汉语句法规则的自动构造方法研究*

周强　　黄昌宁

智能技术与系统国家重点实验室
清华大学计算机科学与技术系，北京 100084

摘要：本文对汉语句法规则的自动构造方法进行了一些探索。通过对汉语句法规律的总结和提炼，提出了一套简单灵活的汉语句法元规则描述体系，包括结构元规则集、标记特征表和中心标记表等部件，在此基础上，构造了一个有效的元规则解释器，取得了较好的实验效果。
关键词：上下文无关语法，元规则，语法构造。

1. 研究动机

在自然语言处理系统中，句法规则起着很重要的作用。因此，句法规则的构造和描述问题一直是计算语言学界研究的一个中心问题。国外在英语方面已做了许多工作，如：G. Carroll, E. Briscoe 和 C. Grover 等在广义短语结构语法(GPSG)框架上开发的句法规则自动构造工具 GDE([CBG91], [BGBC87])等。

但是，中文信息处理面临着的一个很现实的问题，却是如何构造一个合适的句法规则库。汉语作为一种独特的语言，具有许多与英语、日语等其他语言不同的特点。如何解决汉语句法分析中的几个基本问题，如词类的确定、句法描述体系的建立、句法分析方法的选择等，一直是汉语语法研究的焦点。经过几十年的努力，才逐渐找到了一条结合汉语的实际，并借鉴国外语言学的研究成果，来对汉语语言事实进行描述的正确研究道路。但由于各方面条件的限制，目前还没能形成一套适合于计算机分析的比较完整的汉语句法规则形式描述体系。这就给中文信息处理的研究人员带来了很多不便，迫使他们根据不同的研究问题，花费大量的时间和精力，自己去构造所需的句法规则体系。但由于语言学训练的不足，以及所考察的语料规模的限制，使得他们得到的句法规则往往具有很强的领域依赖性，覆盖面比较窄，很难扩展到新的领域以处理新的问题。这种现象造成了很大的资源浪费，严重制约了中文信息处理研究的发展。

本文所进行的汉语句法规则的自动构造方法研究，就是希望在在汉语语言学的理论研究和中文信息处理的实际需要之间搭起一座桥梁。它将通过对汉语句法规律的总结和提炼，形成一套简单灵活的汉语句法元规则描述体系，然后利用一个有效的元规则解释器，可以自动生成所需的句法规则集。由于可以很方便地根据不同的研究问题对元规则描述体系进行适当的调整，从而使它具有较高的适应性，可以在汉语语法推导、自动句法分析和汉外机器翻译等领域的研究中发挥重要的作用。

2. 总体设想

我们的目前研究目标是在现有的词类标记集和句法标记集基础上，自动构造枚举出所有合法的汉语句法规则形式。利用上下文无关文法(CFG)形式化描述体系，即形成这样的语法三元组（VT, VN, R），其中：VT 为终结符集合，即词类标记集；VN 为非终结符集合，即句法标记集；R 为 VT 和 VN 上的一组上下文无关产生式规则，其基本形式为： $A\rightarrow\lambda$, $A\in VN$, $\lambda\in\{VT\cup VN\}_*$。

在汉语的句法描述体系中，短语成分起了承上启下的作用，它具有与汉语句子基本一致的构造方法，这和英语的短语有很大差别。朱德熙先生认为，“ 如果我们把各类词组的结构和功能都足够详细地描写清楚了，那么句子的结构实际上也就描写清楚了，因为句子不过是独立的词组而已。 ”

