自然语言处理导论

张奇 桂韬 黄萱菁

2022.3.2

目 录

1. 绪论

自然语言处理（Natural Language Processing，简称 NLP）主要研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法 [1]。自然语言处理是计算机科学领域和人工智能领域的重要研究方向之一。自然语言处理研究融合了语言学、计算机科学、数学、认知心理学等多学科内容。其研究内容涵盖了从字、词、短语、句子、段落到篇章等不同粒度语言单位，也包含了从处理、理解、生成等不同层面。研究内容涉及的知识点多且复杂。随着自然语言处理的广泛应用以及以深度学习为代表的机器学习算法的快速进步，近年来自然语言处理算法也发展迅速。

本章主要介绍自然语言处理的基本概念和研究内容，并对自然语言处理范式进行总结和介绍。

1.1 自然语言处理基本概念

语言是人类区别于其他动物的本质特性。人类的多种智能也都与语言有密切的关系。人类的逻辑思维以语言为形式，绝大多数的知识也是以语言文字的形式记载和流传。现在互联网上已有超过数十万亿数量的网页资源，而这些网页中的信息大多都是用自然语言描述的。人工智能想要获取知识，就必须理解人类所使用的非精确的、有歧义的、杂乱的语言。

自然语言处理目标就是实现人机之间的有效通信，意味着要使计算机能够理解自然语言的意义，也能以自然语言文本来表达给定的意图、思想等 [1]。前者称为**自然语言理解**（Natural Language Understanding，简称 NLU），后者称为**自然语言生成**（Natural Language Generation，简称 NLG）。需要说明的是，自然语言处理、自然语言理解以及计算语言学这些概念并没有严格统一的定义。本书采用吴立德教授在 1997 年所著的《大规模中文文本处理》中所给出的定义 [1]。无论是自然语言理解还是自然语言生成，目前都是还是开放性问题（Open Problem），通用的高精度高鲁棒自然语言处理系统还没有解决方案，仍然需要长期研究的目标。但是针对特定领域应用，有很多具有自然语言处理能力的系统已经有产业化应用，例如：智能客服系统、机器翻译系统、语音助手、电子邮件筛选、新闻写作等。

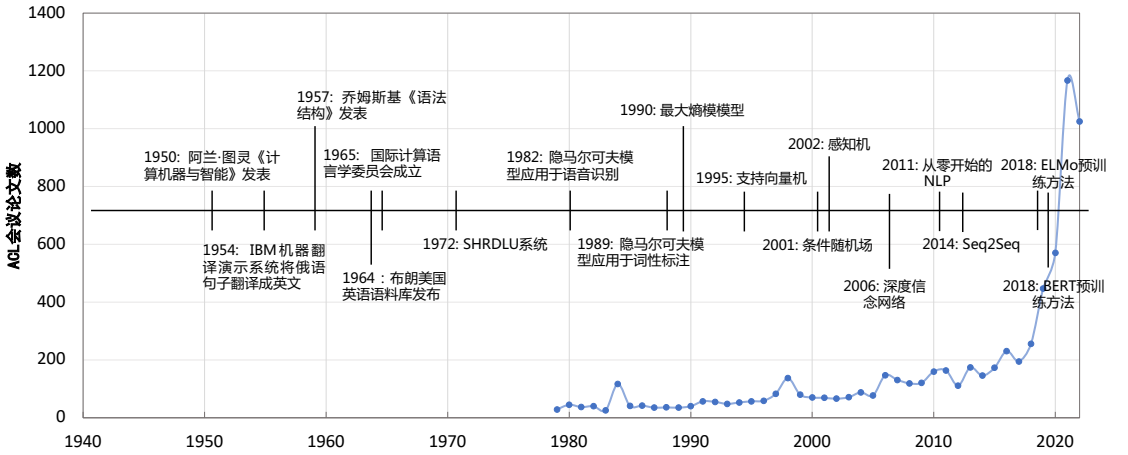

图 1.1　自然语言处理简史时间线

1.1.1 自然语言处理简史

自然语言处理的研究历史可以追溯到 1947 年，当时第一台通用计算机 ENIAC 也才刚刚面世一年，Warren Weaver 就提出了利用计算机翻译人类语言的可能，并于 1949 年发布了著名的《Translation（翻译）》备忘录。1950 年, Alan Turing 发表了著名的具有划时代意义的论文《Computing Machinery and Intelligence（计算机器与智能）》，提出了使用图灵测试（Turing Test）对机器是否具备智能进行评测，既如果一台机器能够与人类展开对话而不能被辨别出其机器身份，那么称这台机器具有智能。1951 年语言学家 Yehoshua Bar-Hillel 在麻省理工学院开始了机器翻译研究。1954 年乔治城大学与 IBM 合作的机器翻译演示系统将 60 多个俄语句子翻译成了英文。研究者们当时期望通过三到五年的时间完全解决机器翻译问题。20 世纪 50 年代初是自然语言处理的萌芽期。自然语言处理的简史的时间线如图1.1所示。大体来看自然语言处理经历了 20 世纪 50 年代末到 60 年代的初创期、20 世纪 70 年代到 80 年代的理性主义时代、20 世纪 90 年代到 21 世纪初的经验主义时代以及 2006 年至今的深度学习时代。

20 世纪 50 年代末到 60 年代, 大量的研究不断涌现, 并且形成了两大流派：符号学派（Symbolic）和随机学派（Stochastic）。以美国语言学家 Noam Chomsky 为代表的符号学派提出了形式语言理论，乔姆斯基于 1957 年发表的《语法结构（Syntactic Structures）》介绍了生成语法的概念，并提出了一种特定的生成语法称为转换语法。开启了使用数学方法研究语言的先河。随机学派则是以 1959 年 Bledsoe 和 Browning 将贝叶斯方法（Bayesian method）应用于字符识字识别问题为代表。试图通过贝叶斯方法来解决自然语言处理中的问题。这期间计算语言学（Computational Linguistics）概念也被正式提出。1962 年美国成立了“机器翻译和计算语言学学会”（Association for Machine Translation and Computational Linguistics）。1965 年国际计算语言学委员会（The International Committee on

