前　言

PREFACE

自然语言生成经过几十年的发展，已经成为人工智能和自然语言处理的重要研究领域。最早的自然语言生成系统采用规则、模板的方法，设计各司其职的模块进行文本生成，其中体现了很多专家设计的词汇、语法、句法甚至语用的语言学知识。统计语言模型则从概率统计的角度提出了语言建模的新思路，将词汇与上下文的依赖关系编码在条件概率中。以深度学习模型为基本架构的现代语言生成模型绝大多数通过端到端训练的方式，能更好地建模词汇与上下文之间统计共现关系，显著地提升了文本生成的性能。特别是以 Transformer 为基础架构的预训练语言生成模型，能够较好地捕获包括词汇、语法、句法、语义等各层面的语言学知识，极大地推动了自然语言生成的进展，生成效果令人惊叹。

技术的进步显著地推动了应用的发展。就自然语言生成而言，机器翻译、摘要生成、故事生成、对话生成、诗歌生成等任务都广泛地应用了以神经网络为基本架构的现代语言生成方法，生成效果相比传统方法进步显著，在许多实际应用场景中大显身手。以神经机器翻译为例，在数据丰富的领域，机器翻译的效果甚至可以媲美人工翻译的效果。Google 新推出的聊天机器人 Meena 采用基于 Transformer 的架构，在某些方面接近甚至超过人类对话的效果。GPT 系列模型甚至可以生成人物角色丰富、故事情节曲折的长文本故事。机器创作，包括强调创新和创意的语言生成任务，如现代诗、歌词、古诗生成等，业已成为人工智能领域广受关注的研究课题，并在一些应用场景中落地，微软小冰甚至出版了机器创作的现代诗歌集。

正因为这样的背景，我们认为系统地总结自然语言生成的算法、模型和技术是十分必要的。通过梳理自然语言生成特有的问题和挑战，我们希望整理、概括和归纳现有的自然语言生成模型、框架和方法，以便我们更好地思考这个领域的现状和未来。而且，目前已有的相关书籍中，还未见以自然语言生成为专题的书籍，这也是我们写这本书的重要原因之一。

本书的写作围绕自然语言文本的概率建模展开。无论是传统的统计语言模型，还是现代神经网络语言模型，都可以归结到一个基本问题，即给定上文如何预测下文。传统的统计语言模型采用符号化的条件概率表，并利用共现次数直接估计条件概率。在基于神经网络的模型中，条件概率通过一个参数化模型来表达。模型容量越大，数据越多，这种参数化模型的优势体现得越明显。从统计语言建模到神经语言建模的发展过程实际是从语言文字的符号表示到向量表示的转变过程。静态词向量、语境化语言表示的建模思路深刻地改变了传统语言表示的计算范式，但其背后恒久不变的思想依然是分布假设：出现在相似上下文中的词是相似的。

本书围绕文本的“条件概率建模”这条主线，从基础模型、优化方法、生成方式、生成机制等多个层面进行介绍。在基础模型方面，介绍了目前主流的循环神经网络和 Transformer 两类模型，并从模型结构、注意力机制等角度分析了两者的区别与联系。在优化方法方面，介绍了变分自编码器中变分优化和生成式对抗网络中的对抗优化方法。在生成方式方面，除了经典的自回归语言生成方式，还介绍了前沿的非自回归生成方式，为文本生成提供了一种新的视角。在生成机制方面，介绍了语言生成中重要的规划机制和知识融入机制，并介绍了具体应用案例。最后系统地整理了语言生成的评价方法，从语言生成到语言评价形成了一个闭环。

本书可作为高等院校计算机科学与技术、人工智能、大数据等相关专业高年级本科生、研究生相关课程的教材，也适合从事自然语言处理研究、应用实践的科研人员和工程技术人员参考。本书的内容对理解现代语言生成模型的原理、优势和弊端将有很大的帮助。需要注意的是，理解本书内容需要具备概率统计、微积分、线性代数、机器学习的基本知识；对于深度学习方面的知识，则要求具备多层感知机、反向传播算法、自编码器等神经网络的基本知识。

本书是清华大学计算机系、人工智能研究院对话式智能 (CoAI) 小组集体努力的成果，也反映了课题组这几年在语言生成上的探索与积累，部分成果也编入了本书中。本书具体分工是：朱小燕负责审校、订正全书内容；黄民烈设计了全书结构，负责撰写和审校全部书稿；黄斐撰写了部分内容。此外，柯沛、关健、计昊哲、邵智宏也参与了部分内容的写作和资料整理工作，顾煜贤、周昊、郑楚杰、吴尘等参与了资料收集、整理工作。另外，如

果没有国内外同行的研究工作，不可能有本书的出版。感谢清华大学人工智能研究院、国强研究院、国家自然科学基金委的支持。最后，感谢家人在写作期间无条件的支持。

由于编写时间仓促，书中难免存在错误、疏漏之处，望读者包涵，请批评指正。

黄民烈

2020 年 8 月 20 日于清华园