第 3 章 自然语言处理

充分利用信息将会给人们带来巨大的收益，而大量的信息以自然语言（英语、汉语等）的形式存在。如何有效地获取和利用以自然语言形式出现的信息？自然语言处理（Natural Language Processing，NLP）就是用计算机对自然语言信息进行处理的方法和技术。

计算机视觉技术大多数属于感知智能的层次，而本章介绍的自然语言处理技术中有相当一部分属于认知智能的层次。由于自然语言中的歧义现象以及认知科学发展的不成熟，因此与计算机视觉领域相比，自然语言处理技术仍亟待发展，一些难题（例如机器翻译、自动问答等）仍未完全解决。

本章介绍自然语言处理技术，主要介绍自然语言处理的概念、自然语言理解和自然语言生成，对自然语言处理的典型应用——机器翻译、问答系统等进行介绍。

3.1 自然语言处理概述

信息同能源、材料一起构成经济发展与社会进步的三大战略资源。信息技术正在推动和改变人类的生产、生活甚至思维方式。信息是无形的，但它可以用语言表达。语言是信息的重要载体之一，是文化的支柱，是人类思维、沟通与交流的工具。语言与经济、文化、教育、社会发展和人类进步有着紧密的关系。

3.1.1 自然语言处理的概念

自然语言处理通常是指用计算机对人类自然语言进行的有意义的分析与操作。自然语言处理的对象，包括字、词、句子、段落与篇章。图 3.1 列出了自然语言处理各级别对应的研究任务。

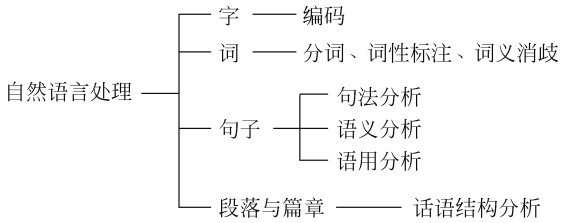


图 3.1 自然语言处理各级别对应的研究任务

自然语言处理的内容包括基础技术、核心技术和应用，如图 3.2 所示。

自然语言处理的研究方法主要包括两类：基于规则的理性主义方法和基于语料库的经

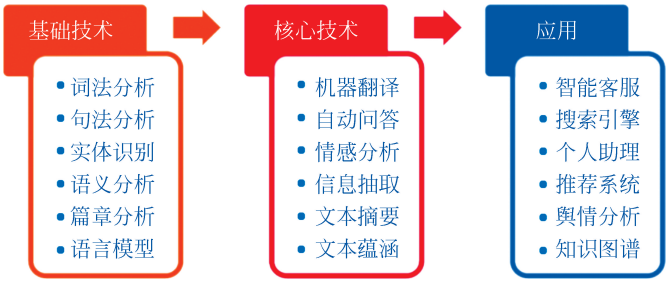


图 3.2 自然语言处理的内容

验主义方法，以下分别进行介绍。

第一类方法是基于规则的理性主义方法。这类方法基于以规则形式表达的语言知识进行符号推理，从而实现自然语言处理。这类方法强调人对语言知识的理性整理，受到了美国语言学家乔姆斯基(Noam Chomsky)主张的人具有先天语言能力观点的影响。以下是基于规则方法的代表性成果：

(1) 基于词典和规则的形态还原、词性标注及分词。

(2) 基于上下文无关文法(Context-Free Grammar，CFG)和扩充的上下文无关文法句法。

(3) 基于逻辑形式和格语法的句义分析。

(4) 基于规则的机器翻译。

第二类方法是基于语料库的经验主义方法。这类方法以大规模语料库为语言知识基础，利用统计学习和深度学习等方法自动获取隐含在语料库中的知识，学习到的知识体现为一系列模型参数，然后基于学习到的参数和相应的模型进行语言信息处理。以下是基于语料库方法的代表性成果：

(1) 语言模型(*N* 元文法)。

(2) 分词、词性标注(序列化标注模型)。

(3) 句法分析(概率上下文无关模型)。

(4) 文本分类(朴素贝叶斯模型、最大熵模型)。

(5) 机器翻译(IBM Model 等)。

(6) 机器翻译(基于人工神经网络的深度学习方法)。

语料库(corpus)是指存放在计算机中的原始语料文本或经过加工后带有语言学信息标注的语料文本。可以把语料库看作一个特殊的数据库，能够从中提取语言数据，以便对其进行分析、处理。第 2 章介绍过计算机视觉领域的部分数据集，而语料库则是自然语言处理(包含语音处理)领域中数据集的专有名称。

语料库具有以下特征：

(1) 语料库中存储了在实际使用中真实出现过的语言材料。

(2) 语料库是以计算机为载体，承载语言知识的基础资源，但并不等于语言知识。

(3) 真实语料需要经过分析、处理和加工，才能成为有用的资源。

实际上，不仅人工智能领域中的自然语言处理研究者使用语料库进行语言的研究和处理，传统的语言学研究者也依赖语料库进行语言学的研究。例如，2019 年 11 月上海外国语

大学就成立了语料库研究院，这是一个校级跨学科实体研究机构。因此，语料库是人工智能领域自然语言处理研究者和语言学研究者共同的研究利器。不同学科的研究彼此之间不再壁垒森严，跨学科的合作研究是当代科学研究的趋势之一。

3.1.2　自然语言处理的发展史

自然语言处理的历史几乎跟人工智能的历史一样长。计算机出现之后就有了人工智能的研究，而最早的人工智能研究就已经涉及了自然语言处理和机器翻译。一般认为，自然语言处理的研究是从机器翻译系统的研究开始的。

一般把自然语言处理的发展过程粗略地划分为萌芽期、复苏期和以大规模真实文本处理为代表的繁荣期。

1. 萌芽期（20 世纪 40 年代至 60 年代中期）

在这个时期，自然语言处理的主流方法是经验主义方法。机器翻译是自然语言处理最早的研究领域。

1947 年，被誉为机器翻译鼻祖的美国数学家韦弗（Warren Weaver）最早提出了机器翻译的概念，并与英国数学家 Andrew Booth 在 1949 年共同提出了机器翻译 4 种可能的实现策略。从冷战的初期开始，当时的美国、苏联等国家展开的英俄互译研究工作开启了自然语言处理的早期阶段。由于早期研究理论和技术的局限，所以当时开发的机器翻译系统技术水性较低，不能满足实际应用的需要。1954 年，美国乔治敦大学与 IBM 公司合作，在 IBM 701 计算机上将俄语翻译成英语，进行了第一次机器翻译的试验。尽管这次试验的文本仅包含了 250 个俄语单词，语法规则也只有 6 条，但它第一次展示了机器翻译的可行性。

1956 年，美国著名语言学家乔姆斯基提出了形式语言和形式文法的概念，把自然语言和程序设计语言放在相同的层面，使用统一的数学方法进行解释和定义。乔姆斯基建立了转换生成文法，使语言学的研究进入了定量研究的阶段。乔姆斯基建立的文法体系仍然是目前自然语言处理中文法分析依赖的体系，也是基于规则理性主义方法的主要理论基础，但它不能处理复杂的自然语言问题。

机器翻译作为自然语言处理的核心研究领域，在这个时期经历了极不平坦的发展道路。第一代机器翻译系统的质量很低，并且随着研究的深入，人们看到的不是机器翻译的成功，而是一个又一个无法克服的困难。在此后一段时间内，机器翻译的研究跌到了低谷。1966 年，由皮尔斯（John R. Pierce）担任主席的 ALPAC（Automatic Language Processing Advisory Committee，自动语言处理咨询委员会）发表了报告，全面否定了机器翻译的可行性。从此，机器翻译进入了长期的低谷期。

2. 复苏期（20 世纪 60 年代后期至 80 年代中期）

在这个时期，自然语言处理领域的主流方法是理性主义方法。人们更关心思维科学，通过建立很多小的自然语言处理系统模拟人的语言智能和行为。在该时期，计算语言学（Computational Linguistics，CL）的理论得到长足的发展。这个时期，自然语言理解领域的发展又可以分为 20 世纪 60 年代以关键词匹配技术为主的阶段和 20 世纪 70 年代以句法-语义分析技术为主的阶段。

