从标注语料库中归纳语法规则：“V+N”序列实验分析*

孙宏林
(北京语言文化大学语言信息处理研究所)

摘要：传统的基于规则的方法由于难以处理真实语料中的复杂现象因而面临严峻的挑战，而基于统计的方法在分析语言结构方面又先天不足。本文提出了一种把这两种方法结合起来的途径，即从大规模标注语料库中归纳语法规则，这样可以充分地缩小语法规则的颗粒度、提高其确定性。本文利用从语料库中归纳出来的14条简单的语法规则用于判断汉语中的“V+N”序列在什么情况下是一个合法的短语，实验得到了较好的结果。研究表明：几乎所有的分析错误和难点都是由于结构歧义造成的。本文从实例的分析中发现了一系列结构歧义现象。
关键词：语料库，句法分析，歧义

Acquiring Grammatical Rules by Induction from Tagged Corpus: A Case Study on “V+N” Sequence in Chinese

Sun Honglin
Center for Language Information Processing
Beijing Language & Culture University,100083,Beijing,China
Email:yyxxs@mail.netchina.co.cn

Abstract: The traditional rule-based method is being faced with great challenges because of its weakness in processing real natural language texts ,and the statistics_based method is deficient in dealing with the structure of languages. This paper presents an approach to combine the two methods ,that is to acquire grammatical rules by a process of induction from tagged corpus. This is manifested by a case study on V+N sequence in Chinese,applying 14 simple rules acquired from the induction process in determining when V+N is a construction and the experiment got good results.The research shows that almost all the errors and difficulties are produced by structural ambiguties,some of which are found at first time from the analysis.
Keyword: corpus, parsing,structural ambiguity

1. 引言

传统的基于规则的方法由于难以处理真实语料中的复杂现象因而面临严峻的挑战，而基于统计的方法在分析语言结构方面又先天不足。如何把这两种方法结合起来，使其相互取长补短，以提高句法分析的水平，是许多人都在探讨的问题。传统的规则系统中使用的规则基本上是语言学家语法知识的总结，尽管有些规则也是从大量语言事实中总结出来的，但由于人对语言的认识方式与计算机处理语言的方式有很大的区别，所以这些语法规则都可以归入基于直觉的（intuition_based)的范畴。基于直觉的语法规则难以满足计算机分析自然语言的需要，这也是产生规则方法缺陷的重要原因。因此一些有见地的语言学家认识到了利用计算机语料库来验证已有的语法规则和发现新的语法规则的重要性([8],[9])。由于较大规模的英语树库（treebank）的建成，近年国外有不少关于从树库中自动获取概率短语结构语法的研究。但由于自动句法分析技术还很不成熟，目前建立大规模高质量汉语树库

* 本项研究得到国家自然科学基金的资助（项目号：69433010）。

的难度还很大。但是，随着自动词性标注技术的实用化，目前已建成了一些汉语的标注语料库（即经过分词和词性标注的语料库）[7]。本文所要探讨的就是从经过分词和词性标注的语料库中获取句法规则的问题。具体来说，就是根据语言事实归纳出这样的规则，用以确定任意上下文中的一个特定词类序列是不是一个语法形式（参见[1]，[6]），即看其是否实例化为一个合法的短语。

为了说明语法归纳的过程，本文具体地探讨了汉语中“动词+名词”（记为“ V+N ”）序列的捆绑规则。如果考虑到语流中任意的“ V+N ”序列，问题会变得非常复杂，这绝非一篇文章所能容纳。因此我们对其进行了限制：（ i ）不考虑 V 和前面成分的关系；（ 2 ）V 限于带名词性宾语的，即我们假定已经知道这里动词是带名词性宾语的，它的后面一定有一个名词性成分作它的宾语（在语料库中动词的标记为 vwn，参见[7]）。在这两个前提下，我们研究在什么条件下 vwn 和 N 应该捆绑在一起。

2. 语法规则的归纳

2.1 归纳语法规则的过程

我们从标注语料库中获取语法规则的过程是：从一个基于直觉的规则集出发，对一个语料样本进行自动分析，然后，由人工对自动分析的结果进行检验，对原来的规则进行修改和补充，然后再用新的规则集自动分析另一个语料样本，再进行人工检验和规则修正，如此循环，直到得到满意的结果为止。

