人工智能前沿技术丛书

Natural Language Processing
A Pre-trained Model Approach

# 自然语言处理

## 基于预训练模型的方法

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预训练模型的方法，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模型的自然语言处理技术。本书包括基础知识、预训练词向量和预训练模型三部分：基础知识部分介绍自然语言处理和深度学习的基础知识和基本工具；预训练词向量部分介绍静态词向量和动态词向量的预训练方法及应用；预训练模型部分介绍几种典型的预训练语言模型及应用，以及预训练模型的最新进展。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，不仅能让读者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**

ISBN 978-7-121-41512-8

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　　印张：20　　字数：422 千字
版　　次：2021 年 7 月第 1 版
印　　次：2021 年 7 月第 1 次印刷
定　　价：118.00 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：（010）51260888-819，syd@phei.com.cn。

# 推荐序

FOREWORD

自然语言处理的目标是使得机器具有和人类一样的语言理解与运用能力。在过去的十年里，自然语言处理经历了两次里程碑式的重要发展。第一次是深度学习的勃兴，使得传统的特征工程方法被摒弃，而基于深度神经网络的表示学习迅速成为自然语言处理的主流。第二次则是 2018 年以来大规模预训练语言模型的应用，开启了基于“预训练 + 精调”的新一代自然语言处理范式。每一次的发展都为自然语言处理系统的能力带来了巨大的进步。与此同时，这些令人欣喜的发展也带给我们很多关于语言智能的更本质的思考。由车万翔等人所著的《自然语言处理：基于预训练模型的方法》一书从预训练模型的角度对这两次重要的发展进行了系统性的论述，能够帮助读者深入理解这些技术背后的原理、相互之间的联系以及潜在的局限性，对于当前学术界和工业界的相关研究与应用都具有重要的价值。

本书包括三部分，共 9 章。书中从自然语言处理与神经网络的基础知识出发，沿着预训练模型的发展轨迹系统讨论了静态词向量、动态词向量，以及语言模型的预训练方法，还深入讨论了模型优化、蒸馏与压缩、生成模型、多模态融合等前沿进展，内容上兼具广度与深度。本书作者车万翔等人研发的语言技术平台 LTP，是国内自然语言处理领域较早、影响力大且仍在不断发展完善的开源平台之一。LTP 的“进化”历程也对应着作者对于自然语言处理不同时期范式变迁的思考与实践——从最初发布时使用的传统机器学习方法，到基于深度学习的多任务学习框架，再到近年来发布的基于预训练模型的统一框架。可以说，本书的问世是作者多年深耕于自然语言处理领域的自然结果。

本书的一大特色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，极具实战经验。书中为代表性的模型提供了规范的示例代码以及实践指导，这对于刚刚进入自然语言处理领域并热爱实践与应用的读者而言是一份难得的学习资源。

本书可以作为计算机科学、人工智能和机器学习专业的学生、研究者，以及人工智能应用开发者的参考书，也适合高校教师和研究机构的研究人员阅读。

**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

自然语言处理被誉为“人工智能皇冠上的明珠”。近年来，以 BERT、GPT 为代表的大规模预训练语言模型异军突起，使问答、检索、摘要、阅读理解等自然语言处理任务的性能都得到了显著提升。《自然语言处理：基于预训练模型的方法》一书深入浅出地阐述了预训练语言模型技术，全面深入地分析了它的发展方向，非常适合人工智能和自然语言处理领域的学习者和从事研发的人士阅读。读者可在较短的时间内了解和掌握其关键技术并快速上手。特此推荐！

**周明**

创新工场首席科学家

微软亚洲研究院原副院长

中国计算机学会副理事长

国际计算语言学会（ACL）主席（2019 年）

预训练语言模型是当今自然语言处理的核心技术。车万翔教授等人所著的本书从基础知识、预训练词向量、预训练模型等几个方面全面系统地介绍了该项技术。选题合理，立论明确，讲述清晰，出版及时。相信每一位读者都会从中获得很大的收获。向大家推荐！

**李航**

ACL/IEEE Fellow

字节跳动人工智能实验室总监

在运动智能和感知智能突飞猛进的发展态势下，以自然语言处理为核心的认知智能已成为人工智能极大的挑战。随着业界对认知智能重视程度的持续提升，基于预训练模型的自然语言处理方法一经提出，便快速席卷了诸多 NLP 任务。本书系统地介绍了该类方法，并配有丰富的实践案例和代码，对于从事 AI 技术研究和相关行业的爱好者而言，是一本不可多得的参考学习佳作！