Computational Linguistics,ICCL)成立,并与当年召开了第一届国际计算语言学大会(The International Conference on Computational Linguistics，COLING)。20 世纪 60 年代还出现了第一个大规模语料库，布朗美国英语语料库（Brown Corpus)，包含来自不同文体的 500 多篇书面文本，超过一百万单词，涉及新闻、小说、科技文等。自然语言处理研究全面开启。

20 世纪 70 年代到 80 年代，更多的工作从不同角度开展了系统的研究，也产生了一系列的研究范式，至今仍对自然语言处理研究起着重要作用。这些范式主要包括：基于逻辑的范式（Logic-based Paradigm）、基于规则的范式（Rule-based Paradigm）和随机范式（Stochastic Paradigm）。1970 年 Colmerauer 等人使用逻辑方法所研制的 Q 系统（Q-system）和“变形语法”（metamorphosis grammar）并在机器翻译中得到应用。以及 1980 年 Pereira 和 Warren 提出的“定子句语法”（Definite Clause Grammar）都是逻辑范式成功应用的范例之一。基于规则的范式是这个时代最典型的模式之一，1972 年研制的 SHRDLU 系统是其中一个代表性工作。该系统模拟了一个玩具积木世界，能够接受自然语言的书面指令（例如：Pick up a big red block.)，指挥机器人移动玩具积木块。1970 年, William A. Woods 提出了扩充转移网络（Augmented Transition Network）用来描述自然语言输入，并用于自然语言处理若干任务中。受到 20 世纪 80 年代初隐马尔可夫模型（Hidden Markov Model）和噪声信道与解码模型（Noisy channel model and decoding model）在语音识别中的成功应用，随机范式也逐渐在自然语言处理任务中展露头角，包括词性标注 [2]、姓名检索 [3] 等。

从 20 世纪 90 年代开始，自然语言处理开启了繁荣发展的时代。自 1989 年机器翻译任务中引入语料库方法之后，这种建立在大规模真实语料上的研究方法将自然语言处理研究推向了新的高度。90 年代后期开始，基于机器学习和数据驱动的方法取代了早期基于规则和基于逻辑的方法，基本成为自然语言处理的标准模式。自然语言处理的各类任务，包括词法分析、词性标注、句法分析、文本分类、机器翻译等全都开始引入机器学习算法。这期间朴素贝叶斯（Naive Bayes）[4]、K 近邻（K-nearest neighbor）[5]、支撑向量机（Support Vector Machine，SVM）[6]、最大熵模型（Maximum Entropy, ME）[7]、神经网络（Neural Network）[8]、条件随机场（Conditional Random Fields）[9]、感知机 [10] 等方法也都在自然语言处理不同任务上进行了尝试并取得了一定的成功。这种以大规模数据为基础进行分析的方法称为“经验主义”（empiricism)。随着数据驱动的方法的发展，大部分关于自然语言处理的理论都大打折扣，特别是数据量的不断增加以及计算能力的不断提高，经验主义方法直到现在也还在主导着自然语言处理处理领域。从当前自然语言处理领域重要会议 EMNLP（Empirical Methods in Natural Language Processing ）的名称和发展也可以看到经验主义的发展过程。

2006 年加拿大多伦多大学教授 Geoffrey Hinton 和他的学生 Ruslan Salakhutdinov 在《科学》杂志上发表了基于深度信念网络（Deep Belief Networks, DBN）以及无监督预逐层预训练结合有监督训练微调的方法解决深层神经网络训练中梯度消失问题 [11]，将神经网络重新拉回到机器学习研究者的视野中。2012 年基于卷积神经网络（Convolutional Neural Network，CNN）网络的 AlexNet 在图像识别领域 ImageNet 竞赛中取得惊人的效果，开启了深度学习在学术界和工业界的浪潮 [12]。

2011 年“从零开始的 NLP”论文引起了极大的关注，深度神经网络可以不使用人工特征的情况下，一个统一的网络架构在词性标注、组块分析、命名实体识别、语义角色标注等任务中都取得了很好的效果 [13]。2014 年 Seq2Seq（序列到序列）的模型 [14] 在机器翻译任务上取得了非常好的效果，并且完全不依赖任何人工特征，推动了神经机器翻译的广泛落地。这种端到端的方式进行编码和解码的方式不仅有效推动了包括生成式摘要 [15]、对话系统 [16, 17] 等在内的其它自然语言生成问题上取得了突破，还应用于自然语言处理中很多任务，包括句法分析 [18]、问题回答 [19]、中文分词 [20] 等。此外，循环神经网络（Recurrent neural network，RNN）[21]、长短时记忆网络（Long Short Term Memory Network, LSTM）[22]、递归神经网络（Recursive Neural Network）[23]、卷积神经网络（Convolutional Neural Network, CNN）[24]、图神经网络（Graph Neural Networks，GNN）[25, 26] 等神经网络模型也都成功应用于自然语言处理各个任务中。

2018 年美国艾伦人工智能研究所（Allen Institute for AI）和华盛顿大学（Washington University）联合发表的论文中提出了名为 ELMo 的上下文相关的文本表示方法，首先利用语言模型或其他自监督任务进行预训练，此后在处理下游任务时，从预训练的网络中提取对应单词的网络各层的词嵌入作为新特征补充到下游任务中，在多个自然语言处理任务上表现非常突出 [27]。此后，深度学习开启了预训练模型（Pre-trained Models，PTM）结合任务微调的新范式。谷歌、OpenAI、微软、清华大学、百度、智源研究院等先后提出了 BERT[28]，GPT[29]，XLNet[30]、ERNIE(THU)[31]、ERNIE(Baidu)[32]、悟道等大规模预训练模型，在几乎所有自然语言处理任务中都取得了非常好的效果，甚至在很多任务的标准评测集合上达到了超越人类准确率的水平。尤其是类似阅读理解、常识推理等任务上有惊人的效果提升。与此同时，预训练模型的规模也越来越大，2018 年谷歌开发的 BERT-Base 模型有 1.1 亿参数，BERT-Large 模型有 3.4 亿参数，到了 2019 年 OpenAI 开发的 GPT-2 模型就达到了 15 亿参数量。2021 年 GPT-3 模型参数量更是达到了 1750 亿，而同年谷歌开发的 Switch Transformer 模型参数量首次超过万亿，达到了 1.6 万亿。在此之后不久，北京智源研究院所发布的“悟道 2.0”模型就刷新了上述记录，模型参数量达到了 1.75 万亿。虽然预训练大模型取得了巨大的成功，但是仍然面临模型鲁棒性亟待提升、超大规模模型如何高效适配下游任务、大模型的理论解释等诸多问题。