我们对规则进行了如下限制：（ 1 ）基本上是基于词类序列的，个别情况下可以具体到词形，不使用任何词典信息。（ 2 ）只往前看“ V+N ”序列右边 1 到 2 个词。

我们首先从标注语料库中抽取了 2000 个“ V+N ”(这里的“ V ”仅限于带名词性宾语的动词，即相当于附录中的 vwn)序列的实例及其所在的句子，然后根据“ V+N ”序列右边第 1 个词和第 2 个词的词性对这些实例进行了排序，以便于验证规则及对结果的分析。归纳规则的过程分为三步：（ 1 ）使用 8 条基于直觉的规则（这些规则只用到右边一个词的词性信息）自动分析了 500 个实例，在与人工分析结果比较之后对原来的规则进行修改，得到 10 条规则。（ 2 ）利用 10 条规则又分析了另外 500 个实例，通过结果分析又对对规则进行了修改，得到了 14 条规则，其中部分规则用到了右边第 2 个词性。（ 3 ）用这 14 条规则对另外 1000 个实例进行了分析，没有增加规则的数目，只是对一些规则进行了一些补充)。

波普（ Popper ）用下面的公式来概括一般科学研究的进展模式：

P1 ——> TT ——> EE ——> P2

这里， P1 表示“问题”（ problem ）， TT 表示“试验性的理论”（ tentative theory ），EE 表示“减少错误”（ error elimination ）， P2 表示“新的问题”[8]。以上显示的语法归纳过程与这一模式是完全一致的。

2.2 “ V+N ”序列的捆绑规则

设“ V+N ”序列所在的句子包含 n 个词（含标点符号）， N 对应的是其中第 i 个词，V 的序号为 i-1 ， N 后的词依次为 Wi+1,Wi+2,...，相应的词性标记为 Ti+1,Ti+2,...。以下是我们通过归纳得到的 14 条规则，用于确定‘ V+N ’是不是一个语法形式，其中每一条规则实际上是判断结果为真的条件，其中 a IN A 表示元素 a 在集合 A 中（词性标记的含义见附录）：

```
R1:  Ti+1=NULL                          /* 句子的长度为 i */
R2:  (Ti+1= ‘ w ’ AND Ti+1!= ‘ wd3 ’ ) OR
     (Ti+1= ‘ wd3 ’ AND Ti+2!= ‘ n ’ AND Ti+2!= ‘ a ’ )
R3:  Ti+1= ‘ y ’                  R4:  Ti+1= ‘ vwn ’
R5:  Ti+1= ‘ vw0 ’ AND Ti+2 IN {NULL, ‘ ag0 ’, ‘ w ’, ‘ vi ’, ‘ vwv ’, ‘ vwn ’}
R6:  Ti+1 IN { ‘ va ’, ‘ vi ’, ‘ vws ’, ‘ vv ’, ‘ vf ’, ‘ vwv ’, ‘ vwj ’ }
R7:  Ti+1= ‘ d ’        R8:  Ti+1= ‘ ur ’          R9:  Ti+1= ‘ usi ’
R10: Ti+1= ‘ c ’ AND Ti+2 IN { ‘ va ’, ‘ vw ’, ‘ d ’, ‘ iv ’ }
```

```
R11: Ti+1=‘usd’ AND Ti+2 IN {‘va’, ‘d’, ‘y’, ‘wd’}
R12: Ti+1=‘ag0’ AND ( (Ti+2=‘w’ AND Ti+2!= ‘wd3’) OR Ti+2=‘y’)
R13: Ti+1=‘ng0’ AND Wi+1 IN {‘时’, ‘时候’, ‘期间’, ‘途中’}
R14: Ti+1=‘nf’ AND Ti+2=‘w’ AND Ti+2!= ‘wd3’
```

3. 实验结果

3.1 封闭测试

应用以上的 14 条规则，我们首先对用以归纳的 2000 个实例进行了自动分析，下表是实验的结果：

规则	识别数	正确数	正确率	规则	识别数	正确数	正确率
R1	13	13	100%	R9	7	7	100%
R2	487	487	100%	R10	10	10	100%
R3	2	2	100%	R11	10	9	90%
R4	50	43	86%	R12	2	2	100%
R5	27	27	100%	R13	6	6	100%
R6	29	29	100%	R14	5	5	100%
R7	22	20	91%				
R8	6	4	67%	合计	675	663	98.4%

