第3章

句法分析

本章重点

- 句法树
- 句法分析方法

本章难点

- 概率分布上下文无关语法

3.1 句法分析概述

3.1.1 句法分析概要

句法分析(Syntactic Parsing 或者 Parsing)是识别句子包含的句法成分要素以及成分之间的内在关系,一般以句法树来表示句法分析的结果。实现该过程的应用称作句法分析器(Parser),根据侧重目标分为完全句法分析和局部句法分析。完全句法分析以获取整个句子的句法结构为最终目的,而局部句法分析仅关注局部成分,依存句法分析属于局部分析法。句法分析也可以分为基于规则的方法和基于统计的方法两类。基于规则的方法一般事先构建专家规则,但在大文本的场景下可能会因为语法规则覆盖度有限性问题而影响处理效果,另外一个缺点是可迁移性一般不高。随着大规模标注树库的出现,基于统计模型的句法分析方法逐渐得到广泛应用。统计句法分析模型基于候选句法树,从各候选句法树中找出最有可能的候选结果,通常选择概率最高的候选树作为最终结果。自然语言处理句法分析目前面临的关键技术问题如下。

(1) 语义消歧:语言中存在很多一词多义的用法,歧义与消歧是自然语言理解中最核心的问题之一,在词语、句子、段落篇章等各个层次都会出现因为语境不同而产生歧义

的现象，消歧是指根据上下文识别正确语义的过程。由于句子一般是由词语组成，词义消歧是句子消歧的基础。

（2）路径优化：句法分析的搜索空间和句子长度存在指数对应关系，因此，在句子长度超过特定阈值时，搜索空间会变得十分庞大，从而降低了处理效率。优化搜索路径，以确保能够在合理时间范围内查找到模型定义最优解，是句法分析的目标。

3.1.2 句法树

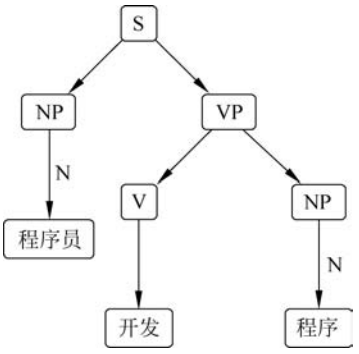

图 3-1 树状图示例

在计算机中，可以用树状结构图来表示文本结构，使用字符 S 代表句子；NP、VP、PP 分别代表名词短语、动词短语、介词短语；N，V，P，M，Q 则分别是名词、动词、介词、数量词和时量词。图 3-1 表示了常见的树状图示例。

表 3-1 列出了清华树库部分句法的功能标记和结构标记。功能标记集主要侧重于对汉语短语进行功能分类。结构标记集则侧重于对不同句法成分内部的结构语义关系进行更加深入的描述。比较常见的结构标记如主谓、述宾、述补、介宾等。

表 3-1 句法标记集示例

序号	功能标记	代表名称	结构标记	代表名称
1	np	名词短语	ZW	主谓结构
2	tp	时间短语	PO	述宾结构
3	vp	动词短语	SB	述补结构
4	ap	形容词短语	DZ	定中结构
5	sp	处所短语	JB	介宾结构
6	bp	区别词短语	AD	附加结构
7	dp	副词短语	SX	顺序结构
8	pp	介词短语	XX	缺省结构
9	mbar	数词准短语	LH	联合结构
10	mp	数量短语	ZZ	状中结构
11	dj	单句句型	CD	重叠结构

表 3-2 列出了根据国家标准得到的部分汉语词类标记集。

表 3-2 汉语词类标记集

序号	代码	代表名称	代码	代表名称
1	a	形容词	nP	指人专名
2	aD	副形词，形容词直接作状语	nS	地点专名
3	b	区别词	v	动词

续表

序号	代码	代表名称	代码	代表名称
4	c	连词	t	时间词
5	d	副词	x	任意字符串
6	dB	否定前副词	y	语气词
7	dD	程度副词	z	状态词
8	dN	否定副词	u	助词
9	p	介词	s	处所词
10	n	名词	i	成语
11	m	数词	e	叹词
12	l	连接语	h	前缀

