中国科学: 信息科学 2023 年 第 53 卷 第 9 期: 1645–1687
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

大模型时代的自然语言处理: 挑战、机遇与发展 †

车万翔[4], 窦志成[11], 冯岩松[1], 桂韬[3], 韩先培[10], 户保田[5], 黄民烈[6], 黄萱菁[2*], 刘康[9], 刘挺[4], 刘知远[6*], 秦兵[4], 邱锡鹏[2], 万小军[1], 王宇轩[8], 文继荣[11], 严睿[11], 张家俊[9], 张民[5,7*], 张奇[2], 赵军[9], 赵鑫[11], 赵妍妍[4]

1. 北京大学王选计算机研究所, 北京 100080
2. 复旦大学计算机科学技术学院, 上海 200438
3. 复旦大学现代语言学研究院, 上海 200433
4. 哈尔滨工业大学计算学部, 哈尔滨 150001
5. 哈尔滨工业大学 (深圳) 计算机科学与技术学院, 深圳 518055
6. 清华大学计算机科学与技术系, 北京 100084
7. 苏州大学计算机科学与技术学院, 苏州 215006
8. 之江实验室, 杭州 311121
9. 中国科学院自动化研究所, 北京 100190
10. 中国科学院软件研究所, 北京 100190
11. 中国人民大学高瓴人工智能学院, 北京 100872

* 通信作者. E-mail: xjhuang@fudan.edu.cn, liuzy@tsinghua.edu.cn, zhangmin2021@hit.edu.cn
† 作者和单位排名不分先后, 同等贡献, 按拼音排序

收稿日期: 2023–04–21; 修回日期: 2023–05–30; 接受日期: 2023–06–03; 网络出版日期: 2023–09–11

摘要 近期发布的 ChatGPT 和 GPT-4 等大型语言模型, 不仅能高质量完成自然语言生成任务, 生成流畅通顺, 贴合人类需求的语言, 而且具备以生成式框架完成各种开放域自然语言理解任务的能力. 在少样本、零样本场景下, 大模型可取得接近乃至达到传统监督学习方法的性能, 且具有较强的领域泛化性, 从而对传统自然语言核心任务产生了巨大的冲击和影响. 本文就大模型对自然语言处理的影响进行了详细的调研和分析, 试图探究大模型对自然语言处理核心任务带来哪些挑战和机遇, 探讨大模型将加强哪些自然语言处理共性问题的研究热度, 展望大模型和自然语言处理技术的未来发展趋势和应用. 分析结果表明, 大模型时代的自然语言处理依然大有可为. 我们不仅可以将大模型作为研究方法和手段, 学习、借鉴大型语言模型的特点和优势, 变革自然语言处理的主流研究范式, 对分散独立的自然语言处理任务进行整合, 进一步提升自然语言核心任务的能力; 还可就可解释性、公平性、安全性、信息准确性等共性问题开展深入研究, 促进大模型能力和服务质量的提升. 未来, 以大模型作为基座, 拓展其感知、计算、推理、交互和控制能力, 自然语言处理技术将进一步助力通用人工智能的发

引用格式: 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展. 中国科学: 信息科学, 2023, 53: 1645–1687, doi: 10.1360/SSI-2023-0113
Che W X, Dou Z C, Feng Y S, et al. Towards a comprehensive understanding of the impact of large language models on natural language processing: challenges, opportunities and future directions (in Chinese). Sci Sin Inform, 2023, 53: 1645–1687, doi: 10.1360/SSI-2023-0113

展, 促进各行各业的生产力进步, 更好地为人类社会服务.

关键词 ChatGPT, 对话式大模型, 大型语言模型, 自然语言处理, 通用人工智能

1 引言

2022 年 11 月 30 日, OpenAI 发布了对话式语言大模型 ChatGPT (chat generative pre-trained transformer)[1)]. 该模型允许用户使用自然语言对话形式进行交互, 可实现包括自动问答、文本分类、自动文摘、机器翻译、聊天对话等各种自然语言理解和自然语言生成任务. ChatGPT 在开放域自然语言理解上展现了出色的性能, 甚至无需调整模型参数, 仅使用极少数示例数据即可在某些任务上超过了针对特定任务设计并且使用监督数据进行训练的模型. 当面对用户所提出的各种文本生成任务时, ChatGPT 在多数情况下可以生成流畅通顺、有逻辑性且多样化的长文本.

ChatGPT 自发布以来引起了广泛的关注, 仅在 5 天内注册用户就超过了 100 万. 据雅虎财经[2)]统计, 在 ChatGPT 推出仅两个月后, 月活跃用户已达 1 亿, 相比之下, 之前一直被认为是用户增长速度最快的消费级应用程序 Tiktok 则花费了 9 个月的时间. 稍后不久, 微软于 2023 年 2 月 8 日推出了新一代 AI 驱动搜索引擎 New Bing[3)], 该引擎将基于 ChatGPT 技术的生成模型与 Bing 搜索深度集成, 创造了对话式搜索的新范式. 2023 年 3 月 14 日, OpenAI 发布了下一代生成式多模态预训练大模型 GPT-4 (generative pre-trained transformer 4)[4)], 它不仅能够理解自然语言文本, 还能够对图片内容进行深度理解, 并且具备比 ChatGPT 更强的问题求解和推理能力, 在多种人类考试和自然语言理解任务中取得了更加优秀的成绩 [1].

长期以来, 自然语言处理任务主要采用监督学习范式, 即针对特定任务, 给定监督数据, 设计统计学习模型, 通过最小化损失函数来学习模型参数, 并在新数据上进行模型推断. 随着深度神经网络的兴起, 传统的统计机器学习模型逐渐被神经网络模型所替代, 但仍然遵循监督学习的范式. 2020 年 5 月 Open AI 发布的首个千亿参数 GPT-3 (generative pre-trained transformer 3) 模型初步展示了生成式模型的强大功能, 其具备流畅的文本生成能力, 能够撰写新闻稿, 模仿人类叙事, 创作诗歌, 初步验证了通过海量数据和大量参数训练出来的大模型能够迁移到其他类型的任务 [2]. 然而, 直到 ChatGPT 的出现, 学术界才意识到大模型对于传统自然语言处理任务范式的潜在颠覆性.

以 ChatGPT 为代表的大型语言模型, 给自然语言处理带来的是威胁、挑战还是新的机遇? 今后的自然语言处理核心任务将采用何种主流范式实现语言理解和生成? 自然语言处理的研究领域将如何延伸? 以大模型为代表的自然语言处理技术将如何引领通用人工智能的发展? 我们就大模型对自然语言处理的影响进行了详细的调研和思考, 试图分析大模型对自然语言处理核心任务带来的冲击和启发, 探讨大模型将加强哪些自然语言处理共性问题的研究热度, 展望大模型和自然语言处理技术的未来发展和应用, 以期回答上述问题.

1) https://chat.openai.com.

2) https://finance.yahoo.com/news/chatgpt-on-track-to-surpass-100-million-users-faster-than-tiktok-or-instagram-ubs-214423357.html.

3) https://www.bing.com/new.

4) https://openai.com/product/gpt-4.

2 背景知识

在探讨大模型给自然语言处理带来的挑战和机遇之前, 首先需要介绍相关的背景知识, 包括自然语言处理的概念和研究历史, 大规模预训练语言模型从语言模型、预训练模型到大模型的技术发展历程, 以及 ChatGPT 和 GPT-4 的基本技术与能力.

2.1 自然语言处理

自然语言处理 (natural language processing, NLP) 的目标是实现人机之间的有效通信, 使得计算机既能够理解自然语言的意义, 也能以自然语言文本来表达意图与思想. 自然语言处理是人工智能领域的重要研究方向, 融合了语言学、计算机科学、机器学习、数学、认知心理学等多个学科领域的知识. 总体而言, 自然语言处理包含自然语言理解和自然语言生成两个主要方面, 研究内容覆盖的粒度包括字、词、短语、句子、段落和篇章等多种层次. 由于语言的复杂性, 高精度、高鲁棒、可解释的通用自然语言处理系统目前还没有成熟解决方案, 仍需进行长期研究.

自然语言处理的研究历史基本上与通用计算机相同, 1946 年第 1 台现代电子数字计算机 ENIAC 问世后, 1947 年 Warren Weaver 便提出了利用计算机翻译人类语言的可能, 并于 1949 年发布了著名的 *Translation* (翻译) 备忘录, 开启了自然语言处理的研究历程. 在大规模语言模型出现之前, 自然语言处理经历了从理性主义到经验主义再到深度学习的 3 个历史阶段. 在这个发展过程中, 也逐渐形成了一些范式, 包括基于规则的方法、基于机器学习的方法和基于深度学习的方法. 基于规则的方法其核心思想是通过使用语言学知识, 如词汇和形式文法等, 来制定规则, 从而完成自然语言处理任务. 基于机器学习的自然语言处理方法主要采用有监督分类算法, 将自然语言处理任务转化为某种分类任务, 在此基础上, 根据任务特性构建特征表示, 并构建大规模的有标注语料进行模型训练. 深度学习方法则通过构建深度模型, 将特征学习和预测模型融合在一起. 通过优化算法, 模型可以自动地学习出良好的特征表示, 并基于此进行结果预测.

2.2 大规模预训练语言模型

语言模型 (language model, LM) 旨在计算给定词序列 $w_1, w_2, \ldots, w_m$ 作为句子的概率分布 $P(w_1, w_2, \ldots, w_m)$. 然而, 联合概率 $P(w_1, w_2, \ldots, w_m)$ 的参数量十分巨大, 因此直接计算它非常困难. 如果将 $w_1, w_2, \ldots, w_m$ 视为一个变量, 则其可能性具有 $|V|^m$ 种, 其中 m 代表句子的长度, $|V|$ 表示词表中词语的数量. 为了减少 $P(w_1, w_2, \ldots, w_m)$ 的模型参数量, 可以利用通常情况下句子序列从左至右的生成过程, 使用链式法则进行分解, 并限制历史长度以降低模型参数量, 从而得到 n 元语言模型 (n-gram language model).

随着深度学习的进步, 利用分布式表示和神经网络的语言模型已成为研究的热点. 2000 年, Bengio 等 [3] 提出了一种使用前馈神经网络来估计 $P(w_i|w_{i-n+1}, \ldots, w_{i-1})$ 的语言模型. 此后, 循环神经网络 [4]、卷积神经网络 [5]、端到端记忆网络 [6] 等神经网络方法都被成功地应用于语言模型建模. 相较于传统的 n 元语言模型, 神经网络方法可以在一定程度上避免数据稀疏问题. 此外, 一些神经网络模型可以突破对历史长度的限制, 从而更好地建模长距离依赖关系.

随着对深度神经网络研究的不断深入, 研究人员们发现单词的嵌入表示对于任务的效果有很大影响, 而通过大规模语言模型可以获得很好的初始单词向量. 虽然语言模型的训练过程采用有监督方法, 但由于可使用原始文本获得训练目标, 因此只需大规模无标注文本即可训练模型. 2018 年, Peters 等 [7] 提出了 ELMo (embeddings from language models) 方法, 该方法设计了多层双向长短期记忆网络, 通过

语言模型任务, 利用大规模语料库来获取更好的单词表示, 在多个自然语言处理任务上获得了良好的效果. 之后, Devlin 等[8] 在 2018 年提出了 BERT (bidirectional encoder representations from transformers) 方法, 该方法基于 Transformer 架构[9] 和掩码语言模型, 通过大规模预训练获得, 并在多个自然语言处理任务中取得了巨大的提升, 如阅读理解、语义匹配等, 引领了大规模预训练语言模型研究的热潮[10].

随着时间的推移, 大规模语言模型不断发展. 在 2020 年, Open AI 发布了 GPT-3 这款生成式大规模预训练语言模型, 其参数量高达 1750 亿[2]. 随后, 谷歌 (Google) 开发的 Switch Transformer 模型的参数量首次超过万亿[11]. 北京智源研究院发布了参数量超过 1.75 万亿的预训练模型 “悟道 2.0”. Meta、百度、华为等公司和研究机构也相继发布了不同的大规模语言模型, 例如 PaLM[12], LaMDA[13], T0[14] 等. 这些大规模语言模型也被称为大模型, 在文本生成、少样本学习、零样本学习、推理任务等方面展现了优异的效果. ChatGPT 和 GPT-4 的发布, 更使得大模型研究进入了大规模应用阶段, 并引发了新一轮的自然语言处理范式的发展.

2.3 ChatGPT 与 GPT-4

由于 ChatGPT 绝大部分的技术细节还没有完全公开, 一些已经公开的研究内容和方法也仍然需要更多时间进行验证. 但总的说来, ChatGPT 的整个发展和技术演进过程可参考文献 [15] 中的图 1. 作为基于 GPT-3.5 架构的一个大型语言模型, ChatGPT 的训练过程包含 3 个主要阶段. 第一个阶段是基础大模型训练, 目的是完成长距离语言模型的预训练, 并使得模型具备代码生成的能力. 第二个阶段是指令微调, 通过给定指令对模型进行微调, 以使其能够完成各种任务. 最后一个阶段是类人对齐, 通过加入更多的人工提示词, 并使用监督学习和基于强化学习的方法, 使得模型的输出更加贴合人类需求. 其中, 基础的大模型预训练给模型带来了语言生成能力, 并且具备了上下文学习 (in-context learning) 能力和世界知识, 包括事实性知识 (factual knowledge) 和常识 (commonsense). 指令微调使得模型能够更好地遵循人的指令, 增强了模型的零样本能力. 类人对齐则使得模型的输出符合人类的期望, 不仅可以忠实地输出对人类有用的结果, 也避免了输出有害内容并拒绝模型知识范围之外的问题.

ChatGPT 发布后不久, OpenAI 随即发布了 GPT-4 模型. 其上下文窗口长度可能从 GPT-3.5-turbo 的 4096 词符提高到 32768 词符. 另外, GPT-4 是多模态模型, 能识别、提取图像信息, 并给出文字反馈. 在演示中, GPT-4 还能根据手绘草图, 快速生成网站代码. GPT-4 模型的语言能力比 GPT-3.5 有较大提高, 在许多学术和专业测试中表现出超过绝大多数人类的水平[1]. 甚至, 根据微软研究院的分析和测评, 无需特殊设计指令, GPT-4 即可回答数学、编程、视觉、药物、法律和心理等问题, 其性能远超 ChatGPT, 几乎达到人类水准, 可被合理地认为是一个早期 (虽然尚不完备) 的通用人工智能系统[16].

ChatGPT 与 GPT-4 都具有很强的可扩展性. 除了与 New Bing 搜索引擎集成之外, ChatGPT 提供了插件服务, 以获取最新信息, 运行计算, 使用第三方服务[5]. 基于 GPT-4, 微软发布了 Copilot 智能助手, 大幅度提升了 Office, GitHub 等生产力工具的智能水平和服务能力.

3 大模型时代的自然语言处理核心任务

自然语言处理包含自然语言理解和自然语言生成两个方面, 常见任务包括文本分类、结构分析 (词

5) https://openai.com/blog/chatgpt-plugins.

法分析、分词、词性标注、句法分析、篇章分析)、语义分析、知识图谱、信息提取、情感计算、文本生成、自动文摘、机器翻译、对话系统、信息检索和自动问答等. 在神经网络方法出现之前, 因为缺乏行之有效的语义建模和语言生成手段, 自然语言处理的主流方法是基于机器学习的方法, 采用有监督分类, 将自然语言处理任务转化为某种分类任务. 在神经网络时代, Word2Vec 词嵌入模型、BERT 等上下文相关语言模型为词语、句子乃至篇章的分布式语义提供了有效的建模手段; 编码器 – 解码器架构和注意力机制提升了文本生成的能力; 相比传统自然语言处理所遵循的词法 – 句法 – 语义 – 语篇 – 语用分析级联式处理架构, 端到端的神经网络训练方法减少了错误传播, 极大提升了下游任务的性能. 不过, 神经网络方法仍然遵循监督学习范式, 需要针对特定任务, 给定监督数据, 设计深度学习模型, 通过最小化损失函数来学习模型参数. 由于深度学习也是一种机器学习方法, 因此从某种程度上, 基于神经网络的方法和基于机器学习的方法并无本质区别.

然而, 不同于通常的深度学习方法, 以 ChatGPT 为代表的生成式大模型, 除了能高质量完成自然语言生成类任务之外, 还具备以生成式框架完成各种开放域自然语言理解任务的能力. 只需要将模型输出转换为任务特定的输出格式, 无需针对特定任务标注大量的训练数据, ChatGPT 即可在少样本乃至零样本上, 达到令人满意的性能, 甚至可在某些任务上超过了特别设计并使用监督数据进行训练的模型. 因此, ChatGPT 对各种自然语言处理核心任务带来了巨大的、不可避免的冲击和影响, 也酝酿着新的研究机遇. 接下来, 针对各种自然语言处理核心任务, 我们将首先介绍其任务需求和主流方法, 然后分析大模型对其主流研究范式所带来的影响, 并探讨未来研究趋势.

3.1 文本分类

3.1.1 文本分类及其主流方法

文本分类是自然语言处理中重要的基础任务, 是指对文本内容根据分类目标预测其类型, 其应用也非常广泛, 比如: 新闻分类、垃圾邮件过滤、情感倾向分析等. 作为一个经典的语言理解任务, 文本分类已经经历了长达数十年的研究. 文本分类通常采用有监督机器学习方法, 通过训练数据构建分类模型. 早期的文本分类方法大多基于统计学习范式, 如朴素贝叶斯 [17]、K- 近邻 [18]、支持向量机 [19]. 尽管在准确性和稳定性上相比简单的基于规则的方法有了很大程度的提升, 这些方法仍然需要花费大量精力在特征工程之上. 与之相对, 深度学习方法能够避免人工设计规则和特征, 自动化地构建包含丰富语义的文本表示. 因此在 2010 年以后, 深度神经网络, 如循环神经网络 [20]、卷积神经网络 [21]、图神经网络等 [22], 逐渐取代统计机器学习方法成为文本分类的主流方法.

近几年, 得益于无监督预训练技术的蓬勃发展, 以 ELMo [7], GPT [23], BERT [8] 为代表的预训练模型实现了从海量文本语料中高质量的文本语义建模, 并极大地提高了文本分类任务的性能. 不过, 这些文本分类算法在处理开放域任务时, 仍然遇到以下主要挑战: 领域或者数据分布发生变化时, 分类准确率会大幅度地降低; 模型训练需要依赖大规模的训练语料; 分类目标需要提前确定, 如果增加分类目标种类, 需要对模型进行重新训练; 针对一些任务, 仍然需要对模型结构, 损失函数等方面进行特殊设计.

3.1.2 大模型时代的文本分类

文本分类任务是受到 ChatGPT 影响较大的自然语言处理任务之一. 特别是文本分类任务所遇到的小样本、目标不确定、领域迁移、需要特殊设计等问题, ChatGPT 都有非常强的性能. 在少样本甚至是零样本的情况下, ChatGPT 都可以取得非常好的分类效果. 针对 GPT-3.5 系列模型的评测结果也可以看到, 在情感倾向分析、词性标注等任务上, 其分类效果甚至都超过了单任务的有监督分类的

性能, 甚至在鲁棒性上相较于此前的方法效果更好 [24]. 通过指令交互和上下文学习, ChatGPT 还可以有效地解决目标不确定、领域迁移、少样本等此前有监督分类算法所面临的难点问题. 此外, 模型结构和损失函数并不需要针对特定任务进行修改. 因此, 大模型的出现对文本分类任务有非常大的影响.

