人工智能前沿技术丛书

Natural Language Processing

A Large Language Model Approach

自然语言处理

基于大语言模型的方法

车万翔　郭　江　崔一鸣◎著

刘　挺◎主审

電子工業出版社

Publishing House of Electronics Industry

北京•BEIJING

内 容 简 介

自然语言处理被誉为“人工智能皇冠上的明珠”。深度学习等技术的引入为自然语言处理技术带来了一场革命，尤其是近年来出现的基于大语言模型的方法，已成为研究自然语言处理的新范式。本书在介绍自然语言处理、深度学习等基本概念的基础上，重点介绍新的基于预训练语言模型和大语言模型的自然语言处理技术。本书包括基础知识、预训练语言模型和大语言模型三部分：基础知识部分主要介绍自然语言处理和深度学习的基础知识、基本工具集和常用数据集；预训练语言模型部分主要介绍语言模型、预训练词向量、预训练语言模型的实现方法和应用；大语言模型部分首先介绍大语言模型的预训练方法，其次介绍大语言模型的适配、应用和评估方法，接着介绍基于预训练语言模型思想的各种延伸技术，最后以 DeepSeek 系列模型为例，介绍大语言模型的最新技术进展。除了理论知识，本书还有针对性地结合具体案例提供相应的 PyTorch 代码实现，让读者不仅能对理论有更深刻的理解，还能快速地实现自然语言处理模型，达到理论和实践的统一。

本书既适合具有一定机器学习基础的高等院校学生、研究机构的研究者，以及希望深入研究自然语言处理算法的计算机工程师阅读，也适合对人工智能、深度学习、大语言模型和自然语言处理感兴趣的学生和希望进入人工智能应用领域的研究者参考。

图书在版编目（CIP）数据

自然语言处理：基于大语言模型的方法 / 车万翔，
郭江，崔一鸣著. --北京：电子工业出版社，2025. 3.
（人工智能前沿技术丛书）. --ISBN 978-7-121-49598-4

Ⅰ. TP391

中国国家版本馆 CIP 数据核字第 20251UG930 号

责任编辑：宋亚东
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：720×1000　1/16　　印张：27.5　　字数：600 千字
版　　次：2025 年 3 月第 1 版
印　　次：2025 年 3 月第 1 次印刷
定　　价：158.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：syd@phei.com.cn。

推荐序

FOREWORD

自然语言处理的目标是使机器具有和人类一样理解与生成语言的能力。自然语言处理技术经历了从理性主义到经验主义的嬗变。经过最近十年左右的发展，自然语言处理在深度学习的框架下迅速演进为基于预训练语言模型的方法。尤其是 2022 年底以来，以 ChatGPT 为里程碑式标志的一系列大型语言模型竞相问世，展现了强大的语言理解、生成和知识推理能力，彻底颠覆了自然语言处理领域的格局，成为自然语言处理乃至整个人工智能领域新的统一范式。

车万翔教授领衔撰写的《自然语言处理：基于大语言模型的方法》一书，以他们 2021 年编写的《自然语言处理：基于预训练模型的方法》为基底，在预训练模型的基础上，融入大量关于最新大语言模型的深入内容，旨在帮助读者深入理解这些技术背后的原理、相互之间的联系及存在的局限性，对于当前学术界和工业界的相关研究与应用均具有重要价值。

本书包括三部分，共 13 章，从自然语言处理与神经网络的基础知识讲起，沿着预训练语言模型的发展轨迹，系统性地探讨了语言模型、预训练词向量和预训练语言模型等方法，继而深入介绍了大语言模型的预训练、适配、应用、评估等关键技术环节。

本书作者长期致力于自然语言处理，尤其是预训练语言模型及大语言模型方法的研究工作，取得了一系列突出的科研成果。本书正是他们多年深入耕耘该领域的成果体现。

本书的鲜明特色之一是含有丰富的实践内容。作者均为活跃在科研一线的青年学者，具有丰富的实战经验。本书针对代表性的模型提供了规范的示例代码和实践指导，是一份宝贵的学习资源，尤其适合那些刚刚迈入自然语言处理领域并热衷于实践与应用的读者学习。

