超级标注：准句法剖析的一种方法*

Srinivas Bangalore
AT & T Labs-Research, USA
Joshi Aravind K.
University of Pennsylvania, USA

1. 引言

本文介绍一种健壮型句法剖析方法——“超级标注”。这种方法把语言学意义上的词项描写跟统计技术结合在一起。我们的观点是，如果词项在局部语境复杂的限制条件下得以详尽描写（超级标注），语言结构的计算方法才可以确定。如此获得的每一词项的描写相对地十分繁杂，对句法剖析器来说，增加了局部歧义的现象，但这种局部歧义可以通过运用从剖析形式的语料库中收集得来的、超级标注出现的概率分布来加以消除。超级标注消歧，事实上成为一种分析形式（准剖析）的表述方式。

在语言学研究中，如何更详尽地描写词项可以有许多方法。

* [原文出处] “Supertagging: An Approach to Almost Parsing” *Computational Linguistics*, 1999, 25: 2, 237 - 265.

我们的想法是，把词项的描写落实在同一描写框架里词项施加限制的元素之上。另外，对每一词项的描写必须详尽地落实到词项本身能出现的各种不同的句法环境当中。当然，这样做会增加句法剖析器局部出现歧义的机会。句法剖析器甚至要在词项描写集整合之前就要决定在词项描写集中，哪一项详尽的描写更适用于句子的解读。解决的办法很明显，就是让句法剖析器负担起整个工作。句法剖析器可能会为所有的描写消歧，并且相对于句子的某项解读，就每一词项选择一项描写。然而，有另一种可供选择的句法剖析方法，以减轻句法剖析器消歧上的工作量。这个想法是，局部地检查词项描写中显示出来的种种限制，以图消除不相容的描写。[①]在消歧过程中，也可以利用语料库的统计信息。

Joshi & Srinivas（1994）首先把这些想法应用于词汇化的树连接语法（Lexicalized Tree Adjoining Grammar：LTAG)。这些技术也可以应用于其他的词汇化语法。本文介绍改良了的超级标注所取得的消歧结果，我们运用一个更大规模的训练语料库和更佳的平滑技术，使得准确性从先前发表的 68％提升到 92％。本文第二节概述健壮型的句法剖析方法；第三节简单介绍种种词汇化的树连接语法；第四节举例说明超级标注消歧的目的；第五节和第六节详细讨论超级标注消歧的种种方法及结果；第七节讨论在句法剖析前进行超级标注消歧所取得的成效；第八节简介应用超级标注输出信息的一个健壮型的、轻量的依存关系分析器；第九节讨论超级标注消歧技术应用于其他词汇化语法的可行性。

2. 相关的方法

近年来，关于自然语言的健壮型句法剖析有了一些探索。宽泛地说，它们可归入两类：一是基于有限状态语法的句法剖析

器，另一是统计句法剖析器。下面简单地介绍这两类方法并给我们的健壮型句法剖析方法定位。

2.1 基于有限状态语法的句法剖析器

有关基于有限状态语法的句法剖析方法，参考 Joshi (1960), Abney (1990), Appelt 等 (1993), Roche (1993), Grishman (1995), Hobbs 等 (1997), Joshi & Hopely (1997) 以及 Karttunen 等 (1997)。他们把语法用作级联式有限状态规则表述的识别程序。规则表述通常靠人工完成。级联中的每一项识别程序提供一项局部的最佳输出。这些系统的输出并不体现为成分结构，多数体现为名词词组及动词词组，一般称为“浅层分析”。浅层分析并不存在短语层面或修饰语层面的附加物。这些句法剖析器经常产生一项输出，这是因为它们当多于一项规则表述在一个特定位置上与输入串列匹配时运用的是最长的启发式匹配来消歧。就目前来说，这些系统当中不用任何统计信息来消歧。语法本身可分为“领域一独立”和“领域一专指”两种规则表述，这意味着转移到一个新的领域时，会导致重写“领域一依存”的规则表述。这种方法作为信息提取系统的预处理器是颇为成功的 (Hobbs等 1995, Grishman 1995).