但是基于大模型的文本分类仍然存在一定的问题, 由于模型参数容量极大, 对于单个或者少数任务来说, 训练成本和使用成本都非常高, 模型计算所带来的延迟也较高. 如何针对特定任务的需求, 结合大模型的少样本学习、领域泛化和任务泛化能力, 对大模型进行有效的缩减是可能的研究方向. 代表性方法包括知识蒸馏、借助小型的学生模型拟合大规模教师模型的中间表示和任务输出、迁移大模型的能力 [25]; 或通过大模型合成训练数据, 供小模型学习 [26] 等.

此外, 在文献 [24] 中, 我们看到大模型依然存在鲁棒性问题, 针对输入的微小变化, 仍然会造成模型分类结果的变化. 例如针对属性级情感分析任务, 在修改了目标情感词之后, 模型分类结果有接近 30% 的下降. 因此, 如何提升大模型在文本分类任务上的鲁棒性也是需要进一步进行研究的问题. 对这一问题的研究可以从如下 3 个方面展开,

(1) 鲁棒性定义. 模型的鲁棒性问题可能在多种不同的场景和角度下暴露出来, 如领域外样本的不鲁棒性 [27]、对抗样本的不鲁棒性 [28]、长尾样本的不鲁棒性 [29]、噪音样本的不鲁棒性 [30]、零样本类别的不鲁棒性 [31] 等, 如何全面客观地评价和度量模型鲁棒性是一个非常重要的课题.

(2) 鲁棒问题识别. 大多数鲁棒问题的识别依赖于人类专家的先验知识和对模型详细的错误分析 [32,33], 因此对于鲁棒问题识别的研究能够更有效地帮助找到模型本身的缺陷.

(3) 鲁棒性提升. 针对已知的模型鲁棒性问题, 如何更加有效地提升模型的鲁棒性. 如数据增强 [34]、模型与训练策略设计 [35]、归纳偏置与先验 [36]、因果干预等 [37].

3.2 结构化预测

3.2.1 结构化预测任务服务于应用型下游任务

传统的结构化预测任务, 如分词 [38]、词性标注 [39]、句法分析 [40] 等, 主要目的是对自然语言进行结构化和抽象的表示, 从而方便后续的分析和处理. 在预训练模型出现之前, 这些任务的结果在很多应用型下游任务中发挥着重要的作用.

例如, 分词作为传统自然语言处理中的第一步, 起到了提高文本可读性和可处理性, 减少歧义和冗余, 增强后续任务的准确性和效率的作用. 在统计学习中, 分词的结果一般以独热向量的形式作为下游任务模型的特征 [41,42], 而在之后的深度学习中, 分词的结果一般被映射到低维稠密的词向量之后再输入模型 [43,44].

词性标注是传统自然语言处理中另一个重要的基本任务, 该任务通过预测句中每个词的词性标签, 起到消除歧义, 提高后续任务准确率和效率的作用. 与分词类似, 词性标注的结果在统计学习中一般也作为特征输入下游任务模型中, 而在深度学习中, 则映射到词性向量后输入模型中.

句法分析是传统自然语言处理中一个具有悠久历史的任务, 该任务对输入文本句子进行分析以得到其句法结构, 从而为下游任务提供更直接和丰富的语义信息. 在统计学习中, 句法分析的结果一般通过词语在句法树上的相对位置信息和路径信息, 为下游任务模型提供额外的特征 [45,46], 例如 “当前词在句法树中父节点的词性” 等. 而在深度学习中, 一般通过将句法树融入神经网络结构中来向模型中注入句法信息 [47,48].

总的来说, 在统计学习和早期的深度学习方法中, 分词、词性标注、句法分析等传统的结构化预测任务起到对文本进行预处理, 从而消除歧义, 为后续任务提供更丰富信息的作用, 但同时由于级联效应, 也不可避免地为应用型下游任务引入了噪声.

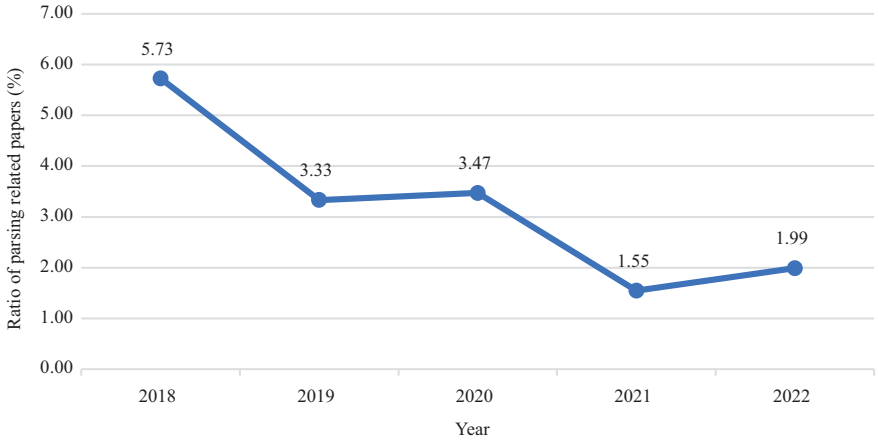

图 1 (网络版彩图) 2018~2022 年 ACL 录用句法分析相关论文占比

Figure 1 (Color online) Ratio of parsing related papers accepted by ACL between 2018 and 2022

3.2.2 结构化预测任务在预训练时代的变化

预训练模型利用大规模无标注语料库进行自监督学习, 从而学习到通用语言知识和表示能力. 随着预训练模型的出现和发展, 传统结构化预测任务受到了严重的冲击. 对于分词任务, 由于预训练模型仍然使用词向量表示文本, 因此分词仍然是预训练模型中重要的预处理步骤, 但为了更好地解决传统分词方法中存在的未登录词问题和数据稀疏问题, 预训练模型中一般都使用 BPE (byte pair encoding) [49], SentencePiece [50] 等方法, 基于训练数据中词片段的出现频率实现词片段级别的分词.

在早期深度学习时代, 人们就对神经网络中词性、句法等信息的必要性展开了探讨 [51]. 而随着预训练模型展现出强大的通用语言能力, 研究者们对模型中显式加入词性、句法等信息的必要性产生了更大的质疑. 一些针对预训练模型进行的探测表明, 预训练模型可以通过大规模的自监督学习从海量文本数据中学习到丰富的语言知识和语言规律, 包括词性、句法和语义等信息 [52]. 例如, Hewitt 等 [53] 发现, BERT 模型输出的词表示向量通过线性变换能反映句中词之间在句法树上的距离. 而 Chi 等 [54] 发现在多语言 BERT 上也存在该现象. 受预训练模型影响, 针对传统结构化预测任务的研究也在减少. 以句法分析为例, 图 1 列出了 2018 年以来历次 ACL (Annual Meeting of the Association for Computational Linguistics) 录用的句法分析相关论文占比, 可以看到, 这类论文数量呈现明显下降趋势.

此外, 自 2022 年 11 月 ChatGPT 被提出以来, 自然语言处理学术界围绕其开展了大量研究, 截至 2023 年 3 月, arXiv 上与 ChatGPT 相关论文共 145 篇, 但其中没有关于词性、句法等结构化预测任务的. 这可能是因为 ChatGPT 采用了生成式模型, 在结构化预测任务上有天生的劣势. 而且其面向的是应用型任务, 从实用主义角度来说如果能使用这种端到端 (end-to-end) 的方法很好地解决应用型任务, 作为中间步骤的结构化预测任务可以省略. 因此, 在大模型时代, 尤其是 ChatGPT 问世后, 词性标注、句法分析等传统结构化预测任务的重要性很可能会逐步降低.

3.2.3 结构化预测任务对于 ChatGPT 的作用

尽管预训练模型和 ChatGPT 在很多自然语言处理任务上表现出了强大的能力, 但这并不意味着传统的结构化预测任务就完全没有作用了. 在一些复杂或特殊的情况下, 这些任务的分析结果还是有一定帮助的. 例如: 在分词方面, 由于中文没有明显的词汇边界, 分词错误可能导致语义理解错误或生

成错误. 在词性标注方面, 由于中文有许多多义词, 词性标注可以帮助确定词语的正确含义和用法. 在句法分析方面, 由于中文语序比较灵活, 句法分析可以帮助确定句子成分和关系.

具体到 ChatGPT, 目前 ChatGPT 在处理某些复杂文本时仍可能因为缺少分词、词性、句法等信息而无法正确理解具有歧义的问题. 例如, 在要求其将 “无线电法国别研究” 翻译为英文时, 由于 ChatGPT 缺乏正确的分词信息, 错误地理解了句子的含义, 导致错误翻译为 “Wireless French Research”. 因此, 当 ChatGPT 面对复杂的问题时, 如果能够调用专门针对结构化预测任务的模型将分词、词性标注、句法分析等结果作为额外信息输入, 可能会提高其效果和稳定性.

3.3 语义分析

3.3.1 语义分析及其典型范式

语义分析是将自然语言表达解析为结构化语义表示的过程. 语义分析可以分为浅层和深层语义分析, 其中浅层语义分析的目标是任务相关的浅层语义结构, 如词义、情感、论点和论元, 深层语义分析的目标是完整的句子语义表示, 如语义解析任务所使用的逻辑表达式、SQL (structured query language)、代码等表示.

语义分析一直是自然语言处理的基础任务 [55], 其核心是如何将表达多样、具有歧义的自然语言表达映射为无歧义的目标语义结构, 上述过程的核心是结构转换和语义映射 [56]. 其中结构转换的核心挑战是自然语言表达的多样性, 也就是同样的语义可以用非常多种不同的句式进行表达; 语义映射的核心挑战是语义表示的领域性, 也就是同样的语义在不同场景和不同的需求下可能会有不同的目标结构, 例如如果底层知识库使用不同的 Schema, 那么同一个句子的 SQL 语义表示会截然不同.

近年来, 随着 BERT, GPT 等预训练大模型的兴起, 大部分语义分析任务都采用了 “预训练大模型 + 任务微调” 的典型范式 [57], 大量工作聚焦于如何设计和高效编码目标端语义结构 [58], 在编码解码过程中加入语义知识和领域知识的引导 [59], 以及在稀缺资源环境中如何高效地构建语义分析模型 [60].

3.3.2 大模型背景下的语义分析

随着 ChatGPT 和 GPT-4 等模型的出现, 语义分析任务受到了多方面的冲击和影响, 包括任务架构的特色性、中间任务的必要性, 以及任务知识的有效性.

(1) 任务架构的特色性. 传统的语义分析研究仍然聚焦于根据不同任务的自身特点设计特色架构, 例如在语义解析任务中, 根据不同目标语义表示 (如 SQL、λ 表达式、AMR (abstract meaning representation) 等) 的特性设计不同的解码模型以充分利用这些语义表示的自身结构特性. 而 ChatGPT 的出现证明了可以在一个通用的生成式架构下完成自然语言处理的诸多不同任务. 与任务特定架构相比, 单一架构构建简单, 可以充分利用任务的多样性, 实现任务间知识的高效共享, 对训练数据的利用更加充分, 从而极大地冲击了传统语义分析的特色架构理念.

(2) 中间任务的必要性. 语义分析任务的主要目的是得到计算机理解语言的中间表示, 如词义消歧、情感分析和论点提取等任务, 这些中间任务的必要性会因为大模型端到端语言理解和生成能力的增强而迅速下降. 例如, 在对话系统中, 情感分析的目的往往是为了生成更符合情境的回复, 而 ChatGPT 已经可以根据上下文的情感生成针对性的回复, 那么情感分析作为中间任务的必要性就会下降.

(3) 任务知识的有效性. 传统语义分析研究中的一个关键点是如何充分利用任务本身的相关知识, 例如将 SQL 等语义表示结构、情感相关的词典等信息通过先验, 外在知识等方式融入模型. 而

ChatGPT 的出现证明了只要模型的规模足够大, 预训练足够充分, 通用架构本身足以捕捉和高效利用任务相关的知识, 而无需专门的特殊设计和算法注入.

基于上述讨论, 我们认为在 ChatGPT 等大模型的背景下, 语义分析的未来发展将呈现出架构通用化、任务归一化、能力按需化的趋势.

(1) 架构通用化. 由于 ChatGPT 证明了充分预训练的通用模型加上有监督微调和基于人类反馈的强化学习 (reinforcement learning from human feedback, RLHF) 等方式可以有效地解决自然语言处理的诸多任务, 未来语义分析任务必然越来越多地放弃任务专用架构, 而是转而研究如何将语义分析任务转换为通用任务形式, 并重点解决上述转换过程中的难点, 如语义结构的自然语言化等.

(2) 任务归一化. 考虑到 ChatGPT 等大模型可以直接理解用户的需求, 生成满足用户需求的回复, 当前大量中间语义分析任务将会被归并到统一的任务形式上. 例如, 未来词义消歧、情感分析、论点提取、语义分析等任务都将被统一为像问答这样的任务形式, 同时不同任务间的知识可以高效地共享.

(3) 能力按需化. ChatGPT 展示了提示生成、指令遵循和上下文学习等强大的交互和控制能力, 未来研究将可以实现按需的语义分析. 一方面, 语义分析的目标和要求可以通过提示语和指令方式进行控制, 从而加强语义分析过程的可控性和精准性; 另一方面, 语义分析模型的能力可以通过上下文学习、用户反馈和提示等方式进行习得和校准, 从而进一步提升了语义分析能力的按需习得和任务自适应.

3.4 知识图谱与文本信息抽取

3.4.1 知识图谱与文本信息抽取的任务和特点

知识图谱 (knowledge graph) 是承载知识的一种表示形式, 通常采用图结构表示实体 (包括物体、事件或抽象概念) 及其之间关系. 图结构中的节点表示实体, 边表示实体之间的关系. 知识图谱在自然语言处理中发挥着重要的作用, 为智能问答、语义检索、机器翻译等任务在给定数据外提供额外的知识, 有助于模型的泛化和性能的提升. 研究者们一直致力于构建大规模知识库. 从早期的 Cyc [61], WordNet [62], HowNet [63], FrameNet [64], 到 DBpedia [65], YAGO [66], Freebase [67], Probase [68], ConceptNet [69], 再到近几年出现的 ATOMIC [70], COMET [71] 等, 研究者们试图从语言、百科、事件、规则等不同侧面构建不同类型的知识图谱. 大数据知识工程时代, 获取知识的方式也从人工标注到众包标注, 再到机器自动抽取和知识关联, 知识库规模和覆盖范围不断增大. 目前已经公开的常识库已经达到上亿甚至超十亿、百亿的规模.

文本信息抽取的任务目标就是从海量文本中抽取出特定语义的信息, 是自动获取知识的一种有效手段. 从任务上分, 包含实体识别、实体消歧、实体关系抽取、事件抽取、事件关系抽取等. 从处理的文本颗粒度分, 包含句子级文本信息抽取、篇章级文本信息抽取和语料库级文本信息抽取等. 从方法上, 包含基于模板规则的文本信息抽取方法、基于统计机器学习模型的文本信息抽取方法、基于深度神经网络的文本信息抽取方法, 以及基于预训练语言模型的文本信息抽取方法等.

3.4.2 ChatGPT 对于知识图谱和文本信息抽取的影响

ChatGPT 的成功无疑给知识图谱相关研究带来了巨大的冲击. 无论是用户主观感受, 还是公开评测, ChatGPT 大模型均展现出强大的需求理解和知识推理能力, 众多研究者认为大模型已经从海量数据中习得各类型的知识, 不少研究者开始质疑知识图谱这一研究方法的可行性, 主要包含如下几个方面:

知识表示: 结构化符号表示 vs. 分布式数值表示. 已有知识图谱奉行符号主义, 经典知识表示理论采用规范化定义的结构化符号 (三元组、树、图、谓词逻辑等) 表示不同知识类型 (例如实体、类别、属性、事件、框架、脚本、规则等). 然而这种表示方法面对实际应用时具有一定局限性. 面对不断出现的知识类型, 过于复杂的知识结构极大地限制了知识表示的范围和应用场景. 更严重的是, 对于那些难以结构化定义, 甚至是 "只可意会不可言传" 的知识 (例如常识) 更加无能为力. 这些缺点使得已有知识图谱尽管具有亿级节点和规模, 面向实际应用时仍然面临所刻画知识类型单一、知识覆盖度低等问题. 相对而言, 以 ChatGPT 为代表的预训练大模型奉行连结主义, 其内部数据采用分布式数值表示 (distributed representation) 的形式, 习得的知识内嵌在海量模型参数中. 相较于传统基于结构化符号的知识表示模型, 分布式数值表示能够突破复杂结构知识难以建模的问题. ChatGPT 的成功使得研究者们开始质疑知识是否有必要一定要构建成结构化符号的形式. Minsky 在文献 [72] 中也提到: 知识可以采用规则、框架、语义网等结构化符号表示形式, 也可以采用自然语言、神经网络等形式来表示. ChatGPT 通过语言模型这一任务 "阅读" 自然语言文本就能习得海量知识, 完全不需要知识结构化建模的过程. 对于下游以神经网络为核心的 AI 应用来说, 围绕大模型构建连续型 (分布式表示) 知识库也许是一种更好的知识库新形式?

以信息抽取为主的知识获取方式受到极大的挑战. 传统知识获取主要以信息抽取为主要手段, 依据预设的知识结构, 从结构化、半结构化、非结构化数据中抽取目标知识 (例如实体识别、关系抽取、事件抽取等). 面对不断涌现的场景和领域, 这一方式只能采用 "看到一类, 定义一类, 构建一类" 的模式构建知识库. 这种挤牙膏式知识库构建手段效率非常低下. 其次, 这一类方法受到知识表示的限制, 只能对于已定义的知识类型进行抽取, 对于那些结构难以定义的知识类型毫无办法. 再次, 已有知识抽取过程包含多个中间子任务 (实体/事件识别、实体/事件消歧、关系抽取、属性抽取等), 这种序列化且多步处理的抽取范式往往会引起错误传递和积累问题, 抽取准确率随着知识结构不断复杂而急剧下降, 需要大量人力进行校验, 保证知识准确性的要求. 相对而言, 以 ChatGPT 为代表的预训练大模型采用无监督语言模型学习, 从海量数据中通过语言模型任务习得参数化的知识, 知识获取完全在连续数值空间进行, 有效减缓了上述基于信息抽取的知识获取范式的诸多问题. 如果知识表示采用分布式数值表示为主, 那么知识获取方式从信息抽取转为语言模型, 将成为一个可预见的趋势.

3.5 情感计算

3.5.1 情感计算的经典方法

情感计算是自然语言处理领域的重要研究方向之一, 其目标是赋予计算机类似于人一样的观察、理解和生成各种情感表达的能力, 它是一个高度综合化的跨学科领域, 涉及计算机科学、心理学、社会学和认知科学等, 通过多学科的深度融合, 分析人与人交互、人与计算机交互过程中的情感特点, 充分设计具有情感反馈的人机交互环境, 从而使人机交互更自然、亲切和生动. 情感计算的概念自 1997 年提出以来, 其相关应用广泛活跃在社交媒体、电商网站、客服系统、智能语音助手等平台. 近年来, 有关情感计算在精神健康 (如抑郁症检测) 等方面的应用也逐渐兴起.

情感计算涉及多项有挑战性的研究任务, 主要包括:

(1) 情感分类. 该任务旨在将文本情感单元分类为若干个情感类别 (如褒义、贬义、中性). 由于情感分类是典型的自然语言分类任务, 在文本情感计算研究的初期, 学者一般使用基于机器学习的分类算法 (如支持向量机) 配合人工特征提取的方法来进行情感分类. 深度学习时代来临后, 基于深度学习的情感分类算法逐渐涌现出来且地位屹立不倒.

