一个基于合一的汉语句法分析器UBCP的实现

栾 浩 (烟台大学计算机系)
黄昌宁 (清华大学计算机系)

【摘要】本文首先介绍了一个基于合一的汉语句法分析器UBCP的模型。然后介绍了UBCP的形式语法体系以及电子词典的组织结构，最后说明了UBCP的基本设计思想和分析结果的输出形式。

UBCP采用伪并行的不确定分析算法来控制整个分析过程，它将语法制导和词汇驱动有机地结合起来，同时由于引入了伪合一算法和预编译等技术，使整个系统的分析效率明显提高。UBCP系统是在SUN SPARCⅡ工作站上用C语言实现。

关键词：

汉语句法分析器，复杂特征集，合一算法。

The Implementation of an Unification-Based Chinese Parser

【ABSTRACT】In this paper, we put forward a basic model for Unification Based Chinese Parser(UBCP). Then we present the formal grammar and expound how to organize the electronic dictionary in UBCP.

UBCP supports a non-deterministic pseudo-parallel algorithm. It combines syntax-directed controling strategy with lexicon-driven controling strategy unitily. By taking pseudo unification, pre-compiling technologies and so on, we have improved the efficiency of UBCP dramatically. UBCP is running on SUN SPARCⅡ workstaion.

KEYWORDS:

Chinese parser, complex feature set, unification algorithm.

一. 综述

UBCP的构成包括两部分：一是描述汉语语言学知识的知识库，它由电子词典和规则库组成；二是相应的分析控制机制，它由含有合一算法的中央控制器和规则库的预编译器组成。这两部分相对独立，从而改善了系统的可维护性。

UBCP的总体结构框图如图1所示，在下图中，用方框表示处理模块，用带双边的方框表示系统资源和输出结果。

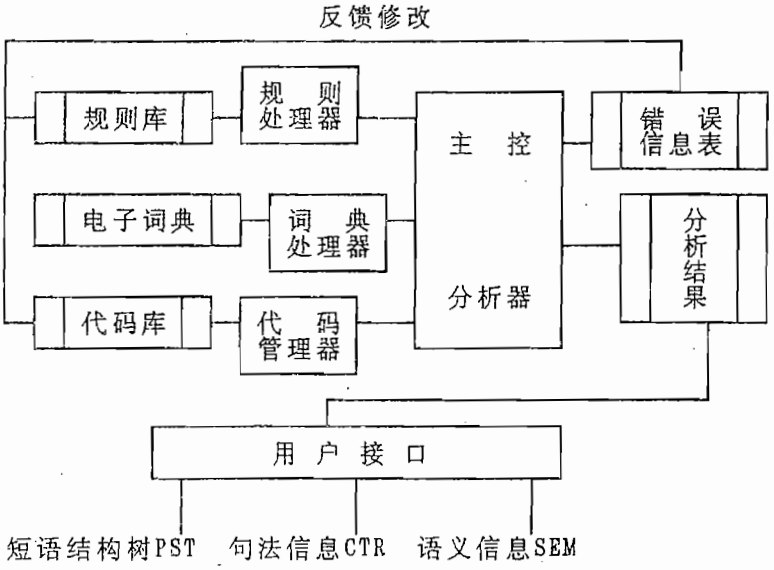


图1. UBCP的总体框图

二. 形式语法体系

考虑到对上下文无关文法存在着高效率的分析算法(如Tomita算法等), 同时参考PATRⅡ文法的表示方法, 在UBCP系统中建立了一种以上下文无关文法为基础, 结合进行上下文约束测试的形式语法体系G, 它可定义为一个六元组: G=(VN, Vt, P, S, AC, AT), 其中, VN, Vt, P; S 与上下文无关文法中相应的符号含义相同, 分别表示非终结符集合、终结符集合、产生式集合以及开始符号, 下面重点介绍新引入的两个符号的含义。

1. 属性约束(ATTRIBUTE CONSTRAINS 简称AC)

属性约束部分的功能是: 充分利用系统中的各类知识资源, 通过对语言单位上下文属性约束描写来体现某短语或句子中各个成分之间的相互制约关系。属性约束是由下面两种形式的若干等式构成的序列:

a. <路径>=<路径>

b. <路径>=原子表达式

其中, <路径>是由形如<f1 f2 ...fn>的属性名组成的, 且<路径>中的属性名fi 既可以是静态属性又可以是动态属性; "原子表达式" 是指原子值和逻运算符'‖'(或)和 '~'(非)构成的表达式。

另外, 任意两个属性约束之间用符号',' 分隔, 它表示两个属性约束之间具有逻辑'与'的关系, 因此, 表达式中的运算符'‖'和'~'再加上',' 便可以构成完整的逻辑运算, 从而保证了任何复杂的属性约束均可以用上述两个等式的组合来加以描述。