按照上述列举的标记规范，现举例说明句子的句法分析，其基本格式为：{<句子序号>{<词语>＋<句法成分标注>}<回车符>}。其中的句法成分标注格式为：[<功能标记>＋<结构标记>…]。

【实例 3-1】 句法标注实例

(1).[dj－ZW[np－DZ 各国/n[np－DZ 友人/n 华侨/n]][vp－ZZ 积极/aD[vp－PO 拥护/v 团结/n]]]

(2).[dj－ZW[np－DZ[mp－DZ 一/N/m 批/qN]员工/n][vp－PO 参加/v 商务班/n]]

3.1.3　常用句法分析相关数据集简介

语料库的句法标注是自然语言处理研究的基础问题，处理目标是对语料文本进行语言句法分析和标注，形成可复用的树库(Tree Bank)语料。目前为止，国内外已经开发完成的大规模常见的树库和句法分析数据集包括英国的 Lancaster-Leeds 树库、美国的 Penn 树库(涵盖中英文)、清华大学句法树库为基础的系列句法分析数据集，以及中国台湾 Sinica 中文树库等。

下面简要分别介绍几种常用树库。

(1) 美国宾夕法尼亚大学英文树库(Penn TreeBank，PTB)。

PTB 从 Wall Street Journal (WSJ)的基准约十万个故事中选取了近 2.5%个用于句法注释。最初的 PTB 句法结构树比较简单，之后标记功能逐渐增加，体现句子中的句法成分，并以建立句法到语义之间的联系为目标。

(2) 美国宾夕法尼亚大学汉语树库(Chinese Treebank，CTB)。

汉语 CTB 与英语 PTB 的标注体系一脉相承，存在交集的部分。涵盖约五十万个单词，因为共享共同的标注框架，在实现英语和汉语的双语信息标注方面具有一定优势。

(3) 中国台湾 Sinica 汉语树库。

Sinica 树库主要处理特点是根据标点对汉语进行处理，对每个处理后的子句再进行句法分析和标注，化整为零从而实现句法树体系。目前共容纳约二十五万个词汇。优点

是因为进行了文本切片处理，最终标注难度得到一定程度下降，缺点是可能存在一定信息丢失。

（4）清华大学句法标注库。

清华大学句法信息标注语料库由一系列子库组成，各子库体现不同的主要功能，包括句法树库（Tsinghua Chinese Treebank，TCT）、功能语块标注库（Functional Chunk Bank，FCB）、基本块标注库（Base Chunk Bank，BCB）、功能块标注库（Functional Chunk Bank，FCB）、依存描述库、句法语义链接库（Syntax-Semantics Linking Bank，SSL）以及句法语义标注库（Syntactically and Semantically Annotated Corpus，SSAC）等。TCT 和 FCB 由人工标注完成，而其他库则通过算法自动提取。

TCT 目前已经标注规模为 100 万汉语词语，涵盖不同文体语料，如文学、学术、新闻和应用等。以标点符号等标记作为句子切分依据，对句法树上的结点提供成分标记和关系标记两种功能。

功能语块标注库主要处理小句层面结构信息，目前完成标注规模约二百万汉语字。一般来说，对于一个句子，主语语块主要描述了句子的陈述对象，即陈述的主题，述语语块则体现相应的动作或行为，宾语语块说明跟动作行为有关的事物。表 3-3 是汉语功能语块标记集的部分汇总。

表 3-3　汉语功能语块标记集

序　　号	语 块 标 记	语 块 内 容
1	S	主语短语
2	P	述语短语
3	O	宾语语块
4	J	兼语语块
5	D	状语语块
6	C	补语语块
7	T	独立语块
8	Y	语气块

【实例 3-2】 功能语块标注实例

1. [S 校长 [P 指出 [O 教育改革进入新阶段。

2. [P 禁止 [O 吸烟。

基本块标注库主要描述句子中直接相邻的、具有特定语义内容的词语序列。约一百万汉语词语，覆盖不同体裁文本。功能块标注库描述了句子的基本架构，是联系句法形式和语义内容的纽带。句法依存描述库描述汉语词在文本句子中可能的句法依存关系，如述宾关系、述补关系以及主谓关系等。句法语义链接库在词汇对层面上建立句法依存关系和事件语义描述的内在联系。句法语义标注库选择文本中体现事件的目标动词，确定其在句子中反映的语义词典的相应内容，形成事件内容的完整描述。

