树库的标注及应用

Anne Abeillé, ed., 2003, Treebanks: Building and Using Parsed Corpora, Kluwer Academic Publishers. (Text, Speech and Language Technology Volume 20)

詹卫东 导读

1 学科背景及本书的定位

树库（Treebank）属于深加工语料库，是语料库语言学和自然语言处理（NLP）技术发展到相对成熟阶段的产物。宽泛而言，语言研究一直以来都离不开“语料”。但从“语料”到现代意义的“语料库”，是从二十世纪五六十年代伴随着电子计算机的应用才开始的，其发展轨迹及趋势有几个明显特点：（1）语料库规模不断扩大，类型不断多样化。（2）标注信息不断丰富。（3）应用范围不断拓宽。这些特点是跟过去半个世纪整个信息社会大环境的飞速变化和 NLP 技术的进步分不开的。计算机存储能力和互联网的加速发展，使得电子化的大规模的自然语言资源越来越容易获得。从上世纪六十年代起步时的百万词级规模到八九十年代的上亿词级规模，再到今天语料库的规模已不再成为人们关心语料库的重点，不难感受到这种惊人的扩容速度。与此同时，语料也从原始形态的生语料库发展到经过多级标注（annotation）的所谓熟语料库。标注的信息从一般的词语形态信息，词类信息等很快发展到了标注句法结构、句法功能、语义角色信息等等。标注词类信息的语料库跟原始语料一样仍然保持着一维串性结构，而标注了句法结构、句法功能信息的语料库则因描述了词语（以及词组）之间的层级组合关系，成为二维的树状结构（Tree Structure），因此这样的语料库就被称为树库。像树库这样的带标语料库的发展还明显得力于 NLP 技术本身发展的推动。这一方面是 NLP 技术的发展需要有树库这样的深加工语料库提供数据支持。另一方面则是由于 NLP 技术的进步反过来大大提高了树库加工的效率，减低了人工成本，使得树库加工成为切实可行的一项工作。从上世纪九十年代开始，NLP 的主流技术从基于规则的方法开始纷纷转向基于统计的方法，在这样的背景下，来自真实语料的语言统计数据逐渐取代以往由人工归纳的语言学专家知识，成为 NLP 应用系统所依赖的主要知识源。在词类标注、句法分析、机器翻译等许多 NLP 技术的相关评测中，基于统计方法的系统都取得了更胜一筹的成绩，从而吸引了更多的研究人员来推进这种数据驱动型 NLP 技术的研究。尽管构建树库是相对成本比较高的语言工程，但受到英语树库的成功鼓舞，从上世纪九十年代中后期开

始，其他语种也陆续启动了树库加工项目。随着机器学习技术在 NLP 领域应用热潮的不断升温，树库的研究和应用也受到越来越多的重视，不但涉及的语种已经扩展到几十个，而且句法标注所依据的理论体系也由生成语法的短语结构语法发展到中心语驱动短语结构语法（HPSG），依存语法（Dependency Grammar）、词汇功能语法（LFG）等等多种理论框架并存的局面（有的树库甚至是把短语结构跟依存关系的标注融合到一块进行标注）。本书出版于 2003 年，距离上世纪 90 年代初英语树库问世已有 10 年。尽管如编者在导言中所说的，树库作为语言资源的一种新形式，本书的多数篇幅是在讨论如何加工树库，有关如何使用树库的篇幅相对较少，但仍然可以说全书内容基本反映了这 10 年间树库研究的整体面貌，是树库研究发展到一定阶段的一个比较全面的总结，起到了承前启后的作用。

2　内容提要

本书正文共 21 章，正文之前有一篇导言（Introduction）。导言是本书编者对全书内容的概括介绍。21 章中有的是专为本书撰写的，有的则是由发表在一些相关会议上的论文改写的。21 章内容分为两大部分：第一部分从第 1 章到第 15 章，讲如何构建树库；第二部分从 16 章到第 21 章，讲如何使用树库。

第 1 章到第 4 章介绍了英语树库的构建。内容分别是美国宾州树库的整体情况介绍，对近 20 年英语树库构建工作的思考，英语语料库（Bank of English）的词汇形态标注、句法标注以及后续的句法功能标注，ICE-GB（国际英语语料库-英国部分）树库的句法结构校对方法。第 5 章和第 6 章介绍了德语树库的构建，分别是德语新闻语料库的句法标注，德语新闻组语料库（USENET）的错误类型标注。第 7 章和第 8 章是两种斯拉夫语族语言树库的构建，第 7 章介绍捷克语树库的构建；第 8 章介绍基于 HPSG 的波兰语句法测试语料库。第 9 章到第 12 是四种罗曼语族语言树库的构建，第 9 章介绍西班牙语树库的开发；第 10 章介绍法语树库的构建；第 11 章介绍意大利语句法-语义树库的构建；第 12 章介绍了一个中世纪葡萄牙语树库的构建。第 13 章到第 15 章是其他语种树库的构建情况介绍，第 13 章介绍了台湾中研院 Sinica 中文树库；第 14 章介绍了日语树库；第 15 章介绍了土耳其语树库。

