汉语言文字学书系
HANYUYANWENZIXUESHUXI

计算语言学视窗

WINDOW TO THE COMPUTATIONAL LINGUISTICS

靳光瑾◉编译

借鉴是创新的金钥匙
介绍国外计算语言学的最新成果

北京广播学院出版社

计算语言学视窗

WINDOW TO THE COMPUTATIONAL LINGUISTICS

靳光瑾◉编译　郑定欧◉审校

北京广播学院出版社

图书在版编目（CIP）数据

计算语言学视窗/靳光瑾编译．－北京：北京广播学院出版社，2003.6
（汉语言文学书系）
ISBN 7－81085－140－3

Ⅰ．计…　Ⅱ．靳…　Ⅲ．数理语言学－研究　Ⅳ．H087

中国版本图书馆 CIP 数据核字（2003）第 046392 号

计算语言学视窗

编　　译：靳光瑾
审　　校：郑定欧
责任编辑：公　田　李钊祥
封面设计：晓强工作室

出版发行：北京广播学院出版社
社　　址：北京市朝阳区定福庄东街 1 号　邮编：100024
电　　话：010－65738557　65738538　　传真：010　65779405
网　　址：http://www.cbbip.bbi.edu.cn
经　　销：新华书店总店北京发行所
印　　刷：北京中科印刷有限公司

开　　本：850×1168 毫米　1/32
印　　张：10.25
版　　次：2003 年 6 月第 1 版　2003 年 6 月第 1 次印刷

ISBN 7－81085－140－3/N·48　　定价：19.00 元

前　言

想出一本集子，打开国际视野，介绍国外在计算语言学方面的研究成果的想法由来已久。着手收集论文是从法国著名计算语言学家莫里斯·格罗斯教授来中国访问时开始的。

2000年夏天，莫里斯·格罗斯教授和克莱蒂安·勒克尔教授应教育部语言文字应用研究所邀请来北京讲学，在学界引起了学者们的关注。莫里斯·格罗斯倡导的词汇－语法学派在法国已经取得很大成果，而且在欧洲带动了一大批语言学家和计算语言学家做了非常扎实的基础研究工作。有关词汇－语法学派的理论方法已由香港城市大学郑定欧博士在《词汇语法理论与汉语句法研究》（北京语言大学出版社，1999）专著中向中国学仁做了介绍。继莫里斯·格罗斯访问后，词汇－语法学派的核心人物，法国蒙纳瓦莱大学艾力克·拉波特教授又于2002年初春来语用所进行学术访问。同年秋季，教育部语信司司长、语用所所长李宇明教授应法国蒙纳瓦莱大学邀请率“教育部语料库考察团”赴法国访问。

译者2001年9月参加了在伦敦召开的“第20届词汇－语法

国际研讨会，以及2002年夏天在英国利物浦大学作学术访问，都接触了大量的计算语言学研究成果，因而更迫切地感到学术交流的重要。目前译者应法国蒙纳瓦莱大学邀请做博士后访问研究。一系列的学术交流活动，促成了这本译文集的诞生。在这里特别要指出的是，这个集子的诞生还得助于郑定欧博士。郑定欧博士是法国词汇－语法基本队伍的成员之一，有着广泛的国际联系。从收集资料到确定篇目，从译文处理到审校，郑博士都提出了不少宝贵的意见。

经过筛选、编排，本集共收论文10篇。以下分别摘要介绍各篇的主要内容：

1．超级标注：准句法剖析的一种方法

该文提出健壮型句法剖析的一种新方法，这种方法把语言学意义上的词项描写跟统计技术结合起来。作者认为，只有词项对其施加限制的元素才能出现在特定的超级标注的框架内。每一词项能够在多少个不同的句法环境下出现，它就能跟多少个超级标注连接上。对句法剖析器来说，增加了局部歧义的现象。然而，这种局部歧义可以通过超级标注出现的概率分布来加以消除。作者曾在词汇化的树连接语法（Lexicalized Tree-Adjoining Grammar：LTAG）的框架里探讨过这些想法。超级标注消歧事实上成为一种分析形式的表述方式，而分析器仅仅需要把个别的超级标注结合起来。

2．基于树库语法的句法剖析研究

该文通过使用宾大树库（context-free grammar：CFG）上下文无关语法对树库进行句法剖析，得出了剖析程序性能的实验研究，并提出紧密相关的理论模式。该项性能明显地受到规则表述和树形变换的影响，而很少受到自上而下或自下而上策略的影响。文章探讨了语法饱和，包括对树库中短语性非终端符号的、彼此紧密相联的成分的分析，句子长度的增加使实际的语法规则

范围有所扩展，并在某些配置中得到超立方的运算复杂度等等情况。

3. 构建现代希伯莱语文本树库

该文描写构建现代希伯莱语文本的第一个树库的过程。基本点是迎合运用自动化的手段降低人工标注成本的需要。文章对于概率的句法剖析器和一个人工标注的、小规模的树库联合使用的有效性进行了探讨。描写了包含 500 句标注了的句子的初始树库，树库剖析的标注方案把形态与句法合二为一。基于应用这些工具进行试验而获得的结果，建立了一套半自动化的程序以扩充树库。

4. 句法、语义论元与基于依存关系的形式语法

该文通过句法关系和主题角色的层级概念构建有关语义论元的句法实现中出现的变异的一种形式化手段，同时构建词汇继承的机制以从个别的“联接类别”获取“配价格式”。把这种形式化构建融入拓扑依存语法（Topological Dependency Grammar: TDG）这一新出现的、词汇化的、基于依存关系的形式语法之中。文章对可以交替地实现为名词性短语或介词短语的论元进行讨论，并模拟主题角色的交替形式。