从 20 世纪 60 年代开始，已经出现了一些自然语言处理系统，用来处理受限的自然语言

子集。这些人机对话系统可以作为专家系统、办公自动化系统以及信息检索系统等的自然语言人机接口，具有很大的实用价值。这些系统大都没有真正意义上的文法分析，而主要使用关键词匹配技术理解输入句子的意思。

1968 年，MIT 的拉斐尔（Bertram Raphael）使用 LISP 语言开发成功语义信息检索（Semantic Information Retrieval，SIR）系统。SIR 系统能够记住用户通过英语告诉它的事实，然后对这些事实进行演绎，并回答用户提出的问题。MIT 的维森鲍姆（Joseph Weizenbaum）在 1966 年设计的 ELIZA 系统能够使用英语和用户进行交谈，这也是第一个聊天机器人（chat robot）。ELIZA 最著名的应用是模拟心理医生和患者进行谈话，如图 3.3 所示。

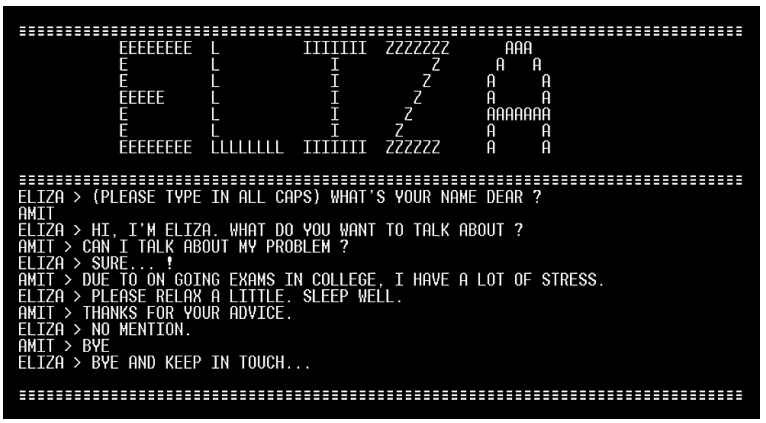


图 3.3 ELIZA 与人谈话

在这些系统中存储了大量包含某些关键词的模式，每个模式都与一个或多个解释（即回答）相对应。系统将当前用户输入的句子与这些模式进行匹配，一旦匹配成功，就得到了这个输入句子的解释，而不再考虑输入句子中那些非关键词对句子意思的影响。因此，基于关键词匹配的理解系统并非真正的自然语言理解系统，它既不懂文法，也不懂语义，仅仅是一种近似匹配技术。这种技术最大的优点是允许输入的句子不一定要遵守规范的语法，甚至可以是文理不通的句子；而其主要缺点则是性能不高，经常会导致错误的分析和回答。

在这个时期，自然语言处理领域取得了很多重要的理论研究成果，包括约束管辖理论、扩充转移网络、广义短语结构语法和句法分析算法等。这些成果为自然语言的自动句法分析奠定了良好的理论基础。在语义分析方面，提出了格语法（case grammar）、语义网络（semantic network）、优选语义学和蒙塔格语法（Montague grammar）等。

自然语言理解研究在句法和语义分析方面的重要进展还表现在建立了一些有影响的自然语言处理系统，在语言分析的深度和难度上有了很大进步。例如，1972 年，美国 BBN 公司伍兹（William Woods）设计了 Lunar 人机接口，该系统第一次允许用户使用日常英语与计算机进行对话，用于协助地质学家查找、比较和评价阿波罗 11 号飞船带回的月球标本的化学分析数据。Lunar 系统的词汇量在 3500 左右。同年，美国斯坦福大学维诺格拉德（Terry Winograd）设计了 SHRDLU 系统，这是一个在积木世界中进行英语对话的自然语言理解系统。该系统把句法、推理、上下文和背景知识灵活地结合起来，模拟一个能够操纵桌子上的积木玩具的机器人手臂，用户通过人机对话方式命令机器人放置积木玩具。系统

通过屏幕给出回答并显示现场的场景，还能提出比较简单的问题。该系统为自然语言的计算机处理做出了巨大贡献。

进入 20 世纪 80 年代之后，自然语言理解的应用研究进一步深入，机器学习研究也十分活跃，出现了许多具有较高水平的实用化系统，其中比较著名的有美国的 METAL 和 LOGOS、日本的 PIVOT 和 HICAT、法国的 ARIANE、德国的 SUSY 等，国内则有由中国软件总公司开发的商品化英汉机译系统“译星”(TRANSTAR)。这些系统是自然语言理解研究的重要成果，表明自然语言处理在理论和应用上均获得了重要进展。

在复苏期，自然语言处理领域取得的研究成果不仅为自然语言理解的发展奠定了坚实的理论基础，而且对目前人类语言能力的研究以及促进认知科学、语言学、心理学和人工智能等相关学科的发展都具有重要的意义。

3. 繁荣期(20 世纪 80 年代后期至今)

从 20 世纪 90 年代开始，自然语言处理的研究人员越来越多地开展实用化和工程化的解决方案研究，经验主义方法得到迅速发展，出现一批商品化的自然语言人机接口，例如美国 AIC 公司的英语人机接口 Intellect 和美国弗雷公司的人机接口 Themis。同时，在自然语言处理研究的基础上，机器翻译也走出了低谷，出现了一些具有较高水平的机器翻译系统并进入了市场，例如美国乔治敦大学的 SYSTRAN 系统。欧洲共同体(欧盟的前身)在 SYSTRAN 系统的基础上实现了英、法、德、西、意、葡等多种语言的互译。SYSTRAN 是基于规则的机器翻译技术的代表性商业化系统(https://www.systransoft.com/)，目前仍然使用很广泛。

这个时期自然语言处理研究的突出标志，是把基于统计的方法引入自然语言处理领域，提出了语料库语言学(corpus linguistics)，并发挥了重要的作用。由于语料库语言学从大规模真实语料中获取语言知识，使得对自然语言规律的认识更加客观、准确，因而受到越来越多的研究者青睐。在 20 世纪 90 年代，随着 Web 的快速发展，语料的获取更加便捷，语料库的规模也越来越大，质量越来越高，语料库语言学的兴起又推动了自然语言处理其他相关技术的快速发展，一系列基于统计模型的自然语言处理系统开发成功了。

基于大规模语料的统计学习方法获得充分发展，结束了基于规则的自然语言处理研究方法一统天下的局面。例如，1983 年，英国语言学家利奇(Geoffrey Neil Leech)领导的研究小组设计了成分似然性自动词性标注系统(Constituent Likelihood Automatic Word-tagging System，CLAWS)，利用已带有词性标记的 Brown 语料库，通过统计模型消除兼类词歧义，对 LOB 语料库(Lancaster-Oslo/Bergen Corpus)约 100 万词的语料进行自动词性标注，准确率可达 96.7%。此外，隐马尔可夫模型(Hidden Markov Model，HMM)等统计方法在语音识别中的成功应用对自然语言处理的发展起到了重要的推动作用。

基于统计的机器学习方法在机器翻译上也取得了成功。IBM 公司的布朗(Peter F. Brown)等人在《计算语言学》(*Computational Linguistics*)杂志发表的《统计机器翻译方法》(1990 年)和《统计机器翻译的数学：参数估计》(1993 年)两篇论文奠定了统计机器翻译(Statistical Machine Translation，SMT)的基础。对 Peter F. Brown 等人建立的模型，一般简称为 IBM 翻译模型。IBM 翻译模型共包括 5 个复杂度依次递增的统计翻译模型。

20 世纪 80 年代以来设想和进行的智能计算机研究也对自然语言理解提出了新的要求，此后又提出了对多媒体计算机的研究。新型的智能计算机和多媒体计算机均要求设计

出更为友好的人机界面，使自然语言、文字、图像和声音等信号都能直接输入计算机。要实现计算机能使用自然语言与人进行对话交流这个目标，就需要计算机具有自然语言能力，尤其是口语理解和生成能力。