2. 属性传递(ATTRIBUTE TRANSFER简称AT)

属性传递用来描述当规则归约成功之后, 新所形成的语法单位的属性。它是两个次

一级语法单位的'合一'结果。即它描述了动态复杂特征集的形成过程，这包括句法成分树、语义网络、语用等动态信息。属性传递是由下面两种形式的若干等式构成的序列：

a. <路径>:=<路径>

b. <路径>:=值

其中，<路径>的概念与AC中类似，只是在此仅仅涉及动态属性名；'值' 是一原子符号，它或者是电子词典中的静态属性值，或者是分析过程中所产生的动态属性值。

综上所述，规则库中每条规则由三部分组成，其形式可描述如下：

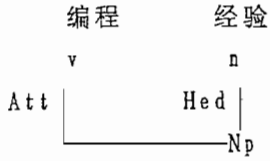
<上下文无关文法 > ：

<属性约束>;

<属性传递>。

3. 规则应用示例

[例] 句法歧义结构的判别：他有一定的<编程经验>。




```
Np ----> v  n;
< v Syn 后名>=可,    **(1)
< v Syn 外内>=内,    **(2)
< n Syn 前动>=可;    **(3)
< Np Ctr Hed>:=<n>,
< Np Ctr Att>:=<v>。
```

注：**(1). 动词可以修饰后跟的名词；(2). 动词必须为不及物动词；(3).名词可以受前面动词的修饰 ；符合以上条件约束的动名组合为定中结构(NP)，而不是述补结构。

在上例中如果单纯地利用上下文无关文法，无法判断'v+n'的序列是NP还是VP,但是如果同时利用句法知识对上下文进行约束就可以排除句法结构的歧义。

三.电子词典

1.词典的组织

UBCP的电子词典每个词条具有如下四部分的信息：

. 词条：描述词形，拼音，音节数，使用领域，使用频率，切分歧义，...

. 语法信息：词性，词性兼类，词法属性，句法属性，...

. 语义信息：语义分类，语义指向， 语义特征，...

. 动词格框架：格名，格约束等。

由于分析的需要，我们还在电子词典中存储了对名词进行语义分类后所得到的语义分类树，该分类树描述了名词之间的层次关系，这是常识知识，如下图所示：

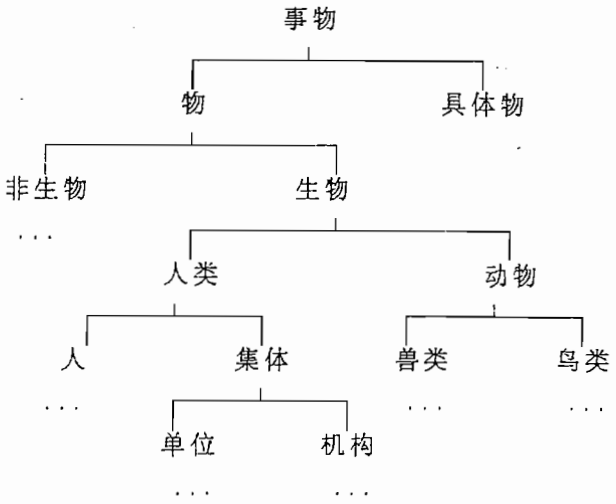


图2. 事物类概念的层次关系

从上图中我们可以知道：如果知道了学生是'人'，则学生就具有其上位概念'人类''生物'、'物'和'事物'所具有的一切属性。

四. 知识库的控制机制

1. 电子词典的控制

在UBCP中，我们利用逻辑合成代替物理合成实现了对其三部电子词典的合并。另外由于UBCP的词典容量为6万词条，为了提高对电子词典的访问效率，我们建立了如下所示的二级索引：

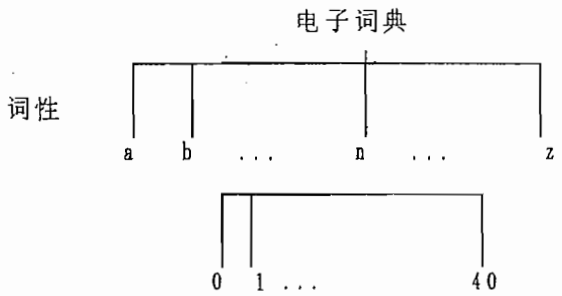


图3 电子词典的索引结构

其中，第一级索引据词性将整个电子词典划分为20个子词典，第二级索引是据词条第一个汉字内码的区号建立的，从而得到第0区到第39区，另外将所有二级汉字放在第40区中。这样一来，在UBCP的电子词典中检索一个词条的平均比较次数仅为6次。

