·专题:ChatGPT 与人工智能技术应用·

ChatGPT 能力分析与未来展望

武俊宏[1†] 赵 阳[2†] 宗成庆[2*]

1. 中国科学院大学 人工智能学院,北京 100049
2. 中国科学院 自动化研究所,北京 100190

[摘 要] 近年来,大语言模型的自然语言处理能力不断提升,尤其近期,聊天生成式预训练模型(ChatGPT)所掌握的“渊博知识”和表现出来的强大对话能力成为举世瞩目的热点话题。ChatGPT 语言理解能力的真实水平如何?与专用模型相比,其性能表现谁居上风?它是否能够成为整个自然语言处理领域的通用模型而取代其它模型,甚至使所有自然语言处理问题得到彻底解决呢?为了回答上述问题,本文对 ChatGPT 在多个自然语言处理任务上的性能表现进行了评估和分析。在此基础上,我们讨论了 ChatGPT 对自然语言处理领域的影响,并对未来的发展进行了展望。

[关键词] 自然语言处理;大语言模型;预训练语言模型;ChatGPT

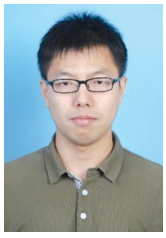
自然语言处理(Natural Language Processing,NLP)是研究如何利用计算机技术对语言文本进行处理、加工和转换的一门学科[1]。由于该学科在理论上面临巨大的挑战,而其技术应用前景极其广泛,因此被誉为人工智能皇冠上的明珠。

自然语言处理技术自 20 世纪 40 年代末期诞生以来经历了 70 多年的曲折历程,先后经历了以符号逻辑方法为主导的理性主义时期、以统计学习理论为基础的经验主义时期和以深度学习方法为驱动的连结主义时期。随着自然语言处理技术的进步,语言建模(Language Model,LM)技术已经经历了从最初的统计语言模型(Statistical Language Model,SLM)发展至神经网络语言模型(Neural Language Model,NLM),再到预训练语言模型(Pre-trained Language Model,PLM)的演进过程[2]。近年来,通过扩展预训练语言模型得到的大模型将语言建模技术推向了一个新的发展高度,其发展速度之快、模型能力之强和通用程度之高,都远远超越任何一个历史时期的任何一种方法,令人瞠目。

大语言模型(Large Language Model,LLM)通常指参数量为百亿级甚至更大规模的神经网络组成

宗成庆 中国科学院自动化所研究员,博士生导师,IEEE Fellow、ACL Fellow、AAIA Fellow、CAAI Fellow 以及 CCF Fellow。主要从事自然语言处理、机器翻译和语言认知计算等研究,发表学术论文 200 余篇,出版专著 3 部、译著 2 部。目前担任中国中文信息学会副理事长、国际计算语言学学会(ACL)候任副主席。曾获国家科技进步奖二等奖、北京市科学技术奖一等奖等奖项。荣获北京市优秀教师、中国科学院优秀导师和李佩教学名师等荣誉。

武俊宏 中国科学院大学人工智能学院博士研究生。主要研究方向为自然语言处理、机器翻译和终身学习等。



赵阳 博士,中国科学院自动化研究所副研究员。主要研究方向为自然语言处理、机器翻译和文本数据挖掘等。作为项目负责人和技术骨干承担多个国家自然科学基金项目、国家重点研发计划项目和特定领域的应用项目。在领域著名期刊和会议上发表论文 30 余篇,出版译著 1 部。目前担任国际学术期刊 *ACM Transactions on Asian and Low-Resource Language Information Processing* 副主编。

收稿日期:2023-07-03;修回日期:2023-07-31

† 共同第一作者。

* 通信作者,Email: cqzong@nlpr.ia.ac.cn

本文受到国家自然科学基金项目(62006224)的资助。

的语言模型，它采用自监督学习方式利用大量未标注数据训练而成。尽管在扩展预训练语言模型时主要是增大模型参数量（模型架构和训练方法基本不变），但这些大规模的预训练语言模型表现出与较小规模的预训练语言模型（例如，330 M 参数的 BERT[3] 和 1.5 B 参数的 GPT-2[4]）不同的行为，并且在解决一系列复杂任务时展现出令人惊讶的能力[2]，这种能力被业界称为涌现能力（Emergent Abilities）[5]。例如，少样本（Few-shot）和零样本（Zero-shot）学习能力，即在给定下游任务时可不依赖任何特定领域的训练数据，而只是通过适当的提示（Prompt）调节模型的行为[6]。随着模型规模的进一步增大，大语言模型在各个任务上的性能也逐渐提高，这一现象被称为规模效应（Scaling Law）[7,8]。目前的研究表明，大语言模型有望成为解决各种任务的通用基础模型，是实现通用人工智能一条可行的希望之路。