自辛顿在2006年提出深度学习技术以来，深度学习最先在计算机视觉领域取得突破性成绩。2011年，微软研究院的邓力和俞栋等人与辛顿合作，创造了第一个基于深度学习的语音识别系统，该系统也成为深度学习在语音识别领域繁荣发展和提升的起点。自2013年提出了神经机器翻译(Neural Machine Translation，NMT)系统之后，神经机器翻译系统取得了很大的进展。神经机器翻译是指直接采用神经网络以端到端方式进行翻译建模的机器翻译方法，一般采用编码器-解码器(encoder-decoder)的结构，更简单直观。NMT中主要使用循环神经网络(Recurrent Neural Network，RNN)结构，并引入了注意力(attention)机制，如长短期记忆(Long Short-Term Memory，LSTM)等。2017年，Google公司瓦斯瓦尼(Ashish Vaswani)等人提出了Transformer模型，该模型完全基于注意力机制，使用注意力实现编码、解码以及编码器和解码器之间的信息传递。基于Transformer这个强大的基础结构，又衍生出了许多强大、复杂的大模型，其中GPT(Generative Pre-Training)和BERT(Bidirectional Encoder Representation from Transformers)是其中两个典型的代表，也是自然语言处理领域中预训练模型的代表，这两个模型在许多自然语言处理的任务上都获得了目前最好的效果。

在深度学习应用于自然语言处理的问题中，机器翻译的进展尤其引人注目，正成为该应用的代表性技术。此外，深度学习还首次使某些应用成为可能，例如，将深度学习成功应用于图像检索、生成式的自然语言对话等。深度学习在自然语言处理中的优势主要在于端到端的训练和表示学习，这使深度学习区别于传统机器学习方法，也使之成为自然语言处理的强大工具。同时，深度学习也面临着一些挑战，例如，缺乏理论基础和模型可解释性，模型训练需要大量数据和强大的计算资源。而深度学习在自然语言处理中也面临一些独特的挑战，如长尾问题、与符号处理的结合以及推理和决策等。可以预见，深度学习与其他技术(包括强化学习、推断、知识等)结合起来，将会使自然语言处理更上一层楼。

3.2 自然语言理解

自然语言理解(Natural Language Understanding，NLU)是指对某种自然语言的文本的真正理解，是自然语言处理的一部分。自然语言理解的研究目标是更好地理解语言和智能的本质，开发实用、有效的语言处理和分析系统。

自然语言理解的应用可以分为两类：基于文本的应用和基于对话的应用。基于文本进行研究的一个非常有意思的领域是故事理解。在这个任务中，自然语言理解系统首先要处理一个故事，然后回答有关这个故事的问题。这和在语文、英语课程中进行的阅读理解测验非常类似，而且这种方法为评价一个自然语言理解系统所能达到的理解深度提供了丰富的手段。基于对话的应用涉及人机交互，也可以使用语音的方式(语音处理技术将在第4章中介绍)。基于对话的自然语言处理应用典型的使用场景包括问答系统、智能客服、教学系统等，2022年年底推出的ChatGPT就属于基于对话的NLP应用。

本节主要介绍基于文本的应用，3.7节将介绍基于对话的应用。

3.2.1　自然语言理解的层次

语言虽然表示为一连串的文字符号或者一串语音流，但其内部实际上是一个层次化的结构。一个句子的层次结构为字→词→句子，而使用语音表达的句子的层次结构为音素→音节→音句，上述的每个层次都受到语法规则的约束。因此，语言的分析和理解过程也应该是一个层次化的过程。

许多现代语言学家把以上过程划分为 5 个层次：语音分析、词法分析、句法分析、语义分析和语用分析。虽然这些层次之间并不是完全相互独立的，但这种层次化的结构确实有助于更好地体现语言的内在结构，也更方便使用计算机完成自然语言理解的任务。

1. 语音分析

在有声语言中，音素是最小的、独立的声音单元。音素是一个或一组音，可以和其他音素相区别。语音分析是根据音位规则，从语音流中区分出一个个独立的音素，再根据音位形态规则找出一个个音节及其对应的词素或词。这部分内容将在第 4 章中进行介绍。

2. 词法分析

词法分析的主要目的是找出词汇的各个词素，从中获得语言学信息，例如 unforgettable 是由 un、forget 和 table 构成的。在英语等语言中，找出句子中的一个个词是很容易的，因为词与词之间有空格隔开。但是要找出各个词素的任务就复杂得多，例如 importable，它可以是 im-port-able，也可以是 import-able，这是因为 im、port 和 import 都是词素。在汉语中则很容易找出一个个词素，因为汉语中的每个字就是一个词素；但是词法分析中词（不是字）是处理的最小单元，而要切分出各个词就不容易了。例如，“我们研究所有东西”，可能是“我们/研究所/有/东西”，也可能是“我们/研究/所有/东西”，这里的“/”用于表示词之间的分隔。

通过词法分析可以从词素中获得许多语言学知识。例如，英语单词词尾中的词素“s”可以表示名词的复数形式或者动词的第三人称单数形式，“ly”一般是副词的后缀，而“ed”通常是动词的过去式与过去分词等，上述信息对于句法分析很有帮助。

3. 句法分析

句法分析是对句子和短语的结构进行分析。在自然语言处理的研究中，主要集中在句法分析上，部分原因在于乔姆斯基的理论贡献。自动句法分析的方法很多，包括短语结构语法、格语法、扩充转移网络、功能语法等。句法分析的对象是一个个的句子，分析的目的是找出词、短语等之间的相互关系以及在句子中的作用等，并以层次结构进行表示。这种层次结构可以反映从属关系、直接成分关系或者语法功能关系。

4. 语义分析

对于自然语言中的实词而言，每个词均用来称呼事务或表达概念。句子是由词组成的，句子的含义与词义直接相关，而不仅仅是词义的简单相加。例如，“我打他”和“他打我”这两句话中的词是完全相同的，但是表达的意思是完全相反的。因此，在进行语义分析时，还需要考虑句子的结构意义。语义分析就是通过分析找出词义、结构意义及其结合意义，从而确定语言所表达的真正含义。在自然语言处理中，语义和语境越来越成为一个重要的研究内容。

5. 语用分析

语用学研究语言符号和使用者之间的关系。具体地说，语用学研究语言所存在的外界环境对语言使用者的影响，描述语言的环境知识以及语言与语言使用者在给定语言环境中的关系。自然语言处理的语用分析更侧重于讲话者/听话者的模型设定，而不是处理嵌入到给定话语的结构信息。研究人员已经提出了一些语言环境计算模型，用来描述讲话者及其目的，以及听话者对讲话者信息的重组方式。构建这些模型的难点在于如何把自然语言处理的各个方面和各种不确定的生理、心理、社会、文化等因素集中在一个完整的模型中。

3.2.2 词法分析

词是最小的能够独立运用的语言单位，因此，词法分析是其他一切自然语言处理问题的基础，会对后续问题产生深刻的影响。词法分析的任务就是：将输入的句子字串转换成词序列，同时标记出各词的词性。这里所说的“字”并不仅限于汉字，也可以指标点符号、外文字母、注音符号和阿拉伯数字等任何可能出现在文本中的文字符号，所有这些文字符号都是构成词的基本单元。从形式上看，词是稳定的字的组合。

词法分析的任务如下：

(1) 形态还原。主要针对英语、德语、法语等。形态还原是指把句子中的词还原成它们的基本词形，例如，把动词的过去式还原为动词原形。

(2) 命名实体识别。识别出人名、地名、机构名等。

(3) 分词。针对汉语等，识别出句子中的词。

(4) 词性标注。为句子中的词添加预定义类别集合中的类别标记。

汉语的词与词紧密相连，没有明显的分隔标志。另外，汉语词的形态变化少，主要靠词序或虚词表示。中文的词法分析任务主要包括分词、未登录词识别、词性标注等。

1. 形态还原

形态还原是指把自然语言中句子里的词还原成基本词形，作为词的其他信息(词典、个性规则)的索引。简单地说，就是把各种时态的单词还原成单词的基本形态。对英语单词进行形态还原，主要是利用给出的规则进行处理。词干提取是抽取词的词干或词根形式，不一定能够表达完整语义。词形还原和词干提取是词形规范化的两类重要方式，都能够达到有效归并词形的目的，二者既有联系也有区别。

2. 命名实体识别

命名实体识别(Named Entity Recognition，NER)，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词、时间等，换言之，就是识别自然文本中的实体指称的边界和类别。早期的命名实体识别方法基本都是基于规则的。后来，基于大规模语料库的统计方法在自然语言处理各个方面取得了不错的效果，一大批机器学习的方法也开始用于完成命名实体识别的任务。

