高等院校计算机类规划教材
国家新闻出版改革发展项目库入库项目
数据科学与大数据技术专业教材丛书

自然语言处理

NATURAL LANGUAGE PROCESSING

双 锴◎编著

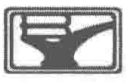
非外借

北京邮电大学出版社
www.buptpress.com

高等院校计算机类规划教材
国家新闻出版改革发展项目库入库项目
数据科学与大数据技术专业教材丛书

自然语言处理

双 锴 编著



北京邮电大学出版社
www.buptpress.com

内 容 简 介

本书着眼于自然语言处理的一些经典研究和前沿应用，重点介绍了深度学习在自然语言处理中的应用。全书共分为11章。第1章概述了自然语言处理的发展过程、难点及应用等。第2～10章从自然语言处理中的基本概念和基础知识出发，逐步介绍了语言模型、分类任务、信息抽取、知识图谱、机器翻译、摘要生成、语言分析这几种典型的基础型和应用型研究任务的发展、算法原理和模型结构以及未来趋势。第11章前瞻性地对时下热门的研究方向进行了分析和讨论。作者对于全书结构和内容都有精心设计，既涵盖科普类知识，避免复杂的公式堆叠，用通俗直白的语言讲解算法的设计思想，又配合有大量的技术性介绍和分析，包括如何利用主流的深度学习框架进行复现，实现了算法和应用的合理结合。

本书的读者对象既包括对该领域不甚了解的普通大众，也包括高校相关专业在校学生以及从事相关领域研究的技术人员。

图书在版编目(CIP)数据

自然语言处理 / 双锴编著. -- 北京：北京邮电大学出版社，2021.8
ISBN 978-7-5635-6385-2

Ⅰ. ①自… Ⅱ. ①双… Ⅲ. ①自然语言处理 Ⅳ. ①TP391

中国版本图书馆CIP数据核字(2021)第104311号

策划编辑：姚 顺 刘纳新 **责任编辑**：刘 颖 **封面设计**：七星博纳

出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路10号
邮政编码：100876
发 行 部：电话：010-62282185 传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：唐山玺诚印务有限公司
开　　本：787 mm×1 092 mm 1/16
印　　张：15.5
字　　数：384千字
版　　次：2021年8月第1版
印　　次：2021年8月第1次印刷

ISBN 978-7-5635-6385-2 **定　　价**：39.00元

大数据顾问委员会

大数据专业教材编委会

前　言

这是一本关于使用深度学习方法解决自然语言处理任务的教材，书中涵盖与自然语言处理相关的多项任务，内容新颖，紧跟技术潮流，几乎覆盖了该领域近几年的全部研究方向。笔者尽量避免使用长篇大论的公式推导和晦涩难懂的理论解释来向读者阐明观点，相反地，在编写时尽量选择通俗易懂的语言并结合生动的应用场景来进行描述。希望读者在阅读过程中不会产生过多理解障碍，最终可以对现代自然语言处理任务形成一个宏观的认识。这是一本既适合高年级本科生又适用于有一定研究基础的研究生入门自然语言处理领域的教材，同时该书也可以供相关领域的工程技术人员参阅。

作为自然语言处理领域普及类教材，本书从自然语言处理基础到深度学习中的应用，再到其他研究热点的介绍及发展趋势的展望，对自然语言处理领域进行了详细且系统的概括和介绍，有助于读者对该领域的研究脉络进行总结和梳理。全书共分为 11 章。第 1 章概述了自然语言处理的发展过程、难点及应用等。第 2～10 章从自然语言处理中的基本概念和基础知识出发，逐步介绍了语言模型、分类任务、信息抽取、知识图谱、机器翻译、摘要生成、语言分析这几种典型的基础型和应用型研究任务的发展、算法原理和模型结构以及未来趋势。第 11 章前瞻性地对时下热门的研究方向进行了分析和讨论。

书中每个章节都提供了一些精选思考题。这些题目不仅可以帮助读者回顾本章的基础知识，更重要的是引导读者对本章的重点内容进行回顾和思考。对于一般课程，这些思考题的深度足以支持。为了给学有余力的读者进一步提升的空间，建议在授课时辅以编程大作业，把知识融会贯通到实践中有利于加深对知识点的理解。

为了保证本书涵盖尽可能广的自然语言处理中的各项任务，同时在讲解相关理论和技术细节时足够清晰直白，本书在编撰时有详有略，因此存在一些无法详细阐述的技术细节。此外，和深度学习相关的自然语言处理研究结论日新月异，书中的相关说法和对应内容很难保证在出版时仍保持业界最新，想要做到知识同步还需要读者及时查阅资料。为了启发读者，本书提供了一些经典或更为专业的扩展阅读材料，便于读者深入了解，读者可扫描相关二维码阅读。书中涉及的英文专业术语，考虑到读者阅读的流畅性将其中一部分直译为中文术语，但绝大多数保持其原始英文表述，一是由于意译的结果不统一易引起歧义，二是有意使读者提前熟悉这些文献中的高频词。