2022 年 11 月底美国 OpenAI 公司发布的聊天生成型预训练模型 ChatGPT① 在世界范围内引发了轰动。该模型是在生成型预训练模型（Generative Pre-training Transformer，GPT）[9] 系列模型的基础之上，通过指令微调（Instruction Tuning）并从调试人员的反馈中强化学习（Reinforcement Learning from Human Feedback，RLHF）[10] 训练建立起来的。ChatGPT 一经发布，立刻成为史上用户增长速度最快的消费应用。同其他大模型相比，由于其采用了指令微调和 RLHF 等技术，ChatGPT 具有更加强大的理解人类用户意图和偏好的能力，既可以根据指令生成高质量的回复，也可以针对不恰当的输入拒绝回答，甚至更正对话中的错误。其超乎寻常的理解和会话能力让部分人认为，ChatGPT 的出现标志着通用人工智能的“奇点”时刻已经到来。

然而，作为通用模型的 ChatGPT，与专用模型的性能对比，其表现如何？以 ChatGPT 为代表的大模型能否成为一个通用模型同时完成所有不同的下游任务？换句话说，通用大模型能否成为 NLP 学科方向的终结者？为回答上述问题，本文评估了 ChatGPT 在机器翻译、文本摘要、情感分析和信息抽取等多个自然语言处理任务上的性能表现，分析了 ChatGPT 与专用模型相比的优势和不足，并对未来 NLP 学科方向的发展进行了展望。

1 ChatGPT 在自然语言处理任务上的性能

自 ChatGPT 发布以来，已有工作评估了 ChatGPT 在自然语言处理任务上的性能。其中部分工作[11,12] 主要关注 ChatGPT 的通用性能，针对大量任务做了简单测试。另一些工作[13-18] 则聚焦于某一具体任务。

为了评估 ChatGPT 在自然语言处理任务上的实际表现，本文选择了 4 种常见的也是典型的自然语言处理任务：机器翻译、信息抽取、文本摘要和情感分析。这 4 项任务既涵盖了语言生成、序列表示和文本分类三项自然语言处理的基础任务，也涉及到不同语言之间的转换。下面依次介绍 ChatGPT 在这些任务上的性能表现。

1.1 机器翻译

机器翻译（Machine Translation，MT）是将一种语言（源语言）自动翻译成另外一种语言（目标语言）的技术，是自然语言处理中最具挑战性的研究课题，其性能表现体现着模型处理跨语言理解、转换和生成的综合能力[1]。Brown 等[9] 和 Ouyang 等[10] 对比了 ChatGPT 模型和商业翻译模型的性能差异。他们的实验结果表明，ChatGPT 在高资源场景下的翻译性能可与最优秀的商业翻译模型相媲美，但在低资源场景下的性能则显著落后。在具体语言上，ChatGPT 更擅长处理目标语言为英语的翻译任务。

为了研究 ChatGPT 多语言翻译的能力，尤其是在中低资源语言翻译方面的能力，本文选取了 Flores-200[19] 的测试集进行评估。该测试集包含 1 012 个句子在 204 种语言上的翻译。为了对 ChatGPT 在不同资源语言上的翻译性能进行分析，本文根据 GPT-3 训练使用的数据集中不同语言所占比例将语言划分为高资源（占比＞0.1%）、中等资源（0.1%＞占比＞0.001%）和低资源（占比＜0.001%）三类，并从每个类别中选择了两种语言测试 ChatGPT 将其翻译成英语（X－＞En）的能力。这 6 种语言是：（1）高资源语言：中文（Zh）、德语（Ge）；（2）中等资源语言：爱沙尼亚语（Et）、立陶宛语（Lt）；（3）低资源语言：僧伽罗语（Si）、尼泊尔语（Ne）。

