《自然语言处理：大模型理论与实践》

赵宇 任福继 陈星延 陈中普 陈珍珠 施龙 张阳 李庆 谢志龙

2024 年

序 言

随着人工智能技术的飞速发展，自然语言处理成为了计算机科学与人工智能领域中不可或缺的关键技术之一。作为一名长期致力于人工智能和自然语言处理研究的学者，我见证了这一领域的迅猛变革，从基于规则的方法到现今基于深度学习的革命性技术，尤其是大模型技术的应用，给自然语言处理领域带来了前所未有的机遇和挑战。

记得 1988-1990 年，我在博士期间开发了一个日中机器翻译系统，用的是 Fortran 语言，在汉字输入输出、字符串处理上费尽了脑筋。去年，我有机会应邀在很多场合演讲大模型和人工智能，提到过目前的大模型在理论上并无多大创新。我列举了国际研究同行发来的我们 1999 年发表的论文《Automatic Derivation of Programs for Image Processing from Natural Language Descriptions》（从自然语言描述自动推导程序），尽管由于周知的数据规模和算力限制，但模型系统成功的理论基础与大模型极其一致。既然目前大模型理论上并无多大创新，那为何现今它具有如此巨大的能力，推动人工智能奔腾向前？我认为就是本书要提到的涌现。

本书以其独特的视角和结构，全面系统地介绍了大模型技术在自然语言处理中的理论与实践。全书内容丰富，涵盖了语言模型的基础知识、大模型的关键技术以及大模型在实际中的应用实践，不仅为初学者提供了详尽的入门指南，也为研究人员和专业人士提供了深入的技术解析和实用的开发案例。

本书从自然语言处理的背景知识出发，逐步引入词向量、统计语言模型、神经语言模型和预训练语言模型，然后详细介绍了大模型的架构、训练方法、应用及评估策略，并通过丰富的实例和习题，帮助读者加深理解和掌握大模型技术的核心内容。特别的是，本书对大模型的本地开发和应用开发部分，提供了实战演练和实用的代码示例，对实际开发工作具有重要的指导意义。

本书的出版正值大模型技术蓬勃发展的关键时期，对于高校本科生、研究生、教学科研人员，以及从事自然语言处理研究和开发的专业人士来说，都是一本不可多得的参考书籍。通过阅读本书，读者不仅可以系统地掌握自然语言

处理的基础知识和前沿技术，还能在实际操作中提升自身的开发和研究能力，探索自然语言处理的广阔前景。

我深信，随着自然语言处理技术的不断发展和应用，未来将会有更多的突破和创新，推动人类与机器之间更加自然和智能的交流方式，开创人工智能更加辉煌的未来。

任福继
日本工程院院士、中国人工智能学会理事会名誉理事长
2024 年 6 月于成都

前 言

自然语言处理（Natural Language Processing, NLP）是计算机科学与人工智能交叉领域中的一门关键技术，其目标是使计算机能够理解、解释、生成人类语言。在当今人工智能时代，NLP 技术已经深刻地渗透到我们日常生活的方方面面，从智能助手、语音识别到机器翻译和文本生成，NLP 正以惊人的速度改变着我们的生活方式。特别的是，2022 年底以 ChatGPT 为代表的大模型技术横空出世，进一步推动了新一代人工智能技术的发展。大模型技术颠覆了自然语言处理领域传统的知识体系。然而，目前以大模型技术为主线介绍自然语言处理知识的教材较为缺乏。基于此考虑，催生了我编著本教材的想法。

本教材以自然语言处理中语言模型为主线，主要内容分为三部分，包括语言模型基础、大语言模型关键技术和大模型实践。在此之前，首先介绍了自然语言处理的背景知识。然后，在语言模型基础部分介绍了词向量、统计语言模型、神经语言模型和预训练语言模型。接着，在大模型理论部分介绍大模型的架构、训练、利用与评估等。最后，在大模型实践部分介绍了大模型的本地开发和应用开发等。

