张倬诚

检索增强生成　大语言模型　机器翻译　长序列建模

中国科学院计算技术研究所

智能信息处理重点实验室

+86-15619026195

zhuocheng_zhang@163.com

zhangzhuocheng20z@ict.ac.cn

GitHub

教育经历

•中国科学院计算技术研究所	**计算机应用技术**	研究生（工学博士）	*2020.9 至 今*
•西安交通大学	**自动化**	本科（工学学士）	*2016.9 至 2020.6*
•西安交通大学	**人工智能（菁英班）**	本科（辅修）	*2016.9 至 2020.6*

主要科研经历

•检索增强生成

– **LevelRAG：层次化多跳混合检索框架**：该方法由一个高层搜索器和多个低层搜索器构成，分别进行复杂问题的检索规划和针对检索器的查询自适应优化。为了进一步利用稀疏检索在精细化搜索方面的优势，我们又提出了一套**基于 Lucene 语法的改写策略**并将其应用于稀疏搜索器。最终我们的方法在各个数据集上都取得了显著的性能提升。

– **FlexRAG：灵活高效的检索增强生成系统**：主导研发 FlexRAG 框架，具备高可复现性、易用性和优异性能，支持文本、多模态和网络访问等多种 RAG 场景。框架提供完整流水线及评估流程，支持异步处理与持久化缓存，并兼容 Hugging Face 生态，便于快速构建 RAG 系统。相关成果已被 **ACL 2025 Demo Track** 接收。

•篇章级机器翻译

– **篇章级机器翻译中的缩放定律**：通过大规模实验发现，过长上下文并不能持续提升翻译质量，且**最优上下文长度与模型规模呈对数关系**。进一步分析发现误差累积是限制篇章翻译性能的重要因素。该研究发表于 **EMNLP 2023 Findings**。

– **篇章级机器翻译中的长度偏差问题及缓解**：针对传统模型在处理不同长度输入时性能不稳的问题，提出**句级滑动窗口、注意力长度缩放与动态训练长度采样**方法，有效缓解篇章翻译的长度偏差问题，使模型可稳定处理任意长度序列。该研究发表于 **EMNLP 2023 Findings**。

项目、竞赛及实习经历

•第 18 界全国机器翻译大会（CCMT2022） 2022.4-2022.5

组织参与汉泰方向翻译测评，应用迭代式数据增强，启发式数据清洗、正则化对比学习等技术，并取得该赛道**第一名**的成绩。

•腾讯，信息安全部，技术工程事业群 2021.9-2022.9

进行篇章翻译的算法研究，组织标注了民族语言到汉语方向的篇章级翻译数据，弥补了该领域的数据缺陷；在算法方面设计了多粒度多头注意力机制，有效提高了篇章级机器翻译模型的翻译质量。

•华为，技术创新部，云与计算事业群 2018.6-2019.9

进行了 CDN 内容感知的缓存算法开发，进行了图片标题生成算法的开发及复现。

个人技能

大模型相关: 检索生成、数据精炼、模型融合、长序列技术

英语能力: TOEFL 101、CET-4 608

编程语言: 熟练掌握 Python，代码风格较好，熟悉 Pytorch、Deep Speed、Transformers、Faiss 等框架

论文发表情况

- **Zhang Zhuocheng**, Yang Feng, and Min Zhang. 2025. FlexRAG: A Flexible and Comprehensive Framework for Retrieval-Augmented Generation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (System Demonstrations). Association for Computational Linguistics.
- **Zhang Zhuocheng**, Feng Yang, Zhang Min. LevelRAG: Enhancing Retrieval-Augmented Generation with Multi-hop Logic Planning over Rewriting Augmented Searchers. arXiv preprint arXiv:2502.18139.
- **Zhang Zhuocheng**, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Addressing the Length Bias Challenge in Document-Level Neural Machine Translation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 11545-11556, Singapore. Association for Computational Linguistics.
- **Zhang Zhuocheng**, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Scaling Law for Document Neural Machine Translation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8290-8303, Singapore. Association for Computational Linguistics.
- Langlin Huang, Shuhao Gu, **Zhang Zhuocheng**, and Yang Feng. 2023. Enhancing Neural Machine Translation with Semantic Units. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2264-2277, Singapore. Association for Computational Linguistics.
- Shaolei Zhang, Qingkai Fang, **Zhang Zhuocheng**, Zhengrui Ma, Yan Zhou, Langlin Huang, ... and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. arXiv preprint arXiv:2306.10968.
- Yuxia Wu, Guoshuai Zhao, Mingdi Li, **Zhang Zhuocheng**, and Xueming Qian. 2023. Reason Generation for Point of Interest Recommendation Via a Hierarchical Attention-Based Transformer Model. IEEE Transactions on Multimedia.