(2) 情感信息抽取. 该任务旨在定义并抽取情感评论文本中有意义的情感信息单元 (如评价词语、

评价对象、观点持有人等). 其目的在于将无结构化的情感文本转化为计算机容易识别和处理的结构化文本, 继而供情感分析上层的研究 (如情感文摘、情感对话) 和应用 (如评论分析、舆情分析) 服务. 情感信息抽取类任务均可看作序列标注问题, 相关经典算法均为基于深度学习的方法. 此外, 情感抽取类任务之间相关度较大, 因此联合抽取建模算法也非常常见和经典.

(3) 情感生成. 该任务是指生成带有情感的文本, 如果说情感分类和情感信息抽取是针对情感的自然语言分析任务, 那么情感生成就是一个从无到有的逆过程, 是一种典型的文本生成任务. 根据需求的不同, 情感生成又可以具体分为用于对话系统的情感回复生成以及情感文摘. 对于情感生成任务而言, 经典算法可总结为如何在 "编码器 – 解码器" 框架中更自然地融入情感因素.

3.5.2 大模型时代的情感计算

ChatGPT 凭借使用了丰富语料的预训练和庞大的参数规模, 在众多自然语言处理任务上取得了巨大成功. 很多学者也在情感计算诸多任务上, 将 SOTA 算法和 ChatGPT 进行了对比, 不难发现 ChatGPT 在情感计算的很多任务上已经取得了出色的成绩, 那么在大模型时代情感计算该何去何从呢?

更具象化的情感对话技术. 目前 ChatGPT 无论在情感对话理解能力还是生成能力方面, 都展现出了令人惊艳的效果. 虽然其情绪识别的能力仍未超过在特定数据集上微调的模型 [73], 但已能够产生高质量的情感回复, 表达人类的共情能力和提供有效的情绪支持.

对于该领域未来的研究方向, 有以下 5 个方面值得探索: (1) 情感对话的研究要深入垂直领域 [74], 与心理学、社会学等人文社会科学深度结合, 引入更多更准确的相关理论概念, 解决垂直领域的实际问题; (2) 情感对话安全性的研究, 尤其是在对用户进行情绪以及心理疏导时, 要保证生成的回复内容符合人类价值观; (3) 研究合理的自动评价指标 [75], 与人工评价对齐, 以高效全面地评估情感对话系统的性能; (4) 在使用 ChatGPT 进行情绪预测时, 往往无法发挥其真实的性能, 造成预测效果偏低. 那么如何使得大模型自身的预测标准, 与实际下游应用需求的标准对齐, 也是一个值得探究的问题; (5) 提升模型的定制化、个性化情绪表达方式, 与更多的用户群体对齐 [76]. 例如, 在表达人类共情能力时, 目前 ChatGPT 的生成内容太过于按照模板且套路化, 这样的共情表达方式显然还未能够与真实的人类共情对话所对齐.

更高融合度和更高效率的多模态情感计算技术. 目前对于多模态情感计算而言, ChatGPT 与 GPT-4 带来的冲击比文本情感分类任务较小. 目前 ChatGPT 仅支持文本输入, 而 GPT-4 除了文本还可以接收单张图片, 但仍无法将语音作为输入. 此外, 直接处理视频类数据有两个难点, 第一点在于数据量较大, 直接使用预训练模型处理音频或者其对应的特征 (语谱图) 都需要大量的计算. 第二点在于如何融合异构数据带来不同的语义信息.

基于此未来方向包括: (1) 基于预训练模型思路: 建立多模态预训练模型. 当前模型如 ChatGPT 和 GPT-4 暂时还无法直接处理音图文三模态数据, 因此需要设计高效的多模态预训练模型架构, 利用无监督数据, 训练多模态预训练模型, 使其可以直接应用到音图文情感分类任务上; (2) 基于预训练模型思路: 扩展现有语言模型 (基于 API), 如 GPT-4, 将语音预训练模型、视频预训练模型、声文预训练模型作为扩展工具, 使得 GPT-4 具有理解时序数据, 理解音频数据的能力, 进而利用预训练模型的信息整合能力进行多模态信息整合; (3) 基于预训练模型思路: 融合多个单模态预训练模型, 学习细粒度多模态表示, 挖掘信息中的多模态情感语义, 对多模态数据进行情感分类; (4) 基于事先提取的模态特征: 设计高效、少参数的多模态融合模型, 一方面解决异构语义空间对齐问题, 另一方面使得不同模态信息之间进行充分融合; (5) 数据集: 当前数据集数据大部分依靠单模态信息就可以解决, 很难基于其

研究多模态信息组合问题, 因此构建相关数据集是一个重要的研究方向.

与领域深度交叉的情感计算技术. 大模型的出现对于通用人工智能技术有了很大的助力, 然而对于具体的应用领域却仍然捉襟见肘. 因此, 现如今大型语言模型的出现, 对深入到领域的情感计算技术而言, 发展机遇大于挑战. 以精神健康领域为例, 大模型背景下, 未来从计算技术角度入手的探索性工作会变少, 甚至会消失, 而专注于用成熟大模型技术结合心理学理论去解决精神健康实际方面的问题反而会更多. 像一些比较泛化的任务, 比如对话抑郁检测这种, 可能不再需要精巧的计算设计.

对于该领域未来的研究方向 (以情感计算和精神健康交叉为例), 有以下 3 个方面值得探索: (1) 复杂任务与大模型结合, 比如心理咨询. 之前由于心理咨询需要长时间建模, 且需要丰富的专业知识, 之前的对话系统没有能力介入, 而大模型技术给了这种复杂场景应用的机会. 该方向属于大模型带来的机遇, 使得不可能成为可能; (2) 面向心理学的人工智能技术: 运用计算技术去验证心理学结论和解释, 甚至发现新的心理学结论. 比如心理学上对抑郁症的研究结论往往基于有限的样本 (十几个或几十个), 能不能通过计算技术得到证实? 通过分析大量数据, 能否发现心理学现象的新解释? 该方向是与大模型无关的, 属于 AI for Science 的延伸; (3) 大模型辅助的精神健康研究: 由于精神健康领域有些数据本身具有时间跨度长、信号弱、领域特征强的特点. 大模型很难覆盖掉这些内容, 必须要由特定设计的模型处理. 虽然如此, 大模型可以提供更强的知识与分析推测能力, 可以提供优秀的辅助工具.

3.6 文本生成

3.6.1 文本生成及其主流方法

文本生成任务是自然语言处理领域的一个重要研究方向, 涉及到从输入数据生成合成文本, 如文章、对话、摘要等. 文本生成技术的应用广泛, 包括聊天机器人、新闻撰写、智能助手等. 在过去的几十年里, 研究者们提出了许多有效的方法来实现文本生成, 主要包括:

(1) 基于规则的方法 [77]**.** 早期的文本生成方法主要依赖于手工编写的规则和模板. 虽然这种方法具有一定的可控性, 但可扩展性较差, 难以应用于复杂的生成场景.

(2) 基于统计的方法 [78]**.** 随着统计学习方法在自然语言处理领域的广泛应用, 基于统计的文本生成方法逐渐流行起来. 这类方法通过对大量文本数据进行统计分析, 计算词汇之间的条件概率, 从而生成新的文本. 例如, n 元语言模型和隐马尔可夫 (Markov) 模型就是这类方法的代表.

(3) 基于神经网络的方法 [8]**.** 近年来, 深度学习在自然语言处理领域取得了巨大的成功, 基于神经网络的文本生成方法也随之兴起. 预训练语言模型通过在大量无标签文本数据上进行无监督训练, 学习到丰富的语言表示. 例如, BERT, GPT 和 RoBERTa 等模型在各种自然语言处理任务上都取得了突破性的成果. 特别是以 GPT 系列模型为代表的生成式预训练模型, 在文本生成任务上展现出了惊人的表现.

3.6.2 大模型时代的文本生成

作为一种大规模预训练生成式语言模型, ChatGPT 在许多文本生成任务上表现出了强大的能力, 对文本生成领域产生了深远的影响.

首先, ChatGPT 的出现提高了生成文本的质量. 与传统的文本生成方法相比, 基于 GPT 的生成模型能够生成更加自然、连贯和富有创意的文本 [79]. 这主要归功于其在大量无标签文本数据上进行预训练的过程, 使得模型能够学到丰富的语言知识. 此外, ChatGPT 的大参数规模和自注意力机制也使其能够捕捉到文本中的长距离依赖关系.

其次, 在 ChatGPT 的影响下, 文本生成任务的应用场景也得到了拓展. 传统的文本生成方法主要应用于机器翻译、文本摘要等任务, 而 ChatGPT 则能够应对更广泛的场景, 包括问答系统、对话生成、创意写作等. 这使得各行各业都可以从中受益, 如新闻撰写、广告创意、客户服务等领域.

我们分析, 以 ChatGPT 为背景, 文本生成任务的主流方法将聚焦于基于大模型的生成, 其发展趋势包括以下 5 个方面:

(1) 更大规模的预训练模型. 随着计算能力的提高和数据量的增长, 未来预训练模型可能会进一步扩大规模 [80]. 这将有助于模型学习更丰富的语言知识和更精细的语义表达, 从而提高文本生成任务的质量. 但同时, 更大规模的模型也将面临更高的计算成本和能源消耗问题, 因此需要在模型规模与计算效率之间寻求平衡.

(2) 可控性和多样性的提升. 现有的文本生成模型虽然能够生成高质量的文本, 但在可控性和多样性方面仍有改进空间 [81]. 未来的研究可能会着重于开发更有效的生成策略和调节参数, 以实现更好的生成效果和更广泛的应用场景.

(3) 有效减轻生成偏见. 现有的预训练模型可能会在生成文本中产生一些不必要的偏见 [82]. 为了解决这个问题, 未来的研究可能会关注如何在预训练和微调阶段有效地消除或减轻模型的偏见, 提高生成文本的客观性和公正性.

(4) 模型压缩与轻量化. 随着模型规模的增加, 模型部署和使用成本也在上升. 未来, 研究者们可能会探索模型压缩和轻量化的方法, 例如知识蒸馏和网络剪枝 [83], 以降低模型的计算和存储需求, 使其在资源受限的设备上也能高效运行.

(5) 跨模态生成任务 [84]**.** 未来的文本生成任务可能会不再局限于纯文本, 而是涉及多种模态信息的生成, 如图像、音频和视频等. 这将推动跨模态理解和生成技术的发展, 为自然语言处理领域带来更广泛的应用前景.

3.7 自动文摘

3.7.1 自动文摘及其分类

自动文摘是一种利用算法自动实现文本分析, 内容提炼并生成摘要的技术, 主要分为两大类: 抽取式摘要和生成式摘要 [85].

抽取式摘要通过从原文中选取关键句子, 将其组合成摘要. 抽取式摘要的主要优点是生成的摘要通常较为简洁, 且能保留原文的信息 [86]. 然而, 这种方法可能会导致摘要的连贯性不佳, 尤其对于句子之间的关联性较弱的情况. 传统的抽取式摘要方法包括基于词频、图论的方法等.

生成式摘要通过生成全新的句子来表达原文的核心思想. 与抽取式摘要相比, 生成式摘要具有更好的连贯性和可读性 [87]. 然而, 这种方法可能会引入错误的信息或产生偏离原文主题的摘要. 近年来, 随着深度学习技术的发展, 生成式摘要已经取得了很大的进步.

3.7.2 ChatGPT 背景下的自动文摘

ChatGPT 具备自动文摘生成能力, 且相比原有方法具有很大优势, 其原因在于: 首先, ChatGPT 在预训练过程中学习了广泛的知识, 使其在特定领域的摘要生成任务中具有更好的效果. 这种知识迁移能力降低了领域特定的摘要生成模型的训练成本和难度, 进一步提高了摘要生成技术的泛化能力. 其次, ChatGPT 的预训练过程包括无监督的自回归训练, 而其微调过程则采用基于指令的有监督学习. 这种结合使得模型在预训练阶段能够充分挖掘文本数据中的潜在信息, 而在微调阶段则能够对特定任

务进行优化, 提高摘要生成的质量. 最后, 虽然神经网络模型通常被认为是黑箱模型, 但 ChatGPT 采用的注意力机制在一定程度上提高了模型的可解释性.

由于其庞大的参数量和复杂的结构, ChatGPT 在摘要生成任务中可能仍存在一定的局限性, 如引入错误信息等. 在 ChatGPT 的背景下, 未来的自动文摘技术将聚焦以下 5 个方面:

(1) 多模态摘要生成 [88]**.** 随着多模态信息处理的普及, 未来摘要生成任务可能不再局限于纯文本数据, 而是融合了图像、视频和音频等多种信息. 通过整合各种类型的数据, 未来的摘要生成系统可以更好地理解原始信息并生成更丰富的摘要内容.

(2) 更强的领域适应性 [89]**.** 虽然 ChatGPT 在多个领域表现出色, 但某些领域可能需要特定的知识和技能. 为此, 未来摘要生成任务将更加关注领域适应性, 通过特定领域的预训练和微调, 以生成更准确和可靠的领域相关摘要.

(3) 可解释性和可控制性的提高. 尽管 ChatGPT 在一定程度上提高了模型的可解释性, 但仍有进一步提升的空间. 未来的摘要生成任务可能会更注重模型的可解释性和可控制性, 以便让用户更好地理解生成摘要的过程, 并根据需求调整生成摘要的风格和内容.

(4) 个性化摘要生成. 未来的摘要生成任务可能会更加关注个性化需求. 通过对用户的阅读习惯, 兴趣和背景知识的建模, 摘要生成系统可以生成更符合用户需求的摘要, 从而提高用户体验.

(5) 摘要评价方法的完善 [75]**.** 当前的摘要生成评价方法主要依赖于与人工摘要的相似性, 这种方法可能无法充分衡量摘要的质量. 未来的摘要生成任务可能会探索更为全面和准确的评价方法, 以便更好地衡量生成摘要的质量和实用性.

3.8 机器翻译

3.8.1 机器翻译及其特点

机器翻译 (machine translation, MT) 是指利用计算机实现从一种自然语言到另外一种自然语言的自动翻译. 过去几十年, 机器翻译的主流方法已经从基于规则的机器翻译、统计机器翻译, 迁移到神经机器翻译. 现有的神经机器翻译方法主要采用以 Transformer 为代表的编码器 – 解码器模型, 基于大规模双语对照的训练数据学习源语言句子到目标语言句子的映射. 在此基础上, 机器翻译面向特定场景和需求, 逐渐拓展至语音翻译、图像翻译、视频翻译、多语言翻译和低资源小语种翻译等任务. 无论是哪一种具体翻译模式, 当前的机器翻译从数据准备, 建模到应用具有如下特点:

(1) 数据准备阶段. 需要大规模 “输入 – 正确输出” 匹配的监督数据, 例如语音翻译中需要源语言语音和对应目标语言的文本这样的标注数据.

(2) 建模阶段. 模型几乎都是采用编码器 – 解码器架构, 其中编码器用于编码源语言语义, 解码器用于解码出目标语言文本.

(3) 训练阶段. 重生成轻理解, 即更加关注目标语言的生成, 不重视生成过程是否基于源语言文本的语义理解.

(4) 评价和优化阶段. 依赖与人类一致性较差的 BLEU (bilingual evaluation understudy) 等自动评价指标, 现有机器翻译模型几乎都采用 BLEU 等自动评价指标进行评价和优化.

(5) 应用阶段. 几乎都是静态使用方法, 即除了模型接收用户待翻译文本并输出翻译结果之外, 用户和模型之间无任何其他形式的交互.

3.8.2 ChatGPT 对机器翻译的影响

ChatGPT 作为一个通用语言处理模型, 对传统机器翻译将可能产生颠覆性影响, 具体表现包括:

(1) 不再纯粹依赖双语对照数据. ChatGPT 主要采用无监督预训练方式, 不再区分双语数据或者单语数据, 也不再区分语言种类, 这种方式极大降低了数据获取的成本, 并且可以学习面向多语言的翻译能力.

(2) 单向生成式解码器框架. ChatGPT 的核心模型是单向生成式解码器, 实际就是基于文本前缀预测下一个词语, 可以利用任何类型的文本数据 (包括代码数据等) 进行训练, 从而能够学习到更加丰富更加通用的语言知识和模式, 具有更好的语言和领域泛化能力.

(3) 更强的上下文理解能力. ChatGPT 的无监督学习方式和巨大的上下文窗口使其可以学习到更丰富的上下文信息, 从而在翻译时具有更强的上下文理解能力, 能够更加准确地翻译含有歧义或复杂结构的句子.

(4) 更具用户满意度的译文结果. 虽然一些对比评测表明即使在资源丰富的汉英两种语言上, ChatGPT 在自动评价指标方面相比当前在线翻译引擎并没有优势, 但是从用户的角度, 由于 ChatGPT 表现出更强的理解能力, 生成的译文更容易让用户接受, 翻译结果似乎看起来更加流畅更加准确, 实际表明 BLEU 等自动评价指标在很多情况下无法准确反映译文质量.

(5) 交互方式的革新. ChatGPT 不再局限于仅接收用户待翻译输入然后输出译文结果, 更擅长遵循用户交互意图进行实时修改和更新, 实现了更加拟人的翻译过程.

ChatGPT 的这些颠覆性表现给传统机器翻译研究范式带来了极大挑战, 这种挑战从数据、模型到评价和应用是全方位的. 从数据角度, 双语对照语料将不再是限制机器翻译能力的瓶颈; 从模型角度, 基于编码器 – 解码器的神经机器翻译框架将不再是标准范式, 考虑通用性和对上下文大跨度的历史信息建模, 单向生成式解码器将受到更多关注; 从评价角度, 以 BLEU 为代表的自动评价指标将不再是评价译文质量的基准, 至少不能是唯一的评价基准, 需要进一步考虑人工评价或者与人工评价一致性较高的评价方法; 从应用角度, 静态输入输出的交互方式将不再是主流, 能够动态修正不断完善的交互方式有望成为更多用户的选择.

从目前大模型发展的角度, 机器翻译的未来技术趋势并不明朗. 基于从统计机器翻译到神经机器翻译的迁移历程, 可对未来机器翻译发展趋势尝试进行猜测, 主要分为 3 个阶段, 每个阶段持续的时间随着技术的发展可能会具有较大差异. 首先是神经机器翻译主导阶段, 基于编码器 – 解码器框架的神经机器翻译在未来几年还将是主要研究范式之一, 特别是在当前百亿千亿大模型对计算资源依赖非常强的前提下, 只有一亿参数规模左右的神经机器翻译在很多场景下仍然是主流选择; 其次是神经机器翻译和大模型的混合阶段, 大模型的一些要素, 例如数据、模型、评价和应用方式, 将逐步影响并融入神经机器翻译模型的研发, 两类方法相互渗透; 最后是大模型主导阶段, 随着大模型对硬件资源依赖性的降低, 大模型将成为大多数研发团队可以掌控的技术, 基于大模型的机器翻译研发将成为主流.

3.9 对话系统

3.9.1 对话系统的经典方法

对话系统是自然语言处理中的经典且最为困难的任务之一. 从任务类型来看, 可以典型地区分为任务型对话系统和开放域对话系统, 前者以完成任务为目标, 类似助理, 包括订餐、订票、订宾馆之类的传统任务; 后者以社交和情感为主要目标, 通过不限领域不限话题的开放域对话实现陪伴和支持的功能.

(1) 任务型对话系统. 通常分成两种技术流派: 一种是流水线型的系统, 主要包括意图理解、对话管理、对话生成、领域知识库等模块, 每个模块解析的结果通过流水线作业的方式串联起来. 端到端的对话系统则将系统看做黑箱, 给定上文直接生成回复, 中间也有一些任务型的特定处理, 比如查询知

识库的结果, 中间得到的信念状态. 端到端的对话系统的兴起得益于预训练模型的成功.