基于机器学习的命名实体识别方法可以分为以下几类：

- 有监督的学习方法。这类方法需要利用大规模的已标注语料对模型进行参数训练。基于条件随机场(Conditional Random Field，CRF)的方法是命名实体识别中最成功的方法之一。

- 半监督的学习方法。这类方法利用标注的小数据集(种子数据)自举学习。
- 无监督的学习方法。这类方法利用词汇资源(如 WordNet)等进行上下文聚类。
- 混合方法。多种模型相结合或利用统计方法和人工总结的知识库。

由于深度学习在自然语言处理中的广泛应用,基于深度学习的命名实体识别方法也获得了不错的效果,识别的主要思路仍然是把命名实体识别作为序列标注任务来完成。

3. 分词

词是语言中最小的能独立运用的单位,也是语言信息处理的基本单位。中文词与词之间没有明显的分隔符,使得计算机对于词的准确识别变得比较困难。因此,分词就成了中文处理中所要解决的最基本的问题,分词的性能对后续的语言处理,如机器翻译、信息检索等,有着至关重要的影响。

自然语言中经常存在歧义现象,中文也是如此。中文分词中的歧义主要包括以下几类:

- 交集型歧义字段,ABC 可以在 B、C 之间切开,也可以在 A、B 之间切开。例如,“和平等”三个字,在“独立/自主/和/平等/独立/的/原则”中是在“和”“平”之间切开的,而在“讨论/战争/与/和平/等/问题”中是在“平”“等”之间切开的。
- 组合型歧义字段,AB 可以切分为 A 和 B,也可以不切分。例如,“马上”二字,在“他/骑/在/马/上”中切分开了,而在“马上/过来”中则不切分,作为一个词出现。
- 混合型歧义。由交集型歧义和组合型歧义嵌套与交叉而成。例如,“得到达”,第一句是“我/今晚/得/到达/南京”,第二句是“我/得到/达克宁/了 ”,第三句是“我/得/到/达克宁/公司/去”,后面的两句话中都包含了“达克宁”这个命名实体。

4. 词性标注

词性(Part-Of-Speech,POS)是词汇基本的语法属性,也称为词类。词性标注(POS tagging)就是在给定句子中判定每个词的语法范畴,确定其词性并加以标注的过程。词性标注的正确与否将会直接影响到后续的句法分析、语义分析,是中文信息处理的基础性课题之一。常用的词性标注模型有 n 元语法模型、隐马尔可夫模型、最大熵模型、基于决策树的模型等,以上方法都属于基于语料库的经验主义方法。

汉语的特点是缺乏严格意义上的形态标志和形态变化,汉语词性标注的困难在于以下几点:

- 汉语缺乏词的形态变化,不能像印欧语系那样,直接从词的形态变化上判别词的类别。
- 常用词的兼类现象严重。兼类词的使用频度高,兼类现象复杂多样,覆盖面广,又涉及汉语中大部分词性的词,使得词类歧义排除的任务困难重重。
- 研究者本身的主观因素也会造成兼类词处理的困难。

词性标注的重点是解决兼类词和确定未登录词的词性问题。未登录词是指没有被收录在分词词表和语料库中但必须切分出来的词,包括各类专有名词(人名、地名、企业名等)、缩写词、新增词汇等。

兼类词是指一个词具有两种或两种以上的词性。根据研究统计,在英文的 Brown 语料库中,有 10.4%的词是兼类词。例如,以下 3 个英文短语中的 back 就是兼类词,其词性分别是形容词、名词和动词:

- the *back* door
- on my *back*
- promise to *back* the bill

汉语兼类词也非常普遍。例如，以下的“锁”分别是动词和名词：

- 把门*锁*上。
- 买了一把*锁*。

以下例句中，“研究”分别是动词和名词：

- 他*研究*人工智能。
- 他的*研究*工作。

在自然语言处理中，有许多任务可以转化为“将输入的语言序列转化为标注序列”以解决问题，称为序列标注(sequence labeling)，例如命名实体识别、词性标注等。简言之，输入是一个序列，输出也是一个序列。

例如，词性标注就是一个典型的序列标注问题。以下是一个英语句子的词性标注结果，如图 3.4(a)所示。

输入序列： It is easy to learn and use Python .

输出序列： PRP VBZ JJ TO VB CC VB NNP .

这里使用了 NLTK 工具，这是一个 Python 构建的高效的平台，用来处理自然语言数据。NLTK 是一个免费、开源社区驱动的项目。NLTK 的词性一共 36 类。

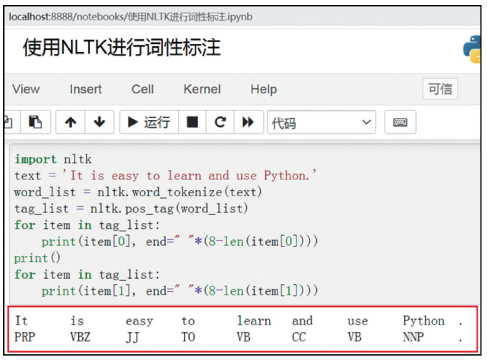
另外，命名实体识别也是一个典型的序列标注问题。以下分别为原句和命名实体识别的结果：

输入序列： 我 爱 北 京 天 安 门

输出序列： O O B E B I E

在输出序列的标记中，B 表示实体的开始，E 表示实体的结束，I 表示实体中间内容，O 表示非实体的单个字。输出序列说明，原句中有两个命名实体，第一个是北京(标注为 BE)，第二个是天安门(标注为 BIE)。

图 3.4(b)中使用 HanLP 分词器进行分词、词性标注和命名实体识别，输出中的 ns 表示命名实体中的地名。输出时把输入序列(语言序列)和得到的输出序列(标记序列)混合在了一起。在许多语料库中，这是一种常用的语料存储方式。



(a) 英语的词性标注

(b) 汉语的词性标注和命名实体识别

图 3.4 英语和汉语的词法分析结果

在词法分析中，有时候会自动过滤某些字或词，这些字或词即被称为停用词（stop words）。这些停用词大都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表。英语和汉语各自有自己的停用词表。停用词和停用词表的概念是由卢恩（Hans Peter Luhn）在 1959 年提出的。

3.2.3　句法分析

句法分析是从句子得到其结构语法的过程。不同的语法形式，对应的句法分析算法也不完全相同。由于短语结构语法（特别是上下文无关文法，Content-Free Grammar，CFG）应用最为广泛，因此以短语结构树为目标的句法分析器研究得最为彻底。很多其他形式语法对应的句法分析器都可以通过对短语结构语法的句法分析器进行改造得到。

形式语言一般是人工构造的语言，是一种确定性的语言。形式语言中的任何一个句子，只能有唯一的一种句法结构是合理的。即使语法本身存在歧义，也往往可以通过人为的方式规定一种合理的解释。例如，程序设计语言中的 if…else if…else，往往都规定 else 是和最近的 if 配对的。而在自然语言中，歧义现象是天然的、大量存在的，这些歧义的多种解释又经常都有可能是合理的。因此，对歧义现象的处理是句法分析器的基本要求。由于要处理大量的歧义现象，自然语言的句法分析器的复杂程度要远高于形式语言的句法分析器。

句法分析的任务是确定句子的句法结构或句子中词汇之间的依存关系。

句法分析主要包括 3 种：完全句法分析、局部句法分析和依存关系分析，其中，前两种句法分析是对句子的句法结构进行分析（也称为短语结构分析），而依存关系分析则是对句子中词汇间的依存关系进行分析。

1. 完全句法分析

在完全句法分析任务中，句子已经完成了词法分析，而句法分析的目的是得到句子的句法结构，通常使用短语结构树表示。可以使用层次分析法将已经进行过词法分析的句子处理为一棵短语结构树。

图 3.5(a)中的表格使用了层次分析法，而图 3.5(b)则是这个句子的短语结构树，IP 代表简单从句，NP 代表名词短语，VP 代表动词短语，ADVP 代表副词短语。

我　弟弟	已经　准备　好了	一切　用品
（主语）	（谓语）	
我　弟弟 （定语）（主语）	准备　好了 （谓语）（补语）	（宾语）
	已经　准备 （状语）（谓语）	一切　用品 （定语）（宾语）