3.2 句法分析方法

句法分析的基本任务是确定句子的语法结构或词汇间的依存关系。句法分析是自然语言处理实现目标的关键环节。句法分析通常分为结构分析和依存关系分析两种。完全句法分析以获取句子整体结构为目标，而局部分析则关注局部成分，依存关系分析属于局部分析。

语法分析的目标是分析语法和句法结构并将其表示为可以理解的信息，包括短语结构和依存句法两种形式。短语结构关注句子短语的层级关系，而依存句法分析不同，着重于句子词语之间的语法关系，通常表述为树形结构。

依存理论认为词语之间存在一定主从关系，具有不等价特征。如果一个词修饰另一个词，则称修饰词为从属词(Dependent)，被修饰的词语称为支配词(Head)，两者之间的关系称为依存关系(Dependency Relationship)。如果将句子中所有词语的依存关系以有向边的方式表述，则得到依存句法树(Dependency Parse Tree)。

语言学家 Robinson 对依存句法树提出以下 4 个约束性的公理。

(1) 有且仅有一个词语(ROOT，虚拟根结点)不依存于其他词语。

(2) 除根结点之外其他单词存在依存关系。

(3) 各单词不能依存于多个单词。

(4) 如果单词 X 依存于 Y，那么位置处于 X 和 Y 之间的单词 Z 只能依存于 X、Y 或 X 和 Y 之间的单词。

这四条公理分别约束了依存句法树根结点唯一性、连通性、无环性和投射性。这些约束对语料库的标注以及依存句法分析器的开发设计创造了基础。按照生成能力，短语结构文法可以划分为四类：无约束短语结构文法，上下文有关文法(Context Sensitive Grammar)，上下文无关文法(Context Free Grammar)和正则文法(Regular Grammar)。上下文无关文法广泛应用于自然语言句法分析，分析算法高效，缺陷是存在歧义问题，这也成为句法分析需要突破的瓶颈难点。下面重点介绍上下文无关文法。

句法分析的评价标准中，比较常见的指标由标记的准确率(Precision)、召回率(Recall)和 $F1$ 值三项组成，参见式(3-1)～式(3-3)。

$$P=\frac{\text{预测正确的短语数量}}{\text{分析得到的短语总数}}\times 100\% \tag{3-1}$$

$$R=\frac{\text{预测正确的短语数量}}{\text{标准树库中短语总数}}\times 100\% \tag{3-2}$$

$$F1=\frac{2PR}{P+R} \tag{3-3}$$

由于语法的解析存在歧义性，因此结果可能导致多种语法树可供备选，从中找出可能性最高的句法树，即概率最大的句法树，是概率分布上下文无关语法 (Probabilistic Context Free Grammar，PCFG)的基本处理逻辑。概率分布上下文无关语法源自上下文无关文法。如图 3-2 和图 3-3 所示，基本结构为树状，并赋予各分支相应的语法规则，各

规则对应各自的实现概率，概率的大小值也可能体现歧义大小。基于语法树，由于各分支相当于完成整个路径选择的子步骤，因此可以将各规则的概率的乘积作为语法树整体的发生概率。自然语言处理中歧义问题比较普遍，特别是复杂的句子可能会有多种不同的句法分析树，通过句法树排歧是一种比较简单直观的方法。

表 3-4 列出了基于概率分布上下文无关语法的结构关系，树形结构参考图 3-2。

表 3-4　概率分布上下文无关语法

上下文无关语法	概率分布	上下文无关语法	概率分布
S⟶NP，VP	1.0	V⟶reached	1.0
VP⟶V，NP	0.8	NP⟶goal	0.2
PP⟶P，NP	1.0	P⟶without	1.0
VP⟶VP，PP	0.4	NP⟶help	0.3
NP⟶NP，PP	0.5	NP⟶instrument	0.5
NP⟶He	0.2	NP⟶doctor	0.6

