

后ChatGPT时代计算语言学和自然语言处理研究方法

潘琪伟¹, 满思聪²

¹新东方英联邦项目, 江苏 南京

²西安欧亚学院人文教育学院, 陕西 西安

收稿日期: 2023年11月2日; 录用日期: 2023年11月16日; 发布日期: 2023年11月27日

摘要

本研究全方位地探索了ChatGPT在自然语言处理各个子领域的应用及其产生的影响, 涵盖了问答系统、对话系统、文本生成、机器翻译, 以及语音识别和生成等方面。研究表明, ChatGPT不仅在上述任务中提升了性能, 开拓了新的研究和应用途径, 同时也为ChatGPT后的计算语言学和自然语言处理领域带来了新的挑战和机遇。本研究详尽地讨论了如何优化理解和生成自然语言的方法, 如何更有效地处理低资源语言和特定领域的问题。此外, 本文还对未来的研究提出了一些建议和期待, 强调了模型解释性和可靠性的重要性, 以及关注模型对社会产生的影响的必要性。期望本研究的发现和建议能为ChatGPT后的计算语言学和自然语言处理研究提供有益的参考和启示。

关键词

ChatGPT, 语言模型, 自然语言处理, 计算机语言学

Research Methods in Computational Linguistics and Natural Language Processing in the Post-ChatGPT Era

Qiwei Pan¹, Sicong Man²

¹Yinglianbang Program, New Oriental School, Nanjing Jiangsu

²School of Humanities and Education, Xi'an Eurasia University, Xi'an Shaanxi

Received: Nov. 2nd, 2023; accepted: Nov. 16th, 2023; published: Nov. 27th, 2023

Abstract

This study explores the application and impact of ChatGPT in various subfields of natural language

文章引用: 潘琪伟, 满思聪. 后 ChatGPT 时代计算语言学和自然语言处理研究方法[J]. 国外英语考试教学与研究, 2023, 5(4): 174-179. DOI: 10.12677/oetpr.2023.54018

processing, encompassing question answering systems, dialogue systems, text generation, machine translation, as well as speech recognition and generation. The findings suggest that ChatGPT not only enhances performance in these tasks and paves the way for new research and application pathways, but also presents new challenges and opportunities for computational linguistics and natural language processing in the post-ChatGPT era. The study discusses in detail how to optimize the understanding and generation of natural language, and how to more effectively handle low-resource languages and specific domains. In addition, this paper proposes suggestions and expectations for future research, emphasizing the importance of model interpretability and reliability, as well as the necessity to pay attention to the social impact of language models. It is hoped that the findings and suggestions of this study can provide valuable references and insights for research in computational linguistics and natural language processing in the post-ChatGPT era.

Keywords

ChatGPT, Language Models, Natural Language Processing, Computational Linguistics

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

在当前信息化时代, 计算语言学和自然语言处理(NLP)已经崭露头角, 成为至关重要的技术。这些技术赋予计算机理解和生成人类语言的能力[1], 从而深刻地改变了人类与世界的互动方式。无论是搜索引擎的查询、社交媒体的情感分析, 还是智能助手的对话, 以及自动翻译等, NLP 的应用已经深入到生活的各个层面。

在 NLP 的发展历程中, ChatGPT 无疑是一个里程碑式的存在。作为一种先进的自然语言生成模型, ChatGPT 的出现极大地推动了 NLP 的研究和应用[2]。其强大的表现力和广泛的应用, 不仅改变了对计算机能力的认知, 也为 NLP 的研究方法和技术提出了新的挑战 and 机遇。“后 ChatGPT 时代”这一术语, 被用来描述 ChatGPT 以后的 NLP 发展阶段[2], 此阶段以强大的语言模型作为基础, 但同时也面临着如何更好理解和利用这些模型的问题。而随着新的技术和应用的出现, 研究方法和工具的更新和扩充成为了必然。

本文意在探讨在后 ChatGPT 时代, 计算语言学和 NLP 的研究方法应该如何发展。将从语言模型、情感分析、文本生成、机器翻译、语音识别和生成[3] [4] [5]等重要的研究方向进行阐述。每个部分将介绍该方向的基本概念和应用, 评估 ChatGPT 对该领域的影响, 并探讨后 ChatGPT 时代的研究方向。通过这一系列研究, 本文旨在提供一个全面而深入的视角, 来理解和应对后 ChatGPT 时代的计算语言学和 NLP 的挑战和机遇。