(2) 开放域对话系统. 由于不限领域和话题, 开放域对话系统一般被建模为给定上文直接生成回复的条件生成过程. 早期在生成模型兴起之前, 一般广泛采用基于检索匹配的方法, 即根据用户输入从一个大的对话语料库检索得到相似对话的回复; 循环神经网络和 Transformer 模型流行后, 开始采用生成方法, 即构建大规模的对话语料训练生成模型. 在开放域对话系统中, 一般通过精细建模知识、情感、个性, 以实现更类人的回复生成.

3.9.2 大模型时代的对话系统

基于 Transformer [90] 的生成模型给类人对话系统的研发带来了新的曙光, 而 ChatGPT 之类的大规模预训练模型则把类人对话系统推向了新的高度. 在整个发展过程中, 开放域闲聊系统其实更占据主流: 2021 年谷歌研发了 Meena [91], 2022 年推出了 LaMDA [13]; Meta 持续研发了 Blender 1/2/3 [92~94]; 中文方面, 百度研发了 Plato 1/2/XL/K [95~98], 清华 CoAI 研发了 CDial-GPT [99], EVA [100, 101], OPD[6)]等. 这些模型都是采用专门收集的对话数据进行训练. 工业界成熟的产品方面, 国外有个性化角色化的对话平台 Character AI, 国内有聆心智能研发的 AI 乌托邦. 在通用任务助理这条线上的主要突破, 才是 2022 年 12 月份的事情, 包括 OpenAI 推出的 ChatGPT、Anthropic AI 推出的 Claude、谷歌推出的 Bard、百度推出的文心一言、复旦大学推出的 Moss、智谱推出的 ChatGLM 等. 在研究和工业发展上, 大模型时代的对话系统研究呈现了新的特点和新的趋势:

(1) 专门的对话数据训练大模型 vs. 通用模型为底座的少量数据精调. 在过去, 开放域对话系统的训练基本上都从零开始用大量的对话数据训练, 比如 Reddit、微博评论等数据. 但在大模型时代, 以通用模型为底座, 辅以少量对话数据进行有监督精调, 从人类反馈中强化学习变成一个新的主要训练方式. 这得益于大规模预训练模型强大的跨领域, 跨话题泛化能力, 只需要少量的数据就可以触发模型的能力.

(2) 任务型对话和开放域对话在技术架构的统一. 过去两种类型的对话系统采用了很不同的技术路线和架构. 大模型时代, 这两种技术路线开始趋同. 尤其是 ChatGPT 所展示的惊人意图理解、上下文建模、语义解析、API 插件能力, 说明在一个端到端的生成框架中, 完全可以做到任务型对话所需要的深度语义理解能力. 这给统一两种对话任务提供模型和架构的基础.

(3) 任务助理和情感社交的统一. ChatGPT 定义为通用任务助理, 设计定位为完成任务, 遵循指令的机器, 有很强的机器属性, 没有情感和同理心. 但作为通用人工智能时代的对话智能体, 除了满足人的信息需求 (功能层面、提高生成力、提升创造力等), 还需要满足人的情感社交的需求 (情感层、需要共情、陪伴、支持等), 即具有拟人的特征. 因此, 未来需要统一功能特征和拟人特征, 才能真正打造类人的对话智能体.

(4) 可控、可配、个性化的对话系统. ChatGPT 展现了一个千人一面的任务助理, 但对系统行为进行个性化配置以适应不同的场景, 需求和用户也非常重要. 对话系统的可控、可配、个性化是未来重要的研究方向: 可控, 指的是生成内容可控, 行为可控; 可配, 指的是用户可以通过少量的 "参数" 输入即可以定制特定的对话系统; 个性化, 指的是可以配置对话系统本身的角色和个性, 以及根据用户的不同, 动态自适应地调整对话系统的行为.

3.10 信息检索

信息检索旨在针对给定的用户查询, 从现有的资源库 (通常为自然语言文本) 中寻找相关的信息

6) http://coai.cs.tsinghua.edu.cn/static/opd/posts/opd_blog/.

资源返回给用户, 以满足用户的信息需求. 作为克服信息过载最重要的技术之一, 信息检索系统已经被广泛应用于多种下游应用任务, 成为人类获取知识信息的重要途径 [102]. 近期, 以 ChatGPT 为代表的大型语言模型 [103] 的出现对于信息获取技术产生了重要影响, 为传统的信息检索研究范式以及应用任务带来了新的契机和挑战.

3.10.1 信息检索发展历程及主流方法

信息检索有着悠长的研究历史, 最早可追溯到 20 世纪 50 年代. 在信息检索发展的早期, 最常用的方法是基于稀疏向量的词项匹配方法, 将查询和文档同时表示为稀疏向量, 并通过倒排索引进行检索, 典型的方法包括 TF-IDF [104] 和 BM25 [105] 等. 随着机器学习技术的发展, 研究者也将这一方法引入了信息检索领域, 基于学习排序 (learning-to-rank) 的检索方法通过设计基于特征的排序函数引入多种人工特征 [106]. 进入到深度学习时代, 基于表示学习的检索方法受到了广泛关注, 核心思想是将查询与文档投影到隐含空间进行表示, 计算二者在隐含空间的语义相似度 [107], 从而消除手工提取特征的步骤. 通常来说, 表示学习方法的效果受限于底层神经网络的性能. 近年来, 预训练语言模型逐渐成为自然语言处理领域的主流方法, 以此为基础的稠密检索方法能够更为全面有效地表示文本中的语义信息, 通过近似最近邻算法实现高效语义检索, 大幅提升了检索效果 [108]. 最近, 学术界又提出生成式检索模型, 放弃了显式的倒排索引结构, 使用预训练语言模型直接端到端地生成相关文档的词符 [109].

3.10.2 大型语言模型对信息检索领域的影响

近期, ChatGPT 展现了强大的对话问答能力, 给自然语言处理领域带来了巨大的震撼, 未来, 以 ChatGPT 为代表的大型语言模型能够为信息检索领域带来哪些改变, 无疑让人更加期待.

大型语言模型为传统知识密集型任务流水线带来性能提升. 在现有的知识密集型任务中, 信息检索作为重要的信息获取技术, 从现有的资源库中获取相关的背景知识. 例如, 在开放域问答中, 通常采用基于 “召回 – 精排 – 阅读理解” 的 3 阶段流水线 [110]: 先从文档语料库中召回相关文档, 再对候选文档进行重排序, 最后使用阅读理解模型得到精确答案. 在之前的方法中, 这 3 个阶段的任务通常使用基于预训练语言模型的方法完成[7)]. 大型语言模型的出现为改善这一流水线中的功能组件带来了希望. 首先, 大型语言模型可以作为流水线中的阅读理解模型, 通过参考问题和相关文档生成更为精准的答案. 例如, 研究者在检索模型后使用 110 亿参数量的大型语言模型作为阅读理解模型, 在知识密集型任务上超越了以往的流水线问答基线的效果 [111]. 此外, 大型语言模型还可以增强检索模块, 例如, 有研究发现在 “检索 – 阅读理解” 的流水线中, 可以用大型语言模型替代检索模型来生成和问题相关的参考文档, 这种生成的参考文档在部分场景下 (例如开放性问题) 比使用检索模型召回 – 精排后的文档更有可能包含正确答案 [112].

将大型语言模型用于数据增强, 在无监督检索场景下效果明显. ChatGPT 在各类语言任务中取得了显著的效果, 其核心原因是其记忆了广泛的世界知识 [103]. 以往的稠密检索模型通常需要大量的领域标注数据进行训练. 然而, 在很多领域的特定任务中 (如金融、生物、法律等领域), 通常很难获得较为充足的监督数据, 对于这类检索方法的领域迁移带来较大的阻碍. 由于大型语言模型编码了广泛的领域知识, 文本任务泛化性强, 可用作稀疏场景下的数据增强技术. 通过适合的任务指令或上下文提示, 大型语言模型能够在特定领域合成大量的伪标签数据, 从而更好地提升检索模型在标注数据稀缺场景的性能 [113]. 例如, 可以基于指令让 ChatGPT 针对给定查询生成假设文档, 从文档语料库中

7) 召回阶段可以拓展为多路方法的归并 (多采用效率较高的模型), 重排序阶段可以泛化到多个阶段 (重在排序性能). 通常来说, 召回阶段对于效率的考虑更多, 重排序阶段更加关注模型性能.

基于最近邻算法找到表示最相近的真实文档, 进一步构造查询 – 文档数据对, 从而达到和有监督微调相当的效果 [114]; 另一方面, 也可以使用 ChatGPT 基于给定文档生成对应的查询, 进行伪标签数据的构造 [115].

信息检索模型可用于大型语言模型的辅助模块. 大型语言模型通过上下文学习、指令微调等技术统一了多种语言任务的范式, 具有强大的通用任务求解能力, 但仍然存在时效性受限, 生成幻觉等问题 [103]. 信息检索技术可以作为大型语言模型的辅助模块 (或技术), 进而改善上述问题. 例如, ChatGPT 可能对于训练数据时间段以外的问题缺乏相应的知识信息, 通过借助信息检索模块可以为 ChatGPT 补充最新的知识. 最近, OpenAI 已经支持在 ChatGPT 中使用检索插件, 能够访问个人或经过许可的组织信息源, 根据用户的查询需求, 从数据源 (如文件、笔记、电子邮件或公共文档等) 中获取相关文档, 作为 ChatGPT 的信息补充. 幻觉问题也是 ChatGPT 等大型语言模型存在的典型问题之一 [116], 通常表现为输出中经常存在 "事实性错误", 而检索增强的大型语言模型能够有效减轻幻觉. 例如, 使用大型语言模型的输出在知识库中检索相关证据文档, 利用这些证据文档对于不正确的事实信息进行修改, 使模型生成更加符合事实的输出 [117].

3.11 自动问答

3.11.1 自动问答的任务和特点

自动问答系统是验证机器理解语言的一项重要任务. 自动问答系统接受用户以自然语言方式提出的各类问题, 通过检索、匹配、推理等手段, 从不同类型、不同结构的数据中获取准确答案. 相对于对话系统的闲聊功能, 自动问答更加偏向于一问一答式的事实获取和推理. 按答案来源的不同, 可分为检索式问答、知识库问答、表格问答、常见问题回答、社区问答、阅读理解等 [118~120].

检索式问答. 答案来源于海量文本库, 问答系统需要通过 "检索"+"抽取" 的方式, 从海量文本语料中获取答案. 相关的评测包括: Text Retrieval Conference (TREC) (QA track) 和 NII Testbeds and Community for Information Access Research (NTCIR) (跨语言问答评测、cross language QA) 等.

知识库问答. 答案来源于结构化的表格或知识库, 问答系统需要将用户自然语言问题解析成结构化的查询语句 (SQL 或者 SPARQL (SPARQL protocol and RDF query language) 等), 并从结构化数据中查询得到所需的答案. 相关的评测任务如 Cross Language Evaluation Forum (CLEF) 组织的基于关联数据的问答评测 (question answering over linked data, QALD) 等.

社区问答. 类似于常见问题回答 (frequently asked question, FAQ), 答案来源于事先编写好的 "问题 – 答案" 对. 问答系统需要从海量问答对中检索出与当前问题语义最接近的历史问题及答案.

阅读理解. 答案来源于用户指定的单篇文档, 问答系统需要具备强大的语言理解能力, 从给定文本中抽取出相对应的答案. 近些年, 随着给定文档的不同, 也包含了其他一些子任务, 例如: 多文档阅读理解、表格文本混合问答等.

3.11.2 ChatGPT 对于自动问答的影响

ChatGPT 的出现很大程度上改变了自动问答的研究现状, 不仅在问答形式, 而且在技术范式上为传统问答任务带来了新的契机和挑战.

强大的基础模型底座将代替定制化的问答模型. ChatGPT 在各类问答任务上取得显著效果, 其核心原因是基于千亿规模的基础模型底座. 在以往问答任务中, 问答模型往往采用定制化的设计方式, 针对不同问答形式、不同推理任务, 设计与之对应的不同模块, 例如: 外部知识获取 [121~123]、知识推理 [124] 等. 这不但增加了问答系统的复杂性, 也降低了模型在不同问答场景间的迁移能力. 相较而言,

ChatGPT 则不再需要额外的特殊设计, 其背后的大模型已经存储了海量知识 [112], 仅仅需要一个基于语言模型的文本生成任务, 就能实现已有不同类型知识匹配和推理模式的隐式调用. Qin 等 [73] 在多个问答和推理任务上通过对比发现, 在零样本条件下, ChatGPT 在部分数据集上已经超过了微调模型. 在面向知识库的复杂问答方面, Tan 等 [125] 的评测发现 ChatGPT 在仅依靠自身知识的情况下, 在 7 个测试集中的 2 个 (WQSP 和 GraphQuestions) 上取得了当前最好的结果, 这也说明了 ChatGPT 自身具备一定的知识储备, 且能够处理复杂问题. 这些都证明大规模的基础模型底座对于问答系统的重要性.

基于语言模型的问答框架使得不同问答形式的边界将变得模糊. ChatGPT 在问答通用性方面取得了显著的效果, 其核心原因是将不同任务建模成语言模型的统一范式, 这使得各类问答形式的边界变得模糊. 在任务统一的框架下, 得益于大型语言模型愈发强大的知识获取、匹配和推理能力, 知识库问答, 常见问题回答等形式的任务可能不再需要与其他问答任务进行区分, 阅读理解形式的问答 [118]、表格问答 [120] 等任务也不再特殊, 都可以在基于语言模型的问答框架下进行. 例如, 给定篇章或者表格, ChatGPT 可以根据用户的提问, 在候选项中选择答案, 或者从原文进行答案抽取, 甚至是利用表格中的数据直接进行计算. 在赋予 ChatGPT 利用其他外部工具能力 [126] 的情况下, 它能更好地完成各类原本需要特别设计的问答形式. 由此可以看出, ChatGPT 正在模糊了各项问答任务形式的界限, 问答形式复杂多样的现状也将随之改变.

上下文学习, 思维链等新技术将改变问答推理的基本范式, 极大提升问答系统的效果. ChatGPT 在上下文中学习的能力势必激发大模型内部已经具备的各种能力, 使得其问答能力得到进一步增强. 和在各领域上不断微调模型相比, 利用上下文学习, ChatGPT 在特定场景下给出小样本示例就能进行问答, 这种问答模式无疑更让人期待未来问答系统的效果. 思维链 (chain-of-thought, CoT) [127] 是指令示范的一种特殊形式, 它通过引发大型语言模型的逐步推理来生成答案. 利用 CoT 微调的模型使用带有逐步推理的指令数据, 其取得的性能显著优于之前的微调模型. 例如, 在需要复杂推理的 GSM8K 数据集上, 通过设计思维链, 8 个示例的提示就能达到很好的效果, 而微调方法则需要利用完整的训练集 [128], 这种鲜明的对比也增强了人们对于 CoT 等新技术的期待. CoT 技术同时增加了模型输出的可解释性 [129], 逐步推理的过程让大型语言模型给出的结果更加可信.

3.12 小结

通过以上分析, 可以发现, ChatGPT 等大型语言模型, 对文本分类、结构分析、语义分析、信息提取、知识图谱、情感计算、文本生成、自动文摘、机器翻译、对话系统、信息检索和自动问答各种核心的自然语言理解和生成任务均产生了巨大的冲击和影响.

ChatGPT 在大规模预训练过程中习得广泛的语言和世界知识, 处理自然语言任务时不仅能在少样本, 零样本场景下接近乃至达到传统监督学习方法的性能指标, 且具有较强的领域泛化性. 这将激励, 促进研究者们打破固有思维方式的樊篱, 学习、借鉴 ChatGPT 等大模型的特点和优势, 对自然语言处理的主流研究范式进行变革, 进一步提升自然语言核心任务的能力, 例如以生成式框架完成各种开放域自然语言处理任务并减少级联损失, 通过多任务学习促进知识共享, 通过扩展上下文窗口提升理解能力, 通过指令遵循和上下文学习从大模型有效提取信息, 通过思维链提升问题拆解和推理能力, 通过基于人类反馈的强化学习实现和人类意图对齐等.

长期以来, 自然语言处理分为自然语言理解和自然语言生成两个领域, 每个领域各有多种核心任务, 每种任务又可根据任务形式、目标、数据等进一步细分, 今后在各种应用任务的主流架构和范式逐渐统一的情况下, 有望进一步得到整合, 以增强自然语言处理模型的通用性, 减少重复性工作. 另一方

面, 基于大模型的强大基座能力, 针对具体任务进行按需适配、数据增强、个性化、拟人交互, 可进一步拓展自然语言处理的应用场景, 为各行各业提供更好的服务.

4 大模型时代的自然语言处理共性问题

在自然语言处理研究领域中, 除了各种核心任务之外, 还有可解释性、公平性、安全性、可靠性、能耗、数据质量和评价等一些共性问题. 这些问题不是某种任务所特有的, 而是广泛存在于各种自然语言理解和生成任务中. 围绕这些共性问题进行针对性研究, 分析其成因和机理, 设计应对措施, 对确保自然语言处理任务的性能、效率、稳定性和领域适用性至关重要.

大模型自身同样存在着自然语言处理的共性问题, 如模型可控性、多样性、鲁棒性和可解释性仍需提升, 训练和使用成本过高, 语言数据质量缺乏保障, 评价方法单一等. ChatGPT 的一项亮点技术是 “与人类意图对齐”, 其目的除了理解用户意图之外, 还需要拒绝不合理的请求, 给出负责的、合乎人类道德准则和伦理规范的答案. 由于大模型的结构复杂、参数庞大、生成过程难以解释, 生成文本时经常面临幻觉生成、错误知识、前后不一致等问题, 人们对于从系统获取信息的准确性无从感知, 给系统的广泛实际应用带来了极大的潜在风险. 因此, 如何提升模型的公平性、无害性、有益性和鲁棒性, 确保大模型拥有正确的价值观, 保障大模型生成内容的信息准确性变得愈发重要.

随着以 GPT-3 为代表的大模型技术逐渐发展, 模型的参数数量、计算时延、训练所需的资源等都在显著增加. 在语言建模能力不断增长的同时, 模型的计算成本与能耗指标也成为当前大模型成功应用的一大门槛.

大规模高质量文本数据资源在模型的构建过程中扮演了极其重要的作用, 训练数据规模越大, 种类越丰富, 质量越高, 所得到的大规模语言模型的性能越好, 而训练数据中的瑕疵数据, 可能会对模型的表现产生负面影响; 相较于以前的单一类型或少数任务驱动的基准评测, 针对大规模语言模型的评测需覆盖的问题场景范围更广, 复杂度更高, 难度也更大, 需要探索更有效合理的任务评价指标.

总之, 这些由大模型所强化的真实需求, 将极大地加强模型分析和可解释性、伦理问题与安全性、信息准确性、计算成本与能源消耗、数据资源和模型评价等各种共性问题的研究热度.

4.1 模型分析和可解释性

基于深度神经网络的自然语言处理模型在学习过程中主要依赖梯度下降方法, 逐步优化大量非线性计算中的模型参数. 模型性能的提升往往伴随着参数数量的增加 (例如 GPT-3 拥有 1750 亿个参数) 以及网络结构的加深 (通过堆叠非线性算子实现) [2]. 随着模型参数和深度的增长, 模型的决策过程变得越来越难以解释. 模型分析和可解释性技术已经成为模型可靠性、可信性、公平性和安全性的基础支撑 [130]. 一方面, 由于深度神经网络中特征表示和模型以向量和参数的形式存在, 人类难以直接理解这些向量和参数背后的信息和意义. 另一方面, 大模型的参数规模和架构导致人类难以理解其复杂性. 因此, 对模型的特征、机理和过程进行解释, 可以增进人们对模型的理解, 有效避免模型的偏见, 增强模型的鲁棒性和性能.