基于上述信息，得出相应句法树的生成概率表示于式(3-4)：

$$P_1 = P(\mathrm{S}) \times P(\mathrm{NP}) \times P(\mathrm{VP}) \times P(\mathrm{V}) \times P(\mathrm{NP}) \times P(\mathrm{NP}) \times P(\mathrm{PP}) \times P(\mathrm{P}) \times P(\mathrm{NP})$$
$$= 1.0 \times 0.2 \times 0.8 \times 1.0 \times 0.5 \times 0.2 \times 1.0 \times 1.0 \times 0.2 = 0.0032 \tag{3-4}$$

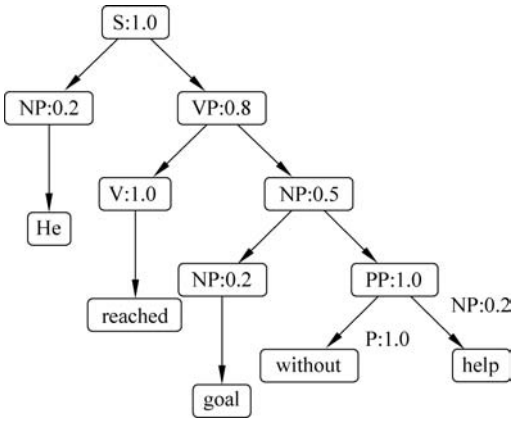

图 3-2　上下文无关概率分布

基于歧义的观点，如果存在另外一种理解导致各规则以及各结点的概率值呈现为如图 3-3 所示结果，根据概率分布上下文无关语法得出该句法树的概率则表示为式(3-5)：

$$P_2 = P(\mathrm{S}) \times P(\mathrm{NP}) \times P(\mathrm{VP}) \times P(\mathrm{VP}) \times P(\mathrm{V}) \times P(\mathrm{NP}) \times P(\mathrm{PP}) \times P(\mathrm{P}) \times P(\mathrm{NP})$$
$$= 1.0 \times 0.2 \times 0.7 \times 1.0 \times 1.0 \times 0.2 \times 1.0 \times 1.0 \times 0.2$$
$$= 0.0056 \tag{3-5}$$

比较两个概率值，第二个句法树的生成概率高，因此选择第二棵句法树作为最终结果。如果存在多种歧义，可以使用类似的方法求出概率最大的句法树。

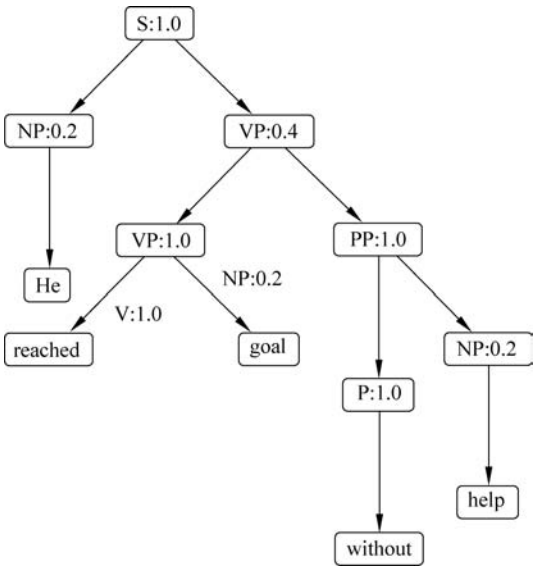

图 3-3　上下文无关概率分布(歧义)

3.3　神经网络句法分析

传统句法分析方法使用了人工标注的特征及其组合,需要前期大量的特征准备工作,人工成本问题受到关注。近年来,随着机器学习和深度学习的逐渐兴起,基于神经网络的句法分析模型开始涌现。神经网络能够对特征信息进行自动建模,具有自主学习能力,可以对特征进行自动优化,避免了大量的手动特征标注工程,并且,基于神经网络的句法分析模型的处理性能一般也优于传统的句法分析模型,因此,开发性能优异的神经网络算法成为近年研究的聚焦点。

神经网络句法分析可以基于前向反馈信息,基于结构化信息,基于搜索或者基于层次化信息模型,不同的分析方法处理方式有所不同,处理效率上可能也存在差异。

3.4　句法分析算法

句法分析过程中,假定字符序列 $S=w_1w_2\cdots w_n$ 和概率上下文无关语法 G,一般情况下,下述三点成为解决问题的关键。

(1) 如何计算由 G 产生 S 的概率 $P(S|G)$?

(2) 如果 S 有多种语法树,如何选择最优值?

