

一个汉语句法分析器的设计与实现*

沈李斌 陈宣 陈玉泉 董天石 陆汝占

(上海交通大学计算机科学与工程系 上海 200030)

摘要: 文章从语法分析的角度出发, 分析了汉语句法处理中的一些关键问题, 进而将汉语的语法语义特征与句法分析技术相结合, 设计并初步实现了一个用于处理汉语的句法分析器。句法分析器的形式语法体系, 采用结合复杂特征集的上下文无关文法。在文法规则的移进和规约时采用了复杂特征的伪合一。句法分析器的分析算法采用了改进的双向图算法、规则分层技术和规则预编译技术。句法分析器的输入使用了多输入压缩技术。

关键字: 自然语言处理 句法分析器 语法

Design and Implementation of a Chinese Parser

Shen Libin Chen Xuán Chen Yuquan Dong Tiansi Lu Ruzhan

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200030)

Abstract: This paper firstly analyzes some key problems in Chinese parser, and then adapts the parsing technique to syntactic and semantic feature of Chinese. A parser designed to process Chinese based on context-free grammar and complex feature set is implemented. The parser employs the technique of pseudo unification, bidirectional chart algorithm, cascaded analysis and pre-compiling to improve its performance.

Keywords : Natural Language Processing Parser Grammar

一、引论

句法分析是自然语言处理中的重要阶段。由于汉语在语法和语义上的特殊性, 因此在这方面还处于研究与探索阶段。由合一文法所衍生出的各种句法分析的算法, 对于处理西方语言非常适用。国内的学者也试图用类似的处理手段来处理汉语^[1,2]。但就效果而言, 尚不及切分技术来得成熟。这在很大程度上是由于, 没有一个大而全的语法体系能对汉语进行精确地描述。事实上, 汉语是一种内涵语言, 有时也称为意合语言^[3]。所以问题在于, 怎样的语法语义框架才能对汉语作最有效的刻画, 进而用以实现有效的句法分析。

我们在现有技术的基础上, 将汉语的语法语义特征与句法分析技术相结合, 设计并初步实现了一个用于处理汉语的句法分析器。考虑到目前语言描述框架上的不成熟, 句法分析器的设计中采用了开放的结构, 力图融合现有的、和可能出现的一些语言描述形式。

* 本文受国家自然科学基金和 863 计划资助

句法分析器的形式语法体系，采用结合复杂特征集的上下文无关文法，并参考了 PATR-II 文法^[4]的表示方式。在语法规则的规约时采用了复杂特征的伪合一^[5]。句法分析器的分析算法采用了改进的双向图算法^[6,7]、规则分层技术和规则预编译技术。句法分析器对输入的分词标注使用多输入压缩技术。句法分析器还提供了对电子词典使用的支持。