5. 动词的词汇表述

作者构建了一个在计算语言学意义上对自然语言处理系统来说用得上的大规模词库。对数量可观的英语动词所显示的句法现象作出解释。从区分三种结果补语入手，每一种以事件的模板出现，有着自己的词汇表述。基于结果补语的数据，开发了一个主要和次要事件模板的框架以及一个管住它们出现的核对机制。该文的理论可以界定两种非次范畴化论元的实现并加以解释。模板理论和核对机制直接对动词的论元交替模式提出解释。

6. 基于论元结构统计分布的动词自动分类

作者认为词汇知识的自动习得对于形形色色的自然语言处理

工作至关重要。动词知识尤其重要，这是一个句子中信息的主要来源，一个行动或状态与其参与者相关联的述语谓项结构（即谁对谁做了什么）。该文中，作者提出观测的学习实验，并使用从大型标注语料库中提取的语言学统计指标来建立分类系统。对有关规则系统的语言运用及其错误进行详细分析，证实所提出的特征得到了与动词的谓项结构相关的属性。计算结果证实了有关词干关系的知识对于动词分类极其重要，且可以通过自动手段从语料库中对其进行收集。接下来论证了一个在更深层次语言学知识的统计技巧的稳定性和可量测性的有效结合。

7．词义辨析的一致标准

该文探讨动词的多义性问题，通过运用相关的句法框架和动词类别来探讨如何简化对不同词义给出定义的工作手段，并强调设立具体标准的重要性，诸如不同的述语论元结构、语义类别限制以及词汇同现现象。

8．从词的关系中识别词汇标记

该文研究一种表层的、可从篇章中自动提取意义相关的、成对的词的组合。自动提取语义上成对的组合，可以直接应用在两个方面：一是在篇章检索中，意义相关的词对可以成为互相替代的查询词；二可应用于语言学的研究，词对可提供对篇章中类属词群的初步描写。

作者试图回答两个问题：第一个问题是关于在一组辨识模式的样本和它们所连接的词对间匹配的紧密程度如何。第二个问题是关于辨识出来的意义关系（如果有的话）的性质。最后总结关于在篇章中词汇模式以及词语关系的作用，并探讨它们在信息技术和语言描写中作为辅助手段的价值。

9．基于语料库的英语复合词研究

语料库的应用迫使语言学家面对一些在案头研究中容易忽略的语言现象，作者在这里汇报一项基于英国报纸的大规模英语语

料库的构词模式研究。通过对新的复合方式的观察指出应用真实语料可能会给研究构词法的理论家和描写语言学家带来问题。文章不仅指出一些在大部头的语法书中没有描写过的模式而且还指出在90年代早期英语里已经常见的某些模式如何突破某些理论著作已成定论的原则。

10. 语料库语言学与词典学

该文作者指出，语料库语言学要成为语言学的一个独立分支，必须具有以下特性：(1) 依附于 de Saussure 和 Firth 所定义的结构主义；(2) 不同于认知语言学，以搭配与意义、话语与意义为例阐述自己的观点。就语料库语言学与词典学的关系而言，作者认为“选择有用的引例似乎比传统的定义对使用者更有帮助”，从而把“引例”和篇章元素联系起来，并且说明，这就是语料库语言学（包括多语语料库语言学）所面临的挑战和是否能取得成就的关键。

译者希望今后能不断推出《计算语言学视窗》续集，争取一两年出一集，给同行们搭个桥，互相切磋，以求计算语言学有更大的发展。

译者

2003年5月于北京

目　录

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—独立”和“领域—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—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—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—重写形式语法，跟属于串列—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。

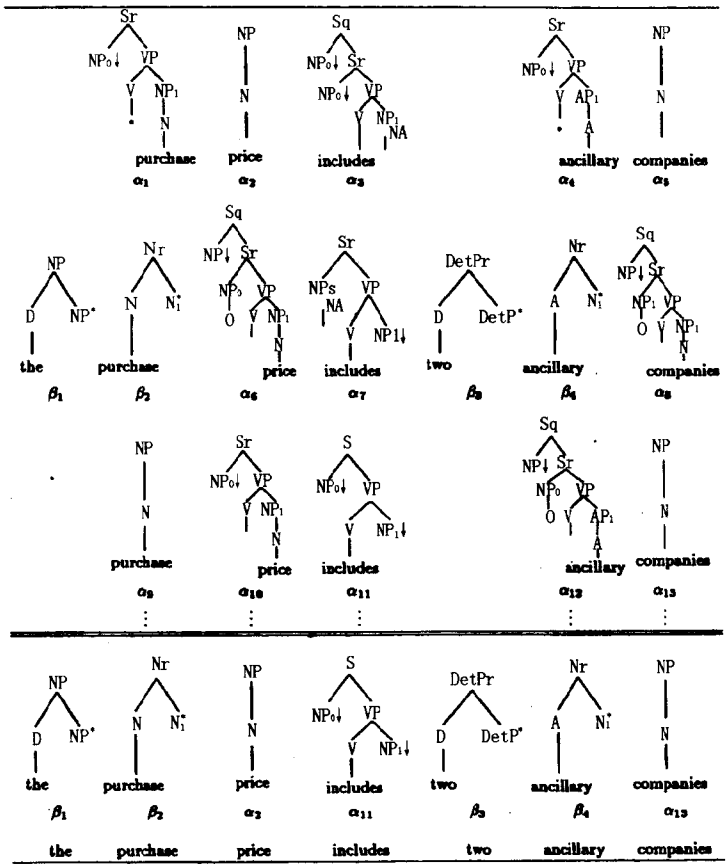

图1　跟句子“the purchase price includes two ancillary companies”词项连接的超级标记选例