在 2000 个实例中，‘V+N’成结构的 822 个，根据以上 14 条规则正确识别了其中的 663 个，另有 12 个识别错误，有 159 个没有识别出来，所以：

准确率 = 正确的识别数/识别的总数 = 663/675 = 98.2%

召回率 = 正确的识别数/‘V+N’结构总数 = 663/822 = 80.7%

准确率和召回率是一对矛盾，准确率越高则召回率越低，反之亦然。我们这里首先考虑的是准确率，因而召回率相对较低。

3.2 开放测试

我们用同样的 14 条规则对另外 1000 个实例进行了自动分析，下表是实验的结果：

规则	识别数	正确数	正确率	规则	识别数	正确数	正确率
R1	10	10	100%	R9	2	2	100%
R2	239	238	100%	R10	3	3	100%
R3	2	2	100%	R11	1	1	100%
R4	35	31	88%	R12	0	0	
R5	8	6	75%	R13	3	3	100%
R6	15	15	100%	R14	2	2	100%
R7	12	9	75%				
R8	3	2	67%	合计	335	324	96.7%

在 1000 个实例中，‘V+N’成结构的 418 个，正确识别了其中的 324 个，另有 11 个识别错误，有 94 个没有识别出来，所以：

准确率 = 正确的识别数/识别的总数 = 324/335 = 96.7%

召回率 = 正确的识别数/‘V+N’结构总数 = 324/418 = 77.5%

开放测试的结果表明，规则已基本收敛，错误类型几乎不再增加。

4. 错误和难点分析

错误分析是指对识别错误的分析，难点分析是指对未识别出的情况的分析。由于开放测试中几乎没有增加新的错误类型，所以下面的分析是基于封闭测试的 2000 个实例的。

4.1 错误分析

在识别出来的 675 例中，有 12 个判断错误。以下分别讨论产生这些错误的原因。

（1） R4 引起的错误(6,括号中为错例数，下同)。例如：

理顺/企业办社会涉及的各种关系　　建立/农民进入市场的有效机制

在以上两例中，N 不是 V 的宾语（V，N 下加下划线，下同），而是后面的主语。在 vwn1+N+vwn2+NP 序列中，可以有两种层次切分：(a) (vwn1 N) (vwn2 NP)，即 N 是 vwn1 的宾语，(vwn1 N) 和(vwn2 NP)构成并列或连动关系；(b)(vwn1 ((N (vwn2 NP))...))，即 N 作后面 VP 的主语，构成的主谓结构成为 vwn1 宾语中的修饰语，比较：

A	B
利用/滩涂养螃蟹	接受/俄解决危机的新建议
前往/兵站检查工作	发展/外商来华投资的良好势头
有/机会参加三峡等大型项目的建设	转变/政府管理经济的职能

在 49 例这类实例中，A 组有 43 个，B 组有 6 个，R4 利用了这种概率的悬殊，但注定了 12.2%的错误率。

（2）由 R7 引起的错误（3）。例如：　达到/历史最高水平

这里，N 不是作 V 的宾语，而是后面 NP 的定语。在 vwn+N+D+VP 中，可以有两种层次切分：(a) (vwn N)(D VP)，构成两个 VP 的并列或连动关系或其他复杂的关系；(b) (vwn (N (D VP)),N 作后面 NP 的定语或 VP 的主语，比较：

A	B
抓经济/不抓其它	干/前人没有干过的事
没有民主/就没有现代化	增强/群众依法维护自身权益的意识

在 22 个这类实例中，A 组 19 个，B 组 3 个，虽然二者的概率比较悬殊，但要得到完全正确的结果光看右边一个词还是不够的。

（3）由 R8 引起的错误（2）。例如：　提供/能源等领域的高新技术

这里，N 不是作 V 的宾语，而是作后面 NP 的定语。“V+N+等”序列有两种层次切分：(a)((V N) 等)，即助词“等”依附在 V+N 构成的 VP 之后；(b) (V (N 等)　，即助词“等”依附在 N 之后。比较：