2.2 统计的句法剖析器

此方法由 IBM 自然语言小组(Fujisaki 等 1989) 开创，而后续的研究中有 Schabes, Roth & Osborne (1993), Jelinek 等 (1994), Magerman (1995), Collins (1996) 以及 Charniak (1997) 这个方法把下列两个问题分开处理：输入串列合格性条件的问题和给该串列指定一项结构的问题。这些系统试图给每一项输入串列都指定某些结构。给输入项指定的结构规则自动从大规模标注语料库中提取，然后，这些规则需要对语言获得合理的

覆盖。由此而产生的一套规则集在语言学意义上说并不透明，而且也不容易更改。词语上和结构上的歧义是通过运用编进规则里的概率信息加以消除。这使系统能够为每一个输入项指定最可取的结构。这些系统的输出项包括成分分析，成分分析的细致程度取决于用来训练系统的树库中的标注的细致程度。

也有一些句法剖析器把概率（加权）信息跟人工语法结合起来使用，如 Black 等（1993），Nagao（1994），Alshawi & Carter (1994) 以及 Srinivas, Doran & Kulick (1995)。他们运用概率信息，首先是用来对句法剖析器所产生的分析形式分级排列，而不是着眼于系统本身的健壮性。

3. 词汇化语法

词汇化语法特别适用于对自然语言语法的规范性描述。词汇在语言形式化的种种模式中起着重要的作用，如：词汇功能语法（Kaplan & Bresnan 1983），广义短语结构语法（Gazdar 等 1985），中心词驱动短语结构语法（Pollard & Sag 1987），组合范畴语法（Steedman 1987），词汇一语法（Gross 1984），（Schabes & Joshi 1991），连接语法（Sleator & Temperley 1991）以及管辖约束语法的某些变体（Chomsky 1992）。句法剖析、词汇语义学以及机器翻译，在此仅举数项，都得益于词汇化。词汇化为在词汇中把句法和语义信息结合起来提供了一个清晰的界面。下面我们联系到局部句法剖析来讨论词汇化的价值以及其他相关问题，并且简略地介绍一下作为词汇化语法这一类别的代表的基于特征的词汇化树连接语法（Feature－based Lexicalized Tree Adjoining Grammar, FB－LTAG）。

FB－LTAG（Joshi, Levy & Takahashi 1975, Vijay-Shanker

1987, Schabes, Abeille & Joshi 1988, Vijay-Shanker & Joshi 1991, Joshi & Schabes 1996）是一种树一重写形式语法，跟属于串列一重写形式语法，如上下文无关语法和中心词语法不同。FB－LTAG 的基本元素称为“基本树”。每一棵基本树在边缘都起码跟一个词项挂钩。跟基本树挂钩的词项称为该树的锚。基本树即为锚的详尽描写，它提供锚所确定的句法和语义（谓语论元）的限制的位置领域。基本树分两种：（a）初始树和（b）辅助树。在面向自然语言的 FB－LTAG 中，初始树为不含递归的简单句的短语结构树，辅助树则含递归结构。基本树由替换及附加两种方法结合而成。把多棵基本树结合起来就成了推导树，而把基本树结合起来以产生句子的分析形式的过程则用推导树来表述。推导树也可以理解为在句子的词之间不带标记弧的依存树。

4．超级标注

词性消歧技术（词性标注器）（Church 1988，Weischedel 等 1993，Brill 1993）的应用经常先于句法剖析，以图消除或大量减少词性歧义。词性标注器都是局部的，因为它们运用有限语境的信息来确定某一个词选择的标记。众所周知，这些词性标注器都比较成功。

在词汇化语法里，如词汇化树连接语法，每一词项起码跟一项基本结构，即基本树挂钩。

词汇化树连接语法的基本结构，通过限定所有的依存元素（而且只有这些依存元素）皆出现于同一的结构当中，把种种依存关系包括长距离的依存关系加以定域化。定域的结果是，一个词项可能（一般来说，十分可能）跟超过一项基本结构挂钩。我们把这些基本结构称为“超级标注”以便和标准的词性标记区分