随着人工智能的飞速发展，自然语言处理也成为研究热点，大量的学术成果和工业应用推动着自然语言处理的进步。虽然身为自然语言处理的研究者，但笔者的时间和精力有限，可能对于自然语言处理的一些问题的理解也不够深入，甚至有些地方可能存在出入，因此书中难免有谬误之处，希望读者指出后一起讨论。

双　锴
于北京邮电大学

目 录

第1章

绪　论

本章思维导图

语言是人类区别于其他动物的本质特性，人类的多种智能都与语言有着密切的关系。人类的逻辑思维以语言为形式传达，人类的绝大部分知识也是通过语言文字的形式进行记载从而流传下来。因而，用自然语言与计算机进行通信，具有明显的理论意义和实际价值，必然成为人工智能的一个核心发展方向。

自然语言处理经历过怎样的发展历程，现在面临什么样的瓶颈？从研究内容和应用角度具体落地为哪些任务？现在最热门的深度学习技术是如何运用在自然语言处理领域的？本章将对这些问题进行简要的概括介绍。

图 1-1 为本章的思维导图，是对本章的知识脉络的总结。

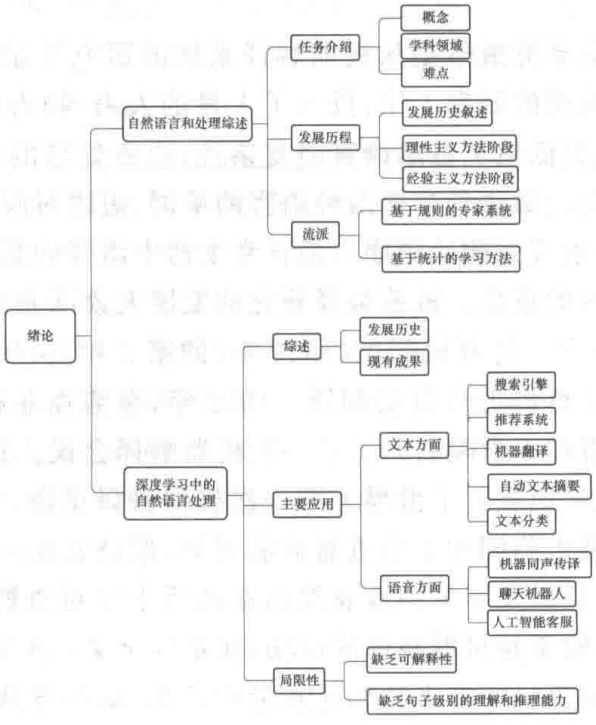


图 1-1　绪论思维导图

1.1 自然语言处理发展

1.1.1 什么是自然语言处理?

(1) 自然语言处理的概念

自然语言处理(Natural Language Processing,NLP)是人工智能和语言学交叉领域下的分支学科。该领域主要探讨如何处理及运用自然语言、自然语言认知(即让计算机"懂"人类的语言)、自然语言生成系统(将计算机数据转化为自然语言),以及自然语言理解系统(将自然语言转化为计算机程序更易于处理的形式)。

所谓"自然语言",其实就是我们日常生活中使用的语言(在这里还包括书面文字和语音视频等),人们熟知的汉语、日语、韩语、英语、法语等语言都属于此范畴。至于"自然语言处理",则是对自然语言进行数字化处理的一种技术,是通过语音文字等形式与计算机进行通信,从而实现"人机交互"的技术。

(2) 自然语言处理的学科领域

自然语言处理是一门多学科交叉的技术,其中包括语言学、计算机科学(提供模型表示、算法设计、计算机实现)、数学(数学模型)、心理学(人类言语心理模型和理论)、哲学(提供人类思维和语言的更深层次理论)、统计学(提供样本数据的预测统计技术)、电子工程(信息论基础和语言信号处理技术)、生物学(人类言语行为机制理论)。

1.1.2 自然语言处理的发展历史

自然语言处理的相关研究最早是从机器翻译系统的研究开始的。20 世纪 60 年代,国外对机器翻译曾有过大规模的研究工作,投入了大量的人力、物力和财力。但是,受到客观历史因素的限制,当时人们低估了自然语言的复杂性,语言处理的理论和技术均不成熟,所以进展并不大。当时主要的做法是存储两种语言的单词、短语对应译法的大词典,翻译时一一对应,技术上只是调整语言的前后顺序。但日常生活中语言的翻译远不是如此简单,很多时候还要参考某句话前后的意思。机器翻译研究的发展大致可分为三个时期。