2. 语言模型简述

语言模型是自然语言处理的基础之一[6], 其核心任务是预测在给定的某些词的情况下, 下一个词出现的概率。这一预测能力使得语言模型能够产生连贯的、符合语言规则的文本, 因此在许多 NLP 任务中都有广泛的应用, 包括但不限于机器翻译、语音识别、文本生成等。

ChatGPT 作为一种基于 Transformer 的大型语言模型, 极大地改变了语言模型的研究和使用[7]。其规模和复杂性大大超过了之前的模型, 使得它能够理解和生成更复杂、更自然的文本, 并且其预训练和微

调方法[8]使得它能够在广泛的任务和领域中进行有效的迁移学习, 而其生成能力和交互性也开启了一种全新的、基于对话的人机交互方式。

在后 ChatGPT 时代, 语言模型的发展趋势可以从几个方面来看。首先, 模型的规模可能会继续增大, 以获取更好的性能和更广泛的知识, 但如何有效地训练和使用这些大模型, 也必然成为重要的研究问题。其次, 微调方法的研究可能会更加重要[9] [10], 微调不仅可以使模型适应特定的任务和领域, 也可以控制模型的行为和输出。如何更好地融合知识, 使模型不仅能生成连贯的文本, 而且能理解和生成具有深度和广度的知识, 也将是一个重要的研究方向。

3. 情感分析

情感分析是自然语言处理领域中的一个关键研究领域, 其核心目标是从文本中提取和识别作者的情绪或观点[11], 此项技术在众多应用中展现出广泛的适用性, 包括但不限于社交媒体监控、在线评论分析以及品牌声誉管理。

作为一种先进的语言模型, ChatGPT 对情感分析的研究和应用产生了深远影响, 其强大的文本理解能力使其能够理解和生成包含情感和观点的复杂文本, 并且通过预训练和微调的策略, ChatGPT 能够从大量文本中学习和提取丰富的情感知识[12], 并将这些知识应用到具体的情感分析任务中。此外, 其卓越的生成能力也使其能够生成具有特定情感的文本, 进而在对话系统、内容生成等领域发挥关键作用。

在 ChatGPT 的后续研究阶段, 情感分析领域的研究趋势可能会朝着多个方向发展。其中, 情感类别的细化和复杂化是一个显著的趋势。相较于现有的正面和负面的二元分类, 未来的研究可能会探索更多的、更细粒度的情感状态, 如喜悦、悲伤、愤怒、恐惧、惊讶、厌恶等, 并尝试理解这些情感状态之间的关系和转变。另一方面, 对情感理解的深化也成为了研究的重点。这不仅包括识别情感, 更包括理解情感的来源、目标和变化过程。例如, 可能需要理解一个人为什么会感到愤怒, 他的愤怒是针对什么, 以及他的愤怒是如何发展和变化的。这可能需要开发新的模型和技术, 以捕捉和理解这些复杂的情感过程。同时, 跨语言和跨文化的情感分析也成为了研究的焦点。由于情感和观点往往与语言和文化紧密相关, 而大部分现有的模型和数据集都是基于特定的语言和文化的, 因此, 需要开发新的模型和数据集, 以理解和处理不同语言和文化中的情感和观点。此外, 伦理问题也引起了研究者的关注。例如, 需要考虑如何在尊重个人隐私和自由表达的同时, 进行有效的情感分析。这可能需要开发新的伦理指南和政策, 以指导情感分析的研究和应用。后 ChatGPT 时代对情感分析的研究和应用提出了许多潜在的方向, 从开发更细粒度的情感类别, 到考虑伦理问题, 都有广阔的研究空间。

4. 文本生成

文本生成是自然语言处理的一个重要任务, 其目标是生成符合人类语言规则、有意义的文本[13]。文本生成在诸多领域都有广泛的应用, 如新闻撰写、故事创作、对话系统、自动回复等。

ChatGPT 在文本生成领域的影响是深远的, 基于 Transformer 的大规模语言模型架构赋予了它强大的文本理解和生成能力。通过大量的预训练, ChatGPT 能够理解复杂的语言结构和含义, 生成连贯、自然的文本[14], 其微调能力也使得它能够适应各种特定的任务和领域, 并且其交互性使得它能够在对话系统和其他交互式应用中发挥巨大的作用。

在 ChatGPT 的后续研究阶段, 文本生成领域可能会展现出以下几个主要的研究趋势: 首先, 对于更自然的文本生成的追求将成为一个显著的研究方向。这涉及到生成的文本更为贴近人类语言的自然习惯, 以及生成更为复杂、富有表达力的文本。这一目标的实现需要对人类语言的复杂性和多样性有深入的洞察, 包括但不限于语法、语义和修辞等各个层面。同时, 也需要研发更为先进的模型和算法, 以更为精确地模拟人类的语言生成过程。其次, 如何更好地控制生成的文本这一议题包括对文本的风格、观点、