开来。需要指出的是，就算一个词具有单一的标准词性，譬如说动词(V)，它通常会有多于一个的超级标注跟它挂钩。既然当句法剖析完成的时候，每一词项只有一个超级标注（假设没有整体性歧义），词汇化树连接语法（Schabes, Abeille & Joshi 1988）需要寻找一大片的超级标注，以在把它们结合起来分析句子之前为每一词项选择正确的超级标注。本文要谈的就是如何运用超级标注消歧的问题。

既然词汇化树连接语法为词汇化语法，我们获得一个新的机会在进行句法剖析之前利用局部信息，如局部词汇依存关系来消除或大量减少超级标注分派带来的歧义。如在标准的词性消歧一样，可以通过 n-元模式运用局部统计信息；n-元模式是基于在一个词汇化树连接语法已分析过的语料库中超级标注的分布。再者，既然超级标注为依存信息编码，也可以利用在一个特定的超级标注和它的附属标记之间的距离分布信息。

需要指出的是，正如标准的词性消歧一样，超级标注消歧也可以由句法剖析器完成。可是，先于句法剖析而进行词性消歧大大减轻了句法剖析器的工作并使整个工作进度加快，而超级标注消歧则更进一步减轻句法剖析器的工作。超级标注消歧之后，差不多已经完成整个句法剖析的工作。这时，句法剖析器只需把各个结构结合起来，即“准剖析”这种方法也可以用于把结构跟句段连接起来。

4.1 超级标注举例

由于具有扩充的位置领域（Extended Domain of Locality, EDL)的特性[②]，词汇化树连接语法把每一词项跟一棵代表着词项在其中出现的每一个句法环境的基本树连接起来。其结果是每一词项都毫无例外地跟多于一棵的基本树相连接。我们把跟每一词项连接的基本结构称为超级标注[③]。图 1 显示跟右列句子每一

词项连接的一些基本树：the purchase price includes two ancillary companies。表 1 提供例句环境，在图 1 当中显示的每一个超级标注都包括在内。

表 1 图 1所显示的超级标记在当中用得上的句法段的例释

超级标记	句法段	示例
α1	名词性谓语	this is the *purchase*
α2	名词短语	the *price*
α3	主题化结构	Almost everything, the price *includes*
α4	形容词性谓语	this is *ancillary*
α5	名词短语	the *company*
β1	限定词	*the* company
β2	名词性修饰语	*purchase* order
α6	名词性谓语/主语提取	what is the *price*
α7	命令式	*include* the share price
β3	限定词	*two* hundred men
β4	形容词性修饰语	*ancillary* unit
α8	名词性谓语/主语提取	which are the *companies*
α9	名词短语	*purchases* have not increased.
α10	名词性谓语	this is the *price*
α11	及物动词	the price *includes* everything
α12	形容词谓语/主语提取	what is *ancillary*
α13	名词短语	*companies* have not been profitable

图 1 显示配置给句子“ the purchase price includes two ancillary companies ”中每一词项的超级标注的初始集。图中超级标注的次序并不重要。图 2 同时也显示为超级标注器配置的最终的超级标注序列，超级标注器利用有关单个超级标注和它们对其他超级标注的依存关系的统计信息（见下面第 6 节）来确定最佳的超级标注序列。被选上的超级标注结合起来以推导出分析形式。如果不用超级标注器，句法剖析器不得不处理整个树集的结合（即