对比测试选用国际上公认的基于词序列（*n*-gram）对比的评价指标 BLEU[20] 和基于句子表示相

① https://chat.openai.com

似度计算的评价指标 COMET[21] 作为评价准则，以谷歌翻译（Google Translate）为比较对象。对比结果如表 1 所示。

观察表 1 所展示的结果可以得到如下两个结论：

（1）总体而言，ChatGPT 的翻译性能逊色于谷歌翻译。在高资源（如中—英、德—英）和中等资源情况下（如爱沙尼亚语—英语和立陶宛语—英语）的翻译性能，ChatGPT 与谷歌翻译模型相差不大。随着资源量逐渐减少，ChatGPT 与谷歌翻译的性能差距逐渐增大，这与已有的对比结论相吻合。

（2）ChatGPT 在低资源语言的翻译中出现了严重的“幻觉翻译（Hallucinatory Translation）”问题，即译文表述流畅，但语义与原文并不一致，属于无中生有的臆想。为进一步探究 ChatGPT 与谷歌翻译的性能差异，我们对僧伽罗语—英语和尼泊尔语—英语这两个低资源语言对的翻译结果进行了样例分析。分析结果发现，ChatGPT 不仅在低资源语言的翻译中出现了幻觉现象，在其它各项自然语言处理任务中均存在不同程度的幻觉。

1.2　信息抽取

信息抽取（Information Extraction，IE）是指从非结构化或半结构化的文本中自动识别抽取出实体、实体属性、实体之间关系以及事件等事实信息，并形成结构化表示的一种文本挖掘技术[22]。Bang 等[11] 对 ChatGPT 在信息抽取任务中的性能做了全面评估，其中包括命名实体识别（Named Entity Recognition，NER）、关系抽取（Relation Extraction，RE）和事件抽取（Event Extraction，EE）三项任务。实验结果表明在这三项任务上，ChatGPT 的性能最高只能达到专门训练出来的最优模型性能的 63.5%、43.0% 和 35.3%。

我们选取国际公开的 CoNLL04[23] 数据集分析 ChatGPT 在关系抽取任务上的表现。结果表明，ChatGPT 在实体关系三元组抽取任务上的 F1 值仅有 24.8%，与最优模型 REBEL[24] 的性能（F1 值为 76.7%）仍有较大差距。人工分析结果表明，ChatGPT 倾向于生成比人工标注更长的文本片段，以更接近人类的语言习惯。同时，ChatGPT 也表现出了引入世界知识的特性，例如对地理位置或组织机构名称缩写进行扩写。为了测试 ChatGPT 所学习到的世界知识对其执行信息抽取任务的影响，我们遵循 Bang 等[11] 中的实验设置，通过调换一对关系中两个实体的位置构建了一组反事实测试样例。测试结果表明，ChatGPT 在调换实体位置后仍能生成正确的关系三元组。这表明 ChatGPT 所包含的世界知识使其在执行信息抽取时更具鲁棒性。

1.3　自动文本摘要

自动文本摘要（Automatic Text Summarization）是利用计算机自动将文本（或文本集合）转换成简短摘要的一种信息压缩技术[1]。文本摘要技术在信息爆炸时代具有重要的应用价值。已有的测试表明，整体而言，ChatGPT 已经能够完成多种文摘任务，但在多数情况下仍然低于现有最好的摘要模型。Kaplan 等[7] 和 Bahri 等[8] 对 ChatGPT 的通用型摘要（Generic Summarization）能力的测试结果表明，ChatGPT 生成的摘要明显不如经过微调后的 BART 模型（以 ROUGE-1[25] 指标值衡量）。Qin 等[12] 进一步研究了 ChatGPT 在查询式文本摘要（Query-based Text Summarization）和要素级文本摘要（Aspect-based Text Summarization）等更多样化任务上的性能。结果表明，在除社交媒体领域以外的三个基准数据集上，ChatGPT 的 ROUGE 分数接近于微调模型，只是在新闻数据集上超过了微调模型。而对于抽取式文本摘要，Jiao 等[13] 的测试表明 ChatGPT 的摘要性能同样低于目前最好的抽取式摘要模型。