1.1.2 自然语言处理的主要研究内容

自然语言处理的研究内容十分庞杂，整体上可以分为基础算法研究和应用技术研究。基础算法研究又可以细分为自然语言理解和自然语言生成。从语言单位角度看涵盖了字、词、短语、句子、段落以及篇章等不同粒度。从语言学研究角度看则涉及形态学、语法学、语义学、语用学等不同层面。此外，由于目前绝大多数自然语言处理算法采用基于机器学习的方法，针对特定的自然语言处理任务，以有监督、无监督、半监督、强化学习等不同的机器学习算法为基础进行构建。因此，自然语言处理研究又与机器学习和语言学研究交织在一起，使得自然语言处理的研究内容涉及范围广，学科交叉度大。

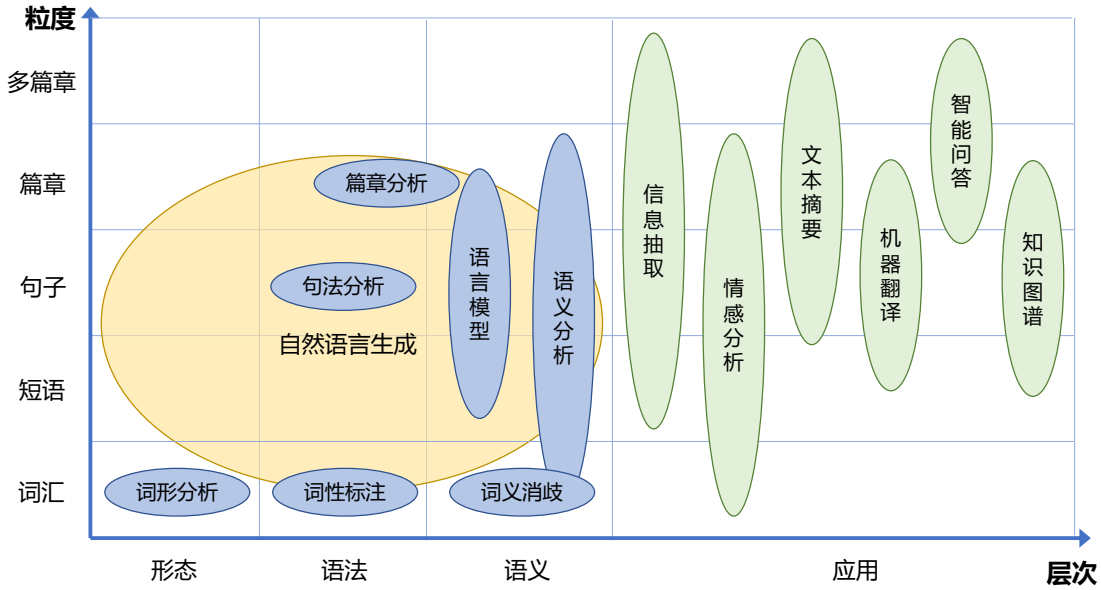

图 1.2　自然语言处理主要研究内容

自然语言处理研究还与语言学术密切相关，语言学研究可以划分为形态、语法、语义、语用等几个层面。形态学（morphology）主要研究单词的内部结构和其构成方式。语法学（syntax）主要研究句子、短语以及词等语法单位的语言结构与语法意义的规律。语义学（semantics）主要研究语言的意义，目标是发现和阐述关于意义的知识。语用学（pragmatics）是从使用者的角度来研究语言，研究在一定的上下文环境下的语言如何理解和使用。在实际的任务中，上述几个层面的问题往往相互还关联，并不能完全独立。语法结构的分析需要词汇形态学的支撑，语法结构也影响着词汇的形态，语法结构和语义也是相关交织，而下上文环境又对语义有重要的影响，因此很多自然语言处理任务并不是完全独立的。但是为了简化任务处理难度，通常处理不同的层面的任务时仍然独立考虑。从自然语言处理研究内容的难度来看，从形态、语法、语义到语用是逐层递增的。目前基于机器学习的自然语言处理算法处理涉及语义的相关任务都较为困难，因此语用层面的自然语言处理算法研究相对较少，大多数的研究集中于形态、语法和语义三个层面。

从语言单元粒度和语言学研究层次两个维度进行归类，自然语言处理主要研究内容如图1.2所示。在单词粒度的研究内容主要包括：词形分析、词性标注、词义消歧，分别针对单词的词性、语法、语义开展研究。句法分析则是主要针对句子根据语法进行结构分析。篇章分析核心是对篇章的连贯性和衔接性进行分析，涉及到篇章级别语法结构，同时也包含部分语义的内容。而语义分析研究则涉及到从词汇、短语、句子到篇章等各个粒度。语言模型主要聚焦于句子粒度，但是也包含部分短语和篇章级别的研究。以上研究内容主要围绕自然语言理解的基础问题开展。除此之

外，自然语言生成则主要研究利用常识、逻辑和语法等知识自动生成文本，涉及形态、语法和语义层面，同时也涵盖从短语到篇章多个粒度。在自然语言处理基础研究内容之上，信息抽取、情感分析、文本摘要、机器翻译、智能问答、对话系统等任务则围绕自然语言处理的应用开展，所处理的语言单元也根据任务特性而不尽相同。

自然语言处理的主要研究内容围绕语言学基础理论，在形态、语法以及语义等层面开展自然语言理解基础算法和自然语言生成基础算法研究。在此基础上围绕自然语言处理的重要应用场景开展一系列的应用技术研究。这些研究内容也已经深度应用于信息检索、虚拟助理、推荐系统、量化交易、智能问诊、精准医疗等众多系统中。