至少为显示出来的 17 棵树）；用超级标注，句法剖析器只需处理 7 棵树的结合。

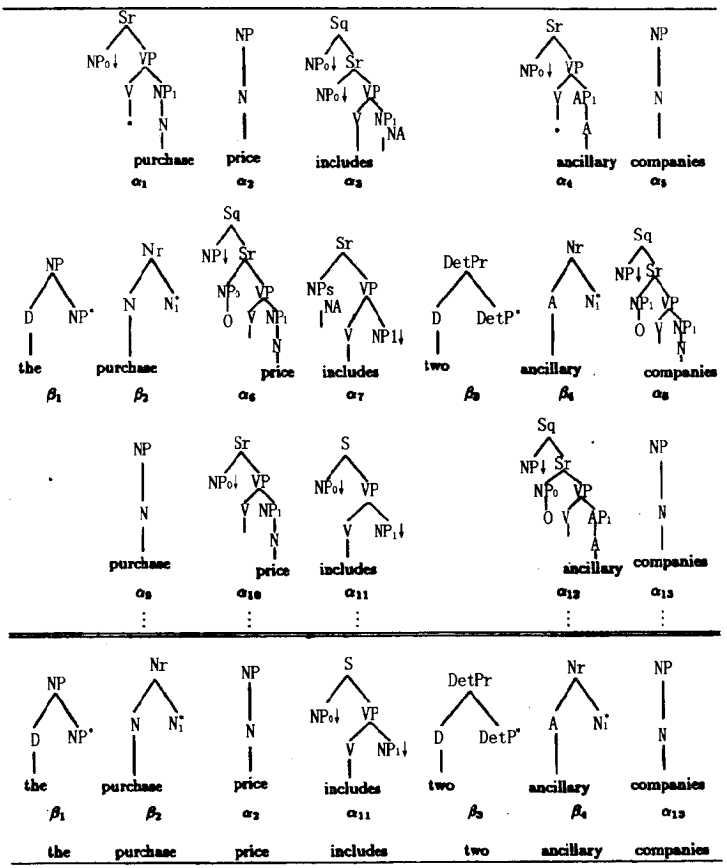

图 1 跟句子“ the purchase price includes two ancillary companies' ，词项连接的超级标记选例

发送	the	purchase	price	includes	two	ancillary	companies.
初始指派	β_1	α_1 β_2 α_9 ⋮	α_2 α_6 α_{10} ⋮	α_3 α_7 α_{11} ⋮	β_3	α_4 β_4 α_{12} ⋮	α_5 α_8 α_{13} ⋮
最终指派	β_1	β_2	α_2	α_{11}	β_3	β_4	α_{13}

图2　句子“the purchase price includes two ancillary companies”超级标注的消歧

5. 运用结构信息来减少超级标注的歧义

我们应该把超级标注的结构看作是为其运用于当中的句法环境提供可容许的限制。某些限制可以在位置上检查出来。下面介绍某些可以用来决定一个超级标注句法环境的可容许性限制④：

——超级标注的跨度。所谓超级标注的跨度，指的是超级标注可覆盖的最小数目的词项。

超级标注的每一个替换位置在输入中至少可以覆盖一个词项。我们可以运用一条简单的规则来消除基于跨度限制的超级标注，那就是，如果超级标注的跨度大于输入串列，那么，超级标注在输入串列的任何分析形式中都无法运用。

——左（右）跨度限制。如果处于锚的左（右）边的超级标注跨度大于超级标注与之挂钩的词项的左（右）边串列的长度，那么，超级标注在输入串列的任何分析形式中都无法运用。

——超级标注中的词项。超级标注可能被消除，如果出现在超级标注边缘的终端并没有出现在输入串列之中。

附带着表示被动结构的固有词项 by 的超级标注在主动句的

分析形式中一律不予考虑。一般地说，这些限制可以用来消除那些其特征不能在输入串列中得到满足的超级标注。譬如说，当输入串列并不包含 wh-词而超级标注则要求一项 wh + NP 时，这个标记就会被除掉。