**胡郁**

科大讯飞执行总裁

# 前言

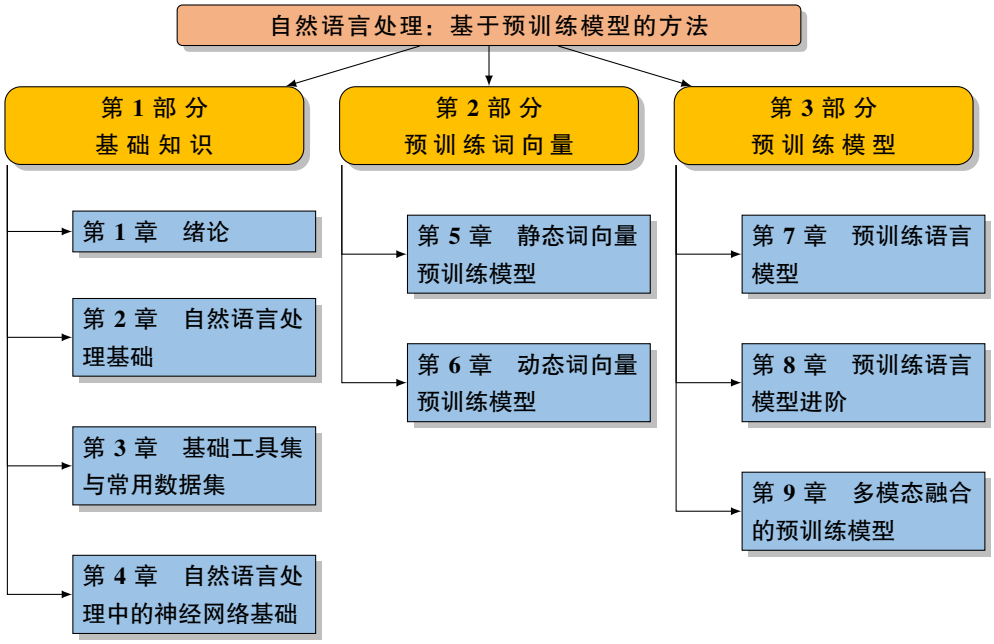
PREFACE

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，为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一系列的变革。但是基于深度学习的算法有一个严重的缺点，就是过度依赖于大规模的有标注数据。2018年以来，以 BERT、GPT 为代表的超大规模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这一缺点，帮助自然语言处理取得了一系列的突破，使得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。那么，预训练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预训练模型的神秘面纱。

## 本书主要内容

本书内容分为三部分：基础知识、预训练词向量和预训练模型。各部分内容安排如下。




**第 1 部分：基础知识。**包括第 2～4 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 章首先介绍两种常用的自然语言处理基础工具集——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，以及以 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：预训练词向量。**包括第 5、6 章，主要介绍静态词向量和动态词向量两种词向量的预训练方法及应用。

第 5 章介绍基于语言模型以及基于词共现两大类方法的静态词向量的预训练技术，它们能够通过自监督学习方法，从未标注文本中获得词汇级别的语义表示。最后提供对应的代码实现。

第 6 章介绍基于双向 LSTM 语言模型的动态词向量的预训练技术，它们能够根据词语所在的不同上下文赋予不同的词向量表示，并作为特征进一步提升下游任务的性能。最后同样提供对应的代码实现。

**第 3 部分：预训练模型。**包括第 7 ~ 9 章，首先介绍几种典型的预训练语言模型及应用，其次介绍目前预训练语言模型的最新进展及融入更多模态的预训练模型。

第 7 章首先介绍两种典型的预训练语言模型，即以 GPT 为代表的基于自回归的预训练语言模型和以 BERT 为代表的基于非自回归的预训练语言模型，其次介绍如何将预训练语言模型应用于典型的自然语言处理任务。

第 8 章主要从四个方面介绍预训练语言模型最新的进展，包括用于提高模型准确率的模型优化方法，用于提高模型表示能力的长文本处理方法，用于提高模型可用性的模型蒸馏与压缩方法，以及用于提高模型应用范围的生成模型。

第 9 章在介绍语言之外，还融合更多模态的预训练模型，包括多种语言的融合、多种媒体的融合以及多种异构知识的融合等。

## 致谢

本书第 1 ~ 4 章及第 9 章部分内容由哈尔滨工业大学车万翔教授编写；第 5、6 章及第 8、9章部分内容由美国麻省理工学院（MIT）郭江博士后编写；第 7 章及第 8 章主要内容由科大讯飞主管研究员崔一鸣编写。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！

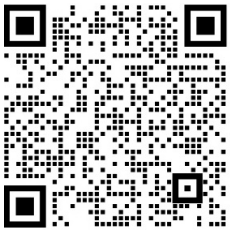
感谢宋亚东先生和电子工业出版社博文视点对本书的重视，以及为本书出版所做的一切。

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

2021 年 3 月

## 读者服务



微信扫码回复：41512

- 获取本书配套代码和习题答案。
- 加入本书读者交流群，与更多读者互动。
- 获取【百场业界大咖直播合集】（永久更新），仅需 1 元。