初创期(1947—1970 年):计算机问世(1946 年)的第二年,英国工程师布斯和美国工程师威弗最早提出了利用计算机进行自动翻译。1952 年,在洛克菲勒基金会的大力支持下,一些英美学者在美国麻省理工学院召开了第一次机器翻译会议。1954 年,《机器翻译》杂志开始公开发行。同年,成功地进行了世界上第一次机器翻译试验。尽管这次试验用的机器词汇仅仅包含了 250 个俄语单词和 6 条机器语法规则,但是它第一次向公众和科学界展示了机器翻译的可行性,并且激发了美国政府部门在随后十年对机器翻译进行大量资助。随着研究的深入,人们看到的不是机器翻译的成功,而是一个又一个它无法克服的局限。第一代机器翻译系统设计上的粗糙所带来的翻译质量的低劣,最终导致了一些人对机器翻译的研究失去信心。有些人甚至错误地认为机器翻译追求的全自动质量目标是不可能实现的。

机器翻译的研究就此陷入低谷。

复苏期(1971—1976年):尽管机器翻译的研究困难重重,但是法国、日本、加拿大等国仍然坚持机器翻译的研究。在20世纪70年代初期,机器翻译又出现了复苏的局面。在这个时期,研究者们普遍认识到,原语和译语两种语言的差异不仅表现在词汇上,而且表现在句法结构上,为了得到可读性强的译文,必须在自动句法分析上多下功夫。通过大量的科学实验,机器翻译的研究者逐渐认识到机器翻译过程本身必须保持原语和译语在语义上的一致,一个好的机器翻译系统应该把原语的语义准确无误地在译语中表现出来。于是,语义分析在机器翻译中越来越受到重视。美国斯坦福大学的威尔克斯提出了"优选语义学",并在此基础上设计了英法机器翻译系统。这个系统的语义表示方法比较细致,能够解决仅仅用句法分析难以解决的歧义现象、代词所指等困难问题,译文质量较高,受到专家学者的一致肯定。

繁荣期(1977年至今):繁荣期最突出的特点是机器翻译研究走上了实用化的道路。在这段时期,出现了一大批实用化的机器翻译系统,机器翻译产品开始进入市场,逐渐由实用化步入商业化。其中,第二代机器翻译系统以基于转换的方法为代表,普遍采用以句法分析为主、语义分析为辅的基于规则的方法和由抽象的转换表示的分层次实现策略。比如加拿大蒙特利尔大学开发研制的实用性机器翻译系统TAUM-METEO就采用了典型的转换方法,整个翻译过程分为5个阶段(英-法翻译):英语形态分析、英语句法分析、转换、法语句法生成和法语形态生成。这个翻译系统投入使用之后,每小时可以翻译6万～30万个词,每天可以翻译1 500～2 000篇天气预报的资料,并能够通过电视、报纸立即公布。TAUM-METEO系统是机器翻译发展史上的一个里程碑,它标志着机器翻译由复苏走向了繁荣。

我国机器翻译的起步并不算太晚,是继美国、苏联、英国之后世界上第四个开展机器翻译研究的国家。早在20世纪50年代机器翻译就被列入我国科学研究的发展规划。一些研究人员还进行了俄汉机器翻译实验,取得了一定的研究成果。我国机器翻译研究的全面开展始于20世纪80年代中期,特别是20世纪90年代以来,一批机器翻译系统相继问世,其中影响力较大的有:中软总公司开发的汉日翻译系统(1993年),中科院计算所研制的IMTEC英汉翻译系统(1992年)等。

在自然语言处理的形成和发展进程中,除机器翻译外,自然语言理解所起到的作用也是不可忽视的。自然语言理解的发展始于20世纪60年代中期机器翻译处于举步维艰之时,到了20世纪70年代初,它的研究已获得了累累硕果。自然语言理解又称"人机对话",就是"让计算机理解自然语言,使计算机获得人类理解自然语言的智能,并对人给计算机提出的问题,通过对话的方式,用自然语言进行回答"。20世纪60年代中期,人们开始由"词对词"的翻译方式逐步转入对自然语言的语法、语义和语用等基本问题的研究,并尝试让计算机来理解自然语言。许多学者认为,判断计算机是否理解了自然语言的最直观方法,就是让人类同计算机对话,如果计算机对人提出的问题能做出有意义的回答,那就证明计算机已经理解了自然语言。最初的"人机对话"系统(或"自然语言理解"系统)的研究工作主要在美国。一般来讲,第一代自然语言理解系统可以分为四种类型:

① 特殊格式系统。根据人机对话内容的特点,采用特定的格式来进行人机对话。

② 以文本为基础的系统。某些研究者不满意在特殊格式系统中种种格式的限制,因为就一个专门领域来说,最方便的还是使用不受特殊格式结构限制的系统来进行人机对话。