句法分析器的总体框架如图 1 所示。

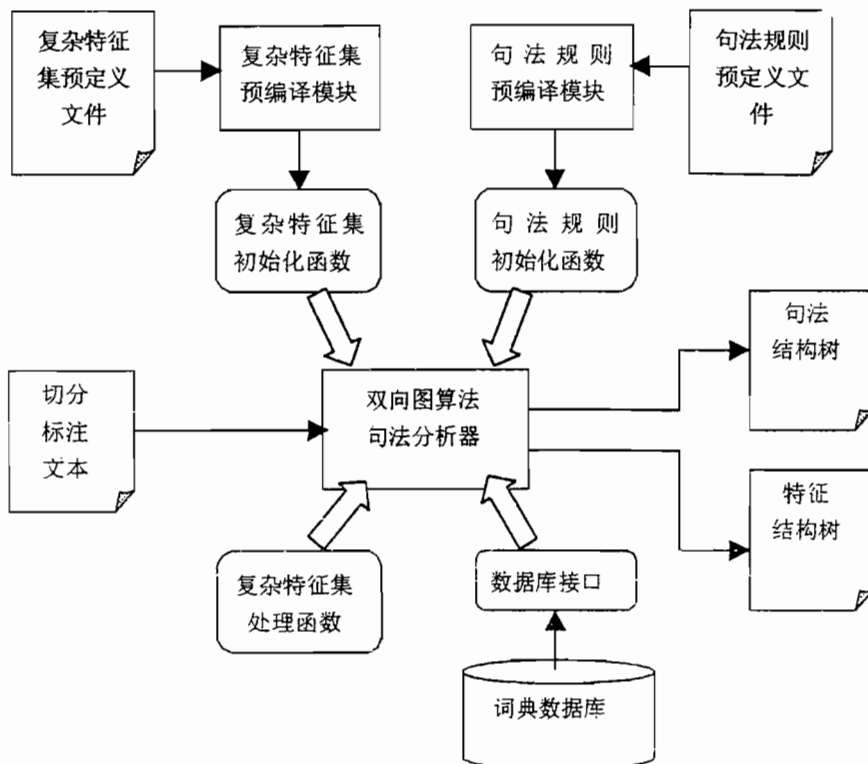


图 1 总体框架

二、形式语法体系

句法分析器的形式语法体系，以上下文无关文法为基础，以复杂特征集的合一作为规则规约的约束测试。我们采用了类似于 PATR-II 的表示方式，并使用伪合一来模拟合一操作。

在句法规则中出现的每一个文法符号都有其特定的复杂特征结构。这些复杂特征结构的定义以一定的格式存放在文本方式的复杂特征集定义文件中。如例 1 是符合定义格式的带嵌套结构的复杂特征结构定义的一部分。

例 1: 动词性短语 VP 的部分复杂特征结构定义

#VP

<KIND OUTER 0 INT>

```

<OBJ OUTER
  <HAS OUTER - BOOL>
  <POS INNER 0 INT>
  <ELEM OUTER NULL VOID> N/A>
<COMP OUTER
  <HAS OUTER - BOOL>
  <ELEM OUTER NULL VOID> N/A>

```

其中，每一个特征由四个域来描述。第一个域是特征名，第二个域是区分外部特征和内部特征。外部特征和内部特征的定义将在下文中详细给出。第三个域是该特征的默认值。最后一个域是特征的类型。例 1 中，特征 KIND 是类型为整数的简单特征，而特征 OBJ 和 COMP 则是复杂特征。

类 PATR-II 的句法规则库是以一定的定义格式保存在文本文件——句法规则定义文件中。每一条规则由三部分组成：上下文无关文法的产生式、特征校验和特征传递。特征校验部分描述了短语中各个成分之间的约束关系，其形式为

```
< 路径 > < 合一算符 > < 值 / 路径 >
```

特征校验的合一算符主要有以下两种：

- == (比较)：通用的对整型、布尔型和字符串的值的比较；
- match (类义匹配)：用于把名词性成分类义同动词某一论元的类义进行匹配；

特征传递为新形成的语法单位设置特征值，其形式为

```
< 路径 > < 赋值算符 > < 值 / 路径 >
```

特征传递的赋值算符有以下三种：

- = (合一赋值)：如果右部是值，则对左部路径所指的原子特征赋值，如果右部也是路径，则使左右两边共享右部路径所指的特征结构，这会引起特征结构的重入；
- <= (复制赋值)：把右部所指的特征结构复制到左边，左右两边的特征结构相同，但并不是共享；
- | (析取赋值)：使右部所指的特征成为左部的一个析取元，这也会引起特征结构的重入。

例 2 是述宾短语产生规则的一种情况的定义示例。不同的述宾短语，它们所对应的语义表达式是不同的，规约所使用的规则也是不同的。

例 2：述宾短语的部分归约规则定义

```
#VP->VP1 NP
```

```

<VP1 HEAD JIASHU> == 2 // VP1 的价为 2
<VP1 HEAD LEIYI OBJ> MATCH <NP HEAD LEIYI> // VP1 与 NP 类义相匹配
<VP1 OBJ HAS> == - // VP1 本身不带宾语
<VP HEAD> <= <VP1 HEAD>
<VP OBJ HAS> = +
<VP OBJ POS> = 0
<VP OBJ ELEM> = <NP>
<VP KIND> = 2

```

其中第 1 至 3 条是特征校验，4 至 8 条是特征传递。

三、句法分析算法

图算法是基于短语结构文法的自底向上的分析方法。在传统的图算法中，活动弧必定是由规则右部第一个文法符号触发的，它的扩展也是严格的由左向右的。双向图算法是对它的改进，使得可以从规则的中间开始往两边分析，从而更适合于自然语言的分析要求。该算法使用触发类，提高了效率。双向图算法也可以有效地表示歧义结构，可以在一次分析过程中同时得到句法组合层次上的不同解释。双向图算法在系统的鲁棒性方面效果也很好。文献^[7]对双向图算法进行了仔细的分析、改进，本文不再赘述。