表 2 显示运用与超级标注歧义（不含结构限制）相关的结构限制导致华尔街日报（WSJ）2012 个句子（48 763 个词）中超级标注歧义缩减的情况[5]。

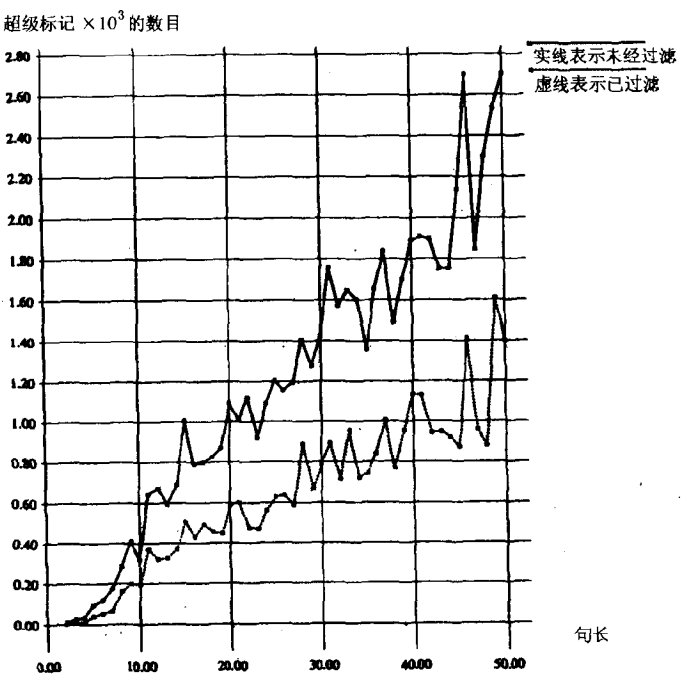

图 3　包含 2 至 50 个词的句子过滤与没过滤的超级标记数目上的比例差百分比

表 2　用上和没用上结构限制的超级标记歧义性

系统	词的总数	超级标记/词的平均数
无结构限制	48,783	47．0
带结构限制	48,783	25．0

这些过滤器对减低超级标注歧义来说，是十分有效的。图 3 的坐标是长度为 2 至 50个词，附带和不附带过滤器的句子标出在句子层面上超级标注的数目。我们在坐标上可以看到，当过滤器用上的时候，超级标注歧义就大大降低。图 4 的坐标显示由于

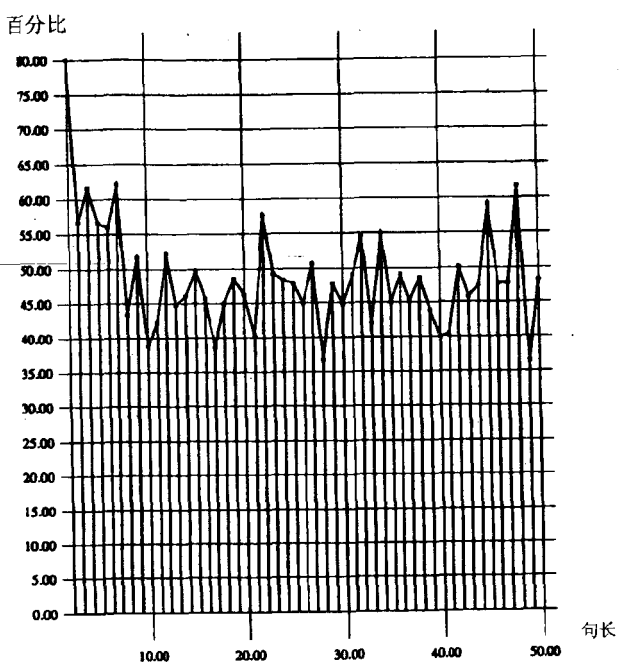

图 4　包含 2至 50 个词的句子过滤与没过滤的超级标记数目上的比较

对包含 2 至 50 个词长度的句子加以过滤而出现超级标注歧义上的比例差。我们看到，超级标注歧义的平均减幅为 50% 左右。这意味着就一个句子来说，仅仅运用超级标注的结构限制，大约 50% 的超级标注歧义在句法剖析之前就可以被消除。超级标注歧义的减缩大大地加速句法剖析器的工作。事实上，在 XTAG 系统里的超级标注歧义是如此之大，以致于句法剖析器如果不用过滤器的话，工作速度会非常明显地减慢。

表 3 列举了运用各种词类的过滤器而获得的超级标注的缩减情况[6]。各种形态的动词最容易引起超级标注歧义的问题，而动词多数的超级标注歧义的问题则来自轻动词和动词助词。过滤器十分有效，它们可以消除 50% 以上的与动词挂钩的超级标注。