* 本项研究得到国家自然科学基金项目资助。

([Zhu85],P74)。汉语句法构成的这一特点，使我们可以从汉语的基本短语结构形式出发，寻找到一条自动构造汉语句法规则的新途径。

汉语短语的基本构造规律可以概括为：

1) 结构完备律：所有短语都可以归结为定中、状中、述补、述宾、连动、联合、主谓等基本结构类型[FX91]。

2) 中心限制律：除主谓结构外，每个短语的外部功能基本上可以由其中心成分确定。

3) 成分位置律：不同句法成分在短语结构组合中所处的位置是由其自身的句法功能确定的，如动词可以处于述宾结构的述语位置，而名词则不行。

以此为基础，我们形成了这样的汉语句法规则自动构造系统的开发设想：

1) 总结了一组句法特征，据此可以很方便地描述不同的结构组合类型，形成结构元规则集。

2) 根据句法功能的不同，为各个标记符（词类标记或句法标记）确定合适的句法特征描述，形成标记特征表。

3) 确定处于结构中心成分位置的句法成分所能体现的外部句法功能标记，形成中心标记表。

这样，以两个标记集（词类标记集和句法标记集）为句法规则生成单元，以三个数据表（句法元规则集、标记特征表和中心标记表）为规则生成控制单元，通过构造一个元规则解释器，就可以自动得到比较完整的符合汉语句法构成规律的规则集。在下面的几节中，将对有关的具体内容进行详细的介绍。

3. 句法功能特征集的提取

表 1 列出了我们目前提取的一组句法功能特征，它主要分为两大部分，一是对某个成分在句法结构中所体现的基本句法功能的描述，如：SUB, PRED, OBJ 等；一是对句法成分本身句法性质的描述，如：Ti, Wei 等。

表 1 句法功能特征描述

特征标记	特征描述
HEAD	能作中心成分
SUB	能作主语
PRED	能作谓语
OBJ	能作宾语
ADN_D	能作直接邻接定语（与被修饰成分间无分隔成分）
ADN_I	能作间接邻接定语（与被修饰成分间有分隔成分）
ADV_D	能作直接邻接状语
ADV_I	能作间接邻接状语
COMP_D	能作直接邻接补语
COMP_I	能作间接邻接补语
SENT	能独立成句
Ti	是体词性成分
Wei	是谓词性成分

这组特征在句法规则自动构造系统中起了重要的作用。一方面，通过这些特征的不同组合，可以很方便地对汉语的基本短语结构进行描述，形成结构元规则，如：述宾结构短语就可以描述为：HEAD & Wei　OBJ；另一方面，利用标记特征表进行信息检索，可以在结构元规则的句法特征描述和构成实际的句法规则的词类（终结符）和句法（非终结符）标记集之间建立有效的联系，从而可以通过对不同标记集的排列组合得到合法的句法结构组合形式。

4. 结构元规则描述体系

以句法功能特征集为基础，并辅之以一些特殊的运算符，我们形成了一套有效的汉语结构元规则描述体系，可以对汉语的基本句法结构进行很好的描述。图 1 介绍了这套元规则的巴科斯范式(BNF)描述，表 2 则列出了部分结构元规则实例。

<结构元规则> :: = <结构描述项> [<分隔标记项>] <结构描述项> [→ <句法标记>]
| <联合结构元规则>
<结构描述项> :: = <句法特征项> | <标记描述项>
<句法特征项> :: = <句法特征> | <句法特征> '&' <句法特征项>
| <标记限制> '&' <句法特征项>
<标记描述项> :: = <标记描述> | <标记描述> '|' <标记描述项>
<标记描述> :: = <词类标记> | <句法标记>
<分隔标记项> :: = <词类标记> | <词类标记> '|' <分隔标记项>
<句法特征> :: = HEAD | SUB | PRED | ……
<标记限制> :: = IsPOST | IsSynT
<词类标记> :: = a | b | c | ……
<句法标记> :: = np | vp | ap | dj | ……
<联合结构元规则> :: = HEAD >> HEAD <<
| HEAD > wD '|' cM HEAD <
| HEAD [w] '*'

图 1 汉语结构元规则的巴科斯范式描述

表 2 部分结构元规则实例

1.	述宾结构元规则：HEAD & Wei OBJ
2.	述补结构元规则：HEAD & Wei COMP_D
3.	定中结构元规则：ADN_I uJDE HEAD & Wei & IsPOST --> np
4.	状中结构元规则：ADV_D HEAD & Wei
5.	主谓结构元规则：SUB PRED & Wei --> dj
6.	联合结构元规则：HEAD > wD | cM HEAD <
7.	整句元规则： SENT wE --> zj