本书既适合计算机科学、人工智能和机器学习专业的学生、研究者及人工智能应用开发者阅读，也适合对大语言模型感兴趣的高校教师和研究机构的研究人员参考。

孙茂松

欧洲科学院外籍院士

清华大学人工智能研究院常务副院长、计算机系教授

推荐语 FOREWORD

大语言模型的训练和推理离不开强大的高性能计算的支持。《自然语言处理：基于大语言模型的方法》一书站在技术前沿，不仅深入探讨了大语言模型的发展历程、核心技术及未来趋势，还详细介绍了如何利用高性能计算系统优化模型训练和部署。通过对本书的学习，读者可以更清晰地把握大语言模型的工作原理，并为未来的研究和应用提供坚实的理论基础。无论是希望深入了解大语言模型的初学者，还是希望将大语言模型应用于实际场景的工程师，本书都将是一本不可或缺的指南。向大家推荐！

廖湘科

中国工程院院士

国防科学技术大学计算机学院教授

近年来，以 ChatGPT 为代表的大语言模型技术迅速崛起，展现出卓越的语言理解、生成及知识推理能力。这些模型能够精准地把握用户意图，实现高效的多轮对话，其回答内容翔实、重点明确，具备高度的概括性、逻辑性和条理性。《自然语言处理：基于大语言模型的方法》一书深入浅出地阐述了大语言模型的技术原理和实现方式，并全面和深入地分析了其发展方向。本书是人工智能领域的学习者和研发人员在短时间内学习、掌握关键技术并快速应用的理想选择。特此推荐！

尼玛扎西

中国工程院院士

西藏大学信息科学技术学院教授

车万翔教授常年聚焦于自然语言处理研究，对该领域具有深刻和独到的见解，研发的语言技术平台（LTP）已成为自然语言处理领域具有广泛影响力的基础技术平台。我与车万翔教授相识多年，常常向他请教人工智能领域的相关问题。他是自然语言处理研究领域不可多得的青年才俊，他在智慧流体力学年度交流会上做的学术报告对我们的研究启发很大。最近喜闻他与合作者即将出版《自然语言处理：基于大语言模型的方法》一书，并先睹了大作初稿。该书既包含大语言模型的基础知识，更包含丰富的实践内容，集成了作者多年研究与实践成果，独具特色。我主要从事 AI for Science 的研

究，大语言模型是 AI for Science 研究的重要工具，为科研人员提供了新的知识获取方式和解决问题的途径。本书不仅能帮助研究人员快速掌握大语言模型的相关技术，还为进一步应用大语言模型解决各类科学问题提供了重要参考。无论是对计算机专业还是其他学科的研究人员而言，本书都是一份不可多得的学习资料。积极推荐大家阅读！