1.1.3 自然语言处理的主要难点

自然语言理解和自然语言生成都是十分困难的任务，这种困难的根本原因是自然语言在各个层面都广泛存在的各种各样的歧义性或多义性（Ambiguity）。一个自然语言文本从形式上是由字符（包括中文汉字、英文字母、符号）组成的一个字符串。由字母或者汉字可以组成词，由词可以组成词组，由词组可以组成句子，进而组成段落、篇章。无论哪种粒度的语言单元，还是从一个层级向上一层级转变中都存在歧义和多义现象。形式上一样的字符串，可以理解为不同的词串、词组串，并有不同的意义 [1]。凯斯和荷勃（Kess, F. J. 和 Hoppe, A. R.）甚至还提出了“语言无处不歧义”的理论 [33]。在某种程度上，我们也可以说自然语言处理基础任务的核心就在于解决歧义问题。

1. 语音歧义

语音歧义（Phonetic ambiguity）主要体现在口语中，是由于语言中同音异义词（Homophone）、爆破音不完全、重音位置不明确等原因造成的。汉字的同音异义现象则更加验证，在汉语中只有 413 个不同的音（节），如果结合声调的变化组合，也仅有 1277 个音（节），而汉字则多达数万个，因此同音字非常多。英语中虽然同音异义词语相对汉语要少得多，但是由于连读、爆破音、重音位置等造成的语音异义也非常常见。

例如：请问您贵姓？

免贵姓 zhang。

这组对话中“zhang”既可以是“张”，也可以是“章”。汉语中同音异义词也有非常多，例如：“chéng shì：城市、程式、成事、城事”、“jìn shì：近视、进士、尽是”、“shǒu shì：首饰、手势”等。

例如：Please hand me the flower. 请把花递给我。

Please hand me the flour. 请把面粉递给我。

这两句话中“flower”和“flour”的发音相同，由同音异义词造成了歧义。类似的情况还包括“see（看见）与 sea（大海）”、“son（太阳）与 sun（儿子）”等。在口语中这些发音相似的单词很可能会引起歧义，对人们的理解造成困扰。

2. 词语切分歧义

词语切分歧义（Word segmentation ambiguity）是在由字符组成词语时的歧义现象。对于英语等印欧语系的语言来说，绝大部分单词之间都由空格或标点分割。但是对汉语、日语等语言来说，单词之间通常没有分隔符。对于这些语言来说，这些连续的字符切分为单词时就会产生歧义。

例如：语言学是一门基础学科。

这门语言学起来很困难。

该例句中“语言学”、“语言”都是词语，在同一个句子中就会出现多种切分方法。这种切分歧义在汉语中普遍存在。我们将在第 2 章详细讨论词语切分歧义的问题以及词语切分的方法。

3. 词义歧义

词义歧义（Word sense ambiguity）是指词语具有相同形式但是不同意义。这种歧义在各种语言中都广泛存在，通常越是常见的词语其词义数量就越多。例如，《现代汉语词典（第七版）》中“打”字，有两个读音“dá”和“dǎ”，分别作为量词、动词和介词，在做为动词时，“打”字有 24 个意项 [34]。

例如：打 dǎ 动词：

(1) 用手或器具撞击物体：～门 |～鼓

(2) 器皿、蛋类等因撞击而破碎：碗～了 | 鸡飞蛋～

(3) 殴打；攻打：～架 |～援

(4) 发生与人交涉的行为：～官司 |～交道

...

英语中存在大量类似的情况，例如，根据 WordNet 3.1 定义，单词“bank”具有名词和动词两种词性，作为名词时具有 10 种词义 [35]。

例如：Bank 名词：

(1) sloping land (especially the slope beside a body of water

“they pulled the canoe up on the bank”

(2) a financial institution that accepts deposits and channels the money into lending activities

“he cashed a check at the bank”

(3) a long ridge or pile

“a huge bank of earth”

(4) an arrangement of similar objects in a row or in tiers

“he operated a bank of switches”

...

我们将在第 4 章中详细讨论词汇的语义歧义问题以及在消除词汇语义歧义的方法。

4. 结构歧义

结构歧义（Structural ambiguity）是由词组成词组或者句子时，由于其组成的词或词组间可能存在不同的语法或语义关系而出现的（潜在）歧义现象。结构歧义有时也称为语法歧义（Grammatical ambiguity）。冯志伟教授在其论文中对结构歧义进行了系统的描述 [36]，其中一些典型的结构歧义如下：

- “VP+ 的 + 是 +NP” 型歧义结构：

  例如：反对 | 的 | 是 | 少数人

  该类型歧义中，VP 是一个双向动词，“VP+ 的”是主语，“是 +NP”是谓语，整个句式是个一个主谓结构。由于主语部分的“VP+ 的”既可以是施事，也可以是受事，因而会产生歧义。这个例子中既可以理解为“提反对意见的是少数人”，也可以理解为“所反对的是少数人”。

- “VP+N1+ 的 +N2” 型歧义结构：

  例如：咬死了 | 猎人 | 的 | 狗

  该类型歧义中，N1 作为 VP 的宾语，述宾结构“VP+N1”加上“的”之后，作为名词 N2 的定语，整个结构是一个定中结构。但是 N1 又可以与“的”结合在一起作为 N2 的定语，构成“N1+ 的 +N2”，这个名词词组作为 VP 的宾语，整个结构构成一个述宾结构。这个例子中既可以理解为“咬死了一只猎人的狗”，也可以理解为“一只把猎人咬死的狗”。

- “N1+ 和 +N2+ 的 +N3” 型歧义结构：

  例如：桌子 | 和 | 椅子 | 的 | 腿

  该类型歧义是由于连词“和”的管辖范围的不同造成的潜在歧义。这个例子中既可以理解为“桌子和（椅子的腿）”，也可以理解为“（桌子和椅子）的腿”。

类似的结构歧义类型有很多，例如：“ADJ+N1+N2”、“VP+ADJ+ 的 +N”等。这些歧义的不同理解会造成不同的句法结构以及语义上的不同。句法的分析主要难度就是解决结构歧义问题。我们将在第 3 章对结构歧义以及如何进行句法分析进行详细介绍。

5. 指代和省略歧义

在由多个句子组成的段落或篇章级别，各种歧义依然存在，例如指代歧义和省略歧义。**指代歧义**（Demonstrative ambiguity）是指代词（如我，你，他等）和代词词组（如“那件事”，“这一点”等）所指的事件可能存在歧义。