本教材主要针对高校本科生、研究生以及教学科研人员，作为教学用书。当然，也适用于计算语言学家、语言学家、数据科学家和 NLP 开发人员等专业人士。考虑到不同读者的学科差异，本书在附录部分介绍了概率论、信息论、机器学习与强化学习等 NLP 交叉学科的基础知识。阅读本教材最好具备 Python 的编程知识。

在写作本教材的过程中，我深切地感受到自然语言处理的迅猛发展。从传统的基于规则的方法到现今基于深度学习的革命性变革，NLP 的前景无疑令人激动。我希望通过本教材，能够为读者提供以语言模型为主线的 NLP 知识体系，并让读者能深入理解大语言模型前沿理论，掌握大语言模型实践技能。希望这本教材成为您学习与实践 NLP 的得力工具，激发您对自然语言处理无尽的好奇心和创造力。

在本书即将付梓之际，我深感荣幸与感激，借此机会向所有在本书编写过程中给予帮助和支持的个人与机构致以最诚挚的谢意！

谨向自然语言处理领域的诸多前辈和专家致以崇高的敬意。我要特别感谢大模型技术相关领域的研究学者，正是你们在 Transformer、GPT、GLM 等模型的发展过程中所做出的杰出贡献，为本书的内容奠定了坚实的基础。你们的无私分享和合作精神，推动了这一领域的飞速发展，并为本书提供了丰富的理论依据和实践经验。

感谢中国人工智能学会自然语言理解专委会主任王小捷教授对本书编写给予的大力支持，并提出了若干宝贵的建议。感谢我们通用人工智能团队以及金融智能与金融工程四川省重点实验室的同仁们，为本书的完成提供了不可或缺的支持。本书初稿完成以后，我们团队自然语言处理课程组老师做了大量的工作。陈珍珠博士校对了第一章的部分内容。陈中普博士和陈星延博士分别校对并修改了第二章的部分内容。陈星延博士校对了第九章全部内容。陈中普博士和张阳博士分别校对了第十章的部分内容。陈中普博士、陈珍珠博士和张阳博士分别编写并校对了第十一章的部分内容。陈星延博士校对了第十二章的全部内容。陈珍珠博士校对了第十三章的全部内容。陈中普博士编写并校对了第十四章和第十五章的部分内容。施龙博士和陈星延博士分别校对了附录 A 预备知识的部分内容。此外，李庆博士和谢志龙博士也参与了本书的部分校对工作。另外，特别感谢韦鳗珍、白芊芊、罗灵、王瑞、刘银峰、许雯婷、钟一、顾添承、刘雅玲、黄浩南、唐川清、郭宇、杨闻博和邓黄怡等同学在文稿编辑、图表绘制和审稿校对等方面所付出的努力，这是本书完成的基础。

本书引用了一些优秀参考文献中的图表、公式和案例等，征求了相关作者的意见并得到了积极支持，在此表示由衷的感谢！

此外，由衷感谢机械工业出版社的辛勤付出，感谢你们在本书编写和出版过程中所展现的专业精神和不懈努力，让本书最终得以面世。

本书的编写得到了国家自然科学基金项目的资助。

最后，对所有期待本书的读者表示感谢，你们的期望和支持激励我不断深入研究。希望本书能为你们带来启发和帮助，并在你们的学术和实践道路上提供有益的参考。

再次感谢所有为本书付出心血和智慧的朋友们！愿本书能为自然语言处理领域的发展贡献绵薄之力，并激励更多的研究者投身于这一充满前景的研究领域！

由于编者水平有限，书中难免有疏漏和不足之处，恳请读者批评指正！如果您发现书中的任何错误或遇到任何问题，可以发送邮件至编者邮箱，真诚期

待您的反馈。

赵 宇
2024 年 6 月

目 录

第一章　绪论

“自然语言处理被誉为人工智能皇冠上的明珠。”