A	B
加强/贸易等领域的合作	引导农民进市场/等方面的问题
参加/三峡等大型项目的建设	(在)依靠群众、扩大民主、健全法制/等方面

（4）由 R11 引起的错误（1）。例如：

保持/社会的持续稳定

这里，N 不是 V 的宾语，而是加上“的”之后构成 NP 作更大的 NP 的定语。“VWN+N+的”是汉语中常见的歧义结构（详见 4.2），当该序列后面出现副词时，有两种可能的切分：(a) (((VWN N) 的) D　，即“的”字结构作主语，“的”后面的副词修饰 VP 产生更大的 VP 作谓语；(b) (VWN ((N 的) D　，即 N 加“的”构成“的”字结构作 NP 的定语，“的”后面的副词修饰 VP 产生更大的 VP 作 NP 的中心语。比较：

A	B
出席大会的/还有一些普通工人	增进/人民的相互了解
学工科的/不懂艺术	维持/香港的进一步发展

4.2 难点分析

由于我们以上的 14 条规则只看到“V+N”序列右边 1 到 2 个词，所以必然会有一些歧义结构难以消歧。我们根据 Ti+1 把不能用这些规则判断的 159 个难点实例分成 10 类，下面对这 10 种类型分别进行分析。

（1）Ti+1 是性质形容词(ag)

投资工业见效快　　为缩小分歧积极斡旋　　维护大局积极进取

在以上三例中，N 作前面动词 V 的宾语，和后面的形容词没有结构关系。但并非都是如此。事实上，当 V+N+ag 序列后面还有别的词语时，V+N 往往不是语法形式。N 和 ag 可能的关系有：

(a)构成主谓结构，如：　赴/经济发达地区挂职学习　　启用/实绩突出的干部

(b)分别作后面 NP 的定语，如：　利用/辽宁丰富的资源　　实现/柬埔寨全面民族和解

(c)构成偏正关系，如：　缓解/资金困难　　关乎/大局稳定

在 32 个该类实例中，V+N 是语法形式的只有 3 个。

（2）Ti+1 是连词(c)或顿号(wd3)

并列连词或顿号（由于顿号与连词的功能相当，故统称连词）连接的成分构成并列关系，并列成分在功能上是一致的，所以 R2 和 R10 可以得到正确的结果。 R2 的意思是：在 V+N+wd3 序列右边的词不是名词或形容词时， V+N 是一个语法形式。反过来说，当其右边的词是名词或形容词时，顿号右边可能是一个 NP，该 NP 和 N 构成并列关系，于是 V+N 不能成为语法结构，如：

A	B
办/实事、好事	发展/经济、科技、教育事业
盗窃/人民币、外币及金银首饰	服从/国家、全局和长远利益

A 组例子中，连词右边的 NP 和 N 构成并列关系，整个并列结构做 V 的宾语。 B 组例子中，并列 NP 做定语，不直接作 V 的宾语。在这两组中， V+N 都不是语法形式。

但是，组成并列结构的成分并不一定是功能一致的，如：

C	D
细川的借款和股票问题	缺水、缺土地、交通闭塞
盲目上项目和开发区热、房地产热的现象	在修建校舍、民办教师转正、师资定向培养等方面

C 组例子中，连词和左边的是 VP，右边的是 NP， D 组例子中，主谓结构和动宾结构并列在一起。所以这就产生了结构上的歧义：(a)连词右边的 NP 和 N 组成并列关系，如 A 组和 B 组；(b) 连词右边的 NP 和由 V+N 构成的 VP 构成并列关系，如 C 组；(c)连词右边的 NP 和后边的 VP 构成主谓结构，主谓结构又和连词前面的 V+N 构成并列关系，如 D 组。

（3） Ti+1 是名词性后缀（ kn ）

V+N+kn 是一个歧义结构，因为名词性后缀既可以依附在 NP 后，又可以依附在 VP 后，即该结构可以有两种层次切分：(a) ((V N) kn)；(b) (V (N kn))。比较：