目前大模型的解释和分析方法可以大致分为以下几类, 包括可视化方法、结构分析方法、行为分析方法和因果干预分析方法. 可视化方法 [131] 的核心是通过可视化特征表示和模型参数中的显著性、关联、规律和趋势, 从而让人快速直观了解模型的工作原理并形成相关假设. 结构分析方法 [132] 通过探测分类器、辅助预测任务、诊断分类器等手段, 分析特定神经网络节点和网络层捕获了什么信息. 行为分析方法 [133] 通过构建针对性的测试数据集, 并通过对测试数据的控制和模型性能的变化来分析模

型背后的机理. 因果干预分析方法 [134] 首先构建模型背后的结构化因果图, 并通过分析和干预结构化因果图来发现模型背后的因果路径以及可能的偏差. 基于上述分析方法, 模型分析和解释可以有效地发现特定决策后面使用的信息, 模型习得和使用的知识, 以及这些知识是如何表示和分布在模型的不同组件中. 同时, 这些方法也可以有效地发现模型可能的偏差, 识别特定类型知识的变化对模型输出的影响, 以及不同的模型架构和优化决策如何影响底层的知识, 从而为模型的公平性和可靠性提供有效的支撑.

模型的可解释性可以分为两个重要的维度, 其一是合理性 (plausibility), 这指的是解释在人类理解层面上的可接受程度; 其二是忠实度 (faithfulness), 这表示解释是否真实地反映了模型的决策过程 [135]. 尽管自然语言处理模型的可解释研究已经取得一些进展, 但仍面临着几个重要的挑战:

缺乏统一且被普遍接受的可解释性的定义和术语. 目前可解释研究的术语还缺乏精确的含义 [136]. 虽然已有的大量公开工作对模型的可解释性进行了研究, 但大多是为了解决某种迫切需求, 缺乏对共性问题的研究. 例如, 模型解释可能是为了解释模型的预测, 也可能是为了理解模型的行为 [137]. 这就导致相关工作无法进行互相对比和借鉴. 此外, 研究上缺少可解释数据集的构建体系, 可解释评测的基准数据集不足, 这使得可解释性研究的进展受到了限制 [138].

缺乏客观完善的评价方法和标准化定量指标. 目前仍缺乏一种被普遍接受的可解释性评估体系, 特别是在自然语言处理领域, 许多解释方法只能依赖于人类的认知来进行定性评估, 难以量化解释方法的性能表现, 难以对同类型的工作进行横向对比 [139]. 除了评估解释的直观性和质量以外, 模型预测的覆盖率 (coverage) 和忠实度也是两个非常重要的维度.

模型的解释性和其他特性之间的关联性缺乏深入的研究. 在设计可解释的深度模型时, 很多工作从直觉上简单地认为模型必须牺牲准确性来换取可解释性, 亦或者可以通过提升可解释性, 带来模型鲁棒性的提升 [140]. 很多已有的研究工作选择先学习一个复杂的黑盒模型, 比如深度神经网络, 然后使用代理模型来提供解释. 然而代理模型的忠实度无法保障, 因为很多时候要么使用一个简单模型来解释一个复杂模型所做的预测, 要么代理模型使用完全不同的推理逻辑来做预测 [141]. 尽管如此, 现实情况是人们仍在使用代理模型, 缺乏自解释方法的探索. 为了建立准确且可解释的模型, 需要深入研究模型的解释性和其他特性之间的关联性.

现有可解释方法的形式仍然有限, 无法满足不同场景的需要. 在自然语言处理研究中典型的可解释方法是通过突出输入特征的方式, 旨在突出阐明 “预测背后的原因” 的特征. 然而, 这种方式仍然存在很大缺陷: 对于算法工程师而言, 突出显示输入文本中的片段可能容易接受, 其可以了解结果决策依据并调试算法的相关信息 [142]. 而对于普通用户而言, 他们可能无法理解规则, 或者可能期望获得更详细可读的信息, 以充分理解模型的输出. 相比较来说, 在呈现解释的多种方式中, 流畅的自然语言尤其具有吸引力, 因为它方便不同背景的用户对模型进行理解, 而不需要数学和信息科学的复杂背景.

同时, 随着 ChatGPT 和 GPT-4 模型的出现和发展, 模型的可解释性和分析研究也将面临新的机遇和挑战. 一方面, 当前的类 ChatGPT 模型仍不完美, 经常面临幻觉生成、错误知识、前后不一致、刻板偏见等问题, 随着 ChatGPT 等模型被用作越来越多智能应用的基座, 任何一个小的偏见、错误和隐私信息泄露都可能会导致极大的风险, 同时模型也可能会面临方方面面的攻击, 如何基于可解释性技术和模型分析技术保障模型的公平性、无害性、有益性和鲁棒性变得愈发重要. 另一方面, 类 ChatGPT 模型使用了许多新的学习方法, 涌现了诸多之前未见的能力, 如思维链推理能力、上下文学习能力、指令遵循能力等, 如何理解和解释这些能力的成因和机制, 对于后续相关模型的能力学习和能力使用, 都是至关重要的新问题. 最后, 类 ChatGPT 模型强大的对话交互能力和自解释能力为大模型的可解释性研究和分析研究提供了新的手段, 使得模型的交互式自然语言解释和分析成为了一个新

的前沿.

4.2 伦理问题与安全性

近年来, 随着深度神经网络研究的不断深入, 特别是 BERT, GPT 为代表的大规模预训练语言模型的广泛应用, 基于深度神经网络的自然语言处理算法在各项任务的评测集合上都取得了非常好的效果. 在一些任务上, 算法的准确率甚至已经超越了人类. 但是不断暴露出来的一系列风险, 也开始引发人们对自然语言处理算法产生信任危机.

考虑以上因素, 在 ACM 于 2018 年发布 "ACM 伦理准则与职业规范"[8), 要求算法必须为人类社会做贡献、避免伤害、诚实正直、公平、无偏见、保护隐私之后, ACL 立即接受了以上规范, 并鼓励对语言处理任务的伦理问题展开研究. 随即, 伦理问题成为了自然语言处理的新兴研究领域之一, 并于 2020 年首次成为了 ACL 会议的投稿主题.

数据偏见与公平性. 发表于 2016 年 NIPS (Annual Conference on Neural Information Processing Systems) 的工作 [82] 发现由于训练语料中的偏见导致训练后的词向量具有性别歧视, "man" 相较于 "woman" 与 "honorable (可敬的)" 的语义距离更近, 而 "woman" 相较于 "man" 则与 "submissive (顺从的)" 的语义距离更近. 2018 年发表于 PNAS (Proceedings of the National Academy of Sciences of the United States of America) 的论文 [143] 使用 20∼21 世纪不同时期的语料训练词向量, 也同样发现不同时期人们对待性别和种族的文化差异, 通过不同时期的语料训练, 从而清楚地反映在词向量中. 2015 年美国芝加哥法院使用的犯罪风险评估系统 COMPAS, 被证明对黑人存在歧视. 黑人被该系统错误地标记为具有高犯罪风险的可能性是白人的 2 倍, 这种错误很可能会使得黑人被法官判处更长的刑期. 上述偏见问题很大程度上都由训练数据中存在的偏见所导致. 绝大多数情况下, 数据偏见都会导致算法受到影响, 从而使得依赖算法的决策形成偏见. 模型公平性等问题在词性标注 [144]、对话系统 [145]、机器翻译 [146] 等各类自然语言处理任务中也十分普遍. 数据驱动的大模型, 仍然存在上述现象. 大模型的训练依赖大量的训练数据, 以 GPT-3 的训练为例, 使用了从 45TB 数据中抽取的超过 500G 数据 [2], 因此针对大模型的公平性和偏见消除仍然是亟待解决的问题.

隐私保护. 随着模型参数量的不断扩大, 深度神经网络模型越来越难直接在手机端侧以及本地服务器运行. 因此, 用户在使用该类服务时, 需要将所有数据提交给云端模型. 这就提高了用户数据, 特别是隐私数据的泄漏风险. 现有的一些研究, 试图从同态加密、输入融合等层面解决云端模型的隐私问题, 但是目前这些研究还是面向 300M∼700M 参数的相对小规模的模型, 在特定任务上开展. 大模型的隐私问题, 相较于之前的小模型更加突出, 现有方法不仅无法在移动设备上运行, 在绝大部分的小规模服务器上也很难应用. 因此, 针对大模型的用户隐私研究是大模型在更广泛领域应用所不可或缺的部分.

大模型的安全性问题. 虽然在 ChatGPT 和 GPT-4 等系统出现之前, 对于大模型安全性的研究一直受到重视, 但 ChatGPT 和 GPT-4 的出现使得这个问题真正地变得迫在眉睫. 因为在 ChatGPT 之前, 语言模型或对话模型实际并没有获得大多数人的信任, 这些模型很难介入到真实人类世界中, 人们也很少真正将它们作为可信赖的工具去使用. 但是 ChatGPT 表现出的高可用性切实扭转了人们的观念, 这些模型开始真正介入到人类社会生活中了, 它们所表达、所传播的价值也很可能会影响到人的观念, 因此对它们安全性的讨论十分重要.

由于大规模语言模型见过海量文本并且学会了人类语言的表达, 这使得它不仅可以产生具有创造性的优质内容, 也可以产生具有强大危害的不安全内容, 如仇恨言论、暴力言论、歧视言论等. OpenAI

8) https://www.acm.org/code-of-ethics.

在研发 ChatGPT 和 GPT-4 等系统时着重考虑了这个问题, 从众多方面努力, 试图缩减其中的有害信息, 并在文本生成阶段加入无害信息判断. 从已有的一些信息看, OpenAI 的做法可能综合了以下 6 种: (1) 从预训练数据中去除有害的文本内容 [1], 如规模庞大的色情内容等; (2) 通过指令微调方法约束模型对不安全问题的回复 [76], 避免模型产生有害回答; (3) 通过奖励器给模型回复进行偏好评分, 并通过强化学习对安全内容加以奖励; (4) 通过分类器给模型回复进行基于安全规则的评价, 并同样使用强化学习方法; (5) 通过人类专家进行对抗测试, 不断地找出安全薄弱环节, 不断地迭代加强模型; (6) 线上实时检测生成回复的有害性, 若意外生成有害回复则提示用户或终止回复. 基于以上多种措施, OpenAI 成功地将 ChatGPT 和 GPT-4 等系统的安全性提升到了十分可观的程度, 在非专业对抗场景下几乎不会生成有害回复, 极大地缓解了安全问题.

未来大模型伦理和安全性的研究, 一方面会集中于被研究较多的伦理方面的问题, 如有害言论、歧视言论等; 另一方面则有可能会更多地关注价值观方面的问题. 前者属于清晰、明确、特征较显著的问题, 而后者则是模糊、隐晦、特征更不显著的问题. 对于价值观安全性的研究, 可能会涉及到意识形态等多方面的内容, 甚至会涉及国家政治安全基础, 具有十分重要的研究意义. 此外, 由于目前所采用的基于 Transformer 的生成式模型依赖训练语料, 可解释性较为缺乏, 对于控制大模型安全性的理论研究和方法研究也会变得十分重要, 现有方法无法保证一定不会生成不安全内容, 而且也没有理论可以确切解释什么样的方法可以严格保证模型安全性, 因此如何从源头的训练数据阶段, 以及在模型输出阶段控制有害内容的生成都是十分值得探索的方向.

4.3 大模型的信息准确性

通过对话的方式获取信息最为贴近人类的习惯, 有望成为人类信息获取的新范式. 虽然对话式语言大模型已在对话内容理解和语言生成方面取得了令人印象深刻的表现 [147], 但由于大模型的结构复杂、参数庞大、生成过程难以解释, 人们对于从系统获得信息的准确性无从感知 [148], 信息的准确性无法得到充分保障, 给系统的广泛实际应用带来了极大的潜在风险 [149]. 因此, 如何保障对话式语言大模型生成内容的信息准确性已成为其广泛应用落地亟须解决的核心科学问题和难题之一. 2023 年 2 月 *Fortune* 杂志的封面文章中, OpenAI 首席技术官 Mira Murati 明确指出: “我们到目前为止一直遵循的研究方向, 目的是解决模型的事实准确性和可靠性等问题. 我们正在继续朝着这些方向努力.”[9)]

人们在使用大模型获取信息时, 经常会遇到机器 “一本正经地胡说八道”, 其中事实错误、内容空洞、逻辑混乱是最为典型的问题 [92]. 图 2 展示了用户与 ChatGPT 对话的例子, 从中可以看出, ChatGPT 准确理解了对话内容, 并进行了流畅回复, 展示了其强大的语言理解和生成能力. 但是其回复内容中存在诸多的信息准确性错误. 图中红色字体部分为事实性错误, 例如, 目前并没有中国人获得诺贝尔经济学奖, 而 ChatGPT 却给出两位经济学家, 且这两位经济学家均不是中国人, 后续关于陈冯富珍的回答也是完全错误的; 蓝色字体标出了逻辑性错误, 例如, 人民没有粮食吃、不去吃肉、通常是因为肉类比粮食更为稀缺、粮食短缺通常说明食物短缺, ChatGPT 使用了错误的推理逻辑, 从营养均衡的角度给出了错误的回答. 当模型错误地使用了知识或使用了错误的知识, 输出了语言层面流畅的回复, 如果这时缺乏相应的机制对信息的准确性进行检验和修正, 那么模型的语言能力越强大, 用户则愈发难以甄别其中的信息准确性风险 [76]. 孔子说: “知之为知之, 不知为不知, 是知也”. 一个可信赖的对话式语言大模型同样应该具有相应的机制对其生成内容的准确性进行检验, 以保障人们获取信息的准确性.

大模型参数规模动辄数百亿甚至更多, 结构复杂, 训练所需的海量数据大多是从互联网获取的用

9) https://www.fortunechina.com/shangye/c/2023-01/31/content_426829.htm.

图 2　(网络版彩图) 用户与 ChatGPT (Feb 13 Version) 聊天的示例

Figure 2　(Color online) One Chinese example of user chatting with ChatGPT (Feb 13 Version) and its English translation

户产生内容, 缺少实时信息, 内容质量参差不齐 [13]. 另一方面, 当前的系统大多基于封闭世界假设, 系统在预先收集的语料上进行训练, 部署使用中不做更新. 而我们所处的真实世界是开放复杂多变的, 新的知识不断涌现, 旧的知识不断更新 [150]. 这种做法不可避免地导致系统准确性问题. 例如, ChatGPT 的训练数据是 2021 年之前的互联网数据, 虽然 OpenAI 设计了相关机制, 对于明显的实时问题拒绝回答, 但是对用户提出的一些没有明显表明的实时性问题, 仍然会产生存在内容空洞或事实性错误的回答, 如图 2 所示, 对于 2023 年上映的电影《满江红》的回答中, ChatGPT 并不具有这方面的知识, 但其仍然进行了回答, 给出了事实完全错误的信息. 相关研究表明, 人类 70% 的知识是在不断实践和交流中获取和更新的 [151]. 如何根据实时多元证据, 对机器生成的不准确内容进行修正, 使系统生成准确内容的能力持续提升, 将具有重要的理论和应用价值.

4.4　计算成本和能源消耗

随着以 GPT-3 为代表的大模型技术的发展, 模型的参数数量、计算时延、训练所需的资源等都在显著增加. 据研究统计 [152], 1760 亿参数的语言模型 BLOOM [153] 的完整训练需要排放大约 24.7 吨二氧化碳, 而 GPT-3 一次完整训练的二氧化碳排放量达到了惊人的 502 吨, 提高了 20 倍. ChatGPT 需要在 GPT-3 的基础上做进一步的指令微调和人类对齐等操作, 还将消耗更多的计算资源并排出更多的二氧化碳. 另外, 微软在搜索中加入 ChatGPT 这类生成式 AI, 会导致每次搜索至少增加 4~5 倍的计算量. 根据国际能源署 (International Energy Agency) 的数据, 数据中心的温室气体排放量已经占到全球温室气体排放量的 1% 左右. 因此, 在模型的理解和生成能力不断增长的同时, 模型的计算成本与能源消耗成为当前大模型成功应用的难点之一.

模型架构效率问题. 当前, 绝大多数的预训练语言模型都采用 Transformer 作为模型框架的骨干, Transformer 结构通过自注意力机制缩短了文本中词语依赖的建模路径, 进而增加了模型在大规模语料上的文本建模能力. 然而基于 Transformer 的预训练模型, 其计算时间复杂度和空间复杂度是和输

入文本长度的平方成正比关系, 架构效率成为了海量数据和长文本建模的重要障碍. 因此, 许多工作分别尝试从降低 Transformer 算子的复杂性和裁剪模型冗余计算的角度设计更高效的模型架构. 在 Transformer 算子重构方面, 设计近似自注意力权重的低秩核函数可以使得 Transformer 的计算复杂度降低为线性时间复杂度 [154]; 限制注意力范围到不同的固定窗口大小可以减少自注意力机制的全局计算量 [155]; 结合全局和局部感受野的混合注意力方法将长序列压缩成少量的局部信息的集合, 可以降低计算复杂度 [156]. 在冗余计算消除方面, 许多研究发现预训练语言模型的本征维度 (intrinsic dimension) 是非常低的, 大量的自注意力矩阵会展现出有限的几组模式 [157], 通过非常少量关键参数的微调就能达到和完整参数空间微调近似的效果 [158]. 例如, 通过仅优化随机投影回完整空间的 200 个可训练参数, RoBERTa 模型就可以在 MRPC 数据集上实现全参数训练 90% 的能力 [159]. 同时, 在机器翻译 [160]、文本摘要 [161] 和语言理解 [162] 等很多自然语言处理任务上, 模型通过裁减算法删掉部分参数或者注意力分支往往能得到更好的预测准确率. 但是, 当前的参数裁剪算法仍然需要在模型训练结束后或者训练过程中进行不断裁剪, 属于事后架构优化, 如何在模型训练开始之前就能对模型冗余计算进行裁剪是一个尚未被解决的问题.

模型训练效率问题. 训练超百亿的大模型, 需要一个强大的软件和硬件基础设施. 但无论是从软件层面还是硬件层面解决训练效率问题, 最终达成的目的都是降低训练时单张 GPU 的显存占用需求或者减少模型的训练时间. 从参数的角度来看, 模型通常以单精度浮点格式 (FP32) 存储每个参数, 如果采用半精度浮点格式 (FP16) 可以在损失极少精度的情况下胜任大部分计算. 半精度浮点格式占用的内存是单精度格式的一半, 但是它也会带来浮点数截断和溢出的问题. 为了解决这个问题, 混合精度训练方法 [163,164] 在模型中同时使用 16 位和 32 位浮点类型, 从而加快运行速度, 减少内存使用. 使用分布式训练的思想可以利用多个计算设备 (如 GPU 或 TPU) 来并行地训练模型. 常见的分布式训练策略有数据并行、张量并行、流水线并行. 数据并行是最常见的并行策略, 数据集会按照设备数量被分割成不同的几个部分, 每个设备上持有一个完整的模型副本. 数据并行的缺点是每个设备都持有整个模型权重的副本, 带来了参数冗余的问题. 零冗余优化器 (ZeRO) [165] 通过对优化器状态、梯度和参数进行划分从而让数据并行策略的效率得到了极大的提高. 张量并行和流水线并行又可以合称为模型并行. 其中, 张量并行是指按照设备数量将一个张量沿特定维度分块, 每个设备都只持有整个张量的一部分 [166~169]. 流水线并行的核心思想是将模型按层分割成若干部分, 每个部分都放在不同的设备上. 但是流水线并行的计算过程中不可避免地会产生一些冒泡时间, 这将会导致每次计算时其余计算资源的浪费 [170], 解决这个问题的方法是更好地设计流水线中前向计算和反向计算的时机 [166]. 以上分布式训练方法在提升训练效率的同时也对设备之间的通讯带来了较高的要求, 通常需要 NVIDIA 公司的 NVLink 和 NVSwitch 技术, 或采用 InfiniBand 网络架构支持.