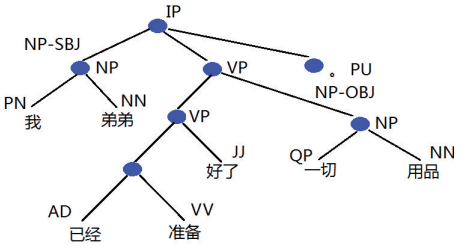


图 3.5　层次分析法和短语结构树

层次分析法枝干分明，便于归纳句型。但是，这种方法会遇到大量的歧义。另外，层次分析法还面临着很多困难：

- 在汉语中，词类与句法成分之间的关系比较复杂，除了副词只能作状语（一对一）之外，其余的都是一对多，即一种词类可以作多种句法成分。

- 词存在兼类。
- 短语存在多义。

在完全句法分析中，乔姆斯基形式文法是很重要的理论。乔姆斯基形式语言理论一共经历了古典理论、标准理论、扩充式标准理论、管辖约束理论、最简理论五个阶段。

乔姆斯基文法用 G 表示形式语法，将其表示为四元组：

$$G=(V_n,V_t,S,P)$$

- V_n：非终结符的有限集合，不能处于生成过程的终点，即在实际句子中不出现。在推导中起变量作用，相当于语言中的语法范畴。
- V_t：终结符的有限集合，只处于生成过程的终点，是句子中实际出现的符号，相当于单词表。
- S：V_n 中的初始符号，相当于语法范畴中的句子。
- P：重写规则，也成为生成规则，一般形式为 $\alpha\rightarrow\beta$，其中 α 和 β 都是符号串，α 至少含有一个 V_n 中的符号。

乔姆斯基根据重写规则的形式，将形式文法分为 4 级：

- 0 型文法（无约束文法）。
- 1 型文法（上下文相关文法）。
- 2 型文法（上下文无关文法）。
- 3 型文法（正则文法）。

句子分析过程是生成过程的逆过程。由于乔姆斯基形式文法中的生成规则是根据语法规则制定的，在分析句子是否由某文法产生的同时就等同于对句子进行语法结构分析。显然，这种方法有很大的局限性：它受到规则的限制。有很多句子无法由规则集生成（例如，“我看你打篮球”），也有很多不合逻辑的句子可以由规则集生成（例如，“我吃冰箱”）。这说明，规则很难具有完备性，同时，符合规则的句子不一定合乎逻辑。

利用概率统计法进行句法分析，主要采用概率上下文无关文法（Probabilistic Content-Free Grammar，PCFG），它是 CFG 的概率拓展，可以直接统计语言学中词与词、词与词组以及词组与词组之间的规约信息，并且可以由语法规则生成给定句子的概率。在自然语言处理领域，如果引入了概率，那么这种方法的作用很有可能是解决歧义现象（消除歧义，简称消歧），因为可以根据概率的大小对可能出现的情况进行选择。

例如，有英文例句如下：

Astronomers saw stars with ears.

首先，给 CFG 的每条句法规则赋予一个概率，这个概率代表了这条规则出现的可能性大小，如图 3.6 所示。对于左端非终结符动词短语 VP 来说，有两条句法规则 VP→V NP 和 VP→VP PP。由于 VP →V NP 更常见，所以经过统计为其赋值 0.7，为另外一条赋值 1－0.7＝0.3，即同一个左端非终结符的语法规则总的概率值为 1。图 3.6 中的概率是为了说明 PCFG，并不是来自语料库的真实概率统计结果。

对于所有可能的句法分析树，计算其整体概率，选择概率最大的作为分析结果。图 3.7 中给出了两棵句法分析树 t_1 和 t_2，一般会选择概率更大的句法树 t_1 作为句法分析的结果。

2. 局部句法分析

相比于完全句法分析要求对整个句子构建句法分析树，局部句法分析仅要求识别句子

S→NP VP	1.0	NP→NP PP	0.4
PP→P NP	1.0	NP→*astronomers*	0.1
VP→V NP	0.7	NP→*ears*	0.18
VP→VP PP	0.3	NP→*saw*	0.04
P→*with*	1.0	NP→*stars*	0.18
V→*saw*	1.0	NP→*telescopes*	0.1

图 3.6　句法规则的概率

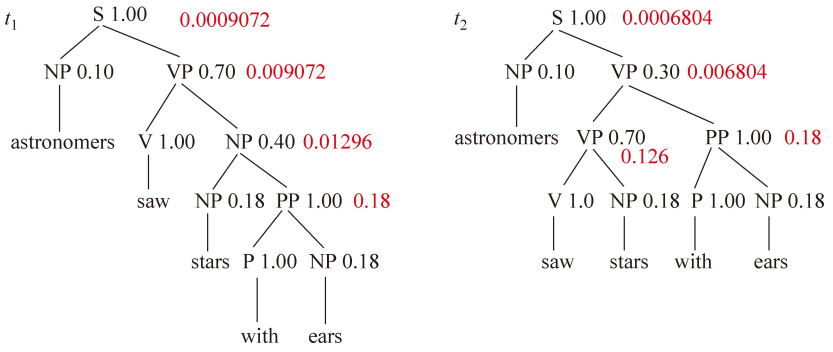


图 3.7　句法分析树的整体概率

中某些结构相对简单的独立成分，如非递归的名词短语、动词短语等。这些识别出来的结构通常被称作语块，语块和短语这两个概念基本可以认为相同。在局部句法分析中，可以将语块的分析转化成序列标注问题来解决。因此，仍然可以使用隐马尔可夫模型进行建模。

3. 依存句法分析

与完全句法分析以及局部句法分析不同，依存句法分析的主要任务是分析出词与词之间的依存关系。现代依存语法理论认为，句法关系和词义体现在词之间的依存关系中。而且，参加组成一个结构的成分（词）之间是不平等（有方向）的，一些成分从属于另一些成分，每个成分最多只能从属于一个成分。而且，哪两个词之间有依存关系是根据句法规则和词义定义的，例如：主语、宾语从属于谓语等。在一句话中，动词是句子的中心，它支配其他成分，而不受其他成分支配。例如，在句子“北京是中国的首都”中，动词“是”是句子的中心，其他成分都依存于它。可以使用有向图和依存树的形式表示依存语法。目前，依存句法分析主要是在大规模训练语料的基础上用机器学习的方法（即数据驱动的方法）得到依存句法分析器。

3.3　语料库和语言知识库

在使用统计方法（包括深度学习技术）的经验主义方法时，一般都需要首先建设大规模、有质量的数据集。计算机视觉领域就是如此，ImageNet 数据集极大地推动了图像分类的研究工作。对于自然语言处理技术来说也是如此，而该领域数据集的特定名称就是语料库（corpus）。

语料库语言学的主要奠基人和倡导者、英国著名语言学家利奇在 1991 年说：“在今天，

仅仅将语料库视为存放语言材料的仓库，是令人无法忍受的观点。新一代的兆亿级的大规模语料库可以作为语言模型的训练和测试手段，来评价一个语言模型的质量；此外，诸如困惑度之类的统计方法也可利用语料库来评估一个语法模型对语料的解释能力。”利奇曾在1983年利用已带有词性标记的Brown语料库对LOB语料库进行了自动词性标注，准确率可达96.7%。

在基于语料库的经验主义方法中，语料库是获取知识的主要来源。语料库中存储的是在语言的实际使用中真实出现过的语言材料，以计算机为载体承载语言知识的基础资源；语料库中的真实语料需要经过加工（分析和处理），才能成为有用的资源。

3.3.1 语料库

根据不同的划分标准，语料库可以分为多种类型。例如，按语种可以分为单语种语料库、双语种语料库、多语种语料库；按照地域可以分为国家语料库、国际语料库；按来源可以分为口语语料库、书面语语料库；按加工方式，单语语料库可分为原始语料库、切分标注语料库、句法树库、语义标注语料库等，双语语料库可分为篇章对齐语料库、句子对齐语料库、词语对齐语料库、结构对齐语料库等。

平行语料库也称平衡语料库，是指其内容来自多种体裁和行业领域，着重考虑的是语料的代表性和平衡性，是与专门语料库相对而言的。平行语料库一般有两种含义。第一种是指同一种语言的语料上的平行，例如，国际英语语料库（International Corpus of English，ICE）一共有20个平行的子语料库，分别来自以英语为母语或官方语言及主要语言的国家，如英国、美国、加拿大、澳大利亚、新西兰等。这些子语料库的平行性表现为语料选取的时间、对象、比例、文本数、文本长度等几乎是一致的。建库的目的是对不同国家的英语进行对比研究。第二种含义是指两种或多种语言之间的平行采样和加工。例如，机器翻译中的双语对齐语料库（句子对齐或段落对齐）就属于这种语料库。