——无名氏

1.1　自然语言处理概述

人类的语言能力是在早期儿童时期发展起来的，并在一生中不断进化。与人类不同，机器无法自然地掌握人类语言理解能力，它们需要通过如人工智能算法或者语言模型的方式，提升其在语言方面的智能。为了探究人类语言的奥秘，研究者们创立了如自然语言处理 (Natural Language Processing, NLP)、自然语言理解 (Natural Language Understanding, NLU)和计算语言学 (Computational Linguistics, CL)等学科，这些学科致力于使用计算方法来分析、理解和生成人类语言，从而构建人机交流的桥梁。随着时间的推移，这些研究领域在技术上取得了显著的进步，推动了人工智能在语言理解和生成方面的技术进步。

NLP、NLU和 CL虽然密切相关，但侧重点有所不同。NLP是一门人工智能领域的交叉学科，涉及计算机科学、人工智能和语言学等多个领域，旨在使计算机能够解释、生成和处理人类自然语言[1]。自然语言处理技术的研究不仅提高了信息处理的效率，而且也极大地拓宽了人机交互的方式，促进了科技创新和社会进步。随着技术的不断发展，NLP在信息检索、智能问答系统、机器翻译等众多应用领域发挥着日益重要的作用。NLU旨在使计算机能够理解和解释人类自然语言的含义和意图，其目标是让计算机能够像人类一样理解和推断自然语言文本或语音输入的意义。CL则更侧重于语言本身的科学研究，试图通过计算机模拟来理解语言的本质和结构，包括语言的生成和理解机制。总

[1]自然语言包括如中文、英文、俄文、阿拉伯文、西班牙语等。不同于自然语言，人们常常还使用各种编程语言等机器语言操纵计算机、使用 SQL 语言操纵数据库，以及使用手势等肢体语言沟通信息。

之，NLP、NLU和 CL都是探索人类语言奥秘、实现人机交流的重要学科，它们共同推动着语言技术的发展和应用。随着技术的不断进步，这些领域之间的界限越来越模糊。在本书中，若无特别说明，将不再区分这三个术语。

语言模型（Language Models, LMs）在推动 NLP、NLU和 CL的发展过程中起到了核心作用。语言模型通过计算一系列词汇在特定上下文中出现的概率，来预测下一个词的可能性。这不仅是实现机器翻译、语音识别等应用的基础，也是让机器能够更自然地与人类进行交流的关键。随着深度学习技术的发展，基于神经网络的语言模型（如 Transformers [168]）极大地推动了这些领域的发展，使得机器在理解和生成自然语言方面的能力得到了质的飞跃。除了技术的进步，数据的积累也对这些领域的发展至关重要。大数据时代的来临为这些领域的研究提供了前所未有的资源。通过分析和处理大量的文本数据，研究者们能够训练更加强大和精准的模型，这些模型能够更好地理解语言的复杂性和多样性。然而，尽管取得了巨大进步，人类语言的复杂性和深层的语义理解仍然是机器面临的巨大挑战。语言不仅仅是词汇和语法的组合，它还包含了文化、情感和隐喻等多层次的含义。因此，未来的研究将需要更深入地探索人类语言的这些方面，以及如何让机器更加准确地理解和生成具有深层含义的语言。

1.2 自然语言处理简史

自然语言处理技术的发展可以追溯到 20 世纪 50 年代初，经历了多个阶段和重要的突破。以下是自然语言处理技术发展的简单概述：

1. 早期阶段（1950s - 1960s）

- 1950 年代早期，阿兰 · 图灵（Alan Turing）提出了著名的 “图灵测试（Turing Test）”，探讨了机器是否能够模仿人类的自然语言交流。
- 1954 年，Georgetown 大学和 IBM 合作，开发了一个名为 “Georgetown-IBM 实验室翻译器” 的系统，用 IBM 701 电脑将 60 多个俄语句子翻译成英文，标志着机器翻译的起步。

除了图灵测试和 Georgetown-IBM 实验之外，这个时期还诞生了一些初步尝试来理解语言的技术，如知名语言学家诺姆 · 乔姆斯基（Noam Chomsky）的生成语法理论，它对后来的语言模型和理解有着深远的影响。