模型推断效率问题. 在生产环境中部署大模型, 通常会遇到推断时延过长的问题. 当模型不能及时为用户提供反馈时, 用户的体验会大幅度下降. 为了提升推断效率, 研究者们提出了多种模型压缩 (model compression) 的方法. 模型量化 (model quantization) [171] 是一种常用的模型压缩方法, 它将高精度浮点数表示的参数压缩成低精度的浮点数, 不仅能减少需要的存储空间, 还能减少神经网络前向过程中所需的计算量. 在自然语言处理领域, 预训练语言模型大多使用 32 位或者 16 位浮点数来表示参数, 而经过量化的模型通常用 8 位, 甚至更少的浮点数来表示 [172,173]. 模型剪枝 (model pruning) 是另外一种压缩模型的常用方法. 这种方法基于模型 “过参数化” 的假设, 将模型的某些 “不重要” 的部分移除, 从而缩小模型. 模型剪枝包含结构化与非结构化两种类型: 结构化剪枝通常移除模型的某些子结构, 例如 Transformer 层或者注意力头 [160,174]; 而非结构化剪枝通常针对模型的权重进行修剪 [175]. 知识蒸馏 (knowledge distillation) [176] 旨在训练一个相对较小的模型 (学生模型) 来学习大型

模型 (教师模型) 的行为, 并作为大型模型的替代. 小模型的层数和参数量通常远小于大模型, 因此推断效率会有大幅度的提升. 在自然语言处理领域, 近年涌现了许多知识蒸馏的工作如 DistillBERT [177], TinyBERT [178], MiniLM [179] 等.

可以预见, 未来大模型训练推理带来的计算成本和能源消耗问题会成为掣肘领域发展的瓶颈之一. 设计更加轻量化的模型架构, 从而减少模型的冗余, 提升模型的架构效率是从根本上优化所需存储和计算资源的一个重要方向. 另外, 提升模型的训练效率、推断效率也是减少资源消耗, 提升用户体验的重要方向. 目前, 许多减少计算成本的工作, 都以一定的性能为代价, 来换取计算效率的提升. 未来, 如何在不损害模型性能的情况下, 更好地设计与应用软硬件基础设施去训练和使用模型, 是学术界和工业界需要共同克服的难题.

4.5 数据资源与模型评价

4.5.1 大模型的数据资源

数据资源是自然语言处理研究的核心组成, 自 20 世纪 90 年代宾州树库开始 [180], 高质量语料库建设极大地促进了语料库语言学的快速发展, 使得自然语言处理研究步入了快车道. 然而, 高质量语料库建设的人工成本和时间成本巨大, 在这一背景下, 研究者探索了基于众包智慧的数据收集和标注思路, 加快了大规模标注数据收集的速度 [181]. 一时间, 不同类型、规模、难度的自然语言处理任务数据集如雨后春笋般涌现出来. 虽然在后续的很多研究中发现, 通过众包机制构建的标注数据集在标注质量等方面存在一定的改进空间, 但它进一步强化了以标注数据为基础的研究模式. 另一方面, 大规模特定任务标注数据集也逐步成为被广泛接受的评价基准, 在公开评测数据集上登顶, 甚至是能否超越人类在该任务数据集上的表现一度成为学术界和工业界的热门话题. 同时, 也需要注意到, 这种模式在一定程度上也限制了自然语言处理研究的发展, 大部分研究工作均关注在一些相对成熟的研究任务上的少数评测数据集上取得更高的成绩 (SOTA), 而或多或少忽略了相关技术在实际应用中所遇到的挑战 [182].

对于预训练语言模型以及后续大规模语言模型发展阶段, 因为在预训练阶段可以通过自监督学习模式去挖掘利用更大规模的无标注高质量文本数据, 在一定程度上缓解了对特定任务大规模人工标注数据集的过分依赖. 无需人工标注数据也可能使模型获得语言运用能力, 这也促使研究者不断突破极限, 去探索构建更大规模的语言模型.

从稍早期的 BERT 系列、GPT 系列、T5 到如今的 Palm, LLaMA, GPT-3.5/4 等大模型, 我们注意到大规模高质量文本数据资源在模型的构建过程中扮演了极其重要的作用, 从早期仅利用了维基百科、小说数据扩展到互联网上爬取的高质量网页、论坛讨论、代码、书籍、科技文献等繁多种类; 训练所需数据规模也从几十亿词猛增到一万亿词. 需要指出的是训练数据规模越大, 质量越高, 所得到的大规模语言模型的性能越好, 其中, 数据质量和多样性的影响尤为突出 [183]. 无法回避的是, 训练数据中的瑕疵数据 (如事实性错误、过时信息), 可能会对模型的表现产生负面影响. 可以预见的是, 互联网上公开可用的数据资源终将被耗尽, 这可能会促使研究者寻找对于已有高质量数据资源更好的利用方式, 探索新的研究范式 [184].

虽然目前大规模语言模型大多依赖高质量无标注训练数据进行预训练, 但并不是在构建过程中完全不需要人工标注数据. 事实上, 对于其中涉及人工参与标注数据的质量要求更高了. 以 ChatGPT 为例, 在强化其上下文学习能力的过程中, 需要人工参与标注高质量的指令, 并在后续强化学习的训练过程中提供针对回复质量的标注, 以支持其策略网络和价值网络的训练. 在使其与人类价值观对齐的模块中, 也采用了类似的基于人类反馈的强化学习模式, 需要标注人员提供高质量的提示标注或反馈

标注, 以及对模型输出进行评分排序. 后续的许多研究中也提到, 这两个模块中所涉及的人工标注数据质量对于最终模型的表现, 特别是与使用体验相关的主观评价密切相关.

一个有趣的比较是, 学术界尝试利用模板等自动化的方法将自然语言处理领域中已有的不同任务数据集转换为指令标注. 尽管其规模庞大, 但最终取得的效果提升却较为有限. 这也从另一个侧面说明, 想要在指令遵循或上下文学习等指令理解方面取得泛化性更好的表现, 仍需要从人类实际应用场景中收集相关数据, 而不能依赖于一些预先定义的任务或模板 [76].

4.5.2 模型评价

自然语言处理领域的模型评价主要是通过检查模型在特定任务相关的测试集上的性能表现来开展的, 例如考察精确率、召回率、F-1、准确率等指标. 但在评价大规模语言模型时, 仅考察某几个现有分类或生成任务数据集已无法完整地体现其真实的语言运用能力. 例如, 可以看到大规模语言模型即便在零样本的情况下, 仍能通过思维链等方式较好地回答复杂推理问题, 若还使用已有某个单一数据集, 已无法全面衡量其真实的问答能力了.

近年来, 学术界和工业级已针对大规模语言模型, 构造了多个更为全面, 多样的语言运用能力评测基准, 涵盖了问答、推理、生成、多语言运用等不同类型. 如斯坦福大学 (Stanford University) 联合多家高校研究机构提出的大型语言模型整体评测框架 (holistic evaluation of language models, HELM) [185], 在超过 40 个场景中从通用性能、鲁棒性、公平性、细粒度语言运用能力等 13 个维度、57 个指标, 对 36 个大规模语言模型进行了较为全面的评测.

相较于以前的单一类型或少数任务驱动的基准评测, 针对大规模语言模型的评测需覆盖的问题场景范围更广, 复杂度更高, 难度也更大. 我们也注意到, 受限于当前评测数据的实际情况, 这种大规模整体评测仍以英语等资源丰富语种的数据集为主, 无法对语言模型在其他资源受限语种上的能力进行更充分的评测.

可以预测的是, 随着工业界和学术界的持续发力, 大规模语言模型的能力将进一步提升, 现有评测基准也可以预见地将被快速超越. 后续研究也应考虑, 如何更为全面地评价大规模语言模型的语言综合运用能力, 例如, 设置更具挑战性的测试项目, 评价其对更复杂场景的驾驭能力; 在人工参与较少的情况下, 衡量其在更贴近实际应用场景的文本生成能力, 以及综合理解运用多模态信息的能力. 更重要的一点是, 需要考虑如何设置测试场景评价目前关注度较少的语言综合运用能力, 也许设置更为多样, 更贴近真实生活的应用场景是一种可能的选择.

5 讨论

第 3 和 4 节探讨了大模型对各种自然语言理解和生成核心任务将带来哪些冲击和影响, 分析了大模型将如何加强自然语言处理共性问题的研究. 本节首先将聚焦大模型自身, 探究如何从模型规模、学习方法、个性化等角度进一步提升大模型的内在能力; 其次, 从工具学习、多模态、具身智能的角度, 讨论如何进一步延伸和扩展大模型的感知、计算、推理、交互和控制能力, 使大模型成为通用人工智能的基座; 最后, 介绍 ChatGPT 等大型语言模型将催生哪些应用场景, 为各行各业带来哪些自然语言处理新应用.

5.1 进一步提升大模型的内在能力

5.1.1 规模提升

在语言模型中, 模型的性能与模型参数量、数据量、训练时长之间存在着一种规律性的关系, 即模型的性能随着参数量、数据量、训练时长的指数级增加而呈现出线性提升, 并且该提升对架构和优化超参数的依赖性非常弱. 这一规律被命名为规模定律 (scaling law), 由 OpenAI 团队于 2020 年首次提出并详细阐述 [186]. 依据规模定律对模型性能的预测, 在现有语言模型的基础上继续增大模型规模和训练数据是进一步提升大模型能力的直接方式.

训练数据. 根据规模定律, 大模型的能力与训练数据量呈对数增长关系 [186], 更大规模、更高质量的训练数据是未来大模型能力突破的关键. 参与 GPT 预训练过程的数据量只有 5 GB 文本 [23], 训练 GPT-2 时数据规模增长到 40 GB [187]. ChatGPT 的训练数据复用了 GPT-3 使用的数据集, 而 GPT-3 的训练数据由大量网页爬取内容、少量书籍内容、百科知识构成, 原始数据文本大小为 45 TB, 大致相当于 4000 亿词符 [2]. 考虑到下一代大模型训练数据可能达到 PB 量级, 数据增长的边际效应和数据规模与构成的取舍便是即将面临的重要挑战. 为了在未来进一步提升大模型的能力, 需要更加聚焦数据质量与结构, 而非盲目追求增加数据规模.

模型架构. 随着硬件能力的提升和模型压缩技术的改进, 继续增大模型参数量仍是可行的. 更大规模的模型能表示更丰富的知识, 理解更复杂的结构, 掌握更抽象的概念, 并在任务中取得更好的性能表现. 对基于 Transformer 架构的改进也是未来增强大模型合理且有希望的方向 [9]. 由于目前大模型通常采用自回归的生成方式, 这些模型普遍存在无法处理连续、结构化输入, 上下文长度受到限制等缺点. 通过改进现有模型架构, 能在一定程度上解决大模型处理多模态数据时遇见的难题, 改善大模型的多轮对话和长文本生成能力.

5.1.2 学习方法

目前类 ChatGPT 模型的强大能力来自于模型对大规模无标注语料的自监督学习, 对人类标注的监督学习, 以及对人类反馈的强化学习, 因此学习方法是大模型能力的重要影响因素. 设计更高效、更可靠的学习算法是进一步提升大模型能力的重要手段.

指令教学. 目前的指令微调方法使得大模型可以理解并遵循人类输入的自然语言指令, 但无法对模型内在参数进行更新从而将用户指令持久化. 例如, 当模型回答错误时, 用户可以纠正模型, 尽管有些情况下模型可以在后续对话中给出纠正过的答案, 但该知识仍未被持久化存储在模型参数中. 通过赋予大模型遵循自然语言指令来自动更新模型知识的能力, 可以显著提升大模型学习新知识, 适应人类社会的能力和效率.

对齐. 目前大模型通常通过基于人类反馈的强化学习技术来对齐人类世界的道德观念, 该方法对于评价模型 (reward model) 具有较高要求, 且存在调优难度大、标注成本高等问题. 目前已有少量研究聚焦于开发更高效的对齐算法, 例如基于人类提供的规则和原则训练 “宪法人工智能” [150], 用以监督其他 AI; 或利用后见重标记提升强化学习的数据利用效率 [188]. 利用大模型已有的遵循指令的能力, 引导模型根据人类给出的标准进行自我评估, 反思输出的合适性, 甚至直接调整输出以满足指令要求是可行的 [189, 190]. 大模型既是系统一又是系统二 [191], 在提高模型的准确性和无害性的同时降低对齐所需的成本.

自我改进. 模型可以通过持续跟踪和分析用户反馈来实现自我改进, 从而更好地适应用户需求和习惯. 这种方法有助于为用户提供更优质的个性化服务. 例如, 模型可以学习用户的常用术语和表达

方式, 与用户进行更自然的交流.

5.1.3 用户建模与个性化

近年来个性化技术被越来越多的学者和企业研究者所重视, 并被广泛用在各种搜索 [192, 193]、推荐和对话系统 [194~196] 中. 在搜索中, 个性化搜索是解决查询词描述力有限、用户查询意图不清晰、查询经常有歧义性等问题的一种重要方法. 在推荐中, 个性化推荐能够根据用户购买行为推测出用户的兴趣爱好, 进而为用户推荐最感兴趣的内容, 最可能购买的物品等. 在对话系统中, 个性化对话可以根据用户的显式兴趣或者隐式兴趣 [196], 生成个性化的回复内容, 同时, 也可以为对话机器人指定个性化的人格 [194, 195]. 总之, 个性化技术体现的是 “以人为中心”, 弥补了传统的以查询、文档等为中心的信息获取系统的缺陷, 避免千人一面.

ChatGPT 的出现将推动个性化技术的升级迭代. 在搜索和推荐方面, 目前的个性化搜索和推荐的基本范式之一是将用户历史行为 (如历史查询词、浏览过的文档、购买过的商品等) 编码成一个或者多个隐式的兴趣向量, 通过计算待排序的网页或者商品与这些用户兴趣向量的相似度来估计用户对待排序网页或者物品的喜好度, 进而生成个性化的排序 [193].

ChatGPT 等高质量的大规模预训练语言模型有潜力更好地对用户历史行为进行建模, 准确地理解用户历史行为之间的潜在语义关联, 进而进行更为准确的用户画像. 事实上, 虽然目前 ChatGPT 不提供个性化的功能, 仅能在一轮对话内部进行上下文学习, 但 ChatGPT 完全可以直接扩展到对整个用户对话历史进行建模, 进而不仅能够提供符合当前对话情景的高质量回复, 也能生成符合用户长期兴趣爱好和特定偏好的内容, 让用户感受到更高程度的关注和满足. 实际上, GPT-4 已经在个性化方面进行了一些有益的尝试. 基于 ChatGPT 实现个性化搜索或者推荐还将有一个独特的优势: ChatGPT 会试图给出对每个内容的解释, 这在一定程度上缓解了之前个性化系统因为结果不可解释导致的用户体验低的问题.

在对话系统方面, ChatGPT 也可以通过预设指令或者提示信息的方式, 根据用户的输入和偏好生成定制化的自然语言回复. 例如, 可以通过提示信息要求 ChatGPT 模仿一个中学生的性格来进行对话, 也可以要求它换成非常严谨和严肃的中年人的风格, 进行角色扮演. 这种灵活的低成本的人格设定和个性化内容生成方式的设定, 可以发掘模型潜在的知识和能力, 使模型的响应更加个性化, 进而提升用户体验, 极大地降低研发个性化聊天机器人或者智能信息助手的成本.

5.2 工具学习

大型语言模型随着参数量和数据集规模的提升, 在少样本和零样本的场景下取得了越来越优秀的表现, 能够生成匹配上下文的流畅的内容, 但仍然不能完美地解决数学计算和逻辑推理等问题, 存在着生成虚假和幻觉内容的缺陷, 并且无法完成语言和认知之外的其他任务. 而事实上, 人类在完成复杂任务时, 除了依赖于语言理解和逻辑推理等认知能力外, 往往也需要借助一些外部工具. 例如, 人类可以使用计算器辅助数学运算, 使用天气预报查询准确的天气信息, 使用 12306 查询和购买火车票, 使用搜索引擎查询互联网信息等. 工具的使用扩充了人类的能力, 让人类能够完成更复杂的任务.

让 ChatGPT 能够与外部工具相交互, 有望不增加模型容量就能解决现在无法解决的任务. 微软的 New Bing 已经为我们展现了可以与搜索引擎交互的语言模型的力量与价值. 例如, 在 New Bing 中查询 “北京今天需要穿秋裤么?”, New Bing 会根据对话意图, 调用天气预报服务, 并根据天气预报反馈的信息, 生成 “风寒效果明显, 因此, 今天穿秋裤是比较合适的. ” 这样有理有据的回答.

5.2.1 工具学习的必要性

让大模型学习使用外部工具, 有如下好处:

(1) 缓解虚假和幻觉问题. 使用工具可以在一定程度上缓解语言模型中存在的生成虚假内容的问题, 避免模型进入幻觉状态. 专业的事情交给专业工具做, 工具返回的结果可以为语言模型提供正确内容的线索, 进而引导模型生成正确的内容. 例如, 使用检索工具, 可以从互联网上检索回相关内容片断 (passage) 并在生成过程中使用.

(2) 理解新知识. 使用工具也可以解决通过固定语料训练的大模型无法回答训练日期之后出现的新知识相关的问题. 目前 ChatGPT 使用的是 2021 年 9 月份的语料, 无法正确回答在此日期之后新出现的事实相关的问题. 使用工具可以动态访问互联网、知识库或者各种服务中的新知识, 进而让历史版本的大模型能够理解和处理新知识.

(3) 控制模型规模. 使用工具可以使知识以 "外挂" 的形式和语言模型进行融合, 避免一味扩大模型容量造成的成本增加.

(4) 规避内部数据泄露风险. 这种方式也可以有效解决在很多数据敏感的场景下数据无法批量参与大模型训练的问题. 例如, 在金融等很多领域, 因为存在隐私泄露等风险, 很多企业不愿意将内部数据用于大模型训练. 在这种场景下, 可以将企业内部数据包装成服务接口, 通过工具学习与语言模型互动.

(5) 连接行为能力. 使用工具也可以支撑通过语言驱动行动的能力, 例如, 让 ChatGPT 通过调用工具的方法和物联网设备、传感器、摄像头等硬件融合, 完成更为复杂的行动任务.

5.2.2 工具学习的研究现状及未来发展趋势

集成工具和外部知识库的思路已经在任务型对话 (例如语言助手、智能音箱) 中广泛使用. 在这些工具中, 广泛使用了基于规则的方法, 或者使用大量的数据来训练模型.

目前通过大模型进行工具学习的研究刚刚起步, 但已经得到了很多研究机构的重视. 例如, Meta 推出了 Toolformer [126], 让语言模型自己学习如何使用外部工具, 例如计算器、问答系统、搜索引擎等. 通过集成这些工具的能力, 语言模型可以在不损失其核心语言能力的同时提高在很多下游任务上的性能. Toolformer 使用特殊的符号来标识 API 调用, 并使用二元分类器来判断是否需要调用 API. 如果需要调用, 则生成 API 名称和参数并发送给外部工具. 然后, Toolformer 接收工具返回的结果, 并将其作为上下文继续生成接下来的文字. 通过这种方式, Toolformer 无缝地将外部工具集成到语言模型, 提升了语言模型的鲁棒性, 扩展了其能力. 微软亚洲研究院推出了 Visual ChatGPT [88], 可以将 ChatGPT 作为处理中心, 来集成一系列不同的视觉基础模型 (例如, Visual Transformer 或者 Stable Diffusion), 完成视觉问答、图片生成、视觉编辑, 以及其他更为复杂的视觉任务. OpenAI 自己也推出了 ChatGPT 插件 (plugins).