第 16 章介绍了树库标注的编码形式。第 17 章和第 18 章讨论如何利用树库进行句法分析评测。第 17 章介绍了构建专门的句法分析评测用树库的方法。第 18 章介绍了用标注依存关系的树库来评价 MINIPAR 句法分析器性能的实验。第 19 章到第 21 章讨论的是从树库中抽取语法知识的问题。第 19 章介绍了直接将树库看作是一部概率语法来进行句法分析的实

验。第 20 章介绍了从树库和 HPSG 规则中抽取词汇化概率树语法的方法。第 21 章介绍了从树库资源生成词汇功能语法（LFG）功能结构（F-Structure）标注语料库的方法。

3 章节内容详细介绍

导言

标注语料库相对于普通的未标注语料库，对自然语言处理和语言学研究的价值更大。本书的 21 篇文章都是讨论标注语料库特别是句法结构标注语料库的，涉及的问题包括：（1）如何选择语料库进行标注？（2）选择标注什么内容？（3）手工标注还是自动标注？用什么辅助工具？采用什么标注格式？（4）如何检索标注语料库？（5）从标注语料库中可以抽取什么语法知识，跟从未标注语料库中抽取知识相比有何优越性？（6）如何利用标注语料库对 NLP 工具比如句法分析器（Parser）或语法检查软件（grammar checker）进行评估？

第 1 章 宾州树库概述

美国宾州大学树库（The Penn Treebank）从 1989 年到 1996 年，历时 8 年，建成约 700 万词的带词性标记语料库和 300 万词的句法结构标注语料库（树库），200 万词的谓词-论元结构标注语料库，160 万词的非流利口语转写带语音标记语料库。本章介绍了宾州树库的三个标注规范：词性标记规范，句法结构标记规范，和非流利口语语音标记规范。此外对语料库加工的方法也做了介绍，词性标注，句法结构标注，非流利口语语音标注都采用的是自动标注和人工校对相结合的方式。自动词性标注先后采用过 PARTS 词性标注程序（Church 1988）[1]和基于转换的错误驱动的词性标注程序（Brill 1993）[2]。句法结构标注采用的是 Fidditch 句法分析器（Hindle 1989）[3]。非流利口语语音标记采用的工具是简单的 Perl 脚本程序，先利用程序将非句子成分自动加上标记，之后再利用一个跟句法结构标注类似的图形界面程序，可以让标注人员方便地手工加注语音标记。作为全世界最早开始的树库加工计划，宾州树库的加工规范，加工方法都带有示范的意义。

---

[1] Kenneth W. Church, 1988, A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, In Proceedings of the Second Conference on Applied Natural Language Processing, 26th Annual Meeting of the Association for Computational Linguistics.

[2] Eric Brill, 1993, A Corpu-based Approach to Language Learning, Ph.D Dissertation, University of Pennsylvania.

[3] Donald Hindle, 1989, Acquiring Disambiguation Rules from Text, In Proceedings of the 27th Annual Meeting of the Association for Computational Linguisitics.

第 2 章 树库构建二十年

通过对过去二十年参与树库加工工作的反思，本章作者提出，真实语料所反映的结构事实跟当代理论语言学秉持的一些假设是相左的。自然语言很难说是一个定义得很完美的系统（well-defined system）。短语结构的分析也并不总是有“正确答案”。从树库中揭示的语言事实是有树库之前所没有认识到的。作者批评了以乔姆斯基为代表的理论语言学缺乏对短语结构进行系统分类的兴趣。一些理论语言学者认为语言远远没有表面上的那么复杂，语言的底层结构可以用有限的规则去定义。这种具有遗传特性的“心理语言机制”是适用于所有自然语言的共性。从语言的“深层结构”这方面来说，英语和捷克语的复杂程度几乎不会超过像 Pascal 或 Java 这样的程序设计语言。但是树库加工的实践经验表明这种理论假设与事实不符。语言中的规则不是法律或化学意义上的规则，它是灵活可变的。语言的使用者可以改变甚至不遵守规则。即使是小规模的树库语料，都可以通过结构的频次统计，得出与以往的理论语言学研究看法不同的结论。比如通过统计 SUSANNE[4]树库中“主语-及物动词-宾语”和“主语-不及物动词”结构的频次发现，后者并不是跟前者地位同等的常用结构（而这却是一般语法书上通行的看法）。在真实语料中，如果谓语动词后面没有带宾语的话，也总是带有其他句法成分。再比如通过统计 CHRISTINE 口语树库可以发现语言结构的复杂性跟年龄的关系更密切，即随着人的年龄增长，所使用语言结构的复杂程度逐渐提高。而不是像之前的研究所声称的那样：结构复杂性跟人所处的社会阶层密切相关。

第 3 章 BOE 英语语料库及后续加工——基于规则句法分析器的功能标注

BOE（Bank of English）英语语料库是由 Collins 出版社资助的国际性英语语料库加工项目。从 1993 年到 1995 年，BOE 语料库中有 2 亿词语料做了词形分析和句法标注。词形分析采用的是一个双层词汇形态分析器，句法标注采用的一个基于英语约束语法（ENGCG）的浅层句法分析器。整个项目的应用目标是将 BOE 标注语料库用于 COBUILD 英语词典第二版的编纂。1995 年 BOE 句法标注的任务结束之后，研究人员对 ENGCG 进行了改进，设计了新的基于规则的功能依存语法分析器（FDG），FDG 考虑了更多的复杂语法现象（如非投射结构、长距离依赖、省略结构和同形成分省略等），比 ENGCG 的语法规则覆盖率更广，