表 1　ChatGPT 的多语言翻译性能

语言对	COMET			BLEU		
	ChatGPT	谷歌翻译	性能比*	ChatGPT	谷歌翻译	性能比
中→英	88.7	89.4	99.3	29.6	38.6	76.7
德→英	90.7	91.0	99.7	43.0	47.3	90.9
爱→英	90.1	91.7	98.3	36.9	45.5	81.1
立→英	86.1	89.2	96.6	32.6	41.3	78.9
尼→英	87.4	93.2	93.8	24.5	51.3	47.7
僧→英	61.3	90.5	67.7	3.5	45.5	7.7

* 表中的“性能比”表示根据相应的计算指标 ChatGPT 的性能与 Google Translate 的性能之间的比值。

考虑到上述对比测试主要集中在英文摘要任务上，本文在中文对话摘要数据集 CSDS[26] 上对 ChatGPT 的摘要能力进行了测试，并与该数据集上目前最好的模型 Fast-RL[27] 进行了对比。采用基于词序列的评价指标 ROUGE 和基于文本表示的评价指标 BERTScore[28] 对生成摘要的质量进行评价，实验结果如表 2 所示。从表中的数据可以看出，ChatGPT 在中文对话摘要上的性能同样不如目前最优的自动文摘模型，这与英文数据集上的测试结果基本一致。另外，我们还分析发现，ChatGPT 生成的摘要平均长度为 189.7 词，远长于人工给出的结果(120.9 词)。尽管我们曾尝试在提示词中加入对摘要长度的限制，但 ChatGPT 并不能遵循给定长度的约束，而且添加的长度限制影响(降低)了生成摘要的质量。在我们的实验中，ChatGPT 更偏向于生成流利度高、叙述详细的文本，这与人们希望的摘要应尽量简洁的要求存在一定的冲突。

1.4 情感分析

情感分析(Sentiment Analysis)是对文本中蕴含的情感、态度、情绪等主观信息进行自动提取、分析、归纳和推理的处理过程[18]，如分析归纳客户评论、社交媒体帖子和新闻文章中的观点、情感和情绪等。Hendy 等[14] 从 4 个方面对 ChatGPT 的情感分析能力进行了具体评估，包括标准评估、极性转换评估和开放领域评估以及情感推断评估。他们利用自动评价指标得到的对比结果表明，在传统的情感分类任务上，ChatGPT 的性能与微调后的 BERT 模型相当，但仍落后于在特定领域内专门训练出来的有监督模型。而在情绪信息抽取任务上，ChatGPT 的准确度相对较低。但是在人工评估中，ChatGPT 在这些任务上的表现并不是太差。在引入极性转换(例如否定或推测后)，ChatGPT 通常能够正确的理解情感极性变化并做出正确预测，而微调的 BERT 模型则不能，这说明 ChatGPT 具有更强的鲁棒性。在开放领域测试中，传统方法在特定领域训练出来的模型通常难以泛化到其它领域，而 ChatGPT 反而展现出了较强的泛化能力。

为了进一步分析 ChatGPT 在情绪信息抽取任务上的表现，我们使用 SemEval-2014[29] 数据集测试了其在要素级情感分析三元组抽取(Aspect-level Sentiment Triplet Extraction)任务上的性能。结果表明，ChatGPT 的性能约为该数据集上最优模型 BDTS[30] 的 64.8%。我们进一步随机选取了 100 个样本进行人工评价。结果表明，ChatGPT 的预测准确率为 42%，与 BDTS 所得到的 68%准确率仍有一定差距。

表 2 ChatGPT 的文本摘要性能

Model	ROUGE-2	ROUGE-L	BERTScore
ChatGPT	17.5	39.1	70.4
Fast-RL	41.4	47.1	79.8

1.5 ChatGPT 能力分析

根据已有专家的测试和本文上述分析不难看出，ChatGPT 作为通用模型在几乎所有的自然语言处理任务上都展示了较好的性能和优势，以至于让很多人感觉到以 ChatGPT 为代表的大模型会很快实现通用人工智能。但是，具体到任何一个专项任务上，如机器翻译、文本摘要和情感分析以及信息抽取等，ChatGPT 的性能表现距离人类理想的通用人工智能技术依然有较大的差距。我们认为，ChatGPT 的主要优势体现在如下两个方面：