李惠

中国科学院院士

哈尔滨工业大学土木学院/计算学部教授

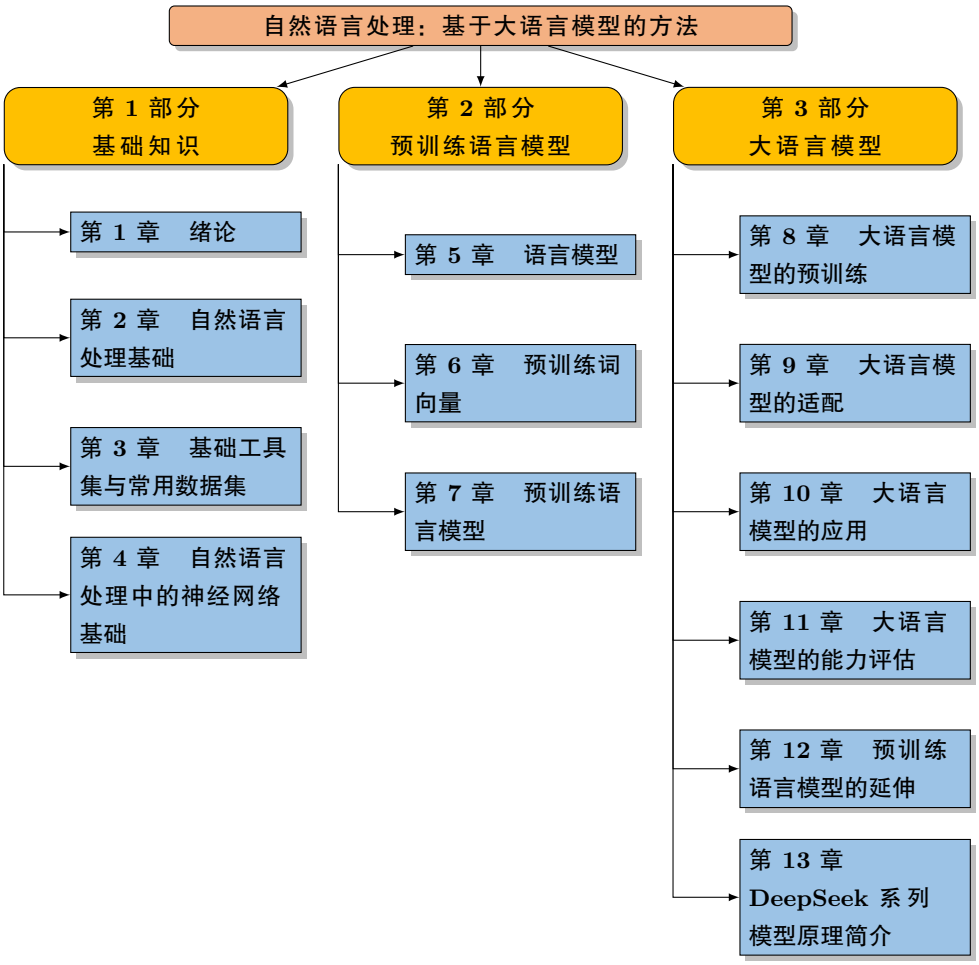
前言 PREFACE

自然语言是人类思维的载体和交流的基本工具，也是人类区别于动物的根本标志，更是人类智能发展的重要外在体现形式。自然语言处理（Natural Language Processing，NLP）主要研究用计算机理解和生成自然语言的各种理论与方法，属于人工智能领域的一个重要的甚至核心的分支。随着互联网的快速发展，网络文本规模呈爆炸性增长，对自然语言处理提出了巨大的应用需求。同时，自然语言处理研究也为人们更深刻地理解语言的机理和社会的机制提供了一条重要的途径，因此具有重要的科学意义。

自然语言处理技术经历了从早期的理性主义到后来的经验主义的转变。近十年来，深度学习技术快速发展，引发了自然语言处理领域的一系列变革。但是基于深度学习的算法有一个严重的缺点，就是过度依赖大规模的有标注数据。2018 年以来，以 BERT、GPT 为代表的预训练语言模型恰好弥补了自然语言处理标注数据不足的缺点，帮助自然语言处理取得了一系列的突破，包括阅读理解在内的众多自然语言处理任务的性能都得到了大幅提高，在有些数据集上甚至达到或超过了人类水平。2022 年底，OpenAI 推出的大语言模型 ChatGPT，以其强大的语言理解、生成及知识推理能力，彻底颠覆了自然语言处理领域的格局，成为自然语言处理乃至整个人工智能领域的统一范式。那么，预训练语言模型以及后来的大语言模型是如何获得如此强大的威力甚至“魔力”的呢？希望本书能够为各位读者揭开大语言模型的神秘面纱。

本书主要内容

本书在《自然语言处理：基于预训练模型的方法》（电子工业出版社，2021）一书的基础上，针对近期自然语言处理领域，尤其是大语言模型方面技术与应用的最新进展，进行了全面的修订和补充。本书主要内容包括三部分：基础知识、预训练语言模型和大语言模型。各部分内容安排如下。