下面对 BNF 的一些描述特点作一下简要说明：

1) 汉语的句法结构组成通常是二元的，即由两个成分组成。但在我们的句法规则描述体系中，对一些结构层次进行了简化处理，允许出现三元结构（详见[ZQd96], [ZY96]），如带助词“的”的定中结构及带助词“得”的述补结构等。BNF 中的<分隔标记项>就是为实现这个目标而设计的。

2) 为了提高元规则的描述能力，我们提供了两个运算符：'&' 和 '|'， 其中'&'运算符作用于句法功能特征，表示处于元规则的某一句法位置上的句法成分必须同时具有所列出的若干句法特征，从而加强了对该位置上的句法成分的限制，保证生成更为准确的句法规则；而'|'运算符则作用于词类或句法标记，表示在元规则的某一句法位置上可以同时出现数个标记，从而提高了元规则的生成能力。

3) 通过两个标记限制符：IsPOST 和 IsSynT，还可以对元规则中不同的句法特征描述所生成的标记集进行进一步的限制，其中'IsPOST'要求它们必须是词类标记，而'IsSynT'则要求它们必须是句法标记。

4) 与其他结构组合相比，汉语的联合结构具有其特有的组成规律：

- 组成成分的数目没有限制，理论上可以是无限的。
- 不同的组成成分之间具有较强的功能相似性，如属于同一词类或句法标记，或者功能类似的标记组。

- 对于多成分联合结构，不同成分之间往往有比较明显的形式标记，如：连词、顿号、逗号等。

为此，我们在 BNF 中构造了三条特殊的元规则，以控制生成汉语的不同联合结构。其中，运算符对：‘>>’和‘<<’ 要求所生成的两个成分具有完全相同的功能标记；而运算符对：‘>’和‘<’只要求所生成的两个成分具有功能类似的标记即可，如 n 和 nbar；另外，通配符 ‘*’ 则用于描述那些具有三个或三个以上并列成分的联合结构。

5. 标记集及其特征描述

5.1 标记集的调整和扩充

标记集的合理设计，是与具体的分析问题密切相关的。我们在进行汉语语料库多级加工和标注研究过程中[ZY95]，依据小标记集的处理思想，曾经设计了一个包含 31 个标记的词类标记集[ZQd96]和 19 个标记的句法标记集[ZY96]，它们在降低自动标注算法和人工校对处理的复杂度以及提高标注结果的一致性方面起了很重要的作用。

现在，为了更好地适应汉语句法规则描述的要求，我们又对其中的词类标记集进行了以下调整和扩充：

1) 删除了三个表征词语类别的词类标记，即：i（成语）、j（简称略语）和 l（习用语），并将有关的词语按其语法功能的不同，分别归入名词、动词等功能词类中。

2) 对一些特殊的虚词，其中主要是助词和连词，增加了特殊的子类标记，如：结构助词“ 的、得、地 ”，并列连词“ 和、与、…… ”等，以便更好地描述它们在句法规则中所起的特殊作用。

3) 对一些具有特殊语法功能的实词，增加了子类标记描述，如：汉语中有一部分动词，其中主要是双音节动词，它们具有以下的特点：① 可以直接修饰名词作定语，② 可以受名词直接修饰，③可以作准谓词性动词的准宾语等，朱德熙先生把它们称为“ 名动词 ”[Zhu79]。由于它们与一般动词的语法功能有很大的差别，因此我们特别设置了一个动词的子类标记 vN 来对它们进行描述。