复杂性等进行控制。例如, 我们可能需要生成的文本具有特定的风格, 如正式或非正式, 或者具有特定的观点。这需要理解和操纵文本的各种属性, 以及这些属性如何影响文本的感知和理解以及开发新的模型和技术, 以更为精细地控制文本的生成。此外, 生成的文本不仅仅是模仿和重复已有的文本, 而是能够生成新的、有价值的内容。这可能涉及到如何让模型理解和生成新的概念、新的观点、新的故事等, 对创新性的理解和度量需要进行更深入的研究, 同时亟待开发能够鼓励和奖励创新的模型和算法。

5. 机器翻译

机器翻译是自然语言处理的一个核心任务[15], 其目标是将一种语言的文本自动翻译成另一种语言的文本。机器翻译在多个领域都有广泛的应用, 包括网页翻译、国际交流、跨语言信息检索等。

ChatGPT 在机器翻译领域中也发挥了重要作用。其基于 Transformer 的模型结构和大规模预训练的方法, 使得它能够理解和生成多种语言的文本, 从而进行高质量的机器翻译[16]。此外, ChatGPT 的微调能力, 使得它能够在特定的翻译任务和领域中提供更准确的翻译。其生成能力, 也使得它能够提供更流畅、自然的翻译结果。

在 ChatGPT 的后续研究阶段, 机器翻译领域的研究趋势逐渐呈现多样化, 更好的上下文理解的追求俨然已是显著的研究方向。翻译过程不仅仅是单个词或句子的含义, 往往更需要理解文本的上下文信息, 因此, 对语言的连贯性和上下文依赖性的深入洞察, 以及能够捕捉和利用这些信息的模型和算法的开发, 都成为了研究的重点。另一方面, 更准确的句法和语义转换也引起了研究者的关注[17], 不同语言的句法和语义结构往往有很大的差异, 而现有的机器翻译模型往往在这方面存在问题, 因此, 理解和操纵句法和语义结构, 以及这些结构如何影响翻译的质量都成了进行翻译质量评估的首要难题, 亟待开发新的模型和技术, 以更为精细地控制句法和语义的转换。而如何更有效地进行低资源语言的翻译也逐渐成为全球研究热点[18], 由于大部分现有的模型和数据集都是基于高资源语言的, 低资源语言的机器翻译仍然是一个挑战, 如何利用有限的资源, 如少量的双语数据或者单语数据, 以及如何利用跨语言的相似性和差异性都需要对低资源语言翻译的特点和困难有更深入的理解并开发能够有效处理这些问题的模型和算法。

6. 语音识别和生成

语音识别是将人类的语音信号转化为文本的技术, 而语音生成则是将文本转化为人类的语音信号[4][5]。这两种技术在许多领域都有广泛的应用, 例如智能助手、自动客服、语音翻译、有声书等。

虽然 ChatGPT 主要是一个文本生成模型, 但它也在语音识别和生成领域产生了影响。对于语音识别, ChatGPT 可以作为后处理系统, 对识别结果进行修正和优化。对于语音生成, ChatGPT 可以生成自然、连贯的文本, 这些文本可以被语音合成系统转化为语音。此外, ChatGPT 的强大的语言理解和生成能力, 也为语音识别和生成提供了新的思路和方法。

语音识别和生成领域的研究内容中, 实现更准确的语音识别已经成为各行业领域的重点, 这包括更有效地处理噪声、方言、口音, 对语音信号的复杂性和多样性的深入理解, 包括语音的声学特性、发音的变异性以及更为先进的模型和算法的研发以更为精确地识别和理解语音信号, 这一领域同时也包括了生成更自然的语音, 并且, 这一领域也必然涉及前文所述情感分析板块, 如何实现模型理解和生成情感和个性化的信息, 例如情感的强度、情感的极性、个性化的风格, 因而使得生成的语音能够根据文本的情感和个性化信息, 具有相应的情感和个性化特征。

7. 结论

经过深入探讨, 本研究发现 ChatGPT 在问答系统、对话系统、文本生成、机器翻译以及语音识别和生成等多个自然语言处理任务中的应用和影响显著。ChatGPT 的出现不仅提升了这些任务的性能, 也开