例如：猴子吃了香蕉，因为它饿了。

猴子吃了香蕉，因为它熟透了。

上述两个句子的前半句完全相同，“它”可以指代“猴子”和“香蕉”，需要具体是后半句的谓词决定指代关系。

省略歧义（Ellipsis Ambiguity）是指自然语言中由于省略所产生的歧义。省略是自然语言中的一种重要的语言现象，尤其在汉语中省略现象非常常见。省略掉一些成分，在绝大部分情况下不会影响句子的表达，但是还是存在一些由于省略造成歧义的问题。

例如：县政府同意乡政府报告。

这个例子中省略了助词“的”，因此使得该句具有两种解释，一个是县政府同意乡政府的那份报告，另外一个是县政府同意乡政府作出报告。

6. 语用歧义

语用歧义（Pragmatic ambiguity）是指由于上下文、说话人属性、场景等语用方面的原因造成的歧义。一句话在不同场合、由不同的人说、不同的语境，都可能产生不同的理解。

例如，由于场景的不同，同样的句子可以有不同的意义。

句子：你知道南京路怎么走吗？

(1) 如果说话人是游客，说话的对象是警察，那么这句话的含义就是问路。

(2) 如果说话人同样是是游客，但是说话的对象换成出租车司机，那么这句话的含义就是询问出租车司机是否可以送他到南京路。

再比如，由于上下文的不同，同样的句子也可以有不同的意义。

句子：女致电男友：地铁站见。如果你到了我还没到，你就等着吧。如果我到了你还没到，你就等着吧!!

这个例子中，同样的句子你就等着吧，前一个的含义是请耐性等待，后一个的含义是你要有麻烦了。

从上述介绍中，可以看到自然语言中存在大量的歧义现象。对人类而言，这些歧义在绝大多数的情况下都可以根据上下文以及相应的语境和场景得到解决。这也就是为什么我们平时使用自然语言交流并没有感知到语言的歧义。但是，为了消解这些歧义，需要使用了大量的知识并进行推理才能完成。而如何表示知识和使用知识、如何完整收集和整理知识以及常识都是极其困难的问题。莫拉维克悖论（Moravec’s paradox）对自然语言处理也依旧适用。也正是由于这些问题，才使得消解歧义是自然语言处理中最大的难点之一。

此外，自然语言并不是一成不变的，而是在动态发展中，存在大量未知语言现象。新词汇、新含义、新用法、新句型等层出不穷 [37]。

例如：新词汇：双碳、双减、绝绝子、社恐、元宇宙

新含义：躺平、打工人、凡尔赛、青蛙、潜水、盖楼

新用法：走召弓虽、YYDS、回忆杀、求扩列、orz

新句型：纠结的说、看书ing、一整个 无语住

这些层见迭出的语言现象对于自然语言处理系统来说也是巨大的挑战。无论是自然语言处理基础任务还是应用系统，如何应对这些未知的情况都是巨大的挑战。

总而言之，自然语言处理的困难来源于非常多的方面，即面临来自于语言本身所不避免的根本性问题，也缺乏通用的语义表示以及语言意义的理论支撑。同时，现阶段自然语言处理算法所依赖的机器学习方法，还存在需要大规模标注数据、跨领域效果差、泛化能力和鲁棒性弱、模型不可解释等诸多问题。也正因此，自然语言处理研究极具挑战，能够称得上“人工智能皇冠上的明珠”。

1.2 自然语言处理的基本范式

自然语言处理的发展经历了从理性主义到经验主义，再到深度学习三个大的历史阶段。在发展过程中也逐渐形成了一定的范式，主要包括：基于规则的方法、基于机器学习的方法以及基于深度学习的方法。这三种范式也基本对应了自然语言处理的不同发展阶段的重点。

1.2.1 基于规则的方法

基于规则的自然语言处理方法的主要思想是通过词汇、形式文法等制定的规则引入语言学知识，从而完成相应的自然语言处理任务。这类方法在自然语言处理早期受到了很大的关注，包括机器翻译在内的很多自然语言处理任务都采用此类方法。甚至目前仍有很多系统还在使用基于规则的方法完成特定任务。基于规则的方法基本处理流程如图1.3所示，主要包含：数据构建、规则构建、规则应用和效果评价等四个部分。

基于规则的方法核心是规则形式定义，其目标是使得语言学家可以在不了解计算机程序设计的情况下，能够容易的将知识转换为规则。这就要求规则描述要具有足够的灵活性并易于使用和理解。规则引擎的目标是高效的利用这些人工定义的大量规则完成特定自然语言处理任务，对输入数据从人工构建的规则库中找到符合条件的规则进行解释执行。这种方式可以使得语言学家不需要编写代码就可以通过数据和效果评价构建规则库，从而提升任务效果。

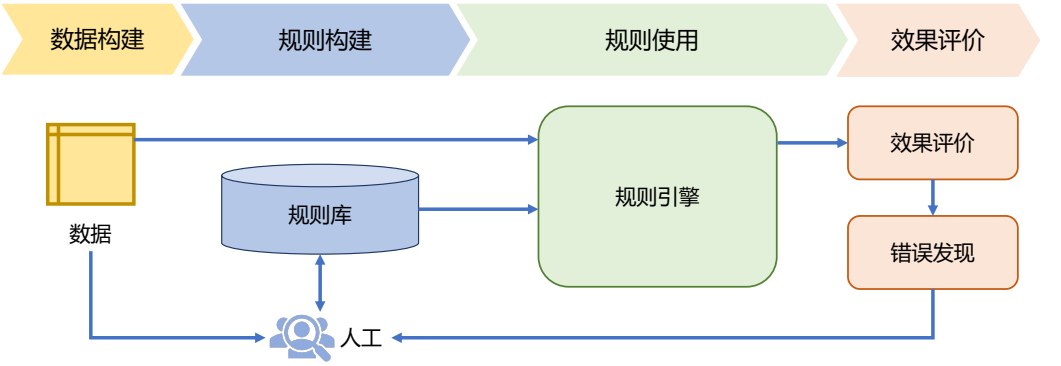

图 1.3 基于规则的自然语言处理算法基本流程

常见的规则包括产生式、框架、自动机、谓词逻辑、语义网等形式。例如，产生式规则是以“IF-THEN”形式构造，表示如果满足条件，则执行相应的语义动作。举例来说，对于机器翻译任务可以构造如下规则库：