本节介绍有代表性的语料库。

1. Brown 语料库

Brown语料库中单词的数量有上百万，以语言研究为导向，它属于第一代语料库。第一代语料库还包括LOB语料库、LLC语料库等。

Brown语料库是第一个计算机存储的美国英语语料库，也是第一个平行语料库。Brown语料库包含100万单词的语料，由美国布朗大学在1963—1964年收集，包括500个连贯英语书面语，每个文本超过2000个单词，整个语料库约1 014 300个单词，用于研究当代美国英语。Brown语料库是英语平行语料库的标准，20世纪80年代构建的英语平行语料库，如LOB语料库及LLC语料库，都遵循了Brown语料库的架构。

2. COBUILD 语料库

COBUILD语料库（Collins Birmingham University International Language Database）由辛克莱尔（John Sinclair）在20世纪80年代建立，其贡献在于它是第一个动态语料库。该语料库由英国伯明汉大学与柯林斯出版社合作完成，规模达2000万词。

3. BNC 语料库

英国国家语料库（British National Corpus，BNC）是目前网络可直接使用的最大的语料

库之一，也是目前世界上最具代表性的当代英语语料库之一。该语料库由英国牛津大学出版社、朗曼出版公司、牛津大学、兰卡斯特大学、大英图书馆等联合开发建立，于 1994 年完成。英国国家语料库词容量超过一亿，由 4124 篇代表广泛的现代英式英语文本构成，其中书面语占 90%，口语占 10%。

该语料库的建立标志着语料库语言学的发展进入一个新的阶段，并在语言学和语言技术研究方面发挥重要作用。

4. 宾州树库

树库(treebank)是一种深加工语料库，可以用来对句子进行分词、词性标注和句法结构关系的标注。树库主要包括以下两类：短语结构树库、依存结构树库，其中，短语结构树库一般使用句子的结构成分描述句子，而依存结构树库则根据句子的依存结构建立。树库的作用主要包括：为自动句法分析器提供数据和平台；为句法学研究提供真实文本标注素材；作为进行句子内部词语义项和语义关系标注的基础。

宾州树库(UPenn Treebank)由美国宾夕法尼亚大学马库斯(Mitchell Marcus)等人在 1989—1996 年历时 8 年建成，包括约 700 万词的带词性标记语料库、300 万词的句法结构标注语料库。宾州树库通过设在宾夕法尼亚大学的语言数据联盟(Linguistic Data Consortium，LDC)组织和发布，LDC 的官网为 http://www.ldc.upenn.edu。

LDC 中文树库是 LDC 开发的中文句法树库，语料取材于新华社和香港新闻等媒体，目前该语料库已发展成为第 8 版，由 3007 个文本文件构成，含有 71 369 个句子、约 162 万个词、约 259 万个汉字。文件采用 UTF-8 编码格式存储。例如，短语“制定了引进外资、加强横向经济联合和对外下放权三个文件”标注后的树状结构信息如图 3.8 所示。

```
(VP (VV 制定)
    (AS 了)
    (NP-OBJ(IP-APP (NP-SBJ(-NONE- * pro * ))
                   (VP (VP (VV 引进)
                           (NP-OBJ(NN 外资)))
                       (PU、)
                       (VP (VV 加强)
                           (NP-OBJ (ADJP(JJ 横向))
                                   (NP (NN 经济)
                                       (NN 联合))))
                       (CC 和)
                       (VP (PP(P 对)
                           (NP(NN 外)))
                       (VP (VV 下放)
                           (NP-OBJ(NN 权))))))
          (QP (CD 三)
              (CLP(M 个)))
          (NP(NN 文件))))
```

图 3.8　示例标注后的树状结构信息

5. 美国国家语料库

美国国家语料库(American National Corpus，ANC)是目前规模最大的关于美国英语

使用现状的语料库，它包括从 1990 年起的各种文字材料、口头材料的文字记录。ANC 的第一个版本包含了 1000 万口语和书面语的美式英语词汇，第二个版本则包含了 2200 万口语和书面语的美式英语词汇。ANC 也通过 LDC 组织和发布。

6.《人民日报》标注语料库

北京大学计算语言学研究所从 1992 年开始进行现代汉语语料库的多级加工，在语料库建设方面成绩显著，先后建成了 2600 万字的 1998 年《人民日报》标注语料库、包含 2000 万汉字和 1000 多万英语单词的篇章级英汉对照双语语料库以及 8000 万字篇章级信息科学与技术领域的语料库等。

《人民日报》标注语料库对《人民日报》1998 年上半年的纯文本语料进行了词语切分和词性标注，严格按照《人民日报》的日期、版序、文章顺序编排。文章中的每个词语都带有词性标记。目前的标记集里除了包括 26 个基本词类标记外，从语料库应用的角度又增加了专有名词(人名 nr、地名 ns、机构名称 nt、其他专有名词 nz)，从语言学角度也增加了一些标记，总共使用了 40 多个标记。后来又推出了 2014 版的《人民日报》标注语料库(大小约 116MB)，可以用来训练词性标注、分词模型、实体识别模型等，如图 3.9 所示。

```
37 新年/t 前夕/f ，/w [国家/n 主席/nnt]/nnt 习近平/nr 通过/p [中国/ns 国际/n 广播/vn 电台/nis]/nt 、/w [中央/n 人民/n 广播/vn 电台/nis]/nt 、/w [中央/n 电视台/nis]/nt ，/w 发表/v 了/ule 2014年/t [新年/t 贺词/n]/nz 。/w
38 这/rzv 是/vshi 习近平/nr 主席/nnt 发表/v 的/ude1 [第/m 一份/mq]/mq [新年/t 贺词/n]/nz 。/w 在/p 这份/r 寥寥/z [数百/m 字/n]/mq 的/ude1 贺词/n 中/f ，/w 我们/rr 读/v 到/v 的/ude1 不/d 只是/d 来自/v [国家/n 领导人/nnt]/nnt 的/ude1 新年/t 问候/vn 与/cc 祝福/vn ，/w 更/d 读/v 到/v 一个/mq 在/p [世界/n 舞台/n]/nz 上/f 步履/n 坚毅/a 、/w 包容/v 自信/a 的/ude1 中国/ns ，/w 一个/mq 在/p 改革/v 大道/ns 上/f 矢志不移/dl 、/w [扬帆/vi 起航/vi]/nz 的/ude1 中国/ns ，/w 一个/mq 在/p [复兴/vn 之/uzhi 路/n]/nz 上/f 努力/ad 让/v 社会/n 更加/d [公平/a 正义/n]/nz 、/w 让/v [人民/n 生活/vn]/nz 更加/d 美好/a 的/ude1 中国/ns 。/w
39 “/w [70/m 多/a 亿/m 人/n]/mq [共同/d 生活/vn]/nz 在/p 我们/rr 这个/rz 星球/n 上/f ，/w 应该/v 守望相助/vl 、/w 同舟共济/vl 、/w [共同/d 发展/vn]/nz 。/w ”/w 这/rzv 是/vshi [中国/ns 政府/nis]/nt 对/p [世界/n 和平/n]/nz 发展/vn 、/w 包容/v 共/d 进/vf 的/ude1 期待/vn 。/w 同/p 处/n 一个/mq 地球村/n ，/w 中国/ns 这个/rz “/w 国际/n 公民/n ”/w 是/vshi 国际/n 责任/n 的/ude1 践行者/nz ，/w 也/d 是/vshi [各国/rzs 人民/n]/nz 追梦/nz 的/ude1 支持者/n 。/w 今年/t 时值/v 甲午战争/nz 爆发/v 两甲子/m ，/w 曾/d 饱受/v 百年/mq 屈辱/n 的/ude1 [中国/ns 人民/n]/nz ，/w 深知/v 和平/n 生活/vn 的/ude1 宝贵/a ，/w 不畏/v [复兴/vn 之/uzhi 路/n]/nz 的/ude1 艰辛/an 。/w 以史为鉴/vl ，/w 面向/v 未来/t ，/w “/w [中国/ns 人民/n]/nz 追寻/v 实现/vn 中华民族/n 伟大/a 复兴/vn 的/ude1 中国/ns 梦/n ，/w 也/d 祝愿/v [各国/rzs 人民/n]/nz 能够/v 实现/vn 自己/rr 的/ude1 梦想/n 。/w ”/w
40 “/w 2014年/t ，/w 我们/rr 将/d 在/p 改革/vn 的/ude1 道路/n 上/f 迈出/v 新的/a 步伐/n 。/w ”/w 这/rzv 是/vshi [中国/ns 政府/nis]/nt 对/p 新/a 一年/mq 全面/ad 深化改革/v 的/ude1 有力/a 表态/vi 。/w 习近平/nr 主席/nnt 将/d 2013/m 定义/n 为/p 对/p 国家/n 和/cc 人民/n “/w 很/d 不/d 平凡/a 的/ude1 一年/mq ”/w 。/w 我们/rr 战胜/v 了/ule 各种/rz 困难/an 和/cc 挑战/vn ，/w 也/d 切实/ad 感受到/nz 了/ule 身边/s 的/ude1 各种/rz 细微/a 变化/vn 。/w 这/rzv 一年/mq 注定/v 是/vshi 载入/v [中国/ns 改革/vn]/nz 史册/n 的/ude1 标志性/n 一年/mq 。/w 全面/ad 改革/v 号角/n 吹响/v ，/w 未来/t 发展/v 蓝图/n 绘/v 就/d ，/w 又/d 一个/mq 美丽/a 的/ude1 “/w 春天/t 故事/n ”/w 正在/d 中国/ns 大地/n 谱写/v 。/w “/w 一分/t 布置/v ，/w 九分/t 落实/v ”/w ,/w 无论/c 多么/d 美好/a 的/ude1 改革/vn 路径/n ，/w 只有/c 付诸/v 实践/vn 才能/n 得到/v 检验/vn 、/w 评估/vn 和/cc 收获/n 。/w 过去/vf 一年/mq ，/w 中国/ns 社会发展/v 中/f 那些/rz 深入人心/vl 的/ude1 新气象/n ，/w 让/v 我们/rr 有/vyou 理由/n 相信/v ：/w 作为/p 全面/ad 深化改革/v 元年/t 的/ude1 2014/m ，/w 中国/ns 的/ude1 改革者/nnd 们/k 会/v 有/vyou 更大/d 作为/p ，/w 公众/n 对/p 改革/vn 的/ude1 参与/vn 将/d 更加/d 积极/ad 。/w
41 “/w 让/v 老百姓/n 过/uguo 上/f 更加/d 幸福/a 的/ude1 生活/vn ，/w 还有/v 大量/m 工作/vn 要/v 做/v 。/w ”/w 这/rzv 是/vshi [中国/ns 政府/nis]/nt 对/p 人民/n 的/ude1 庄严/a 承诺/vn 。/w 犹/d 记得/v 新/a 一届/mq [中央/n 政治局/nis]/nto 常委/nnt 同/p [中外/b 记者/nnt]/nz 见面/vi 时/ng ，/w “/w 人民/n 对/p [美好/a 生活/vn]/nz 的/ude1 向往/vn ，/w 就是/v 我们/rr 的/ude1 [奋斗/vi 目标/n]/nz ”/w ，/w 习近平/nr 的/ude1 深情/n 阐述/v 激起/v
```