需要指出的是，加入了复杂特征集之后，图算法中原有的弧压缩算法^[8]就不能直接使用了。我们在复杂特征集中引入了外部特征的概念，并对含有相同外部特征的弧使用弧压缩算法，从而解决了这个问题。复杂特征的外部性与内部性是针对分析过程中而言的。复杂特征集中的外部特征指在上层分析中的将要访问到的特征；而内部特征正相反，在初始化被赋值后，以后的分析不会再访问到它们，它们的存在只是为了记录一定的信息，对句法分析以后进一步处理可能是有用的。特征是外部的还是内部的是在规划特征集结构时预先确定的，通常外部特征记录的是在上层分析中需要用到的信息，而内部特征则记录一些仅与本特征集相关的信息。事实上区分特征的外部性与内部性对于伪合一算法中的比较特征集和复制工作最小化都有重要的意义。

算法中另一个难点问题是在归约中的文法歧义。首先用一个例子来说明这种情况。如果有两条规则（只表示出短语结构文法部分）：

规则 1: $VP \rightarrow DP VP_1$

规则 2: $VP \rightarrow VP_1 NP$

前者表示用副词性短语修饰动词性短语构成状中结构，后者则是典型的动宾结构。对于 DP VP NP 结构的短语，用上述两条规则归约成一个 VP 时，显然可以通过两种途径。这是一种文法中的歧义，对于人工程序语言来说，如果有这种情况，可以通过改写文法来解决。但对于自然语言，句法规则是表示一定的语法意义的，如上述的状中、动宾结构，这样修改可能造成意义的混乱，也可能引起文法其他部分的改动。

理论上，在基于合一的文法分析中，这两种情况有时是需要同时保留的。这里句法规则中的文法符号实际表示一个复杂特征结构。在通过两种途径归约出 VP 时，复杂特征集的名是相同的，但它们的特征却可能有所区别，这样的两个“VP”实际是两个不同的结构，因此都需要保留。然而，通常无论先扩展成状中结构还是先扩展成动宾结构，最终得到的特征集应该是一样的。我们可以对特征集进行比较，以确定两种情况是否都需保留。联系到外部特征与内部特征的概念，只有外部特征对以后的分析是有效的，因此，只需要对特征集中的外部特征进行比较即可。

另一种用来处理文法歧义的方法就是规则分层。规则分层技术源自于一种称为多层有限状态自动机的快速部分句法分析技术。该算法已用于英语等多种语言的句法分析之中。Abney 于 1990 首次用该方法实现了 CASS (Cascaded Analysis of Syntactic Structure) 系统，未加入任何其它方法就可轻松地达到 85% 以上的准确率^[9]。但将多层限自动机应用于汉语句法分析仍存在一定的困难，这主要归因于汉语本身的特点：句法灵活；短语、句子的构造规则相似，之间没有明显的界限。以前在这方面的研究很少。文献^[10]对汉语规则的分层技术作了考察，给出了一个汉语句法的多层有限状态自动机表示。我们的系统中采用了相似的短语规则的层次，使得句法分析的效率得到改善。

四、电子词典

电子词典中记录了每个词相应的词法、语法与语义的全方位的信息。这些信息以数据库的方式进行组织。每类词性都对应于一张表，词的每一个特征则对应于表中的属性。分析器在读入输入词时根据词典数据库中的信息初始化复杂特征集，分析过程是以这些最底层的复杂特征集为基础的。

词典数据库是整个句法分析过程的依据，它的完备性与正确性，对分析器来说是至关重要的。可喜的是近年来，电子词典的建设，特别是语法信息词典的建设取得了丰硕的成果，为句法分析的提供了强有力的支持。在语义词典的选择上，我们主要借鉴了《同义词词林》^[11]的语义分类，并以此为标准，构建了动词语义格约束的电子词典。该词典中每个动词义项有一定的价数，表明了该动词义项可支配的必有成分的个数，另外它所要求共现的每个价位也对名词的语义类型有一定的限制，这是通过给每个价位设置相应的匹配语义类型集合来描述的。对于句中一个可充当论元的名词性成分，将其语义类型与述语动词的配价模式中各价位的语义类型集合的元素进行匹配，便可以判别该名词性成分在句中充当的语义角色。

可以相信，随着系统中语法、语义词典的不断完善，句法分析的效果也将进一步优化。

五、示例

由于语法语义信息库等因素的限制，目前仅进行了封闭测试，效果良好。1998年国家863项目办公室组织进行汉英机器翻译系统性能评测时使用了400句例句。以下给出的示例，来自于句法分析器对这些例句的处理。