表 3 词类与过滤器对超级标记歧义性的影响的对应表

词类	没用过滤器的超级标记平均值	用上过滤器的超级标记平均值	超级标记歧义下降比例
VBP	516.5	250.0	51.6
VB	435.8	224.9	48.4
VBD	209.0	100.7	51.8
VBN	188.2	74.7	60.3
MD	167.2	121.0	27.6
VBZ	165.1	71.6	56.6
VBG	100.7	49.8	50.5
RP	34.5	30.9	10.5
IN	24.3	20.9	14.0
JJS	23.8	12.7	46.9
WRB	23.1	14.3	38.2
JJR	22.7	14.2	37.7
JJ	21.7	13.5	37.9

词类	没用过滤器的 超级标记平均值	用上过滤器的 超级标记平均值	超级标记歧义 下降比例
,	20.0	10.7	46.6
NN	19.8	10.7	46.0
NNS	17.0	10.5	38.6
NNP	15.0	10.2	31.9
NNPS	15.0	10.2	32.1
LS	15.0	15.0	0.0
FW	15.0	15.0	0.0
-RRB-	15.0	10.7	28.4
-LRB-	15.0	12.3	18.0
RBR	14.9	9.5	36.3
RBS	14.9	6.1	59.2
CC	14.8	3.4	76.9
EX	14.0	5.8	58.7
CD	13.3	9.9	25.8
TO	11.3	10.8	4.5
PRP	10.7	5.3	50.2
UH	10.0	3.0	70.0
RB	10.0	5.3	46.4
"	6.0	3.2	47.7
:	5.5	3.2	42.1
PDT	5.4	4.9	9.0
WP	4.6	2.9	35.8
WP$	4.0	1.8	36.2
DT	3.9	3.1	21.8

词类	没用过滤器的超级标记平均值	用上过滤器的超级标记平均值	超级标记歧义下降比例
PRP $	3.8	2.9	22.2
.	3.0	1.0	65.4
POS	2.5	2.1	13.9
WDT	1.2	1.1	5.5

结构限制在减少超级标注歧义方面是有效的，句法剖析器的搜索空间很大。在下面几节当中，介绍超级标注消歧的随机方法和基于规则的方法。

6. 模式、数据、试验和结果

在讨论超级标注消歧的种种模式之前，回顾一下这项工作的时间进程。我们这样做，不仅是为了展示从 1994 年报告过的早期工作至今所取得的进展，也是为了解释我们之所以选择某些超级标注消歧模式的理据。在下面的小节中简单地总结一下早期的工作。

6.1 早期工作

Joshi & Srinivas（1994）运用三元模式和依存模式为超级标注消歧进行了试验。三元模式为测试集里 68% 的词给出正确的超级标注。这个三元模式是在词类对、超级标注对而不是在词对、超级标注对上受训练；至于词类、超级标注的一对是来自 5000 句（华尔街日报）的句子的 LTAG 推导，并在 100 句（华尔街日报）句子测试过。在后续的工作中，使用了一个规模更大的训练库，并加入了平滑技术而明显地改进了三元模式的性能。

在第 6.5 节会详细讨论这个模式以及它应用于一系列语料库中的效果。在第 6.2 节，我们简略地说明在早期工作中提到的超级标注的依存模式。

6.2 依存模式

在消除超级标注歧义的 n-元模式中，在 n-词窗口以外出现的超级标注彼此之间的依存关系不能合并。如果窗口的大小并不受到预设约束的话，这种限制可以克服，相反要维持每一个超级标注的依附标记之间距离的概率分布。我们用一种很明显的方法来界定超级标注之间的依存关系，即：一个超级标注依存于另一个超级标注，如果前者替换后者或附加于后者。这样，替换本身和超级标注的尾节可以看成是指定该超级标注的依存要求。一个超级标注依附于另一个超级标注的概率是从以推导结构标注的句库中收集得来的。既然给出为每一词项而设的超级标注集以及超级标注对之间的依存信息，依存模式的目标是计算出跨越整个串列上最像样的依存连接。算出来的依存连接就是超级标注序列，一个词项一个序列，再附上依存信息。