2. 规则库的控制

UBCP的规则库按上下文无关文法部分划分为规则块，即将上下文无关文法部分相同

的规则放在一个规则块中，然后在规则块内根据规则的优先级由高到低进行排序，如下图所示：

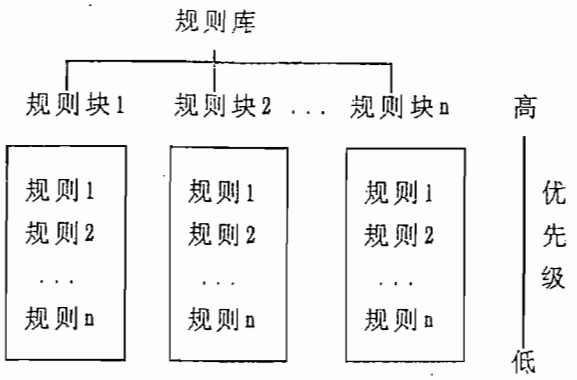


图4. 规则库的控制

规则的优先级分为静态优先级和动态优先级。静态优先级的计算可以按照这样的原则：一是根据大量的语料进行统计计算，使用频率高的规则优先级更高；二是同一个规则块中，属性约束部分多的规则优先级更高。动态优先级的计算是在分析过程中进行的，为此我们建立了一个自适应线性链表，并遵循'使用频率越高的规则优先级越高'的原则来动态地调整该链表。

五. 中心分析器的设计思想及其输出形式

1. 设计思想

UBCP的分析器主要由两部分组成：①以Tomita算法为基础，按上下文无关文法部分实现语法制导翻译，从而保证了UBCP系统的分析效率；②在Tomita算法的基础之上引入了伪合一算法，实现了句法成份之间的上下文约束测试和动态复杂特征集的产生。

在UBCP中，借助于图结构栈实现了伪并行进行不确定分析，所以在分析过程中，如果某种分析出错时不需要回溯，而只需要撤消相应的分析进程。这样一来既保证了UBCP的分析效率，同时在句子具有固有歧义的情况下，可以并行给出多种分析结果。

2. 分析结果

UBCP的分析算法是一个全解的算法，它按照句子的歧义输出所有可能的分析结果，得到一棵共享紧缩森林(无歧义时则为一棵短语结构树)；另外还输出与各结点相关的动态复杂特征集DCFS，在DCFS中包括如下几种动态信息：

. 句法信息：中心、主语、谓语、宾语等。
. 语义信息：得到一个语义网。
. 语用信息：时态、体态、语态等。

在SUN SPARCⅡ的X-WINDOW环境支持下，UBCP实现了一个用户友好的输出界面。它可以形象地显示UBCP所得到的中间结果中各概念之间的层次关系和动态特征信息。

六. 结束语

综上所述，UBCP系统具有如下的特点：

1. 语法制导和词汇驱动相结合。以高效的Tomita算法为基础，按上下文无关文法进行语法制导分析，同时结合采用词汇驱动技术，引入复杂特征集和合一运算；

2. UBCP的分析过程中没有回溯，图结构栈技术和合一算法的引入使得伪并行分析过程中的薄记工作量极大简化，从而保证了UBCP的时空效率；

3. UBCP不仅可以借助于复杂特征集和合一运算解决简单的语义歧义问题，还可以利用上下文信息和知识库中的常识进行推理分析。

目前UBCP系统正处于进一步改善阶段，以期使其成为一个高效的汉语句法分析系统。

本文的研究得到国家八五科技攻关项目的基金支持。

参考文献

1. 朱德熙，《语法问答》， 商务印书馆， 1985.
2. 黄昌宁，《复杂特征集和合一运算》，自然语言理解学会第三届学术讨论会，1988.
3. 陈火旺，《程序设计语言编译原理》，国防工业出版社，1984.
4. 周宾，吴立德，《一个汉语句法、语义分析系统》， 中国中文信息学会成立十周年学术讨论会论文集，1991.
5. 俞士汶等，《现代汉语语法电子词典的概要与设计》，中文信息处理1992 国际会议论文集，1992.
6. 李冬、陈志明，《规则描述语言及汉语的句法规则体系》， 中文信息处理1992 国际会议论文集，1992.
7. Gazdar, G. ,G.K. Pullum, and I.A. Sag, 'Genneralized Phrase Structure Grammar', Oxford,England: Blackwell Publishing and Cambridge, Mass, Harvard University press, 1985.
8. Shieber,S.,'An Introduction to Unification-Based Approach to Grammar', CSLI Lexture Notes Series, Center for the Study of Language and Information, Standard, California, 1986.
9. Tomita,M. and J. G. Carbnell, ' The Universal Parser Architecture for Knowledge-Based Machine Translation', Proc. of 10th IJCAI, 1987.
10. Tomita,M. and K.Knight, ' Pseudo Unification and Full Unification', Tech. Report, Center for Machine Translation, Carnegie-Mellon University, 1988.