启了新的研究和应用方向。

然而, 后 ChatGPT 时代的计算语言学和自然语言处理也面临着众多的挑战和机遇。这包括如何更好地理解 and 生成自然语言的挑战, 如更优秀的上下文理解、更准确的句法和语义转换、更自然的语音生成、更优秀的情感和个性化表达等。此外, 如何更有效地处理低资源语言和特定领域的问题也是一个重要的挑战。然而, 这些挑战也带来了机遇, 例如, 通过解决这些挑战, 可以开发出更强大、更通用的自然语言处理系统, 这将对许多领域产生深远的影响。

对于未来的研究, 本研究提出了几点建议和期待。首先, 期望能有更多的研究关注上述的挑战, 以推动自然语言处理的发展。其次, 期望能有更多的研究关注模型的解释性和可靠性, 以提高模型的可用性和可信度。此外, 也期望能有更多的研究关注模型的社会影响, 以确保模型的使用能带来积极的社会效果。

本论文的主要贡献在于提供了一个全面的视角, 系统地分析了 ChatGPT 的应用和影响, 以及它对未来研究的启示。期望这些发现和建议能为后 ChatGPT 时代的计算语言学和自然语言处理研究提供有价值的参考。总结来说, 本研究为理解 ChatGPT 在自然语言处理领域的应用和影响, 以及为后 ChatGPT 时代的研究提供了深入的洞见和实用的建议。

参考文献

- [1] Ponte, J.M. and Croft, W.B. (2017) A Language Modeling Approach to Information Retrieval. *ACM SIGIR Forum*, **51**, 202-208. <https://doi.org/10.1145/3130348.3130368>
- [2] Ding, N., Qin, Y.J., Yang, G., et al. (2023) Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nature Machine Intelligence*, **5**, 220-235. <https://doi.org/10.1038/s42256-023-00626-4>
- [3] Haleem, A., Javaid, M. and Singh, R.P. (2022) An Era of ChatGPT as a Significant Futuristic Support Tool: A Study on Features, Abilities, and Challenges. *Bench Council Transactions on Benchmarks, Standards and Evaluations*, **2**, Article ID: 100089. <https://doi.org/10.1016/j.tbench.2023.100089>
- [4] Kuhn, R. and De Mori, R. (1990) A Cache-Based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 570-583. <https://doi.org/10.1109/34.56193>
- [5] Brown, P., Cocke, J., Pietra, S.A., et al. (1990) A Statistical Approach to Machine Translation. *Computational Linguistics*, **16**, 79-85.
- [6] Brown, P.F., Della Pietra, V.J., Desouza, P.V., et al. (1992) Class-Based n-gram Models of Natural Language. *Computational Linguistics*, **18**, 467-480.
- [7] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention Is All You Need. In: *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, Curran Associates, New York, 5990-6008.
- [8] Devlin, J., Chang, M.W., Lee, K., et al. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.
- [9] Ding, N., Qin, Y.J., Yang, G., et al. (2023) Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nature Machine Intelligence*, **5**, 220-235. <https://doi.org/10.1038/s42256-023-00626-4>
- [10] Chung, H.W., Hou, L., Longpre, S., et al. (2022) Scaling Instruction-Fine-Tuned Language Models.
- [11] Wang, Z.Z., Xie, Q.M., Ding, Z.X., et al. (2023) Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. <https://arxiv.org/pdf/2304.04339.pdf>
- [12] Peters, M., Neumann, M. and Iyyer, M. (2018) Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, 2227-2237. <https://doi.org/10.18653/v1/N18-1202>
- [13] Mauldin, M.L. (1984) Semantic Rule Based Text Generation. *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Annual Meeting of the Association for Computational Linguistics*, Stanford, 2-6 July 1984, 376-380. <https://doi.org/10.3115/980491.980568>
- [14] Zhang, T.Y., Ladhak, F., Durmus, E., et al. (2023) Benchmarking Large Language Models for News Summarization.
- [15] 宗成庆. 统计自然语言处理[M]. 第2版. 北京: 清华大学出版社, 2013.
- [16] Ouyang, L., Wu, J., Jiang, X., et al. (2022) Training Language Models to Follow Instructions with Human Feedback. *36th*

Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, 28 November-9 December 2022, 27730-27744.

- [17] Cohen, J.P., Luck, M. and Honari, S. (2018) Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference*, Granada, 16-20 September 2018, 529-536. https://doi.org/10.1007/978-3-030-00928-1_60
- [18] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y. and Fung, P. (2023) Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, **55**, Article No. 248. <https://doi.org/10.1145/3571730>