(1) 强大的通用处理能力。以 ChatGPT 为代表的大模型能够通过人类指令执行任何用户希望完成的自然语言处理任务，而且性能表现都在上乘，尽管大部分情况下都不及目前最优的专用模型，但其通用的人工智能能力足以让人们刮目相看。无论其宽广的知识面和“渊博”的知识储备，还是规范、流畅的语言表达能力，均已超出人们的想象，甚至超越一般人的表现。其处理(翻译)语言的种类之多、并行回复用户和问题类型的数量之大，更是让专用模型和人类所望尘莫及。

(2) 准确的用户意图理解能力和“随机应变”的交互能力。ChatGPT 几乎能够准确理解和把握人类用户的意图，且能够根据人类的指令和上下文进行自然流畅的人机交互，可随时根据用户的问题和反馈修改模型自身的输出，其看似缜密的推理过程和滴水不漏的应答能力都是已有模型所未能做到的。尽管有时候它也会胡说八道，但其表现仍然一本正经。

正如上面所述，ChatGPT 等大模型的研究和使用所面临的问题和挑战也是显而易见的：

(1) 技不如人：在垂直领域和专项任务上 ChatGPT 的性能不如目前最优的专用模型。

(2) 无中生有：ChatGPT 容易引发的“幻觉”影响了其输出的忠实度和简洁性，由此产生的臆想结论和事实性错误极易以假乱真，混淆视听。

(3) 厚多薄寡：由于在训练 ChatGPT 时不同语

言的样本比例严重不平衡，导致 ChatGPT 在完成多语言处理（翻译）任务时，存在明显的语言敏感的性能差异。

（4）价值趋同：由于 ChatGPT 在训练时需要借助于调试人员的反馈强化学习，实现模型学到的知识与调试人员的标准和要求之间的对齐，因此模型建立的价值观、意识形态和社会伦理观极易受调试人员的影响，而不同国家、不同民族和不同文化的价值趋向是不同的，因此，模型很难很好地处理多元价值观问题。

（5）隐私泄露：在训练大模型时需要大规模的多样化训练样本，而这些样本中难免存在涉及个人隐私的信息，这些信息一旦被模型使用，极有可能产生隐私泄露问题，对相关人员造成伤害。

除了上述问题之外，如何判断被大模型使用的知识和数据是否被侵权，有效保护知识和数据持有者的合法权益；如何界定 ChatGPT 等大模型生成内容的知识产权，建立合情、合理的知识产权保护法规；如何制定大模型使用的明确规定，既充分发挥大模型的强大能力，而又不破坏应有的学术诚信体系等等，都是大模型研究和使用无法回避的问题。

另外，大模型建立和维护的昂贵成本是制约技术落地的重要因素。据半导体研究公司 SemiAnalysis 估计，训练一次有 1 750 亿参数的 GPT-3 基础模型所需要的最低费用约为 84 万美元[①]。而 ChatGPT 是在 GPT-3 模型的基础上经过反复的试错迭代得到的，其开发成本据估约为 500 万美元[②]。OpenAI CEO Sam Altman 也曾在社交媒体上表示，ChatGPT 每与用户互动一次约需数美分[③]。对于拥有亿级月活跃用户规模的情况而言，资金投入量将是一个极为庞大的数字。设想一下，一个特定的用户，尤其是某个特定领域或行业的用户，是需要一个知识面宽泛却在解决本领域问题时表现并非最优的系统，还是更愿意有一个针对性强、性能优越的专用系统呢？

2　NLP 技术未来展望

正如前文所述，尽管大模型并不完美，但看起来前景光明，于是针对大模型的研究正如火如荼。以下问题是当前研究人员关注的热点，或将是未来很长时期 NLP 领域研究的问题：

（1）模型通用性和专用性的均衡方法以及通用领域和垂直领域的权衡问题。问题描述如前文所述。

（2）大模型的轻量化方法。ChatGPT 等大语言模型在实际应用中存在计算和存储资源消耗过高的问题。为了解决这一问题，模型的轻量化方法，如模型压缩和推理加速，成为了研究的重要方向。模型压缩旨在减少大语言模型的参数量和模型规模，以降低模型在部署和推理阶段的计算、存储开销，从而提高模型的推理效率，使大语言模型更广泛地应用于边缘设备、移动终端和实时应用场景。