例3：今天(T)我们(PR)学(V)了(AU)第(M)三(M)课(N)

语法分析树：DJ(TP(T),DJ(NP(PR),VP(VP(VP(V),AU),NP(MCP(MCP(M),MCP(M)),NP(N))))))

特征结构树：

<DJ	<MODIFIER ...>，
<KIND 主谓>，	<OBJ
<SUBJ	<HAS +>，
<HEAD ...>，	<POS 0>，
<PRED	<ELEM ...>，
<KIND 述宾>，	<MODIFIER ...>>>
<HEAD ...>，	

上面的示例中，特征结构树所表示的是，顶层语法成分DJ的部分复杂特征集。由于篇幅原因，以下的示例仅给出语法分析树。

例4：被子(N)叠(V)得(USDF)整整齐齐(A)

语法分析树：DJ(NP(N),VP(VP(V),USDF,AP(A)))

例5：她(PR)笑(V)着(AU)表示(V)她(PR)的(USDE)谢意(N)

语法分析树：DJ(NP(PR),VP(VP(VP(V),AU),VP(VP(V),NP(NP(NP(PR),USDE),NP(N))))))

例6：老虎(N)比(P)猫(N)大(A)得(USDF)多(A)

语法分析树：DJ(NP(N),AP(PP(P,NP(N)),AP(AP(A),USDF,AP(A))))

例 7: 爸爸(N)瘦(A)了(AU)一(M)点(Q)

语法分析树: DJ(NP(N),AP(AP(AP(A),AU),MP(MCP(M),Q)))

例 8: 她(PR)生(V)了(AU)一(M)个(Q)胖墩墩(A)的(USDE)男孩(N)

语法分析树:

DJ(NP(PR),VP(VP(VP(V),AU),NP(MP(MCP(M),Q),NP(AP(AP(A),USDE),NP(N))))))

六、结语

汉语的句法分析是一项比较困难的工作,同时也是非常有意义的。国内许多计算语言学者都在这方面进行着艰苦的研究。我们在这方面的工作也是一个探索,要想一步就把工作做得十分完美是不现实的。因此,系统的开放性成为我们设计的一大目标,我们努力地吧句法分析器做成易于修改与扩充的。

从系统的实现方案来看,分析器的许多部分都是开放的。句法规则与复杂特征结构的定义文件也可以进行文本方式的修改,然后由预编译子系统连入整个系统中;词典数据库可以通过数据库管理系统方便的进行扩充;规则的分层方案是可编辑的。分析器的开放性将有助于我们改进我们的系统,以得到更完善的句法分析器。系统的开放性使系统具有很好的可移植性。通过对句法规则的修订,句法分析器可以被用来处理某些基于特殊文法的文本,如用于汉语词典文本的处理。

目前,我们已经实现了上述的句法分析器的初步功能,并正在进一步优化其分析效果。系统的完善是一个不断积累的过程。需要进一步改进的包括一个完善的基于短语结构文法的语法规则体系、完备的语法语义特征框架及其词汇知识库、以及相应的语料库的建设。

参考文献

- [1] 栾浩, 基于合一的汉语句法分析系统的研究与实现, 语言信息处理专论, 黄昌宁、夏莹编, 清华大学出版社、广西科学技术出版社, 1996。
- [2] 沙新时、吴立德、周斌, 基于合一语法的通用句法分析器: 设计与实施, 中文信息学报, 1993, 第7卷第2期。
- [3] 鲁川, 汉语的意合网络, 语言文字应用, 1998, 第2期。
- [4] Shieber, S. M.. The design of a computer language for linguistic information. Proc. COLING, 362-366. 1984.
- [5] Tomita, M. and K. Knight, Pseudo Unification and Full Unification, Tech. Report, Center for Machine Translation, CMU, 1988
- [6] 吴立德, 大规模中文文本处理, 复旦大学出版社, 1997。
- [7] 沈李斌、陆汝占, 双向图分析器的改进, JSCL'99。
- [8] Allen, James, Natural Language Understanding, 2nd Ed., The Benjamin/Cummings Publishing Com., 1995.
- [9] Abney. S., Partial Parsing via Finite-State Cascades. In: Proceedings of the ESSLLI '96 Robust Parsing Workshop, 1996.
- [10] 张益民, 基于混合方法的中文文本解释研究, 上海交通大学博士论文, 1998。
- [11] 梅家驹, 同义词词林, 上海辞书出版社, 1983。