[4] SUSANNE 是“Surface and underlying structural analysis of natural English”（自然英语的表层和基础结构分析）的首字母缩写（参考 http://www.grsampson.net/RSue.html）。有关 CHRISTINE 口语树库名称的来历请参考 http://www.grsampson.net/RChristine.html。

在错误率没有增加的情况下，FDG 产生的分析结果中歧义数量仅是 ENGCG 的四分之一。小规模的测试显示，FDG 比当时最好的基于统计的句法分析器的效果更好。

第 4 章 完成句法标注语料库：从单文本遍历式校对到跨文本定向式校对

本章介绍了英语国际语料库（ICE）中英国英语部分（GB）的 100 万词语料库加工成树库的过程。特别是对人工校对的方法进行了深入讨论。传统的人工校对是对每个文件中的每个句子逐句进行校对（我们称之为单文本遍历式校对），这种方式的缺点是工作量非常大，同时容易造成前后不一致的问题。本章提出了一种基于结构的校对方法，即对跨文本的树库语料进行检索得到同类标注问题的实例，然后由校对人员来做统一的修改（我们称之为跨文本定向式校对），这样同类问题的处理方式就是前后一致的。不过，虽然可以带来更好的一致性，但相应的辅助工具实现难度更大，多人同时校对时还需要解决管理问题，此外程序还会出现误报的假性错误。传统的单文本遍历式校对方式则没有这些问题。

第 5 章 德语新闻语料库的句法标注

本章介绍了德语新闻语料库（NEGRA）的句法标注。NEGRA 包含超过 2 万个句子（35 万词）。句法标注框架采用的是上下文无关的短语结构和依存语法结构的混合体，在标注句中短语结构类型（如 NP，VP 等）的同时，还标注了短语之间的依存关系（如主语、宾语等）。跟一般上下文无关文法不同的是，NEGRA 的句法树标注允许树分支交叉，这是跟德语词序灵活的特点相适应的，在德语实际语料中存在着大量的“未完成”结构（incomplete structure），允许分支交叉，以及倾向采用更为扁平的多分支树结构（相对于二分支树结构）来进行标注，是 NEGRA 的特点。NEGRA 的标注过程大致分为 3 个阶段，分别是（1）预处理；（2）基于图形界面的人机交互式标注；（3）通过比较确定最终标注结果，即每个句子都有两个人分别独立进行标注，然后进行比较，再决定最终标注结果。在其中第 2 个环节的交互式标注中，NEGRA 利用基于统计的词性标注软件和句法分析软件，对每个句子的多种分析结果给出概率评分，并高亮显示程序认为不可靠的分析结果。在计算机界面上呈现分析结果时采用分段逐步推进的方式。这些手段都显著地提高了标注效率。

第 6 章 德语新闻组语料库的错误类型标注

德国政府资助的 FLAG 项目旨在开发德语受限语言（Controlled Language）和语法检查技术。为此，项目组收集了互联网新闻组（USENET）电子邮件语料约 12 万句。在确定了

一个含 16 种错误类型的标记集后，先由人工在纸上对这 12 万句进行了标注，即把句中的各类错误（如形态错误、句法结构错误、句法语义选择错误、拼写错误等等）加上标记，共标了约 6 万句。然后在计算机标注工具[5]的辅助下，将纸上标注的 6 万句中的 14,492 句录入计算机并核验是否符合标注要求，形成了一个带错误类型标记的新闻组语料库。具体标注信息包括每句的错误数，错误位置，错误范围等等。对错误类型的初步统计显示，真实电子邮件语料中的错误 83%是纯粹的拼写错误，语法错误的比重不高，仅为 16%。FLAG 的工程实践表明，先在纸上快速标注真实语料中的错误，再利用计算机标注工具制成最终的带错误类型标记的语料库，是行之有效的语料库标注方法。

第 7 章 布拉格依存树库的构建：一个三层标注方案

布拉格依存树库（PDT）分三个层级对捷克语语料进行标注。第一层是形态标注；第二层是表层句法标注，采用依存语法架构。第三层是语义标注，仍然采用依存树的表达形式，理论框架则是所谓的功能生成描述（Functional Generative Description）。PDT 的语料来源是捷克语国家语料库（CNC）。语料分布为：普通报纸语料占 60%，经济新闻和分析占 20%，科学杂志语料占 20%。项目计划从 1996 到 2004 年，标注第一层的语料规模要达到 180 万词，标注语义信息的语料规模要达到 100 万词。本章成稿时，项目完成了约三分之二的标注任务。所有三层标注均采用 SGML 语言作为格式规范。第一层形态标注对句子中的每个实例都标注两个信息：词形（lemma）和一组形态范畴特征值（MTag）。第二层表层句法标注要给出句子的依存关系树。在列举了依存语法树需要遵循的基本原则后，本章还介绍了 PDT 对依存语法理论不易处理的一些特殊语法现象（如“并列结构”）的标注策略。在积累了一定量的表层句法标注语料后，就可以借助 Collins 的词汇化概率语法分析器进行训练然后对未标注语料进行依存关系标注，正确率可以达到 80%。第三层语义标注要在句子的依存语法树基础上继续在树节点标签上标注更多的语义和情态范畴属性。PDT 设定了约 40 个语义功能项（比如行动者/承受者，效果，使因，等等），此外还有时态、数、性、比较级等等语法情态属性项的标注。