4) 对标点符号进行了进一步分类，突出了具有不同子类标记的标点符号在句法规则中的不同描述作用。

这样，就形成了一个包含 41 个标记的新的词类标记集，其中的大类标记和子类标记形成了一个层次描述体系，详见附录 1。另外，附录 2 列出了目前所用的句法标记集。

实际上，这样的标记集调整和扩充操作可以针对不同的应用问题不断进行，使最终形成的句法规则集能更好地满足实际的要求。

5.2 标记的句法特征描述

利用表 1 中的句法特征，对标记不同标记的句法成分各自所具有的基本句法功能进行描述，就得到了不同标记的句法特征描述项，它们可以进一步组合成一个标记特征表。例如，对表征形容词的词类标记 a，可以有这样的句法特征描述： HEAD, PRED, ADN_D, ADN_I, ADV_D, ADV_I, COMP_D, COMP_I, SENT, Wei, !OBJ，表示形容词可以处于句法结构的中心成分位置，可以作主语、谓语、宾语、定语、状语、补语等，可以独立成句，是一个谓词性成分。值得注意的是，其中还使用了一个特殊的操作符：搭配限制符 ‘!’，它用于标识同某个标记不能搭配的句法特征。在这个例子中， ‘!OBJ’就描述了汉语形容词的一个重要句法功能：不能带宾语。它可以控制标记 ‘a’ 不在述宾结构元规则：HEAD & Wei OBJ 的 HEAD 位置上出现，从而排除了不合汉语语法的“ a OBJ → ap ”形式的句法规则出现的可能性。

标记特征表的这种结构使它可以很方便地根据不同应用问题的实际需要进行适当的调整。例如，如果需要在句法规则中增加更多的子类信息描述，则只需在标记特征表中增加一些新的子类标记的句法特征描述项即可；另外，即使需要采用新的标记集，也只需对标记特征表的相应内容进行调整，而不必修改结构元规则集，就可以获得所需的句法规则描述。这样就大大提高了目前的元规则集描述体系的灵活性和适应性。

6. 中心成分句法功能的传递和继承

利用中心限制律确定规则左部的句法标记，主要可以通过以下两个途径：1) 标记提升，即将处于中心成分位置的词类标记提升为句法功能类似的规则句法标记；2) 标记传递，即将处于中心成分位置的句法标记向上传递成为整个规则的句法标记。

由于我们的短语分类体系[ZY96]采用了与汉语词语分类体系[YSW93]相类似的“按照语法功能分布”进行分类的思想[Zhu79]，因此可以很容易地建立两者之间的句法功能对应关系，如表 3 所示。据此，就可以方便地形成表 4 所示的中心标记表。

另外，对于一些特殊的句法规则，如：以动词为中心成分的定中结构名词短语规则：np → r uJDE v（他们的学习），以及不太容易确定其中心成分的主谓结构规则等，还可以在结构元规则中通过显性句法标记描述来确定规则左部标记（参见图 1 的 BNF 描述），并设置其优先级高于中心限制律。这样，就可以形成如下的简单的规则左部标记确定流程：

- 检查元规则中是否有显性句法标记描述，若有，则以此为规则左部标记；
- 否则，检查中心成分标记是否在中心标记表中出现，
- 若是，则据此确定规则左部标记；
- 否则，以中心成分的句法标记作为规则的左部标记。

表 3 词类标记和短语标记的句法功能对应关系

词类标记	短语标记	相类似的句法功能
v	vp	在句子中主要作谓语
a,z	ap	在句子中可以作谓语、补语和定语
b	bp	在句子中只能作定语
d	dp	在句子中一般作状语
n	np	在句子中可以作主语、宾语，但一般不能作谓语，不能作“在/到”的宾语
t	tp	在句子中可以作主语、宾语，可以作“在/到”的宾语，并能用“什么时候”提问
s	sp	在句子中可以作主语、宾语，可以作“在/到”的宾语，可以用“哪儿”提问
m,q	mp	在句子中可以作主语、宾语、定语和补语

表 4 确定汉语句法规则左部标记的中心标记表

规则左部标记	中心成分标记
np	np nbar n r vN
vp	vp vbar v
ap	ap abar a z
sp	sp s f
tp	tp t f
bp	b
dp	d