IF 源语言主语 = 我 THEN 英语译文主语 =I

IF 英语译文主语 =I THEN 英语译文 be 动词为 am/was

IF 源语言 = 苹果 AND 没有修饰量词 THEN 英语译文 =apples

条件判断中也可以结合正则表达式，增强规则的泛化能力。再比如，可以根据英语的词典，构造有限状态自动机（Finite State Automaton，FSA）进行英语单词的拼写检查。除此此外，非确定有限状态自动机（Nondeterministic Finite Automaton，NFA）、有限状态转录机（Finite State Transducers，FST）还广泛应用于词法分析、词性标注、句法分析、机器翻译等众多方面。

基于规则的方法从某种程度上可以说是在识图模拟人类完成自然语言处理任务时的思维过程。这类方法主要优点是直观、可解释、不依赖大规模数据。利用规则所表达出来的语言知识具有一定可读性，不同的人之间可以相互理解。规则分析引擎通过规则库所得到的分析结果，也具有很好的可解释性。所使用的规则给出了系统所做出的判断的依据。规则库的构造也能够完全不依赖于大规模的有标注数据，可以仅根据人类背景知识进行构建。但是，基于规则的方法也有明显的缺点，主要包括覆盖率差、大规模规则构建代代价大、难度高。人工构建规则可以较为容易处理常见现象，但是对于复杂的语言现象难以描述。由于语言现象的复杂性，使得基于规则方法整体覆盖率很难提升到非常高的程度。并且，规则库达到一定数量之后维护困难，新增加的规则与已有规则也容易发生冲突。不同人对于同一问题的解决思路的不同，也造成了大规模规则库中规则的不一致性，从而使得维护难度进一步提高。

1.2.2 基于机器学习的方法

基于机器学习的自然语言处理算法绝大部分采用有监督分类算法，利用大规模的有标注训练语料完成模型训练。因此，其基本处理流程如图1.4所示，通常分为四个步骤：数据构建、数据预处理、特征构建以及模型学习。

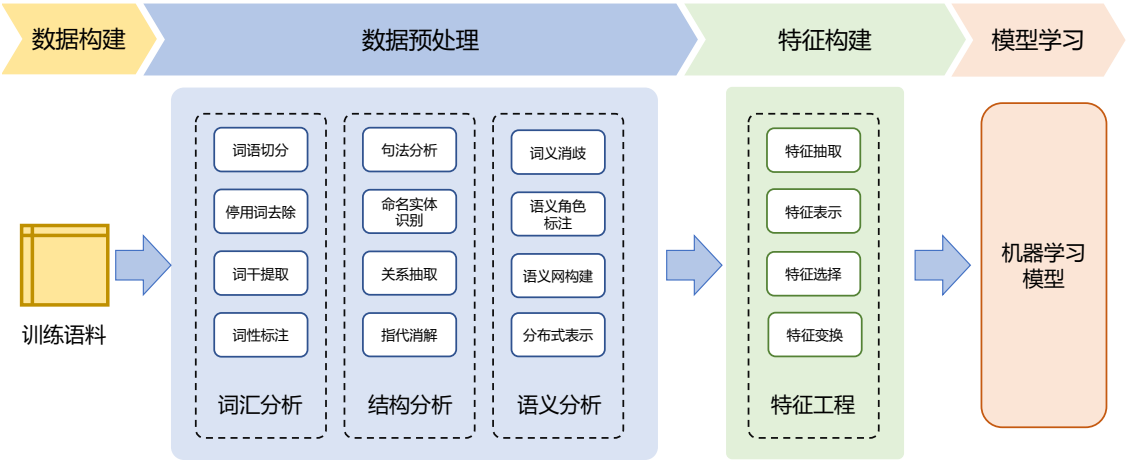

图 1.4 基于机器学习的自然语言处理算法基本流程

(1) 数据构建阶段主要工作是针对任务的要求构建训练语料，也称为语料库（Corpus）。随着自然语言处理研究的不断发展，很多任务都有公开的基准测试集合（Benchmark），可以方便的用来进行模型训练以及进行模型之间的横向对比。针对没有公开数据的任务，也可以采用人工标注的方法构建训练语料。

(2) 数据预处理阶段主要工作是利用自然语言处理基础算法对原始输入，从词汇、句法、结构、语义等层面进行处理，为特征构建提供基础。根据所处理语言和针对任务的不同，采用不同的模块和流程。对于汉语通常首选需要进行分词，对于英语通常要进行词干提取和单词的规范化。在此之后，根据特征构建的需求，还可能需要进行词性标注、句法分析、语义角色标注等。

(3) 特征构建阶段从要工作是针对不同任务从原始输入、词性标注、句法分析、语义分析等结果和数据中提取对于机器学习模型有用的特征。例如，针对属性级情感倾向分析任务，需要根据目标属性，根据句法分析结果提取该属性在对应句子中的评价词等信息。特征定义一般都是由人工完成，根据经验选取适合的特征，这项工作又被称为特征工程（Feature Engineering）。由于针对自然语言任务构建的特征通常维数非常高，又非常稀疏，因此还会利用特征选择算法降低特征维度。也可以通过特征变换，根据人工设计的准则进行有效特征提取，例如：主成分分析、线性判别分析、独立成分分析等。

(4) 模型学习阶段主要工作是根据任务，选择合适的机器学习模型，确定学习准则，采用相应的优化算法，利用语料库训练模型参数。机器学习模型有很多类型，从不同的维度可以分为：分类模型、回归模型、排序模型、生成式模型、判别式模型、有监督模型、无监督模型、半监督模型、弱监督模型等等类别。需要根据任务的目标以及特性选择适合的模型。学习准则是机器学习模型中重要的因素，包括 0-1 损失函数（0-1 Loss Function）、平方损失（Quadratic Loss Function）、交叉熵损失函数（Cross-Entropy Loss Function）、Hinge 损失函数（Hinge Loss Function）等。针对所选择的模型和学习准则需要选择相应的优化算法，包括梯度下降（Gradient Descent Method）、牛顿法（Newton method）、拟牛顿法（Quasi Newton method）、随机梯度下降（Stochastic Gradient Descent, SGD）等。机器学习三要素模型、学习准则、优化算法的选择都会对算法的效果产生影响。此外，模型中通常包含一些可以调整的超参数（Hyper-parameters），也需要通过实验和经验进行选择。

