，提高了系统的可维护性和可利用性。但汉语分析系统本身不但是一项应用软件工程，更重要是中文信息处理的基础工程。目前UCAS系统也在不断完善之中，准备通过使用，不断总结经验，对汉语分析系统设计和实现机制进一步改造，提高系统的分析效率。

参考文献

[1] 王宝库，张东茉，姚天顺，"一种基于复杂特征集的汉语分析器的设计"，<<1992 International conference on Chinese Information Processing(I)>> October 26-28,1992

[2] 郭松，王宝库，"汉语分析中的扩展PS&U规则描述语言及其解释机制"，<<中文电脑国际会议'94>>，1-4 June 1994.

[3] 李东，陈志明，"规则描述语言及汉语的句法规则体系"<<1992 International Conference on Chinese Information Processing>>

[4] 朱靖波，王宝库，候正茂，"一种基于优化图操作的自然语言分析算法—— SOC 算法"，<<中文电脑国际会议'94>>，1-4 June 1994.

[5] 朱靖波，王宝库，姚天顺，"一种基于优化图操作的汉语分析机制的研究与实现"，<<东北大学学报>>，自然科学版，电脑专辑增刊，vol 16，1995