第 1 部分：基础知识，包括第 1～4 章，主要介绍自然语言处理和深度学习的基础知识、基本工具集和常用数据集。

第 2 章首先介绍文本的向量表示方法，重点介绍词嵌入表示。其次介绍自然语言处理的三大任务，包括语言模型、基础任务和应用任务。虽然这些任务看似纷繁复杂，但是基本可以归纳为三类问题，即文本分类问题、结构预测问题和序列到序列问题。最后介绍自然语言处理任务的评价方法。

第 3 章首先介绍三种常用的自然语言处理基础工具集——tiktoken、NLTK 和 LTP。其次介绍本书使用的深度学习框架 PyTorch。最后介绍自然语言处理中常用的大规模预训练数据。

第 4 章首先介绍自然语言处理中常用的四种神经网络模型：多层感知器模型、卷积神经网络、循环神经网络和以 Transformer 为代表的自注意力模型。其次介绍模型的参数优化方法。最后通过两个综合性的实战项目，介绍如何使用深度学习模型解决一个实际的自然语言处理问题。

第 2 部分：预训练语言模型，包括第 5 ~ 7 章，主要介绍语言模型、预训练词向量以及预训练语言模型的实现方法及应用。

第 5 章首先介绍语言模型的基本概念，其次介绍经典的 N 元语言模型及现代的神经网络语言模型的概念和实现方法，最后介绍语言模型的评价方法。

第 6 章介绍词向量的基本概念，以及静态词向量和动态词向量两类预训练词向量的方法及其在自然语言处理任务中的应用。

第 7 章首先介绍基于大规模文本预训练的语言模型，其次重点介绍预训练语言模型的三种基本结构及代表性的预训练语言模型，最后介绍预训练语言模型的应用场景和方法。

第 3 部分：大语言模型，包括第 8 ~ 13 章，首先介绍大语言模型的预训练方法，其次介绍大语言模型的适配、应用及评估方法，最后介绍基于预训练语言模型思想的各种延伸技术。

第 8 章首先以几种经典的开源大语言模型为例，介绍大语言模型的两种基本结构，其次介绍大语言模型预训练过程中的若干关键技术，最后介绍大语言模型的并行训练策略。

第 9 章介绍在将大语言模型应用于具体的现实任务或领域时所需的适配技术，包括基于提示的推断、多任务指令微调、基于人类反馈的强化学习、典型的参数高效精调方法、模型压缩方法，以及大语言模型的中文适配方法等。

第 10 章介绍如何将大语言模型有效应用于各种应用场景，包括在常见任务中的应用方法、利用大语言模型生成指令数据以用于大语言模型的精调、大语言模型的量化与部署、本地化开发与应用、利用大语言模型进行工具调用及实现自动化等方法。

第 11 章介绍大语言模型的能力评估方法，包括通用领域及任务评估、特定领域及任务评估、模型对齐能力评估、大语言模型的评价方法等。

第 12 章介绍预训练语言模型的延伸技术，包括多语言的预训练模型及其在跨语言任务上的应用、代码预训练模型、多模态预训练模型，以及基于大语言模型实现的具身预训练模型。

第 13 章以 DeepSeek 系列模型为例，介绍大语言模型的最新技术进展，包括 DeepSeek 系列模型的技术原理、模型架构优化和基于强化学习获得的推理能力学习等。

致谢

本书第 1 ~ 5 章及第 12 章由哈尔滨工业大学车万翔教授编写；第 6、11 章由美国麻省理工学院（MIT）郭江博士后编写；第 7、8 、10 章由科大讯飞北京研究院副院长崔一鸣编写；第 9 章及第 13 章由三位作者联合编写。全书由哈尔滨工业大学刘挺教授主审。

本书的编写参阅了大量的著作和相关文献，在此一并表示衷心的感谢！

感谢宋亚东先生和电子工业出版社博文视点对本书的重视，以及为本书出版所做的一切。

由于作者水平有限，书中不足及错误之处在所难免，敬请专家和读者给予批评指正。

车万翔

2025 年 2 月