在融合大模型和外部工具方面, 未来需要解决的问题和研究内容主要有:

- **大模型与外部工具融合的方式与模式.** 例如, 是微调 ChatGPT 这样的大模型, 还是在预训练阶段就考虑工具学习, 将工具的使用直接嵌入到预训练阶段.
- **多种工具数据的交互和联合使用.** 目前 Toolformer 在每个对话中只包含一种工具的使用. 在很多场景下, 可能需要在一个对话中使用多种工具. 如何建模工具之间的关联关系是未来研究需要解决的重要问题.
- **如何生成工具学习的数据.** 在 Toolformer 中, 使用大模型本身来生成微调数据, 并附加了分类

任务. 是否有更好的数据标注或者训练数据生成方式?

• **更便捷的工具集成方式.** 如何通过最小的成本添加一个新工具, 高效快速低成本地学习出工具调用的参数以及融合工具的结果来生成内容, 并适应多场景?

5.3 多模态与具身智能

ChatGPT 的相关能力已经超越传统自然语言处理的范畴, 例如可以生成和修改代码、制作网页、设计游戏等, 其出现打开了通往通用人工智能的大门. 大数据、大模型、强算法的研究范式逐渐从自然语言处理领域蔓延至语音、视觉、多模态和机器人等领域, 慢慢形成大一统的研究范式. 下面从多模态和具身智能两方面进行讨论.

多模态. 语言大模型的成功不断推动着多模态大模型的研究和发展. 最近, 微软提出了能够感知语音、语言和视觉信息的多模态大模型 Kosmos-1 [197], 通过统一框架将不同模态进行语义编码, 并利用互联网中的多模态数据从零开始训练. Kosmos-1 模型可以遵循用户指令, 执行上下文学习, 在多个多模态标准测试集上获得了优异性能. OpenAI 最近发布的 GPT-4 是一个真正的多模态大模型, 可以准确理解图文多模态的输入, 在高等物理习题等非常复杂的图文任务上表现出卓越的性能, 如果再配合 OpenAI 的语音识别开源模型 Whisper, 能够同时处理图文音 3 种模态的内容. 当前, 多模态大模型主要聚焦图文音多模态输入的感知和理解, 生成还是以文本为主导. 未来, 多种模态 (甚至任意模态或信号) 输入, 多种模态输出的大模型将受到越来越多的关注, 相信很快可能就会出现超越多模态的物联网大模型的出现.

具身智能. 随着 ChatGPT 和 GPT-4 等大模型在虚拟环境中的语言处理、代码生成和多模态理解等领域不断取得成功, 研究者们开始探索让大模型实现具身智能, 从而感知现实世界并与现实世界进行交互. 其中大模型与机器人的交互是实现具身智能的重要途径. 最近几个月见证了大模型与机器人交互的快速发展. 微软提出了 ChatGPT for Robotics 方法 [198], 采用语言大模型扮演感知、规划和决策中枢, 通过提示语设计与人机对话的方式实现任务分解与指令生成. 提示语是语言大模型 ChatGPT 生成准确的机器人执行代码的关键, 除了机器人可执行的底层库函数描述告诉语言大模型可以调用的指令集合之外, 提示语还包括任务的描述和约束, 例如抓取物体的形状和重量、现实世界的环境描述、机器人和物体的当前状态等.

人机对话使得语言大模型 ChatGPT 通过主动询问的方式明确任务目标和环境等信息. 在这种方法中, 语言大模型压力大, 缺乏现实世界的多模态感知, 缺乏与机器人和环境的深度互动和反馈. 谷歌提出了 PaLM-E 模型 [199], 设计了统一的多模态具身大模型, 通过环境感知, 机器人状态和任务描述完成任务分解与指令生成. PaLM-E 模型的核心架构思想是将连续的, 具身的观察结果, 比如场景图像、机器人和物体的状态估计, 以及其他传感器信号, 通过 ViT 等连续信号的编码和映射模型转换为与语言词汇嵌入空间相同维度的向量序列, 并将其注入到预先训练好的语言模型 PaLM 的语言嵌入空间中, 形成统一的多模态具身大模型. PaLM-E 模型不再需要精心设计的提示语, 而是通过连续信号编码模型实现真实世界环境的感知, 通过语言编码模型实现目标任务的理解, 通过语言与连续信号的交互实现任务规划和指令生成, 通过机器人动作执行后的状态和环境反馈更新 PaLM-E 模型的连续信号输入, 从而不断优化和生成正确的指令序列, 直至任务完成. 最大的 PaLM-E 模型由 5400 亿参数的语言大模型 PaLM 和 220 亿参数的视觉大模型 ViT 组合而成, 最后通过端到端的训练方式完成大模型对机器人的准确规划和控制, 从而让大模型实现了具身智能. 未来, 具身智能的发展将有望实现大模型, 机器人与人的有机统一, 并有望能够实现自主进化.

5.4 基于大模型的自然语言处理新应用

ChatGPT 等大型语言模型具有强大的自然语言理解、问答、对话、翻译和生成能力, 这些能力将激发出政府、企业与个人更多的应用需求, 催生更加广泛的应用场景, 更大程度地节约人力并提高工作效率.

首先, ChatGPT 能够进一步推动自然语言生成技术在媒体出版、电子商务、政府办公等典型行业的应用, 增强新闻撰写、自动文摘、公文生成、智能客服、商品标题与描述生成等 AIGC 应用任务上的应用效果, 大幅提升用户的使用体验和满意程度.

其次, ChatGPT 的出现能够满足以前技术无法应对的应用需求, 给包括教育、医疗、科研、法律、金融等在内的各行各业带来全新的应用机会, 例如 (包括但不限于):

• 会话式信息搜索. 当前搜索引擎仅能返回跟查询相关的网页列表, 无法理解用户的复杂, 深层意图, 也无法将精准答案总结提供给用户, 而新一代搜索引擎则能利用大模型的能力实现与用户的多轮会话, 理解用户的真实信息需求, 并直接返回精准答案信息, 大幅提升用户信息获取的效率和满意度. 微软 Bing 搜索已经集成了 ChatGPT 能力, 并进入公测阶段, 相信不久以后将能开放给全体用户使用.

• 智能小说创作. 利用大模型进行部分类型的文学作品大纲或部分段落的自动撰写或辅助撰写已成为可能, 特别地, 业界可以利用 ChatGPT 技术帮助撰写情节相对简单的网络小说.

• 虚拟医生. 电影《超能陆战队》中的大白掌握了全面的医学知识, 能够帮助和救治人类. ChatGPT 通过阅读大量文献资料也掌握了丰富的医学知识, 因此通过 ChatGPT 技术有望构建可信赖的虚拟医生, 通过与人类的交谈并结合相关身体指标检查结果, 对人类提供治疗和康复建议.

• 虚拟心理咨询师. 当今社会上心理不健康或亚健康的群体日渐增多, 心理咨询师的数量已远远不够, 很多患者也不愿意面对人类心理咨询师透露个人隐私. 利用 ChatGPT 技术则可实现虚拟心理咨询师, 帮助人类排忧解难, 解决心理和情绪上的问题, 提升人类的幸福指数.

• 虚拟教师. 教师的工作在于传道授业解惑, 对知识进行传播, 同时对学生的学习情况进行反馈和总结. 利用 ChatGPT 技术可以构建虚拟教师, 更好地普及教育以及实现教育公平. 相比人类教师, 虚拟教师拥有更多的知识, 能够与学生进行实时互动, 并及时提供反馈.

• 智能科研助手. ChatGPT 有能力对科技文献进行解读, 解释科学概念, 并总结科研进展, 也能帮助进行论文、演讲稿、项目申请书的撰写和准备, 因此利用 ChatGPT 技术能够构建智能科研助手系统, 在科研过程的众多环节对科研人员提供帮助, 提升科研人员的科研效率.

• 智能投资顾问. 利用 ChatGPT 技术阅读掌握大量金融相关材料, 对金融形势和股票趋势进行研判, 为用户提供投资建议.

• 虚拟律师. 利用 ChatGPT 技术阅读大量法律相关材料, 掌握全面的法律条文和知识, 实现对各类案件、冲突、争议的深度理解和智能研判, 为客户提供专业法律意见, 草拟、审查法律文书, 协助参与诉讼, 调解或者仲裁活动.

6 总结与展望

综上所述, ChatGPT 等大型语言模型, 对传统自然语言处理核心任务产生了巨大的冲击和影响. 这些核心任务普遍遵循监督学习范式, 需要针对特定任务, 给定监督数据, 设计和定制机器学习和深度学习模型. 相比之下, 利用 ChatGPT 完成自然语言处理任务, 不仅能在少样本、零样本场景下接近乃至达到传统监督学习方法的性能指标, 且具有较强的领域泛化性.

虽然如此, 面对大型语言模型所带来的冲击, 研究者们完全无需产生 “自然语言处理已经不存在了” 等悲观情绪. 首先, ChatGPT 等对话式大模型, 并非横空出世, 而是沿着神经语言模型的发展路线, 利用海量算力, 基于大规模高质量文本数据所实现的大型全注意力模型. 未来研究者们能够将大模型作为研究方法和手段, 更能够学习, 借鉴生成式无监督预训练、多任务学习、上下文学习、指令遵循、思维链、基于人类反馈的强化学习等大型语言模型的特点和优势, 进一步提升自然语言核心任务的能力.

大模型为自然语言处理带来了架构通用化、任务统一化、能力按需化、模型定制化等变化趋势. 今后在各种自然语言理解和生成任务的主流架构和范式逐渐统一的情况下, 一方面, 各种自然语言处理任务有望进一步得到整合, 以增强自然语言处理模型的通用性, 减少重复性工作; 另一方面, 基于大模型的强大基础能力, 针对具体任务进行按需适配、数据增强、模型压缩与轻量化、跨模态和多模态融合, 加强自然语言处理模型方法的可控性、可配性、领域适应性、多样性、个性化和交互能力, 将进一步拓展自然语言处理的应用场景.

大模型时代的自然语言处理, 存在算法模型的可解释性、公平性、安全性、可靠性、能耗、数据质量和评价等一些共性问题, 这些问题也是妨碍大模型能力提升和服务质量的主要因素. 未来, 针对模型分析和可解释性、伦理问题与安全性、信息准确性、计算成本与能源消耗、数据资源和模型评价等各种自然语言处理共性问题的研究将越来越深入.

自然语言处理是人工智能的重要组成部分, 是人工智能从感知智能上升到认知智能的主要手段. ChatGPT 的出现, 已经打开了通向通用人工智能的大门. 未来, 以大模型作为基座, 利用工具学习、多模态融合、具身智能拓展其感知、计算、推理、交互和控制能力, 自然语言处理技术将进一步助力通用人工智能的发展, 促进各行各业的生产力进步, 更好地为人类社会服务.

参考文献

1 OpenAI. GPT-4 Technical Report. 2023

2 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of the Advances in Neural Information Processing Systems, 2020. 1877–1901

3 Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. In: Proceedings of the Advances in Neural Information Processing Systems, 2000. 932–938

4 Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association, Makuhari, 2010. 1045–1048

5 Pham N Q, Kruszewski G, Boleda G. Convolutional neural network language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016. 1153–1162

6 Sukhbaatar S, Weston J, Fergus R, et al. End-to-end memory networks. In: Proceedings of the Advances in Neural Information Processing Systems, 2015. 2440–2448

7 Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations. In: Proceedings of the North American Chapter of the Association for Computational Linguistics, 2018. 2227–2237

8 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, 2019. 4171–4186

9 Lin T, Wang Y, Liu X, et al. A survey of transformers. AI Open, 2022, 3: 111–132

10 Qiu X P, Sun T X, Xu Y G, et al. Pre-trained models for natural language processing: a survey. Sci China Tech Sci, 2020, 63: 1872–1897

11 Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. 2021. ArXiv:2101.03961

12 Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. 2022. ArXiv:2204.02311

13 Thoppilan R, Daniel D F, Hall J, et al. Lamda: language models for dialog applications. 2022. ArXiv:2201.08239

14 Sanh V, Webson A, Raffel C, et al. Multitask prompted training enables zero-shot task generalization. In: Proceedings of the 10th International Conference on Learning Representations, 2022

15 Ye J J, Chen X T, Xu N, et al. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. 2023. ArXiv:2303.10420

16 Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. 2023. ArXiv:2303.12712

17 Maron M E. Automatic indexing: an experimental inquiry. J ACM, 1961, 8: 404–417

18 Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theor, 1967, 13: 21–27

19 Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning, Chemnitz, 1998. 137–142

20 Dieng A B, Wang C, Gao J, et al. TopicRNN: a recurrent neural network with long-range semantic dependency. In: Proceedings of the International Conference on Learning Representations, 2017

21 Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014. 1746–1751

22 Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 7370–7377

23 Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

24 Chen X T, Ye J J, Zu C, et al. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. 2023. ArXiv:2303.00293

25 Haidar M A, Anchuri N, Rezagholizadeh M, et al. RAIL-KD: random intermediate layer mapping for knowledge distillation. In: Proceedings of Findings of the Association for Computational Linguistics, 2022. 1389–1400

26 Rashid A, Lioutas V, Ghaddar A, et al. Towards zero-shot knowledge distillation for natural language processing. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021. 6551–6561

27 Wang J, Lan C, Liu C, et al. Generalizing to unseen domains: a survey on domain generalization. IEEE Trans Knowl Data Eng, 2022, doi: 10.1109/TKDE.2022.3178128

28 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2015. ArXiv:1412.6572

29 Zhang Y F, Kang B Y, Hooi B, et al. Deep long-tailed learning: a survey. 2021. ArXiv:2110.04596

30 Natarajan N, Dhillon I S, Ravikumar P K, et al. Learning with noisy label. In: Proceedings of the Advances in Neural Information Processing Systems, 2013. 1196–1204

31 Wang J D, Hu X X, Hou W X, et al. On the robustness of ChatGPT: an adversarial and out-of-distribution perspective. 2023. ArXiv:2302.12095

32 Tu L, Lalwani G, Gella S, et al. An empirical study on robustness to spurious correlations using pre-trained language models. Trans Assoc Comput Linguist, 2020, 8: 621–633

33 Geirhos R, Jacobsen J H, Michaelis C, et al. Shortcut learning in deep neural networks. Nat Mach Intell, 2020, 2: 665–673

34 Feng S Y, Gangal V, Wei J, et al. A survey of data augmentation approaches for NLP. In: Proceedings of Findings of the Association for Computational Linguistics, 2021

35 Yaghoobzadeh Y, Mehri S, Combes R T D, et al. Increasing robustness to spurious correlations using forgettable

examples. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021. 3319–3332

36 He H, Zha S, Wang H. Unlearn dataset bias in natural language inference by fitting the residual. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP, 2019. 132–142

37 Balashankar A, Wang X, Packer B, et al. Can we improve model robustness through secondary attribute counterfactuals? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021. 4701–4712

38 Sproat R, Shih C, Gale W, et al. A stochastic finite-state word-segmentation algorithm for chinese. Comput Linguist, 1996, 22: 377–404

39 Márquez L, Rodríguez H, Carmona J, et al. Improving POS tagging using machine-learning techniques. In: Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999

40 Nivre J. An efficient algorithm for projective dependency parsing. In: Proceedings of the 8th International Conference on Parsing Technologies, 2003. 149–160

41 Wan X J. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008. 553–561

42 Tian W, Liu W J, Ma T H. Study and implementation of Chinese intelligent question answering system based on restricted domain. In: Proceedings of the 3rd International Conference on Genetic and Evolutionary Computing, 2009. 217–220

43 Dong L, Wei F R, Zhou M, et al. Adaptive multi-compositionality for recursive neural models with applications to sentiment analysis. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014. 1537–1543

44 Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014. 615–620

45 Agrawal S, Siddiqui T J. Using syntactic and contextual information for sentiment polarity analysis. In: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, 2009. 620–623

46 Aliod D M, Gardiner M. Answerfinder: question answering by combining lexical, syntactic and semantic information. In: Proceedings of the Australasian Language Technology Workshop, 2004

47 Cheng J P, Kartsaklis D. Syntax-aware multi-sense word embeddings for deep compositional models of meaning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015. 1531–1542

48 Bastings J, Titov I, Aziz W, et al. Graph convolutional encoders for syntax-aware neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017. 1957–1967

49 Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016

50 Kudo T, Richardson J. Sentencepiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018. 66–71

51 Marcheggiani D, Frolov A, Titov I. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. In: Proceedings of the 21st Conference on Computational Natural Language Learning, 2017. 411–420

52 Zhou H, Zhang Y, Li Z H, et al. Is POS tagging necessary or even helpful for neural dependency parsing? In: Proceedings of the 9th CCF International Conference on Natural Language Processing and Chinese Computing, 2020. 179–191

53 Hewitt J, Manning C D. A structural probe for finding syntax in word representations. In: Proceedings of the

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4129–4138
54 Chi E A, Hewitt J, Manning C D. Finding universal grammatical relations in multilingual BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 5564–5577
55 Kate R J, Wong Y W, Mooney R J, et al. Learning to transform natural to formal languages. In: Proceedings of the 20th National Conference on Artificial Intelligence, 2005. 1062–1068
56 Herzig J, Berant J. Decoupling structure and lexicon for zero-shot semantic parsing. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018. 1619–1629
57 Dong L, Lapata M. Coarse-to-fine decoding for neural semantic parsing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. 731–742
58 Chen B, Sun L, Han X P. Sequence-to-action: end-to-end semantic graph generation for semantic parsing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. 766–777
59 Zhao C, Su Y, Pauls A, et al. Bridging the generalization gap in text-to-SQL parsing with schema expansion. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 5568–5578
60 Sherborne T, Lapata M. Zero-shot cross-lingual semantic parsing. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 4134–4153
61 Lenat D B. CYC: a large-scale investment in knowledge infrastructure. Commun ACM, 1995, 38: 33–38
62 Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1998
63 Dong Z D, Dong Q. Hownet and the Computation of Meaning. Singapore: World Scientific, 2006
64 Baker C F, Fillmore C J, Lowe J B. The Berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 1998. 86–90
65 Auer S, Bizer C, Kobilarov G, et al. DBpedia: a nucleus for a web of open data. In: Proceedings of the 6th International Semantic Web Conference, 2007. 722–735
66 Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, 2007. 697–706
67 Bollacker K, Cook R, Tufts P. Freebase: a shared database of structured general human knowledge. In: Proceedings of the 22nd National Conference on Artificial Intelligence, 2007. 1962–1963
68 Wu W T, Li H S, Wang H X, et al. Probase: a probabilistic taxonomy for text understanding. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2012. 481–492
69 Speer R, Chin J, Havasi C. Conceptnet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017. 4444–4451
70 Sap M, Le Bras R, Allaway E, et al. Atomic: an atlas of machine commonsense for if-then reasoning. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, 2019
71 Bosselut A, Rashkin H, Sap M, et al. COMET: commonsense transformers for automatic knowledge graph construction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 4762–4779
72 Minsky M L. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. AI Mag, 1991, 12: 34
73 Qin C W, Zhang A, Zhang Z S, et al. Is ChatGPT a general-purpose natural language processing task solver? 2023. ArXiv:2302.06476
74 Xu C W, Guo D Y, Duan N, et al. Baize: an open-source chat model with parameter-efficient tuning on self-chat data. 2023. ArXiv:2304.01196
75 Wang J, Liang Y L, Meng F D, et al. Is ChatGPT a good NLG evaluator? A preliminary study. 2023.