7. 句法规则自动构造工具及其实验结果

图 2 显示了目前的句法规则自动构造工具的总体结构图。在元规则解释器的控制下，通过检索标记特征集，可以将元规则集中的句法特征描述转化为不同的标记集，从中可以经排列组合形成

不同的规则右部组合，然后利用规则标记描述或中心标记表确定相应的规则左部标记，从而形成完整的句法规则。下面给出元规则解释器的具体处理流程：

1. 取一条结构元规则；
2. 检索标记特征表，将句法特征描述转化为标记集；
3. 排列出一个标记组合，形成规则右部结构；
4. 确定规则左部标记（利用上节介绍的方法）；
5. 将一条完整的句法规则加入规则集中；
6. 排列完了吗？若是，则转 7；否则转 3；
7. 还有未处理的元规则吗？若有，则转 1；否则转 8；
8. 输出句法规则集。

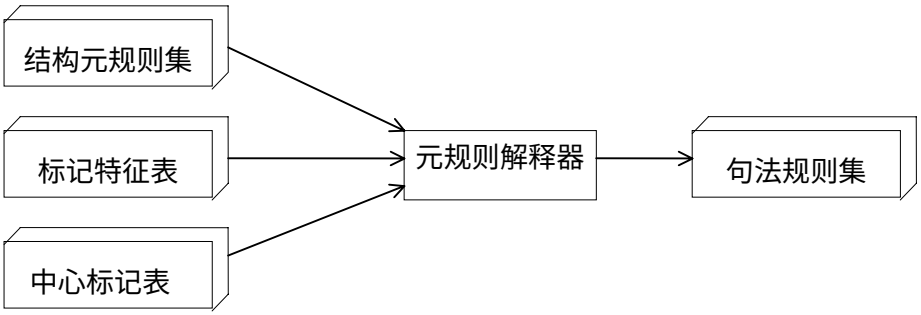

图 2　汉语句法规则自动构造工具的总体结构图

利用这个工具进行汉语句法规则的自动构造实验，我们得到了表 5 所示的实验数据。附录 3 给出了自动构造产生的部分句法规则实例。

表 5 句法规则自动构造工具的实验数据

词类标记数目	句法标记数目	句法特征数目	元规则数目	自动生成的句法规则数目
41	19	13	24	958

利用这个自动构造的句法规则集，并辅之以经人工总结和树库统计得到的特殊规则形式，我们形成了一个比较完整的汉语概率型上下文无关语法（PCFG）初始规则集，它在汉语 PCFG 自动推导研究中发挥了重要的作用。

参考文献

[BGBC87] E. Briscoe, C. Grover, B. Bogurraev, & J. Carroll. (1987). “A formalism and environment for the development of a large grammar of English” *Proceedings of the 10th International Joint Conference on Artificial Intelligence, 703-708, Milan, Italy.*

[CBG91] J. Carroll, E. Briscoe, & C. Grover. (1991). “A development enviroment for large natural language grammars” *Technical Report 233,* Computer Laboratory, Cambridge University, England.

[FX91] 范晓 (1991). 《汉语的短语》，商务印书馆

[YSW93] 俞士汶 (1993) “信息处理用现代汉语词语分类体系”，北大计算语言所内部资料，其摘要刊登于《中国计算机报》1994.5.31. 第 81 版.

[Zhu79] 朱德熙. (1979). 《语法讲义》. 商务印书馆

[Zhu85] 朱德熙. (1985). 《语法答问》. 商务印书馆

[ZQ96] 周强. (1996). “一个汉语短语自动界定模型”，《软件学报》第 7 卷，增刊，315-322

[ZQd96] 周强 (1996). “汉语语料库的短语自动划分和标注研究”，博士学位论文，北京大学计算

机系，1996.6.
[ZY95] 周强，俞士汶. (1995). “ 一个人机互助的汉语语料库多级加工处理系统 CCMP ”，陈力为，袁琦主编，《计算语言学进展与应用》，清华大学出版社，50-55。
[ZY96] 周强，俞士汶. (1996). “ 汉语短语标注标记集的确定 ”，《中文信息学报》, 10(4), 1-11.
[ZZ96] 周强，张伟. (1996). “ 一个汉语改进的短语自动界定模型 ”，In *Proc. of ICCC'96, Singapore, June 4-7,* 75-81.