（3）大模型的终身学习与高效微调。语言是一个动态的领域，新词汇、新概念和新语言现象不断出现。为使 ChatGPT 等大语言模型能够适应不断变化的数据和任务，探索持续学习和高效微调方法至关重要。通过持续学习，大语言模型可以从新数据中学习，并更新自身的知识库，以更好地理解和生成新的语言内容。高效微调则能够将大语言模型的通用知识与特定任务的要求相结合，提高模型在特定任务上的性能。对于持续学习和高效微调的探索，将使大语言模型更好地适应变化的语言数据和任务要求，以提高模型的性能和适应性，满足人们对于新的语言内容的需求。

（4）大模型的可解释性与可控性。ChatGPT 的发展也引发了对模型可解释性和可控性的关注。由于 ChatGPT 的训练基于大规模数据，其生成结果可能受到不当或有害内容的影响。因此，如何确保模型生成的内容符合伦理和准则成为了研究和探讨的重点。研究模型的可解释性旨在把控模型的决策过程和内部机制，以帮助研究者和用户更好地把握模型生成结果的原因和逻辑。可控性研究则是为了实现对模型生成内容和风格的有效控制，限制模型生成含有不当偏见、敏感甚至虚假信息的内容或冒犯性言论等，从而确保模型生成的内容更加合乎伦理准则，也更加真实可靠。

（5）与其他学科领域的交叉融合。ChatGPT 作为一种强大的自然语言处理模型，不仅局限于解决自然语言处理领域内的问题，而且可以为其他学科领域的交叉研究提供有力支撑。由于模型学习到的知识来自巨大的样本空间，先验知识之丰富、各种要素组合关系之复杂、因果关系推断之千奇百怪，远远

① https://www.semianalysis.com/p/the-ai-brick-wall-a-practical-limit
② https://lambdalabs.com/blog/demystifying-gpt-3
③ https://twitter.com/sama/status/1599671496636780546

超出人的想象，这种超乎寻常的能力完全可以为特定学科领域（如生物医学、制药、化学等）提供重要帮助，包括提出问题、预测结论和找到重要发现等，真正让 AI 为科学研究建立功勋。

（6）大模型的产业化应用。在自然语言处理的理论方法研究中，研究者主要利用实验室收集标注的数据进行模型训练和测试，而这些数据和方法往往与产业化实际应用中的情况有一定的隔离和差距。ChatGPT、DALLE-2[①]、Stable Diffusion[②]等一系列人工智能生成内容（Artificial Intelligence Generated Content，AIGC）产品所取得的空前成功表明，聚焦真实世界的实际问题比在学术界建立的简单数据集上比拼性能更为重要。因此，未来工作应该更加聚焦于弥补大语言模型与实际应用场景之间的差距，包括探索多模态的人机交互模式，研发工具学习[31]（Tool Learning）技术，建立模型即服务（Model as a Service，MaaS）生态等。

ChatGPT 为探索通用 AI 蹚出了一条希望之路，但笔者认为，它未必是唯一之路，甚至未必是一条最佳道路。一方面，在理论上大模型本身并没有太大的创新，只是神经网络对传统 n 元文法模型（n-gram Model）的再现和扩展，成功的原因更多地来自于大数据、大算力和大量人工的工程投入，而如何建立具有更强泛化和推理能力，规模更小、人工投入更少的中小规模的智能 NLP 模型，仍然是学界研究和探索的目标，当然这种研究可以借鉴大模型成功的经验。

另一方面，在应用上，正如前文所述，传统 NLP 方法和针对特定领域、特定任务研发的专用模型无论在建造成本和性能表现方面，还是系统的安全性、可靠性、可扩展性等方面，都有其独特的优势，而且系统部署简单，尤其对于特殊应用领域，例如涉及公共安全和国防安全的领域，不便于上网的应用场景等，具有广阔的应用前景。由此可见，NLP 学科方向决不会因为 ChatGPT 的出现而销声匿迹，学术界也不会因为研发的大模型而砸掉自己的饭碗。而且作者根据 NLP 过去 70 多年的发展经验坚定地认为，随着技术的进步，未来一定会出现比大模型性能更优、规模更小、成本更低的 NLP 新模型，新模型突破甚至抛弃大模型的范式并非没有可能。

