Yixuan Wang

wyxstriker | ✉ yixuanwang@ir.hit.edu.cn | ☎ (+86)130-0852-0003

Summary

I am currently a second-year graduate student (expected to graduate in 2025) at HIT@SCIR, advised by Prof. Wanxiang Che. Previously, my main research interests focused on deep learning for natural language generation (NLG), e.g., grammatical error correction. Recently, my main research interest has been in the study of efficient inference for large language models, in particular speculative decoding.

Education

Harbin Institute of Technology 2022 - present

– Master's Degree in Computer Science and Technology

– Advisor: Prof. Wanxiang Che

– Assessment: 91.01 (7/259)

– Honor: **National Scholarship** (2023), Academic Scholarship (First Class 2022)

Harbin Institute of Technology 2018 - 2022

– Bachelor's Degree in Computer Science and Technology

– Assessment: 93.59 (13/184)

– Honor: People's Scholarship (2018, 2019, 2020)

Publications

[1] **Yixuan Wang***, Xianzhen Luo*, Fuxuan Wei, Yijun Liu, Qingfu Zhu, Xuanyu Zhang, Qing Yang, Dongliang Xu, Wanxiang Che. ”Make Some Noise: Unlocking Language Model Parallel Inference Capability through Noisy Training”. Submitted to EMNLP 2024.

- We propose a noisy training framework Make Som Noise as an alternative to supervised fine-tuning. It can significantly enhance the parallel inference of the model without significant task performance loss.

[2] Xianzhen Luo, **Yixuan Wang**, Qingfu Zhu, Zhiming Zhang, Xuanyu Zhang, Qing Yang, Dongliang Xu, Wanxiang Che. ”Turning Trash into Treasure: Accelerating Inference of Large Language Models with Token Recycling”. Submmitted to AAAI 2025.

- We propose a retrieval-based train-free speculative decoding method Token Recycling. that requires no training. It constructs efficient draft tokens by turning historical decoding information into usable adjacency matrix. Experiments show that it far outperforms existing retrieval-based methods, achieving close to 2x speedup for LLM inference.

[3] **Yixuan Wang**, Baoxin Wang, Yijun Liu, Qingfu Zhu, Dayong Wu, Wanxiang Che. ”Improving Grammatical Error Correction via Contextual Data Augmentation”. ACL 2024 Findings.

- We combine grammatical errors from different samples as features and use a pre-trained language model to generate richer contexts for these structured grammatical error templates to achieve data augmentation.

[4] **Yixuan Wang**, Baoxin Wang, Yijun Liu, Dayong Wu, Wanxiang Che. ”LM-Combiner: A Contextual Rewriting Model for Chinese Grammatical Error Correction” LREC-COLING 2024.

- We propose a model-based rewriting soft ensemble method to alleviate the issue of over-correction in grammatical error correction.

[5] Zhongyuan Wang, **Yixuan Wang**, Shaolei Wang, Wanxiang Che. ”Adaptive Unsupervised Self-training for Disfluency Detection”. COLING 2022 Oral.

- We propose an adaptive self-training framework that, compared to simple filtering selection, performs re-weighted learning on all samples based on the confidence scores from a discriminator.

Projects

Data Construction for Huozi LLM Mar. 2023 - May. 2023

I am primarily responsible for the pre-training and instruction fine-tuning dataset construction for the Huozi large language model. For pre-training data, I referenced the cc_net framework to perform deduplication and quality assessment on the collected high-quality Chinese internet text. For instruction fine-tuning dataset construction, I leveraged scripts from Flan to implement the conversion of various NLP task datasets to the SFT dataset, and organized the completion of the construction of the Huozi model's SFT instruction data.

Text Security Assistant for Southern Daily Sep. 2022 - Sep. 2023

As the primary developer, I completed the development of the text security assistant system for Southern Daily. The main functionalities of the system include general text correction, correction of entities such as place names and time, and correction of collocations between various types of entities. This project primarily involves techniques such as word segmentation, part-of-speech tagging, and named entity recognition based on the BERT model, as well as text generation using the BART model.

Competition

2023-07	Achieved the 1st Award on Chinese Essay Fluency Evaluation track 1&2&3, CCL2023.
2022-09	Achieved the 1st Award on UMETRIP-QA track 2&3, CCL2022.