A	B
违反纪律者	扩大/贸易额
造血型	具有/时效性

如果后面还有别的词语的话， N+kn 构成的 NP 还有可能作后面 NP 的定语，如：

作为/企业界的代表	加强/房地产业的管理
关注/苗头性的问题	下达/指令性扶贫任务

（4） Ti+1 是方位词（ nf ）

方位词也被称为“后置词”，因为它经常依附在别的结构之后构成一个方位结构。方位词依附的对象可以是 NP，也可以是 VP，因此在 V+N+nf 中就可能产生结构歧义： nf 可能依附在 N 后，也可能依附在(V+N)后，比较：

A	B
处于困境中	进行/理论上的创造
引进外资中	开展/政府间经济合作

在 A 组中， V+N 是一个语法形式，在 B 组中， N+nf 是一个语法形式，而 V+N 则不是一个语法形式。

（5） Ti+1 是普通名词(ng0)或专有名词(np)

“ V+N1+N2 ”是一个歧义结构，它有两种不同的层次划分：(a) ((V N1) N2)，即 V+N1 构成动宾结构作定语；(b) (V (N1 N2))，即 N1+N2 构成 NP 作 V 的宾语，比较：

A	B
化解矛盾工作	化解/矛盾纠纷
联系会员制度	输出/技术项目
建设国家方面	维护/国家主权

如果考虑到后面还有别的词语的话， N1+N2 还可能是更大 NP 的一部分，这个更大的 NP 作 V 的宾语，如：

制定/专业法和行业法规	加快/产业结构和产品结构的调整
小看/农村治安问题	建立/现代企业制度

另外， N2 和 N1 有可能还没有结构关系，如：

天增岁月人增寿	病魔无情人有情

（6） Ti+1 是介词(p)

V+N+P 不是一个语法形式，介词是“前置词”，它不可能后附， P 后一定有其宾语。

但是，介词结构可以和“的”构成“的”字结构作定语，又可以独立作状语，因而造成了V+N后有两种可能可能：或者为NP，或者为VP，整个结构于是有两种可能的层次划分：(a) (V (N NP))；(b) ((V N) VP)。比较：

A	B
强化/中央对税收的集中统一管理	考察干部以速度为政绩
恢复/中国在关贸总协定中的缔约国地位	代黄菜向人民法院提起上诉
发挥/市场在资源配置中的基础性作用	有机会在地方选举中获胜

在B组中，V+N是一个语法形式，而在A组中，V+N不是一个语法形式。

（7）Ti+1是代词(ra)

代词的功能比较复杂，这里只涉及做定语的代词。V+N+ra不是一个语法形式，我们把由ra作定语的NP记为RP，则V+N+RP有两种可能的层次划分：(a) ((V N) RP)，这是一个复指结构，RP复指V+N；(b) (V (N RP))，即N作RP的定语，N+RP加“的”之后作更大的NP的定语，比较：

A	B
改造内阁这个问题	解决/萨拉热窝这场危机
救新郎其事	借鉴/世界各国的文明成果
取代自民党这一共同目标	取得/社会各阶层的支持

A组中，V+N是语法形式，在B组中，V+N不是语法形式。这里，如果RP既可以在语义上复指N，又可以复指V+N，则会产生语义的歧义，如：

发展生产力这个根本点

（8）Ti+1是不带宾语的动词(vw0)

V+N+V+的+N是一个歧义结构，它至少可以有两种切分：(a) (((V N1) V) 的) N2)；(b) (V (((N1 V) 的) N2)。比较：

A	B
勇斗歹徒受伤的人员	控制/工资增加的水平
来大陆旅游的台胞	跟踪/经济运行的态势
进村调查的科长	接受/欧盟批准的计划

在A组中，V+N1是语法形式，在B组中，V+N1不是语法形式。

（9）Ti+1是作NP修饰语的动词(vtp)

V+N1+V+N2也是一个歧义结构，它有两种可能的层次划分：(a) (V (N1 (V N2))); (b) ((V N1) (V N2))。比较：

A	B
分流/农村剩余劳动力	驻波黑维和部队
反对/北约空袭计划	反政府武装组织
提高/资金使用效益	建设国家领导核心

在A组中，V+N1不是语法形式，在B组中，V+N1是语法形式

（10）Ti+1是助词“的”或“之”

这一类即N右边是助词“的”或“之”的情况，在253个实例中，“的”的实例为251个，“之”的实例只有2个，在“的”的实例中，有99个是V+N成结构的例子，在“之”的实例中有1个是V+N成结构的例子。由于“的”和“之”的性质相同，所以放在一起分析，以下用“的”的地方也适用于“之”。