通过整体流程可以看到，基于机器学习方法的自然语言处理算法需要针对任务构建大规模训练语料，以人工特征构建为核心，针对所需的信息利用自然语言处理基础算法对原始数据进行预处理，并需要选择合适的机器学习模型，确定学习准则，以及采用相应的优化算法。整个流程中需要人工参与和选择的环节非常多，从特征设计到模型，再到优化方法以及超参数，并且这些选择通常依赖经验，缺乏有效理论支持。也使得基于机器学习的方法需要花费大量的时间和工作在特征工程上。开发一个自然语言处理算法的的主要时间消耗在数据预处理、特征构建以及模型选择和实验上。此外，对于复杂的自然语言处理任务需要在数据预处理阶段引入很多不同的模块，这些模块之间需要单独优化，其目标并不一定与任务总体目标一致，其次多模块的级联会造成错误传播，前一步错误会影响后续的模型，这些问题都提高了基于机器学习的方法实际应用的难度。

1.2.3 基于深度学习的方法

深度学习（Deep Learning）方法通过构建有一定“深度”的模型，将特征学习和预测模型融合，通过学习算法使得模型自动的学习出好的特征表示以及预测结果。基于深度学习方法的自然语言处理算法基本流程框架如图1.5所示。与传统机器学习算法的流程相比，基于深度学习方法的流程简化很多，通常仅包含数据构建、数据预处理和模型学习三个部分。同时，在数据预处理方面也大幅度简化，仅包含非常少量的模块。甚至目前很多基于深度学习的自然语言处理算法可以完全省略数据预处理阶段，对于汉语直接使用汉字做为输入，不提前进行分词，对于英语也可以省略单词的规范化步骤。

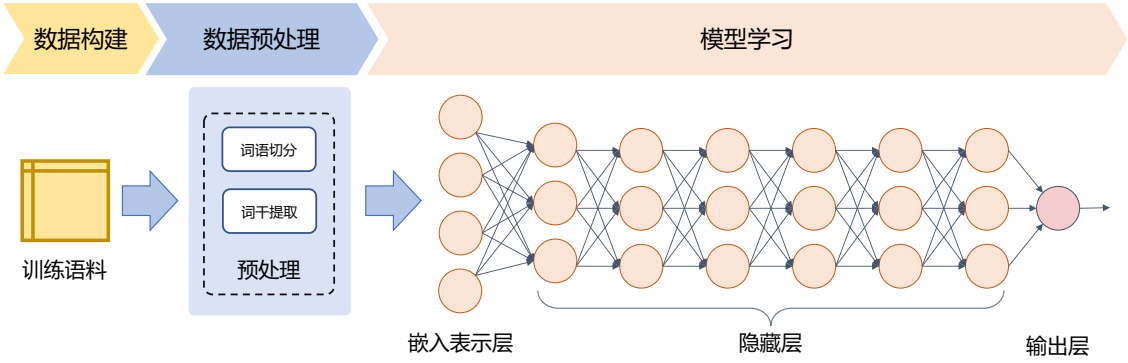

图 1.5　基于深度学习的自然语言处理算法基本流程

深度学习是机器学习的一个子集，通过多层的特征转换，将原始数据转换为更抽象的表示。这些学习到的表示可以在一定程度上完全代替人工设计的特征，这个过程也叫做**表示学习**（Representation Learning）。与基于特征工程的方法所通常采用的离散稀疏表示不同，深度学习算法通常使用分布式表示（Distributed Representation），特征表示为低维稠密向量。分布式表示通常需要从底层特征开始，经过多次非线性转换得到。由于深层结构可以增加特征的重用性，从而使得表示能力指数级增加。因此，表示学习的关键是构建具有一定深度的多层次特征表示 [38]。随着深度学习研究不断深入和计算能力的快速发展，模型深度也从早期的 5 到 10 层增加到现在的数百层。随着模型深度的不断增加，其特征表示能力也不断增强，从而也使得深度学习模型中的预测部分更加简单，预测也更加容易。

自 2018 年 ELMo 模型 [27] 提出之后，基于深度学习的自然语言处理范式又进一步演进为预训练微调范式。首先利用自监督任务对模型进行预训练，通过海量的语料学习到更为通用的语言表示，然后根据下游任务对预训练网络进行调整。这种预训练范式在几乎所有自然语言处理任务上都表现非常出色。预训练模型在模型网络结构上可以采用 LSTM、Transformer 等具有较好序列

建模能力的模型，预训练任务可以采用语言模型、掩码语言模型（Masked Language Model）、机器翻译等自监督或有监督方式，还可以引入知识图谱、多语言、多模态等扩展任务。自 2018 年以来有非常多的相关研究，也仍然面临模型鲁棒性提升、模型可解释性等诸多问题亟待解决。第**??**节将对预训练模型进行详细介绍。

需要特别说明的是，虽然以上三种范式来源于自然语言处理的不同发展阶段，有明显的发展先后顺序，并且在大部分自然语言处理任务的标准评测集合中基于深度学习的方法都好于基于机器学习的方法，更优于基于规则的方法很多。但是，这三种范式各有利弊，在实际应用中需要根据任务的特点、计算量、可控制性以及可解释性等具体情况进行选择。

1.3 本书的内容安排

本书共分为 14 章，主要包含三个部分：第一部分主要介绍自然语言处理的基础技术，包括词汇处理、句法分析、语义分析、篇章分析和语言模型；第二部分主要介绍自然语言处理的一系列核心技术，包括信息抽取、机器翻译、情感分析、文本摘要、知识图谱；第三部分主要介绍基于机器学习的自然语言处理模型的鲁棒性和可解释性问题。本书章节安排如图1.6所示。

图 1.6 本书章节安排

第 2 章到第 6 章将从词汇、句法到篇章三个不同粒度的语言单位，从形态、结构到语义三个不同语言层面，对自然语言处理的基础技术进行介绍。第 2 章主要介绍语言学中词相关的基本概念，以及词语规范化、中文分词、词性分析等词汇分析主要任务和相关算法。第 3 章主要介绍语言