第 8 章 一个标注 HPSG 特征结构的波兰语句法分析测试语料库

本章介绍了一个波兰语书面语句子测试集：BRG（波兰语直译即语法分析数据库）。作为句法分析测试句集，BRG 不仅包含合语法的句子（193 句），还包含不合语法的句子（147

[5] 在评估了 DiET 和 Annotate 两个均支持图形界面的标注工具的特性后，项目组选择了前者作为辅助工具。

句)。句子并不是来自真实语料，而是根据语言学评测分析的需要人为设计的。句子按照复杂程度分为基本句型、复杂句型、特别复杂句型三个类别。每个句子都以人工方式标注了是否正确（其中不合语法的句子都有对应的合语法的句子），包含哪些语法现象（所有语法现象可以在一个索引表中查到，共计 264 种），以及采用 HPSG 的特征结构框图（AVM）标记句中的语法成分及语法属性。在 BRG 的图形界面上，句子可以呈现其短语结构树形式以及 AVM 特征结构。BRG 可以评测波兰语形式语法的覆盖率，其设计出发点是针对基于 HPSG 的语法框架，不过已经实现的 BRG 也可用于评估基于其它形式化语法框架（如定子句语法，依存语法）的波兰语语法系统。

第 9 章 西班牙语树库句法标注方案及开发工具

本章介绍为构建西班牙语新闻语料树库所开发的规范和工具。该树库中包含来自新闻报纸和杂志的 1500 个句子，共计 22,695 个词。这 1500 句中一部分是孤立的单个句子，没有上下文。另一部分是在自然段落中的，有上下文的句子。树库标注规范涉及的问题包括语言单位的识别以及信息标记方式（西班牙语树库参照了宾州树库的标记模式）。每个树节点标签由“语类范畴”（CAT），“语言成分字符串”（String）以及后面带有的一组特征描述（Feature）组成。除一般性的通用标记框架外，本章还介绍了西班牙语树库中一些特殊结构（比如“Se”字结构）的标记方式。西班牙语树库的开发借助了项目组自己开发的工具和一些公开资源，分为两类：（1）标注工具，包括形态句法标注系统和语块识别软件（chunker）。（2）查错工具，包括图形化的树结构编辑工具，特征核查工具，短语结构规则生成工具等。

第 10 章 构建法语树库

本章介绍了一个法语树库加工的情况。该树库来自法语新闻语料，100 万词规模。加工过程分为词汇层标注和句法结构层标注两个阶段。词汇层加工流程为：首先对原始文本进行断句处理，然后进行词性标注、形态标注、未登录词标注、合成词标注，等等，词汇信息标注的多个环节都在自动处理的基础上进行了人工干预，并将得到的正确标注结果记录到词库中，这样可提升词库对后续语料的处理能力。句法结构层加工流程为：对经过词汇信息标注处理的文本，首先基于手工规则进行浅层分析，然后对得到的初步结果进行人工校对，再基于手工规则和法语配价词典进行功能标注，得到树库文本，经过人工校验后得到最终版本。本章还较为详细地介绍了法语树库的语法功能标记集，并讨论了一些具体结构类型的标注方

式（比如非连续成分结构、并列结构等等）。

第 11 章 构建意大利语句法语义树库

本章介绍了一个意大利语句法语义树库（Italian Syntactic-Semantic Treebank，ISST）。ISST 包含四个层级的标注信息：形态句法层、短语结构层、语法功能关系层、词汇语义层。语料规模达 30 万词，其中 21 万词为平衡语料，9 万词为金融领域语料。形态句法层的标记集从 16 个基本词类标记扩展到 31 个子类标记，再加上形态句法特征标记，共 236 个标记。短语结构成分（共 22 类）的标注跟语法功能关系（共 9 种关系）的标注是分别独立进行的。短语结构成分先由浅层句法分析器自动识别，然后由人工修改。语法功能标注则采用依存关系模型，描写词语之间的依存关系。词汇语义层标注主要针对名、动、形以及多词复合表达式等实义性语言单位。具体标记包括三种信息：（1）从意大利语词网（ItalWordNet）中获取的词的义项；（2）词的比喻义、惯用法、旧词新意、专名等；（3）标注操作相关信息（如记录标注者信息）。除标注内容外，本章还介绍了用于标注的软件工具，以及将 ISST 用于意-英机器翻译系统测试其效能的情况。