图 3.9 《人民日报》标注语料库(2014 版)中的语料

7. 联合国平行语料库

联合国平行语料库(1.0 版)由已进入公有领域的联合国正式记录和其他会议文件组成。这些文件多数都有联合国 6 种语言(英、法、俄、汉、阿拉伯、西班牙)的文本。该语料库当前版本包含 1990—2014 年编写并经人工翻译的文字内容，包括以语句为单位对齐的文本。创立语料库既是为了表明联合国对多种语言并用的承诺，也是因为统计机器翻译(Statistical Machine Translation，SMT)在大会和会议管理部各笔译处和联合国统计机器翻译系统 Tapta4UN 中的作用越来越大。

联合国平行语料库的目的是提供多语种的语言资源，帮助相关各界在机器翻译等各种

自然语言处理方面开展研究并取得进展。为了方便使用，该语料库还提供了现成的特定语种双语文本和六语种平行语料子库。

8. 欧洲议会平行语料库

欧洲议会平行语料库是从欧洲议会的会议记录里抽取出来的，是目前互联网上可免费获取的非常规范的平行语料库。该语料库的时间跨度为 1996—2006 年，目前这个语料库还在继续扩建。

第 3 版的欧洲议会平行语料库包括 11 种语言的单语语料库和 10 对双语语料库，其中单语语料库主要用于语言模型的训练，双语语料库主要用于统计机器翻译中翻译模型的训练。欧洲议会平行语料库中 11 种语言的语料如图 3.10 所示。

Danish: det er næsten en personlig rekord for mig dette efterår .
German: das ist für mich fast persönlicher rekord in diesem herbst .
Greek: πρόκειται για το προσωπικό μου ρεκόρ αυτό το φθινόπωρο .
English that is almost a personal record for me this autumn !
Spanish: es la mejor marca que he alcanzado este otoño .
Finnish: se on melkein minun ennätykseni tänä syksynä !
French: c ' est pratiquement un record personnel pour moi , cet automne !
Italian: e ' quasi il mio record personale dell ' autunno .
Dutch: dit is haast een persoonlijk record deze herfst .
Portuguese: é quase o meu recorde pessoal deste semestre !
Swedish: det är nästan personligt rekord för mig denna höst !

图 3.10　欧洲议会平行语料库中 11 种语言的语料

3.3.2　语言知识库

语言知识库在自然语言处理的研究中具有重要的作用。词汇知识库、句法规则库、语法信息库和语义概念库的各类语言知识资源，都是自然语言处理技术赖以建立的重要基础。本节对一些具有代表性的语言知识库进行简要介绍。

语言知识库包含了比语料库更广泛的内容。广义上来说，语言知识库可分为两种类型：第一种是词典、规则库、语义概念库等，其中的语言知识表示是显性的，可使用形式化结构进行描述；第二种语言知识存在于语料库之中，每个语言单位（主要是词）的出现，其意义、内涵、用法都是确定的。语料库是文本的集合，也是语句的结合，其中的每一个语句都是线性的非结构化的文字序列，其中包含的知识都是隐含的。语料加工的目的就是要把隐含的知识明确化，以便计算机能够学习和使用。

下面将对具有代表性的语言知识库进行简要介绍。

1. WordNet

WordNet 是由美国普林斯顿大学米勒（George A. Miller）领导开发的英语词汇知识库，是一种传统的词典信息与计算机技术以及心理语言学等学科有机结合的产物。WordNet 从 1985 年开始建设，目前已经成为国际上非常有影响的英语词汇知识库，其官网为 https://wordnet.princeton.edu。

WordNet 与同义词词林类似，使用同义词集合（synset）作为基本的构建单位来组织。但不同的是，WordNet 不仅是用同义词集合的方式列出概念，而且同义词集合之间是以一

定数量的关系类型相互关联的。这些关联关系包括同义关系、反义关系、上下位关系、整体与部分关系、继承关系等。在这些语义关联关系中，同义关系是最基础的语义关系，也是WordNet组织词汇的方式。为了尽量使语义之间的关系明晰、易于使用，WordNet中没有包含发音、派生形态、词源信息、用法说明、图示等。

由此可见，WordNet是一个按语义关系网络组织的巨大词库，使用多种词汇关系和语义关系组织表示词汇的知识。词形式和词义是WordNet源文件中的两个基本组成部分，其中词形式用规范的词形表示，词义则用同义词集合表示。词汇关系是两个词形式之间的关系，而语义关系是两个词义之间的关系。