学中句法基本概念，以及成分句法分析算法、依存句法分析算法。第 4 章主要介绍语义学和语义表示的基本概念和语义和知识的表示方法，以及词义消歧、语义角色标注等语义分析主要任务和相关算法。第 5 章主要介绍篇章结构基础理论和基本概念，以及话语分割、话语分析和指代消解等篇章分析的主要任务和相关算法。第 6 章主要介绍语言模型基本概念，以及 N 元语言模型、神经语言模型以及预训练语言模型的常见算法。

第 7 章到第 12 章主要介绍自然语言处理支撑各种应用的核心技术。第 7 章主要介绍信息抽取的基本任务和相关算法，包括命名实体识别、关系抽取和事件抽取。第 8 章主要介绍机器翻译的基本概念和常见方法，包括基于统计和基于神经网络的机器翻译方法。第 9 章主要介绍情感倾向分析基本概念和主要任务，包括文档、句子、属性三个不同粒度的分析算法。第 10 章主要介绍智能问答基本任务和分析算法，包括阅读理解、表格问答、社区问题、开发问答等。第 11 章主要介绍文本摘要的相关任务和基本算法，包括生成式文本摘要、抽取式文本摘要等。第 12 章主要介绍知识图谱相关概念和基本任务，包括知识表示学习、知识图谱构建和知识图谱应用。

第 13 章和第 14 章将针对基于机器学习模型的自然语言处理算法所面临的模型鲁棒性问题和可解释性问题进行讨论。第 13 章主要介绍自然语言处理模型鲁棒性的基本概念，以及文本攻击方法、文本防御方法以及模型鲁棒性评价基准。第 14 章主要介绍自然语言处理模型的可解释性问题，主要包括解释性分析工具和可解释自然语言处理。

此外，还需说明特别强调的是自然语言处理中很多任务都转换为了机器学习问题，因此很多机器学习算法可以应用于多个自然语言处理任务。比如，条件随机场模型（Conditional Random Fields，简称 CRF）可用于中文分词，也可以用于词性标注，还可以用于命名实体识别。为了避免重复，我们仅在第**??**节详细介绍了如何使用线性链条件随机场模型进行中文分词，在词性标注、命名实体识别等章节选择了不同的算法进行介绍。需要读者朋友能够融会贯通，在本书学习结束时对特定机器学习模型可以适用于哪些自然语言处理任务有清晰的了解。

参考文献

[1] 吴立德. 大规模中文文本处理[M]. 复旦大学出版社, 1997.

[2] Kupiec J. Augmenting a hidden markov model for phrase-dependent word tagging[C]//Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989. 1989.

[3] Oshika B, Machi F, Evans B, et al. Computational techniques for improved name search[C]//Second Conference on Applied Natural Language Processing. 1988: 203-210.

[4] Sahami M. Learning limited dependence bayesian classifiers.[C]//KDD: volume 96. 1996: 335-338.

[5] Yang Y. Feature selection in statistical learning of text categorization[C]//Proc. 14th International Conference on Machine Learning. 1997: 412-420.

[6] Vapnik V. The nature of statistical learning theory[M]. Springer science & business media, 1999.

[7] Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification[C]//IJCAI-99 workshop on machine learning for information filtering: volume 1. Stockholom, Sweden, 1999: 61-67.

[8] Nakamura M, Maruyama K, Kawabata T, et al. Neural network approach to word category prediction for english texts[C]//COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics. 1990.

[9] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001. 2001: 282-289.

[10] Collins M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms[C]//Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002). 2002: 1-8.

[11] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786):504-507.

[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[13] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(ARTICLE):2493-2537.

[14] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.

[15] Nallapati R, Zhou B, dos Santos C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond[J]. CoNLL 2016, 2016:280.

[16] Qiu M, Li F L, Wang S, et al. Alime chat: A sequence to sequence and rerank based chatbot engine[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017: 498-503.

[17] Lei W, Jin X, Kan M Y, et al. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1437-1447.

[18] Konstas I, Iyer S, Yatskar M, et al. Neural amr: Sequence-to-sequence models for parsing and generation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 146-157.

[19] Yin J, Jiang X, Lu Z, et al. Neural generative question answering[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016: 2972-2978.

[20] Shi X, Huang H, Jian P, et al. Neural chinese word segmentation as sequence to sequence translation [C]//Chinese National Conference on Social Media Processing. Springer, 2017: 91-103.

[21] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model.[C]// Interspeech: volume 2. Makuhari, 2010: 1045-1048.

[22] Sundermeyer M, Schlüter R, Ney H. Lstm neural networks for language modeling[C]//Thirteenth annual conference of the international speech communication association. 2012.

[23] Irsoy O, Cardie C. Deep recursive neural networks for compositionality in language[J]. Advances in neural information processing systems, 2014, 27.

[24] Johnson R, Zhang T. Effective use of word order for text categorization with convolutional neural networks[C/OL]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015: 103-112. https://aclanthology.org/N15-1011. DOI: 10.3115/v1/N15-1011.

[25] Velivcković P, Cucurull G, Casanova A, et al. Graph attention networks[C]//International Conference on Learning Representations. 2018.

[26] Gui T, Zou Y, Zhang Q, et al. A lexicon-based graph neural network for chinese ner[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 1040-1050.

[27] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers): volume 1. 2018: 2227-2237.

[28] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[29] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8):9.

[30] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.

[31] Zhang Z, Han X, Liu Z, et al. Ernie: Enhanced language representation with informative entities[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1441-1451.

[32] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.

[33] Kess J F, Hoppe R A. Ambiguity in psycholinguistics[M]. Benjamins Amsterdam, 1981.

[34] 中国社会科学院语言研究所词典编辑室. 现代汉语词典（第 7 版）[M]. 商务印书馆, 2019.

[35] Miller G A. Wordnet: a lexical database for english[J]. Communications of the ACM, 1995, 38(11): 39-41.

[36] 冯志伟. 论歧义结构的潜在性[J]. 中文信息学报, 1995, 9(4):14-24.

[37] 宗成庆. 统计自然语言处理[M]. 清华大学出版社, 2013.

[38] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8):1798-1828.

索引