第 12 章 构建中世纪葡萄牙语树库

本章是对古代语言进行词法形态分析、词性标注和句法结构标注的一次尝试。通过使用针对当代葡萄牙语开发的词法分析工具、词性标注工具和层叠式浅层句法分析工具，以及相关的语言学资源（主要是词法规则知识库 LKB），作者对中世纪葡萄牙语（主要是应用文文本，如遗嘱，赠品清单，法律文本等）进行了词法分析、词性标注和浅层句法结构的自动标注。本章的探索表明，在同一个语言不同时期的变体之间，或者不同语言之间，如果其相似度足够高，就可以利用已经标注好的树库资源，从中获取相关语言知识，来帮助加工新的树库语料。

第 13 章 SINICA 中文树库

Sinica 中文树库是跟宾州中文树库差不多同时开始构建的最早的中文树库之一。Sinica 中文树库在设计时主要考虑了三个问题：最大资源共享、最小结构复杂度和最恰当的语义信息标注。为了实现最大资源共享，Sinica 树库从已经有一定影响的经过分词和词性标注的达 500 万词规模的 Sinica 平衡语料库中选材。提出最小结构复杂度标准，则是为了保证树库标注的结构信息可以在不同的语言学理论背景下仍然能够共享。在以标注句法信息为主的树库

中标注多少语义信息，是 Sinica 树库设计者考虑的第三个主要问题。Sinica 的方案是标注跟谓词论元关系密切相关的那部分语义信息。这是通常认为与结构的句法表现联系最紧密的语义关系知识。Sinica 中文树库已经标注完成 38,725 棵汉语结构树，包含了 239,532 个单词，题材涉及到政治、旅游、体育、财经和社会等多个领域。

第 14 章 构建日语句法结构标注语料库

本章介绍了从 1996 年到 2000 年历时 4 年构建一个日语树库的情况。语料来源是《每日新闻》的新闻文本。建成的树库规模达到 4 万句。树库标注利用了形态分析器 JUMAN（其词典规模为 20 万词）和依存关系分析器 KNP（包含 600 条语法规则），并在标注过程中对工具本身也进行了改进。KNP 标注的是基于短语（而不是词）的依存语法关系，包括谓词与附加语之间的关系，并列关系和同位关系三种类型。对于如何采用依存语法体系描述日语中的一些特殊结构、并列结构、从句结构等，本章也做了说明。

第 15 章 构建土耳其语树库

本章介绍了构建一个土耳其语树库所采取的语言知识表示框架，并结合分析土耳其语的特点说明了为什么要选择这样的知识表示框架。土耳其语属阿尔泰语系，大量使用复合黏着词，不适合采用跟宾州树库类似的从封闭标记集中选择标记来描述词汇信息的方式。土耳其语树库将词的内部结构表示为一串屈折语素组（像一串珍珠一样），组与组之间由派生成分隔开。对 85 万词的土耳其语新闻语料的统计表明，平均每个复合词包含的屈折语素组不超过两个，并不像一般人想象的那么复杂。句法结构关系的标注采用依存关系模型，即标记由派生边界隔开的屈折语素组之间的依存关系（目前确定的关系集合包含 10 种依存关系）。按照上述表示框架加工完成的土耳其语树库已经达到了 1500 句的规模，预计在辅助标注工具的帮助下，树库规模可以扩充到 2 万句。

第 16 章 句法标注的编码形式

标注语料库因语言学理论背景的差异、标注形式的不同，使得不同语料库之间很难进行比较、融合、用于支持开发可重用的编辑和处理工具，为此，设计一个通用的语料库标注形式体系，就成了非常重要和非常迫切的一项任务。本章提出了一个抽象的标注模型，可以用于不同的标注层次（包括形态-句法标注、结构标注、指代标注等）。这个模型基于语料库编码标准（CES），并采用 XML（可扩展标记语言）规范，可称之为 XCES。XCES 很好地实

现了标注内容和标注形式的分离，这样，不同的语料库在设计时，可以专注于自己的标注内容的设计，而在实现时则可以采用的统一的形式表达框架。从而极大地方便了不同语料库之间的转换、共享、以及工具的重用。

第 17 章 句法分析器评测

树库可以用于评价句法分析器（parser）的性能。传统句法分析评测的代表软件 PARSEVAL 将自动分析结果中的短语结构块跟标准答案（树库）中的短语结构块进行比对，根据括号交叉程度来评估自动句法分析的性能好坏。这种评价方法对树库的要求比较低，只需要做了结构标注的树库就可以用来进行句法分析评测。但缺点也很多，包括树库中标注的结构过于扁平（比如宾州树库就倾向于扁平的树结构）使得对错区分度不高，以及无法对基于依存关系语法模型的句法分析器进行评价等等问题。本章提出了一种新的树库标注方案，即设计了一个层次化的语法关系（GR）标注集，对来自 SUSANNE 语料库的 500 个句子（1 万词）进行了语法关系标注（平均每句标注了 9.72 个语法关系），建成了一个句法分析评测树库，并用该树库做了句法分析性能评测实验。这种评测方法的优点是独立于不同的语法理论框架，既可以评价基于短语结构的句法分析器，也可以评价基于依存关系的句法分析器，此外，可以区分一个分析器对不同的语法知识的处理能力，评价的颗粒度更细。