(3) 如何调整 G 的规则概率参数,使得 $P(S|G)$实现最优化?

可以通过向内算法、维特比算法以及向外算法解决上面的三个问题。

3.5 句法分析工具

基于统计的句法方法在句法分析中具有重要的作用。目前在开源中文句法分析器中比较具有代表性的有 Stanford Parser 和 Berkeley Parser。前者基于因子模型，后者基于非词汇化分析模型。

Berkeley Parser 是由伯克利大学开发的句法分析器，基于 PCFG 的句法分析，目前支持的语言包括英文、中文等语言。分析器的输入形式可以文件为单位，分析完成后得到句法分析结果，支持文本、图像输出以及多线程分析，但分词功能需事先借助外部分词工具来进行分词，再将经过预处理的分词结果作为句法分析器的输入。https://github.com/slavpetrov/berkeleyparser 提供了 Jar 包和模型下载。

StanfordParser 是由斯坦福大学开发的开源句法分析器，是基于概率统计句法分析的应用程序，目前支持中文和英文等多种语言文法。它是一个高度优化的概率上下文无关文法和依存分析器，以宾州树库（Penn Treebank）作为训练数据，支持句法分析树、分词和词性标注文本、短语结构树等输出。内置分词工具、词性标注工具。

3.6 斯坦福句法分析实例

使用斯坦福句法分析器进行中文句法分析需要提前安装 jieba 和 NLTK 库，可以使用 pip install jieba 和 pip install nltk 命令进行安装，句法分析程序 Jar 包的下载地址可以参考斯坦福大学提供的下述互联网网址 https://nlp.stanford.edu/software/lex-parser.shtml#Download，目前最新版本为 4.2。此外，句法分析前需要事先安装 Java 的 JDK 应用程序并且在操作系统的环境变量中完成配置，Java 包可访问 Oracle 提供的下载地址 https://www.oracle.com/java/technologies/javase-downloads.html。

假定对象分析文本为“当今世界正经历一场百年大变革。”，利用斯坦福句法程序分析得到的句法结构如图 3-4 所示。

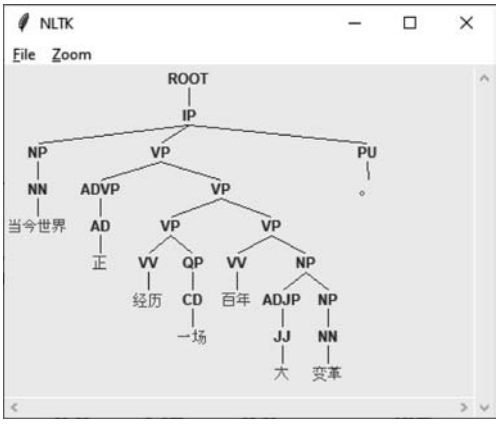

图 3-4 基于斯坦福句法的分析结果

小结

本章主要介绍了传统句法分析以及神经网络句法分析的基本概念，通过实例介绍了概率分布上下文无关语法的实际应用。

关键术语

句法分析、神经网络句法分析、概率分布上下文无关语法

习题

1. 描述句法分析的定义。
2. 描述句法分析的分类。
3. 列举基于规则句法分析的缺点。
4. 简述自然语言处理句法分析目前面临的主要技术难点。
5. 列举常见的结构标记。
6. 列举常用句法分析相关数据集。
7. 清华句法依存描述库包含哪些主要句法依存关系？
8. 简述清华句法语义链接库的功能。
9. 简述完全句法分析的主要目的。
10. 简述局部分析的主要目的。
11. 依存分析属于完全句法分析还是局部句法分析？
12. 按照生成能力，短语结构文法可以如何划分？
13. 成分句法分析的评价标准包括哪些指标？
14. 描述概率分布上下文无关语法。
15. 如果概率分布上下文无关语法存在多种概率树，最后选用哪种作为结果？
16. 概率分布上下文无关语法需要解决哪三个问题？
17. 为解决概率分布上下文无关语法的问题，通常会用到哪些算法？
18. 简要分析传统句法分析的缺点。
19. 简要分析基于神经网络的句法分析优点。
20. 简要说明伯克利句法分析器和斯坦福句法分析器的主要特征。