3 结 语

ChatGPT 作为一种具有强大能力的预训练模型，对于自然语言处理领域的发展带来了深远影响，引领了新的研究范式，创造了新的发展机遇。同时，ChatGPT 的缺陷也为 NLP 研究留下了极大的探索空间。

值得说明的是，ChatGPT 之所以被如此关注，是因为其强大的通用性和与之前同类技术相比超乎寻常的性能表现，而与人的实际要求相比，尤其是针对具体任务的高标准要求，它还有相当大的差距。而且我们也必须清楚地认识到，目前我们所掌握的对于 ChatGPT 的了解大多来自 OpenAI 的博客[③]，究竟还有多少更深层、更具体的技术细节我们不曾了解？自 ChatGPT 发布以来，国内有数十个“大模型”相继发布，这些模型既有发布团队独立研发的，也有在 LLaMA[32]、Vicuna[33]等开源大模型基座上改造而成的；既有为垂直领域构建的，也有面向通用领域搭建的；既有参数量为几十亿的稠密模型，也有万亿参数量的混合专家系统（Mixture of Expert，MoE）。发布者自然各抱初衷，“蹭热度”也情有可原，但重要的是，如何沉下心来，明确目标，坚持开展有理想、有情怀的创新研究，做出超越 ChatGPT 及其更高版本的中国的大语言模型！这是我们的使命。少一点炒作，少一点忽悠，扎扎实实地做好中国的事情，是我们应有的态度和品格。

参 考 文 献

[1] 宗成庆. 统计自然语言处理. 第 2 版. 北京：清华大学出版社，2013.

[2] Zhao WX, Zhou K, Li JY, et al. A survey of large language models. (2023-03-31)/[2023-06-21]. https://arxiv.org/pdf/2303.18223.pdf.

[3] Devlin J, Chang MW, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. (2018-10-11)/[2023-06-21]. https://arxiv.org/pdf/1810.04805.pdf.

[4] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9.

① https://openai.com/dall-e-2

② https://stability.ai/stablediffusion

③ https://openai.com/blog

[5] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. (2022-06-15)/[2023-06-21]. https://arxiv.org/pdf/2206.07682.pdf.

[6] Wei J, Bosma M, Zhao VY, et al. Finetuned language models are zero-shot learners. (2021-09-03)/[2023-06-21]. https://arxiv.org/pdf/2109.01652.pdf.

[7] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. (2020-01-23)/[2023-06-21]. https://arxiv.org/pdf/2001.08361.pdf.

[8] Bahri Y, Dyer E, Kaplan J, et al. Explaining neural scaling laws. (2021-02-12)/[2023-06-21]. https://arxiv.org/pdf/2102.06701.pdf.

[9] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020: 1877—1901.

[10] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022: 27730—27744.

[11] Bang YJ, Cahyawijaya S, Lee N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. (2023-02-08)/[2023-06-21]. https://arxiv.org/pdf/2302.04023.pdf.

[12] Qin C, Zhang A, Zhang Z. et al. Is ChatGPT a general-purpose natural language processing task solver?. (2023-02-08)/[2023-06-21]. https://arxiv.org/pdf/2302.06476.pdf.

[13] Jiao WX, Wang WX, Huang JT, et al. Is ChatGPT a good translator? A preliminary study. (2023-01-20)/[2023-06-21]. https://arxiv.org/pdf/2301.08745.pdf.

[14] Hendy A, Abdelrehim M, Sharaf A, et al. How good are GPT models at machine translation? a comprehensive evaluation. (2023-02-18)/[2023-06-21]. https://arxiv.org/pdf/2302.09210.pdf.

[15] Han RD, Peng T, Yang CH, et al. Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors. (2023-05-23)/[2023-06-21]. https://arxiv.org/pdf/2305.14450.pdf.

[16] Yang XJ, Li Y, Zhang XL, et al. Exploring the limits of ChatGPT for query or aspect-based text summarization. (2023-02-16)/[2023-06-21]. https://arxiv.org/pdf/2302.08081.pdf.