“V+N+的+X”是一个歧义结构，朱德熙先生早在60年代初就发现了这种歧义结构([1])，它可以有两种不同的层次构造：(a) (V ((N 的) X))；(b) (((V N) 的) X)。这里的X表示“的”后面的成分既可以是NP，也可以是VP。比较：

A	B
维护/职工的利益	牵动全局的大事
减轻/农民的负担	解决危机的建议
遵/父母之命	要官之风

显然，B组中的“V+N”是一个语法形式，A组中的“V+N”不是一个语法形式。这类歧义结构有时实例化之后仍不能消除歧义，如著名的例子“咬死了猎人的狗”就

属于此类([1],[2],[5])，下面是几个同类的例子：

有创造性的思维　　留恋过去的感情

当工程师的丈夫　　包围萨拉热窝的炮兵阵地

下表分别列出了以上 10 类的实例总数及其中 V+N 成结构的实例数（用规则已抽出的实例除外）：

类型	实例总数	V+N 结构数	类型	实例总数	V+N 结构数
1	32	3	7	26	10
2	57	5	8	8	4
3	17	3	9	50	11
4	18	3	9	39	4
5	626	16	10	253	100

5. 结论

我们从关于“ V+N ”序列捆绑的 8 条简单规则出发，通过 2000 个实例的归纳得到了 14 条简单的规则，用以确定“ V+N ”是不是一个合法的短语，准确率达到 98%左右，召回率达到 80%左右。由于我们只用到了右边 1-2 词的词类信息，如果使上下文再大一些，则有望进一步提高性能。实验表明，从标注语料库中归纳语法规则是一条切实可行的路子。

通过对错误和难点的分析，我们发现几乎所有的分析错误和难点都是由词类序列的结构歧义造成的。通过对这些实例的分析，我们发现了一系列歧义结构，深入研究这些歧义结构对于自然语言处理以及语言学的研究都具有重要的意义([3],[4])。研究表明，从语料库中归纳语法规则既是建立自然语言形式语法体系的有效途径，也是深化语法研究的重要途径。这种方法还可以应用在信息抽取、文本分类等应用领域。

参考文献

[1]朱德熙(1962)，句法结构，《中国语文》 8-9 期。

[2]朱德熙(1980)，汉语句法中的歧义现象，《中国语文》 2 期。

[3]孙茂松，黄昌宁(1989)，汉语中的兼类词、同形词类组及其处理策略，《中文信息学报》 Vol.3,No.4.

[4]钱树人(1993)，歧义、系统歧义和语境，《中文信息学报》 Vol.7,No.2.

[5]冯志伟(1995)，论歧义结构的潜在性，《中文信息学报》 Vol.9,No.4.

[6]马真，陆俭明(1996)，“名词+动词”词语串研究，《中国语文》 4 期。

[7]孙宏林等(1996)，“现代汉语研究语料库系统”概述，计算机时代的汉语和汉字研究，罗振声、袁毓林主编，清华大学出版社。

[8]Geoffery Leech,1992. Corpora and theories of linguistic performance,in Directions in Corpus Linguistics,(ed.) by Jan Svartvik,Berlin:Mouton de Gruyter,105-122.

[9]Geoffery Leech,Roger Garside,1991.Running a grammar factory:The production of syntactically analysed corpora or treebanks,in English Computer Corpora,(ed.) by Stig Johanson et al.,Berlin:Mouton de Gruyter.

附录　　规则中用到的词性标记注释

n	名词	np	专名	ng0	普通名词		
nf	方位词	va	助动词	vi	系动词		
vf	形式动词	vv	“来/去”+VP	vtp	V 作 NP 修饰语		
vw0	V 不带宾	vwn	V 带体词宾语	vwv	V 带谓词宾语		
vws	V 带小句宾语	vwj	V 带兼语宾语	vwc	V 作补语		
ag	性质形容词	ra	代词作定语	p	介词		
d	副词	c	连词	usd	助词“的”		
usz	助词“之”	uss	助词“所”	ur	其它助词		
y	语气词	kn	名词后缀	iv	谓词性成语		
w	标点符号	wb	标号	wd	点号	wd3	顿号