第 18 章 基于依存语法的句法分析器（MINIPAR）评测

本章提出了一个基于依存关系的句法分析器评测方法。这种方法的要点是：句子的分析结果表示为一个依存关系列表。表中每个依存关系由<词，范畴，中心成分，依存关系>四元组表示。用于评测的树库（标准答案）中的每个句子都表示为若干个这种四元组。句法分析器自动分析产生的结果也同样是这样的四元组。评测时只要依次对比每个四元组，标记正确或错误即可，最后的评测结果由基于正确数和错误数计算得到的准确率、召回率以及 F-分值来表示。本章还介绍了用这种方法对 MINIPAR 分析器[6]在 SUSANNE 树库（500 篇文章）上的分析性能进行评测实验的情况。结果显示 MINIPAR 可以覆盖 SUSANNE 树库中的 79% 的依存关系，准确率可以达到 89%。

[6] MINIPAR 系统的词库来自 WordNet，有 13 万词条。语法规则是手工编制的一个有 35 个节点和 59 个连接组成的依存关系网。有关 MINIPAR 系统的细节介绍请参阅 Berwick et al. 1991, Principle-based Parsing: Computation and Psycholinguistics. Kluwer Academic Publishers.

第 19 章 从树库中抽取概率型语法

“数据驱动的句法分析”（Data-Oriented Parsing，DOP）将一个已经标注好的树库直接作为一部概率语法来使用。对一个待分析的句子，通过搜索树库中存在的子树进行合并，从而得到该句子的分析树结果。分析结果的可能性大小依赖结果中的子树在树库中的共现频次。通常情况下 DOP 模型对分析时所选用的子树没有大小以及复杂度的限制，这不可避免地使得每次进行句法分析时所查找的候选子树集合过于庞大而造成冗余。但是如果限制了候选子树的大小及复杂度，是会降低分析性能，还是有可能提高分析性能？这是理论上和计算上都很值得探讨的问题。通过对子树集合施加不同的限制（包括是否考虑重叠子树，考虑子树的大小，考虑子树的叶子节点是否含具体词语，考虑子树频次，考虑子树中的非中心词节点，等等），可以模拟实现包括概率上下文无关文法（stochastic context-free grammars）以及概率词汇化语法（stochastic lexicalized grammars）在内的不同的概率语法模型，从而也可以对这些概率语法模型进行比较。本章作者使用了两个树库语料（一个是包含 750 句的小规模的航空旅游信息系统 ATIS 树库，一个是包含约 5 万句的华尔街日报 WSJ 树库）做了对子树施加各种限制进行句法分析的实验，结果显示几乎所有的对子树候选集合的限制，都会或显著或轻微地降低 DOP 句法分析的精确度。

第 20 章 从树库和 HPSG 语法中抽取词汇化概率树语法的方法

词汇化概率树语法（SLTG）比一般的概率上下文无关文法描述的分布和层次信息更丰富，而且允许对不同类型的树结构进行循环重组，使得基于 SLTG 的句法分析效果更好。从树库中抽取 SLTG 意味着语法树来自语言的真实使用，从 HPSG 语法中抽取 SLTG 则使得语法能处理的语言现象跟领域无关。把从这两个来源获取的 SLTG 融合到一起，可以提高语法的覆盖面，使 SLTG 分析器能更好地适应新领域的句法分析。从树库获取 SLTG 的主要操作是从根节点开始递归遍历整棵树，在语言学的树结构分解原则指导下，获取 SLTG 的基础树、左辅助树、右辅助树、中间树[7]等。本章所遵循的树分解原则是简单的中心词驱动的分解原则（Head-driven decomposition principle），即将树上的非中心节点子树剪枝去掉。所得子树自动带有词汇锚点（叶子节点是具体词语或词语的词性标记等）。本章用于获取 SLTG 的树库有两个，一是宾州树库，一是德语 NEGRA 树库，后者首先被转换为宾州树库格式后，再用来抽取 SLTG。从 HPSG 获取 SLTG 的方法也是类似的，首先用 LinGO 项目中的英语 HPSG