Joshi & Srinivas (1994) 提出这个超级标注模式以后却没有继续试验，主要有两个原因。首先，缺乏一个大规模的、其推导结构已被 LTAG 分析过的语料库，而这样的语料库对于可靠地评估这个模式的种种参数是必需的。我们现在正在建立一个大规模的、已被 LTAG 剖析过的《华尔街日报》语料库，当中每一句子已为正确的推导所标注。不用依存模式进行超级标注的第二个原因是，超级标注的目标在于探讨局部技术甚至在句法剖析开始之前给超级标注消歧可达到什么样的应用程度。但是，依存模式跟整体句法剖析基本相似而失去超级标注的作用。

6.3 附有平滑技术的 n-元模式

我们通过把平滑技术加进模式之中，通过在一个更大规模的训练语料库基础上训练该模式，提高了三元模式的效能。同时，我们也提出一些有关超级标注消歧新的模式。上述几点，我们在这一节中详加讨论。

为了训练和测试超级标注消歧的模式，我们利用了两个数据集。第一个集子是通过对下列的语料库进行剖析而收集起来的：华尔街日报[7]、IBM手册和 ATIS 语料库。ATIS语料库的建立是基于作为 XTAG 系统(Doran 等 1994) 的一部分开发出来的覆盖面广的英语语法（Doran 等 1994)。这些语料库的每一个句子都会附上正确的推导，而正确的推导则来自由 XTAG 系统产生的全部推导。

第二个是规模更大的数据集。它是通过把宾大树库中（华尔街日报）的句子剖析而加以改造后收集起来的。目标是把句子中的每一词项跟一个超级标注连接起来，而句子的短语结构分析形式业已给出。这个过程包含着一些基于局部树语境的经验知识。经验知识利用有关词的支配节点的标记（如，父母、祖父母和曾祖父母），其兄弟节点的标记（左和右）及其父母的兄弟节点的标记的信息。图 5 显示这种改造结果的一个例证。应该注意的是，这种改造并非完美无缺，其准确度仅是就一阶近似值而言。这是往往由于改造时的错误以及由于在宾大树库分析形式中某些信息的短缺，如缺乏把附加补足语跟介词短语论元区分出来的信息。就算改造过的超级标注语料库可以进一步改善，语料库就其目前的形态而言在改进超级标注模式的性能方面已是一个十分有用的资源。下面加以讨论。

```
((“S”
    (“NP－SBJ”(“NNP”“Mr.”(“NNP”“Vinken” ))
```

```
("VP"("VBZ""is")
 ("NP-PRD"
  ("NP"("NN""chairman"))
  ("PP"("IN""of")
   ("NP"
    ("NP"("NNP"Elsevier")("NNP""N.V."))
    (",",")
    ("NP"("DT""the")("NNP""Dutch")("VBG""publishing")("NN"
"group"))))))
 (".""."))
```

Mr.//NNP//B_Nn	（名词修饰语）
Vinken//NNP//A_NXN	（中心名词）
is//VBZ//B_Vvx	（助动词）
chairman//NN//A_nxON1	（谓语名词）
of//IN//B_nxPnx	（名词－附加介词）
Elsevier//NNP//B_Nn	（名词修饰语）
N.V.//NNP//A_NXN	（中心名词）
,//,//B_nxPUnxpu	（同位逗号）
the//DT//B_Dnx	（限定词）
Dutch//NNP//B_Nn	（名词修饰语）
publishing//VBG//B_Vn	（分词动词，名词性修饰语）
group//NN//A_NXN	（中心名词）
.//.//B_sPU	（句子标点符号）

图 5 （华尔街日报》句子“Mr. Vinken is chairman of Elesvier N.V., the Dutch publishing group”的短语结构树和从短语结构树获取的超级标记

6.4 一元模式

利用结构信息把不能应用于输入串列的任何分析形式的超级标注过滤掉，减少了超级标注的歧义性，但很明显并不能完全加以消除。消除跟每一词项挂钩的超级标注歧义性方法之一，是依据词汇优先关系来把超级标注排序；词汇优先关系是指词项与超