Research of the Automatic Construction Methods for Chinese Context-Free Grammar

Zhou Qiang, Huang Changning

The State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science, Tsinghua University, Beijing 100084
zhouq@s1000e.cs.tsinghua.edu.cn

ABSTRACT: In this paper, we explore methods to construct Chinese syntactic rules automatically. Base on Chinese syntactic principles, we propose a simple and flexible descriptive system for Chinese syntactic meta rules, which includes structural meta rule set, tag feature list and head tag list. Then, an efficient meta rule explainer is constructed to explain the information in the system. Current experiments show encouraging results.

KEYWORDS: Context-Free Grammar(CFG), Meta Rule, Grammar construction.

附录 1 词类标记集

标记代码	标记名称及描述	标记代码	标记名称及描述
a	形容词	s	处所词
b	区别词	t	时间词
c	连词	u	助词
cM	中置连词，如：和、与、…	uJDE	结构助词“的”
d	副词	uJDI	结构助词“地”
e	叹词	uJDD	结构助词“得”
f	方位词	uD	动态助词“着、了、过”
g	语素字	uS	助词“所”
gN	名词性语素字	uSD	助词“似的、一样”
gV	动词性语素字	v	动词
gA	形容词性语素字	vN	名动词
gT	时间词性语素字	w	标点符号
h	前缀	wD	停顿符号“逗号、顿号”
k	后缀	wE	句子结束符“句号、问号、感叹号”
m	数词	wLB	左标号，包括“单引号、双引号、左括号、…”
n	名词	wRB	右标号，包括“单引号、双引号、右括号、…”
ngp	指人的专有名词	wM	引句标志：“冒号”
o	象声词	x	非语素字
p	介词	y	语气词
q	量词	z	状态词
r	代词		

附录 2 句法标记集

标记代码	标记名称及其实例
1). np	名词性短语，如：我们买的，漂亮的帽子
2). nbar	名词性准短语，如：工人们，资本主义
3). vbar	动词性准短语，如：看了一看，学过
4). vp	动词性短语，如：给他一本书，去看电影
5). abar	形容词性准短语，如：高兴高兴，红了
6). ap	形容词性短语，如：特别安静，更舒服一点
7). dp	副词性短语，如：虚心地，非常非常
8). pp	介词短语，如：在北京，被他的老师
9). bp	区别词性短语，如：大型中型小型
10). tp	时间词性短语，如：战争初期，周末晚上
11). sp	处所词性短语，如：村子里，中国内地
12). mbar	数词准短语，如：一千三百

13). mp	数量短语，如：两三天，这群
14). dj	单句句型，如：她态度和蔼　那时候，天气还很冷
15). fj	复句句型，如：如果他愿意，我就陪他去看看
16). zj	整句，如：你去不去?　火又盛，烟又大。
17). jq	句群，如：救命啊！救命啊！
18). dlc	独立成分
19). yj	直接引语

附录 3 部分自动构造的句法规则

下面列出了一部分自动生成的汉语句法规则，其描述格式为：

{<规则左部标记〉{<规则右部结构组合>}}

ap

a *	a a	a abar
a ap	a cM a	a cM abar
a cM ap		

dj

n a	n abar	n ap
n v	n vbar	n vp
n wD a	n wD abar	n wD ap
n wD v		

fj

a wD fj	abar wD dj	abar wD fj
ap wD dj	ap wD fj	c wD fj
d wD fj	dj wD abar	dj wD ap
dj wD vbar	dj wD vp	dp wD fj

np

bp uJDE n	dj uJDE	dj uJDE a
f uJDE n	m n	

vp

c vbar	c vp	c wD v
c wD vbar	c wD vp	d v
d vbar	d vp	d wD
d wD vbar		