[7] 辅助树（auxiliary tree）的最左或最右叶子节点和根节点相同。中间树（wrapping tree）是被其他节点包围的子树。

语法（该语法包含 7000 个语法类型，覆盖面很广）对测试语料进行自动分析获得基于 HPSG 的标注树库，然后再从树库中获取 SLTG。

第 21 章 从树库资源到词汇功能语法的功能结构标注语料库

在语料上标注功能信息或基本的谓词-论元关系信息对 NLP 应用有重要意义，但直接标注的工作量很大，即便采用自动分析的方法产生初始结果再由人工校对，也将面临从大量歧义结果中选择正确结果的困难。如何从已经标注了基本句法结构信息的树库开始，自动添加功能结构（F-structure）信息，是本章的中心议题。本章介绍了两种方法来完成这个任务。一种方法是先从树库中抽取 CFG 规则，然后人工给出一组基于正则表达式的标注规则[8]，由程序根据标注规则，自动将 CFG 规则附加上功能结构描述。第二种方法是借助一个为机器翻译系统设计的表达式改写系统（term rewriting system），根据系统给定的形式语言（term representation language），由人工定义从短语结构树到功能描述的映射规则，然后自动对树库中的树结构进行节点（及其父子关系）匹配，对匹配成功的树结构（即 c-成分结构），附加上相应的功能结构（f-功能结构）描述。本章作者选取了美联社新闻语料（AP）树库中的 100 句对第一种方法进行了实验，选取了 SUSANNE 树库中的 166 句对第二种方法进行了实验。结果显示，用这两种方法对经过短语结构标注的树库进行扩展，标注更加丰富的功能结构信息，是可行的。第二种方法如果限定树结构深度为 1，则可以获得跟第一种方法相同的效果。第一种方法需要借助人工编写的标注规则，这样的规则除可以用于本章所描述的任务外，还可以直接作为词汇功能语法（LFG）的语言知识库使用。

4 评论及对读者的建议

二十世纪五六十年代乔姆斯基创立转换生成语法，为语言学吹响的理性主义号角响彻至今，让扛着经验主义大旗的语料库语言学跟语言学的主流一直以来被动地保持着距离。即便是在语料库语言学伴随着技术进步得到长足发展的今天，语料库在理论语言学研究中的地位也并没有实质性的改变。正如本书第二章的作者 Geoffrey Sampson 在反思过去二十年从事语料加工研究工作时所指出的那样，理论语言学认为在丰富多样的“语言表现”（performance）之下，才是语言学真正的研究对象——即跟人的心智密切关联的“语言能力”（competence）。

[8] 标注规则形如 L>R@A。其中 L 是 CFG 规则的左部，R 是 CFG 规则的右部。如果一条 CFG 规则的左部和右部分别跟 L、R 匹配，则该规则就可以添加 A 代表的功能结构。

这个语言能力可以被定义为有限的若干条“规则”，而这样的规则在很大程度上是生物遗传机制的一部分，对于所有的自然语言都是普遍存在的。语料库语言学的支持者，特别是多年面对真实语料进行树库构建的研究者则很难认同这样的观点。Sampson 甚至叫板说，任何一条被认为可靠的语言学规则，拿到真实语料库中去检验，都可以找到违反该规则的实例[9]，而且这样的实例并不是“语言表现”中偶发性的错误。

语料库，特别是像树库这样的深加工语料库，对于计算语言学和研究价值和对于自然语言处理技术的实用价值，在今天已经是没有争议的了。从事计算语言学、自然语言处理、语言工程研究的读者可以从本书中了解上世纪九十年代到本世纪初这十年间有关树库加工的方方面面，包括一般可以选择什么样的处理流程、使用哪些辅助工具、对语言结构进行标记时需要注意哪些特殊语法现象、什么样的校对方法更有效、从加工好的树库中可以提取哪些语言知识、如何用树库来评价句法分析器的效果，等等。这些实践经验的总结对于推动今后的树库研究工作都有借鉴意义。但是，除了在这些工程层面的参考价值之外，本书是否还有更多的理论层面的价值呢？显然，如何认识语料库跟理论语言学研究之间的关系，对于理论语言学者和语料库语言学者都仍然是一个问题。Sampson 的反思更多地是强调了理论语言学界不愿面对真实语料的“固执”，但是，想象一下，一个从事语言学研究的读者可以从本书中看到什么呢？难道他会轻易地就被树库研究的魅力所征服，从对语言现象进行内省式的演绎思辨转而一头扎入语料库的汪洋大海去对语言知识做归纳式的探索吗？恐怕没有那么简单。尽管像 Sampson 这样的树库研究者喜欢诟病语言学理论研究的先验假设（aprioristic theorizing）无法解决真实语料的实际问题，但通观全书，不难看到，树库标注中从词汇层次的标注，到句法语义层次的标注，其基本知识架构无不来自理论语言学的研究成果。相反，语料库语言学却并没有提出自己独有的关于自然语言知识的理论体系。构建树库，只是在理论语言学的现有研究成果（比如短语结构文法、依存语法等）指导下，到真实语料中去做“重复性”的标记工作。至于说理论研究的成果无法覆盖全部的语言事实，并不能构成对理论语言学研究的全盘否定。在语言工程实践中，树库加工者也往往只是以工程师的方式（即打补丁的方法）在解决问题，还没有从理论高度提出全新的基于语料库（包括树库）的语法知识表示体系来。本书对语料库跟语言学理论研究之间的关系问题涉及甚少，但读者对此却不可不察。如果带着这个问题去读书中章节，有可能在获得工程上的实用价值外，还可以引起更

[9] Sampson 在这里举了一个英语中有关反身代词（reflexive pronoun）的规则的例子，即小句中跟它前面的成分共指（co-referential）的反身代词，不能做小句的主语。但是 LOB 语料库中有英国哲学家伯特兰·罗素（Bertrand Russell）的一句话违反了这个规则。