WordNet中词汇的组织方式如图3.11所示。

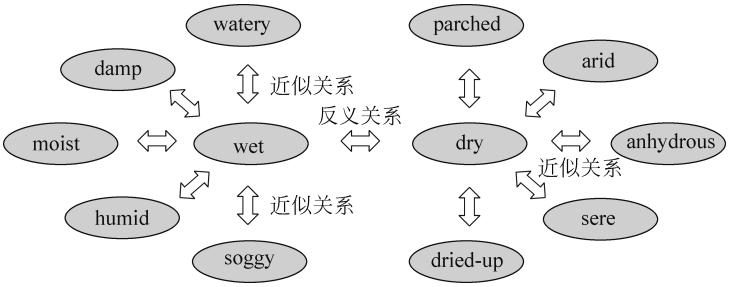


图3.11　WordNet中词汇的组织方式

2. 北京大学综合型语言知识库

北京大学计算语言学研究所俞士汶教授领导建立的综合型语言知识库CLKB覆盖了词、词组、句子、篇章各单位和词法、句法、语义多个层面，从汉语向多语言辐射，从通用领域深入到专业领域。CLKB是目前国际上规模最大并获得广泛认可的汉语语言知识资源，其中的《现代汉语语法信息词典》是一部面向语言信息处理对的大型电子词典，收录了8万个汉语词语，在依据语法功能分布完成的词语分类的基础上，又根据分类进一步描述了每个词语的详细语法属性。

《现代汉语语法信息词典》以复杂特征集、合一运算理论为基础，采用“属性-属性值”的形式详细描述了词语的句法知识，并使用关系数据库技术把“属性-属性值”的描述形式转换为数据库表的字段和值，如表3.1所示。

表3.1　《现代汉语语法信息词典》示例

词　语	词　类	同　形	拼　音	备　注	…
挨	v	A	ai1	触，碰，靠近	
挨	v	B	ai2	遭受，忍受	
保管	v	1	bao3guan3	保存	
保管	v	2	bao3guan3	担保	
报告	n		bao4gao4	书面文件	
报告	v		bao4gao4	发表讲话	

续表

词　语	词　类	同　形	拼　音	备　注	…
别	d		bie2	不要	
别	v	A	bie2	分离	
别	v	B	bie2	附着，固定	

在表 3.1 中，属性“词语”“词类”“同形”是主要的描述信息，其中的“同形”用于对同一词类的同形词的不同义项在粗粒度上进行区分，如果某个词在读音和词类均相同的情况下义项不同，那么“同形”的值使用 1、2、3 等数字进行区分。当“同形”的值使用 A、B、C 进等字母进行区分主要有以下两种情况：第一种情况是读音不同；第二种情况是词类相同但词义不同。

除了《现代汉语语法信息词典》之外，综合型语言知识库还包含现代汉语多级标注语料库。该多级标注语料库是在对《人民日报》语料的基础上进行词语切分和词性标注建立的大规模现代汉语基本标注语料库（规模达 6000 万字）的基础上，以《现代汉语语法信息词典》和《现代汉语语义词典》为参考，加注不同粒度的词义信息之后形成的。基本标注语料库中的命名实体都使用相应标记进行了标注。

3. 知网

知网（HowNet）是机器翻译专家董振东领导创建的汉语语言知识库，是一个以汉语和英语中词语代表的概念为描述对象，以揭示概念与概念之间以及概念具有的属性之间的关系为基本内容的常识知识库。

知网作为一个知识系统，是一个意义的网络，它反映了概念的共性和个性。图 3.12 展示了知网中的多层语义关系网络，其中重点反映了概念之间和概念的属性之间的各种关系。

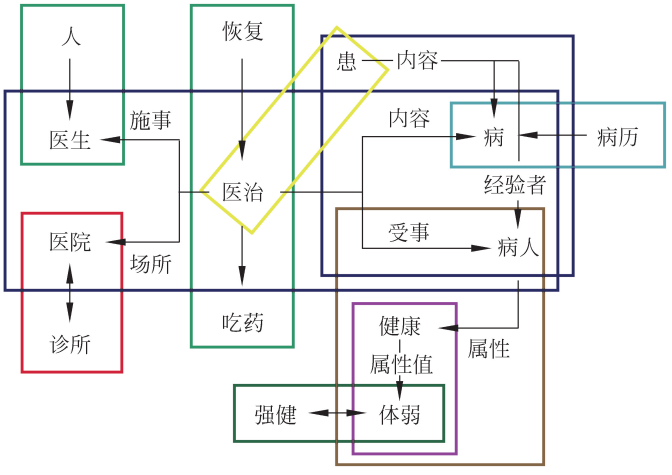


图 3.12　知网中的多层语义关系网络

通过对各种关系进行标注，知网把这种知识网络系统明确地教给了计算机，进而使知识对计算机来说成为可利用、可计算的。在知网中，一共定义了 16 种语义关系，并且这些语义

关系是用户借助于《同义、反义以及对义组的形成》自主构建的。知网是一个知识系统，而不是一部语义词典。知网使用了概念与概念之间的关系以及概念的属性与属性之间的关系并形成了一个网状的知识系统，这是知网与其他树状词汇数据库的本质区别。

总体来说，知网是一个具有丰富内容和严密逻辑的语言知识系统，它作为自然语言处理领域，尤其是中文信息处理技术重要的基础资源，在实际应用中发挥着越来越重要的作用。知网可以应用于词汇语义相似度计算、词汇语义消歧、命名实体识别和文本分类等许多具体问题和任务上。

3.4 语言模型

语言模型(Language Model，LM)在自然语言处理中占有重要的地位，尤其在基于统计模型的语音识别、机器翻译、中文分词、句法分析等研究中应用广泛。目前主要采用的是 n 元(n-gram)语法模型，这种模型构建简单、直接，但由于数据缺乏而必须采用平滑方法进行处理。本节主要介绍 n 元语法的基本概念。由于 n 元语法模型是一个马尔可夫链，因此首先来看马尔可夫链。

3.4.1 马尔可夫链

俄罗斯数学家、圣彼得堡数学学派代表性人物安德列·安德列维奇·马尔可夫(Andrei Andreyevich Markov，1856—1922)在 1906—1912 年提出了马尔可夫链。马尔可夫链(也称为马尔可夫过程)是一个典型的随机过程。

考虑一个随机变量的序列 $X=\{X_0,X_1,X_2,\cdots,X_t,\cdots\}$，这里的 X_t 表示时刻 t 的随机变量，$t=0,1,2,\cdots$。各个随机变量的取值集合范围相同，称为状态空间，表示为 S。随机变量可以是离散的，也可以是连续的。由随机变量组成的序列就构成了随机过程。

假设在时刻 0 的随机变量 X_0 遵循概率分布 $P(X_0)=\pi_0$，称为初始状态分布。在某个时刻 $t\geqslant 1$ 的随机变量 X_t 与前一个时刻的随机变量 X_{t-1} 之间有条件概率分布 $P(X_t|X_{t-1})$，如果 X_t 只依赖于 X_{t-1}，而不依赖于更早的随机变量，这一性质称为马尔可夫性，即

$$P(X_t|X_0,X_1,X_2,\cdots,X_{t-1})=P(X_t|X_{t-1}),\quad t=1,2,3,\cdots \tag{3-1}$$

马尔可夫性的直观解释是：未来只依赖于现在(假设现在是已知的)，而与过去无关，也称为无后效性。这个假设在很多应用和情况下是合理的。具有上述马尔可夫性的随机序列 $X=\{X_0,X_1,X_2,\cdots,X_t,\cdots\}$ 就称为马尔可夫链或马尔可夫过程。条件概率分布 $P(X_t|X_{t-1})$ 称为马尔可夫链的转移概率分布，简称转移概率。转移概率决定了马尔可夫链的特性。

如果转移概率分布 $P(X_t|X_{t-1})$ 与时刻 t 无关，即

$$P(X_{t+s}|X_{t-1+s})=P(X_t|X_{t-1}),\quad t=1,2,3,\cdots;s=1,2,3,\cdots \tag{3-2}$$

则称该马尔可夫链为时间齐次的马尔可夫链，本书中提到的马尔可夫链都是时间齐次的。以上定义的是一阶马尔可夫链。相应地，二阶马尔可夫链是满足式(3-3)的随机过程序列：

$$P(X_t|X_0,X_1,X_2,\cdots,X_{t-1})=P(X_t|X_{t-2},X_{t-1}) \tag{3-3}$$

同理，可以扩展到 n 阶马尔可夫链，满足 n 阶马尔可夫性：

$$P(X_t|X_0,X_1,X_2,\cdots,X_{t-1})=P(X_t|X_{t-n},\cdots,X_{t-2},X_{t-1}) \tag{3-4}$$