[17] Zhang HP, Liu X, Zhang JW. Extractive summarization via ChatGPT for faithful summary generation. (2023-04-09)/[2023-06-21]. https://arxiv.org/pdf/2304.04193.pdf.

[18] Wang ZZ, Xie QM, Ding ZX, et al. Is ChatGPT a good sentiment analyzer? A preliminary study. (2023-04-10)/[2023-06-21]. https://arxiv.org/pdf/2304.04339.pdf.

[19] Costa-jussà MR, Cross J, Çelebi O, et al. No language left behind: scaling human-centered machine translation. (2022-07-11)/[2023-06-21]. https://arxiv.org/pdf/2207.04672.pdf.

[20] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002: 311—318.

[21] Rei R, De Souza JGC, Alves D, et al. COMET-22: Unbabel-IST 2022 submission for the metrics shared task// Proceedings of the Seventh Conference on Machine Translation. Abu Dhabi: Association for Computational Linguistics, 2022: 578—585.

[22] 宗成庆,夏睿,张家俊. 文本数据挖掘. 第 2 版. 北京：清华大学出版社, 2022.

[23] Roth D, Yih W. A linear programming formulation for global inference in natural language tasks// Proceedings of the Eighth Conference on Computational Natural Language Learning. Boston: Association for Computational Linguistics, 2004: 1—8.

[24] Cabot PLH, Navigli R. REBEL: Relation extraction by end-to-end language generation// Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana: Association for Computational Linguistics, 2021: 2370—2381.

[25] Lin CY. Rouge: a package for automatic evaluation of summaries// Proceedings of the Workshop on Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004: 74—81.

[26] Lin HT, Ma LQ, Zhu JN, et al. CSDS: A fine-grained chinese dataset for customer service dialogue summarization// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021: 4436—4451.

[27] Chen YC, Bansal M. Fast abstractive summarization with reinforce-selected sentence rewriting// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 675—686.

[28] Zhang T, Kishore V, Wu F, et al. BERTScore: evaluating text generation with BERT. (2019-04-21)/[2023-06-21]. https://arxiv.org/pdf/1904.09675.pdf.

[29] Pontiki M, Galanis D, Papageorgiou H, et al. Semeval-2015 task 12: aspect based sentiment analysis// Proceedings of the 9th International Workshop on Semantic Evaluation. Denver: Association for Computational Linguistics, 2015: 19—30.

[30] Zhang YC, Yang YF, Li YH, et al. Boundary-driven table-filling for aspect sentiment triplet extraction// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics, 2022: 6485—6498.

[31] Qin YJ, Hu SD, Lin YK, et al. Tool learning with foundation models. (2023-04-17)/[2023-06-21]. https://arxiv.org/pdf/2304.08354.pdf.

[32] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. (2023-02-27)/[2023-06-21]. https://arxiv.org/pdf/2302.13971.pdf.

[33] Chiang WL, Li Z, Lin Z, et al. Vicuna: an open-source chatbot impressing gpt-4 with 90% * chatgpt quality. (2023-04-14)/[2023-06-21]. https://vicuna.lmsys.org.

Analysis of ChatGPT's Capabilities and Future Prospects

Junhong Wu[1†] Yang Zhao[2†] Chengqing Zong[2*]

1. *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing* 100049

2. *Institute of Automation, Chinese Academy of Sciences, Beijing* 100190

Abstract In recent years, the natural language processing capabilities of large language models have been continuously improving. Particularly, the profound knowledge and powerful conversational abilities exhibited by the ChatGPT have become a globally prominent topic of interest. Questions arise regarding the true level of language understanding capacity possessed by ChatGPT and how its performance compares to specialized models. Can it become a universal model in the entire field of natural language processing (NLP), replacing other models, and even completely solving all NLP problems? To address these questions, this paper conducts a meticulous evaluation and analysis of the performance of ChatGPT across multiple natural language processing tasks. Furthermore, the impact of ChatGPT on the field of natural language processing is discussed, and future developments are anticipated.

Keywords natural language processing; large language model; pre-trained language model; ChatGPT

(责任编辑 崔国增 姜钧译)

† Contributed equally as co-first authors.

* Corresponding Author, Email: cqzong@nlpr.ia.ac.cn