多思考。

5 延伸阅读指南

在经过了头十年的发展后，树库的研究继续在深度和广度两个方向拓展。标注的语言知识从句法结构拓展到语义层面的命题结构（Propbank），再到篇章层面的指代关系（co-reference resolution），近年来又出现多语对齐的树库（Parallel Aligned Treebank），以及将树库资源跟其他语言知识资源（如语义本体知识）融合到一起形成的多层标注的综合型语言知识资源库（如 OntoNotes）。这些发展都极大地丰富了当代语料库语言学的内容，对目前以数据驱动和机器学习理论模型为依托的 NLP 技术的进步产生着显著而深远的影响。与此同时，研究人员也比以往更加重视树库工程与语言学理论之间的内在联系。从 2002 年开始的“树库与语言学理论国际研讨会”（Workshop On Treebank and Linguistic Theories，简称 TLT）到 2012 年将召开第十届会议。会议的主题涉及到树库研究的各个方面，除传统的面向 NLP 的应用外，现在也有很多有关树库在语言研究和教学中的应用方面的讨论。此外，两年一届的“语言资源与评估国际研讨会”（Language Resource And Evaluation Conference，简称 LREC）从 1998 年开始，一直都是以语言资源建设和应用为主题，其中有很大一部分是有关树库的研究工作。除专题性会议外，计算语言学领域有影响的国际会议（如 COLING，ACL 年会等）都可以看到相当数量的跟树库有关的研究论文。通过阅读这些反映前沿研究进展情况的会议论文集，可以了解到树库研究的最新动态。

本书对树库加工工作的介绍，大多数是以宏观框架性内容的描述为主，如果要深入了解树库加工各个环节的细节情况，需要具备当代形式化语言学理论知识和 NLP 核心技术知识作为基础。如果此前对语言知识的形式化表达理论和相关 NLP 技术（包括词法层面的形态分析，中文的自动分词，词性标注，句法层面的结构分析等）了解不多，则需要阅读相关文献来充实这方面的知识。

下一节主要就是针对上面谈到的这两方面的延伸阅读需求给出的参考文献：一部分是有关 NLP（特别是其中的句法分析技术）、语料库语言学和形式化语言知识表达基础理论的；另一部分是反映树库研究的一些最新进展情况的。

6 延伸阅读文献

6.1 NLP 计算模型以及句法分析技术

Daniel Jurafsky & James H. Martin, 2000, 2008, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall Inc..

Steven Bird, Ewan Klein, and Edward Loper，2009，Natural Language Processing with Python, O'Reilly Media, Inc.. Chapter 3, 8-11.

刘挺、马金山，2009，汉语自动句法分析的理论与方法，《当代语言学》2009 年第 2 期。

6.2 语料库语言学

黄昌宁、李涓子，2002，《语料库语言学》商务印书馆。

俞士汶、段慧明、朱学锋、孙斌，2002，北京大学现代汉语语料库基本加工规范，《中文信息学报》2002 年第 5、6 期。

Graeme Kennedy, 1998, An Introduction to Corpus Linguistics, Addison Wesley Longman Limited.

Anke Ludeling & Merja Kyto, eds., 2008, Corpus Linguistics: An International Handbook, Mouton de Gruyter.

6.3 树库加工及应用

詹卫东，2000，《面向中文信息处理的现代汉语短语结构规则研究》，清华大学出版社。

周强，2004，汉语句法树库标注体系，《中文信息学报》2004 年第 4 期。

周强，2007，汉语基本语块描述体系，《中文信息学报》2007 年第 3 期。

王跃龙、姬东鸿，2009，汉语树库综述，《当代语言学》2009 年第 1 期。

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer, 2005, The Penn Chinese Treebank: Phrase structure annotation of a large corpus, In Natural Language Processing II(2): pp.207-238, Cambridge University Press.

Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao, and Kuang-Yu Chen., 2000, Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface, In *Proceedings of the Second Chinese Language Processing Workshop*, pp.29-37, HongKong.

Sameer Pradhan, Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel, 2007a, OntoNotes: A Unified Relational Semantic Representation, In *International Journal of Semantic Computin*g, Vol.1, No.4, pp405–419.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla, 2007b, Unrestricted Coreference: Indentifying Entities and Events in OntoNotes, In *Proceedings of the IEEE International Conference on Semantic Computing*(ICSC), September, 17-19, 2007.

Ralph Weischedel, Eduard Hovy, Martha Palmer, Mitch Marcus, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue, 2011, OntoNotes: A Large Training Corpus for Enhanced Processing. In Joseph Olive, Caitlin Christianson and John McCary, eds., *Handbook of Natural Language Processing and Machine Translation*, Springer.

Kiyotaka Uchimoto, Yujie Zhang, Kiyoshi Sudo, Masaki Murata, Satoshi Sekine, and Hitoshi Isahara, 2004, Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications, In *Proceeding of the Workshop on Multilingual Linguistic Ressources MLR2004*. August 28th, 2004, University of Geneva, Switzerland.

Xuansong Li, Stephanie Strassel, Stephen Grimes, Safa Ismael, Xiaoyi Ma, Niyu Ge, Ann Bies, Nianwen Xue, and Mohamed Maamouri, 2010, Parallel Aligned Treebank Corpora at LDC: Methodology, Annotation and Integration, In *Workshop on Annotation and Exploitation of Parallel Corpora* (AEPC), University of Tartu, Estonia.