ArXiv:2303.04048

76 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. In: Proceedings of the Advances in Neural Information Processing Systems, 2022. 27730–27744

77 Mauldin M L. Semantic rule based text generation. In: Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Annual Meeting of the Association for Computational Linguistics, 1984. 376–380

78 De Novais E M, Tadeu T D, Paraboni I. Improved text generation using n-gram statistics. In: Proceedings of the Advances in Artificial Intelligence-IBERAMIA, 2010. 316–325

79 Zhang T Y, Ladhak F, Durmus E, et al. Benchmarking large language models for news summarization. 2023. ArXiv:2301.13848

80 Mustafa B, Riquelme C, Puigcerver J, et al. Multimodal contrastive learning with limoe: the languageimage mixture of experts. 2022. ArXiv:2206.02770

81 Zhao T C, Zhao R, Eskenazi M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017. 654–664

82 Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of the Advances in Neural Information Processing Systems, 2016

83 Liang K J, Hao W T, Shen D H, et al. Mixkd: towards efficient distillation of large-scale language models. In: Proceedings of the International Conference on Learning Representations, 2021

84 Liu Z L, Yu X W, Zhang L, et al. Deid-GPT: zero-shot medical text de-identification by GPT-4. 2023. ArXiv:2303.11032

85 Hsu W T, Lin C K, Lee M Y, et al. A unified model for extractive and abstractive summarization using inconsistency loss. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. 132–141

86 Chen X Y, Gao S, Tao C Y, et al. Iterative document representation learning towards summarization with polishing. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018. 4088–4097

87 Chen X Y, Chan Z M, Gao S, et al. Learning towards abstractive timeline summarization. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019

88 Wu C F, Yin S M, Qi W Z, et al. Visual ChatGPT: talking, drawing and editing with visual foundation models. 2023. ArXiv:2303.04671

89 Shen Y L, Song K T, Tan X, et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face. 2023. ArXiv:2303.17580

90 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems, 2017

91 Adiwardana D, Luong M T, So D R, et al. Towards a human-like open-domain chatbot. 2020. ArXiv:2001.09977

92 Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021. 300–325

93 Komeili M, Shuster K, Weston J. Internet-augmented dialogue generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 8460–8478

94 Shuster K, Xu J, Komeili M, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. 2022. ArXiv:2208.03188

95 Bao S Q, He H, Wang F, et al. PLATO: pre-trained dialogue generation model with discrete latent variable. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 85–96

96 Bao S Q, He H, Wang F, et al. PLATO-2: towards building an open-domain chatbot via curriculum learning. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021. 2513–2525

97 Bao S Q, He H, Wang F, et al. PLATO-XL: exploring the large-scale pre-training of dialogue generation. In: Proceedings of the Findings of the Association for Computational Linguistics: AACL-IJCNLP, 2022. 107–118
98 Bao S Q, He H, Xu J, et al. PLATO-K: internal and external knowledge enhanced dialogue generation. 2022. ArXiv:2211.00910
99 Wang Y D, Ke P, Zheng Y H, et al. A large-scale chinese short-text conversation dataset. In: Proceedings of the 9th CCF International Conference on Natural Language Processing and Chinese Computing, 2020. 91–103
100 Zhou H, Ke P, Zhang Z, et al. EVA: an open-domain chinese dialogue system with large-scale generative pre-training. 2021. ArXiv:2108.01547
101 Gu Y, Wen J, Sun H, et al. EVA2.0: investigating open-domain Chinese dialogue systems with large-scale pre-training. Mach Intell Res, 2023, doi: 10.1007/s11633-022-1387-3
102 Zhao W X, Liu J, Ren R Y, et al. Dense text retrieval based on pretrained language models: a survey. 2022. ArXiv:2211.14876
103 Zhao W X, Zhou K, Li J Y, et al. A survey of large language models. 2023. ArXiv:2303.18223
104 Aizawa A. An information-theoretic perspective of TF-IDF measures. Inf Process Manage, 2003, 39: 45–65
105 Robertson S E, Walker S, Jones S, et al. Okapi at TREC-3. In: Proceedings of Overview of the 3rd Text Retrieval Conference, 1995. 109
106 Liu T Y. Learning to rank for information retrieval. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010
107 Guo J F, Fan Y X, Ai Q Y, et al. A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 2016. 55–64
108 Qu Y Q, Ding Y C, Liu J, et al. RocketQA: an optimized training approach to dense passage retrieval for open-domain question answering. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021. 5835–5847
109 Tay Y, Tran V Q, Dehghani M, et al. Transformer memory as a differentiable search index. 2022. ArXiv:2202.06991
110 Zhu F B, Lei W Q, Wang C, et al. Retrieving and reading: a comprehensive survey on open-domain question answering. 2021. ArXiv:2101.00774
111 Izacard G, Lewis P, Lomeli M, et al. Few-shot learning with retrieval augmented language models. 2022. ArXiv:2208.03299
112 Yu W H, Iter D, Wang S H, et al. Generate rather than retrieve: large language models are strong context generators. 2022. ArXiv:2209.10063
113 Dai Z Y, Zhao V Y, Ma J, et al. Promptagator: few-shot dense retrieval from 8 examples. 2022. ArXiv:2209.11755
114 Gao L Y, Ma X G, Lin J, et al. Precise zero-shot dense retrieval without relevance labels. 2022. ArXiv:2212.10496
115 Bonifacio L, Abonizio H, Fadaee M, et al. Inpars: data augmentation for information retrieval using large language models. 2022. ArXiv:2202.05144
116 Bang Y J, Cahyawijaya S, Lee N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. 2023. ArXiv:2302.04023
117 Gao L, Dai Z, Pasupat P, et al. Attributed text generation via post-hoc research and revision. 2022. ArXiv:2210.08726
118 Liu S, Zhang X, Zhang S, et al. Neural machine reading comprehension: methods and trends. Appl Sci, 2019, 9: 3698
119 Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014. 615–620
120 Zhu F B, Lei W Q, Huang Y C, et al. TAT-QA: a question answering benchmark on a hybrid of tabular and textual content in finance. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. 3277–3287

121 Yasunaga M, Ren H, Bosselut A, et al. QA-GNN: reasoning with language models and knowledge graphs for question answering. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021. 535–546
122 Xu Y C, Zhu C G, Xu R C, et al. Fusing context into knowledge graph for commonsense question answering. In: Proceedings of the Association for Computational Linguistics: ACL-IJCNLP, 2021. 1201–1207
123 Khashabi D, Min S, Khot T, et al. UNIFIEDQA: crossing format boundaries with a single QA system. In: Proceedings of the Association for Computational Linguistics: EMNLP, 2020. 1896–1907
124 Dua D, Wang Y Z, Dasigi P, et al. DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 2368–2378
125 Tan Y M, Min D H, Li Y, et al. Evaluation of ChatGPT as a question answering system for answering complex questions. 2023. ArXiv:2303.07992
126 Schick T, Dwivedi Y J, Dessí R, et al. Toolformer: language models can teach themselves to use tools. 2023. ArXiv:2302.04761
127 Wei J, Wang X Z, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022. ArXiv:2201.11903
128 Fu Y, Peng H, Sabharwal A, et al. Complexity-based prompting for multi-step reasoning. 2023. ArXiv:2210.00720
129 Creswell A, Shanahan M, Higgins I. Selection-inference: exploiting large language models for interpretable logical reasoning. 2022. ArXiv:2205.09712
130 Belinkov Y, Glass J. Analysis methods in neural language processing: a survey. Trans Assoc Comput Linguist, 2019, 7: 49–72
131 Karpathy A, Johnson J, Li F F. Visualizing and understanding recurrent networks. 2015. ArXiv:1506.02078
132 Hupkes D, Veldhoen S, Zuidema W. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. J Artif Intell Res, 2018, 61: 907–926
133 Tenney I, Xia P, Chen B, et al. What do you learn from context? Probing for sentence structure in contextualized word representations. 2019. ArXiv:1905.06316
134 Cao B X, Lin H Y, Han X P, et al. Can prompt probe pretrained language models? Understanding the invisible risks from a causal view. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 5796–5808
135 Jacovi A, Goldberg Y. Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 4198–4205
136 Lipton Z C. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue, 2018, 16: 31–57
137 Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. ACM Comput Surv, 2019, 51: 1–42
138 Danilevsky M, Qian K, Aharonov R, et al. A survey of the state of explainable AI for natural language processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020. 447–459
139 Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst, 2021, 32: 4793–4813
140 Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017. ArXiv:1702.08608
141 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell, 2019, 1: 206–215
142 Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier.

In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 1135–1144

143 Garg N, Schiebinger L, Jurafsky D, et al. Word embeddings quantify 100 years of gender and ethnic stereotypes. In: Proceedings of the National Academy of Sciences, 2018. 3635–3644

144 Garimella A, Banea C, Hovy D, et al. Women's syntactic resilience and men's grammatical luck: gender-bias in part-of-speech tagging and dependency parsing. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 3493–3498

145 Dinan E, Fan A, Williams A, et al. Queens are powerful too: mitigating gender bias in dialogue generation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 8173–8188

146 Stanovsky G, Smith N A, Zettlemoyer L. Evaluating gender bias in machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 1679–1684

147 Li H. Language models: past, present, and future. Commun ACM, 2022, 65: 56–63

148 Glaese A, McAleese N, Trębacz M, et al. Improving alignment of dialogue agents via targeted human judgements. 2022. ArXiv:2209.14375

149 Bai Y T, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. 2022. ArXiv:2204.05862

150 Bai Y T, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from ai feedback. 2022. ArXiv:2212.08073

151 Liu B. Learning on the job: online lifelong and continual learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 13544–13549

152 Luccioni A S, Viguier S, Ligozat A L. Estimating the carbon footprint of bloom, a 176b parameter language model. 2022. ArXiv:2211.02001

153 Scao T L, Fan A, Akiki C, et al. Bloom: a 176b-parameter open-access multilingual language model. 2022. ArXiv:2211.05100

154 Peng H, Pappas N, Yogatama D, et al. Random feature attention. In: Proceedings of the International Conference on Learning Representations, 2021

155 Roy A, Saffar M, Vaswani A, et al. Efficient content-based sparse attention with routing transformers. Trans Assoc Comput Linguist, 2021, 9: 53–68

156 Guo Q P, Qiu X P, Liu P F, et al. Star-transformer. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 1315–1325

157 Kovaleva O, Romanov A, Rogers A, et al. Revealing the dark secrets of bert. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. 4365–4374

158 Su Y S, Wang X Z, Qin Y J, et al. On transferability of prompt tuning for natural language processing. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022. 3949–3969

159 Aghajanyan A, Gupta S, Zettlemoyer L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. 7319–7328

160 Michel P, Levy O, Neubig G. Are sixteen heads really better than one? In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019. 14037–14047

161 Baan J, ter Hoeve M, van der Wees M, et al. Understanding multi-head attention in abstractive summarization. 2019. ArXiv:1911.03898

162 Zheng R, Rong B, Zhou Y, et al. Robust lottery tickets for pre-trained language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 2211–2224

163 Micikevicius P, Narang S R, Alben J, et al. Mixed precision training. In: Proceedings of the 6th International Conference on Learning Representations, 2018

164 Baboulin M, Buttari A, Dongarra J, et al. Accelerating scientific computations with mixed precision algorithms. Comput Phys Commun, 2009, 180: 2526–2533

165 Rajbhandari S, Rasley J, Ruwase O, et al. Zero: memory optimizations toward training trillion parameter models. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2020. 1–16

166 Narayanan D, Shoeybi M, Casper J, et al. Efficient large-scale language model training on GPU clusters using megatron-LM. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2021. 1–15

167 Xu Q F, Li S G, Gong C Y, et al. An efficient 2D method for training super-large deep learning models. 2021. ArXiv:2104.05343

168 Wang B X, Xu Q F, Bian Z D, et al. 2.5-dimensional distributed model training. 2021. ArXiv:2105.14500

169 Bian Z D, Xu Q F, Wang B X, et al. Maximizing parallelism in distributed training for huge neural networks. 2021. ArXiv:2105.14450

170 Huang Y P, Cheng Y L, Bapna A, et al. GPipe: efficient training of giant neural networks using pipeline parallelism. In: Proceedings of the Advances in Neural Information Processing Systems, 2019. 103–112

171 Stock P, Joulin A, Gribonval R, et al. And the bit goes down: revisiting the quantization of neural networks. In: Proceedings of the 8th International Conference on Learning Representations, 2020

172 Zafrir O, Boudoukh G, Izsak P, et al. Q8BERT: quantized 8bit BERT. In: Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, 2019. 36–39

173 Zhang W, Hou L, Yin Y C, et al. Ternarybert: distillation-aware ultra-low bit BERT. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020. 509–521

174 Fan A, Grave E, Joulin A. Reducing transformer depth on demand with structured dropout. In: Proceedings of the 8th International Conference on Learning Representations, 2020

175 Gordon M A, Duh K, Andrews N. Compressing BERT: studying the effects of weight pruning on transfer learning. In: Proceedings of the 5th Workshop on Representation Learning for NLP, 2020. 143–155

176 Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015. ArXiv:1503.02531

177 Sanh V, Debut L, Chaumond J, et al. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019. ArXiv:1910.01108

178 Jiao X Q, Yin Y C, Shang L F, et al. Tinybert: distilling BERT for natural language understanding. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP, 2020. 4163–4174

179 Wang W H, Wei F R, Li D, et al. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Proceedings of the Advances in Neural Information Processing Systems, 2022. 5776–5788

180 Marcus M P, Santorini B, Marcinkiewicz M A. Building a large annotated corpus of english: the Penn treebank. Comput Linguist, 1993, 19: 313–330

181 Callison B C. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2009. 286–295

182 Tedeschi S, Bos J, Declerck T, et al. What's the meaning of superhuman performance in today's NLU? 2023. ArXiv:2305.08414

183 Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res, 2020, 21: 5485–5551

184 Villalobos P, Sevilla J, Heim L, et al. Will we run out of data? An analysis of the limits of scaling datasets in machine learning. 2022. ArXiv:2211.04325

185 Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. 2022. ArXiv:2211.09110
186 Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. 2020. ArXiv:2001.08361
187 Radford A, Wu J, Child R, et al. Better language models and their implications. OpenAI Blog, 2019. https://openai.com/research/better-language-models
188 Zhang T J, Liu F C, Wong J, et al. The wisdom of hindsight makes language models better instruction followers. 2023. ArXiv:2302.05206
189 Shinn N, Labash B, Gopinath A. Reflexion: an autonomous agent with dynamic memory and self-reflection. 2023. ArXiv:2303.11366
190 Ganguli D, Askell A, Schiefer N, et al. The capacity for moral self-correction in large language models. 2023. ArXiv:2302.07459
191 Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011
192 Dou Z C, Song R H, Wen J. R. A large-scale evaluation and analysis of personalized search strategies. In: Proceedings of the 16th International Conference on World Wide Web, 2007. 581–590
193 Zhou Y J, Dou Z C, Wen J. R. Encoding history with context-aware representation learning for personalized search. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020. 1111–1120
194 Qian Q, Huang M L, Zhao H Z, et al. Assigning personality/profile to a chatting machine for coherent conversation generation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018. 4279–4285
195 Wu C H, Zheng Y H, Mao X X, et al. Transferable persona-grounded dialogues via grounded minimal edits. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021. 2368–2382
196 Ma Z Y, Dou Z C, Zhu Y T, et al. One chatbot per person: creating personalized chatbots based on implicit user profiles. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021. 555–564
197 Huang S, Dong L, Wang W, et al. Language is not all you need: aligning perception with language models. 2023. ArXiv:2302.14045
198 Vemprala S, Bonatti R, Bucker A, et al. ChatGPT for robotics: design principles and model abilities. Microsoft Auton Syst Robot Res, 2023, 2: 20
199 Driess D, Xia F, Sajjadi M. S, et al. PaLM-E: an embodied multimodal language model. 2023. ArXiv:2303.03378

Towards a comprehensive understanding of the impact of large language models on natural language processing: challenges, opportunities and future directions†

Wanxiang CHE[4], Zhicheng DOU[11], Yansong FENG[1], Tao GUI[3], Xianpei HAN[10], Baotian HU[5], Minlie HUANG[6], Xuanjing HUANG[2*], Kang LIU[9], Ting LIU[4], Zhiyuan LIU[6*], Bing QIN[4], Xipeng QIU[2], Xiaojun WAN[1], Yuxuan WANG[8], Jirong WEN[11], Rui YAN[11], Jiajun ZHANG[9], Min ZHANG[5,7*], Qi ZHANG[2], Jun ZHAO[9], Xin ZHAO[11] & Yanyan ZHAO[4]

1. *Wangxuan Institute of Computer Technology, Peking University, Beijing* 100080*, China*;
2. *School of Computer Science, Fudan University, Shanghai* 200438*, China*;
3. *Institute of Modern Languages and Linguistics, Fudan University, Shanghai* 200433*, China*;
4. *Faculty of Computing, Harbin Institute of Technology, Harbin* 150001*, China*;
5. *School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen* 518055*, China*;
6. *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084*, China*;
7. *School of Computer Science and Technology, Suzhou University, Suzhou* 215006*, China*;
8. *Zhejiang Lab, Hangzhou* 311121*, China*;
9. *Institute of Automation, Chinese Academy of Sciences, Beijing* 100190*, China*;
10. *Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China*;
11. *Gaoling School of Artificial Intelligence, Renmin University of China, Beijing* 100872*, China*

* Corresponding author. E-mail: xjhuang@fudan.edu.cn, liuzy@tsinghua.edu.cn, zhangmin2021@hit.edu.cn

† Equal contribution, sorted alphabetically

Abstract Recently released large language models (LLMs), including ChatGPT and GPT-4, have exhibited remarkable proficiency in tasks involving natural language generation and understanding. These models possess the capacity to generate language that is both fluent and coherent, catering effectively to human requirements. Additionally, they excel across a spectrum of open-domain natural language understanding tasks. In scenarios characterized by limited available data (few-shot and zero-shot), LLMs can attain performance levels comparable to and occasionally surpass, those achieved by conventional supervised learning techniques. Moreover, LLMs exhibit strong domain generalization capabilities, leading to significant impacts on traditional natural language processing (NLP) tasks. This comprehensive survey exhaustively investigates and analyzes the impact that LLMs have on NLP. The objective is to delve into the challenges and opportunities presented by the integration of LLMs into core NLP tasks. The survey discusses the aspects of NLP research that stand to gain greater potency due to the advent of LLMs. Moreover, it anticipates the future trajectory of LLMs and NLP technology. Through our analysis, it becomes apparent that substantial strides are yet to be made in the realm of NLP during the era defined by LLMs. We suggest that researchers can use LLMs as research methods and tools, learn from the characteristics and advantages of LLMs, change the mainstream research paradigm of NLP, integrate scattered and independent NLP tasks, and further enhance the ability of core NLP tasks. Additionally, we endorse the pursuit of detailed research addressing prevalent issues such as interpretability, fairness, security, and information accuracy. This will serve to enhance the capabilities and quality of service furnished by LLMs. In subsequent years, by establishing LLMs as a foundation and extending their capabilities in perception, computation, reasoning, interaction, and control, NLP technology will further expedite the evolution of artificial general intelligence. Consequently, it will drive advancements in productivity across diverse industries, thereby better serving society.

Keywords ChatGPT, chat generative pre-trained transformer, large language models, natural language processing, artificial general intelligence