Zhehao ZHANG

(+86) 136-2971-2586 • zhehao_zhang@outlook.com • GitHub • Personal Website

Education

Shanghai Jiao Tong University (SJTU)(# 3 Best in China on US News 2023) 09/2019 – 07/2023 (Expected)

Bachelor of Engineering in Artificial Intelligence **(Zhiyuan Honor Degree)** Shanghai, China

GPA: 88.51/100 (3.73/4.0) (3-rd year: 91.87/100)

- Awards: Zhiyuan honor program scholarship (2019-2022, 5%)
- Selected courses: Natural language processing (94 points), Reinforcement learning (94 points), Data structure [Honor] (92 points), Knowledge representation and reasoning (97 points), Intelligent speech recognition (92 points), Data mining (91 points).

Publication

[1] **Zhehao Zhang**, Handong Zhao, Tong Yu, Shuai Li. 2023. Exploring Soft Prompt Initialization Strategy for Few-shot Continual Text Classification, Currently Under Review for *The 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*

[2] **Zhehao Zhang**, Jiaao Chen, Diyi Yang. 2023. Mitigating Biases in Hate Speech Detection from A Causal Perspective, Currently Under Review for *The 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*

Research Experiences

Stanford University, **Social and Language Technologies (SALT) Lab** Jun 2022 – Jan 2023

Research Assistant, Advisor: Prof. Diyi Yang Stanford, CA

- Utilized grammar induction and mutual information to search for biased grammar patterns on hate speech detection datasets. Analyzed the spuriousness of biased patterns, discovered identity biases using causal interference, and then proposed a method to mitigate such biases using a Multi-Task Intervention (MTI) and Data-Specific Intervention (DSI) based on the confounders.
- Validated the effectiveness of the method by running experiments and analyzing the results across nine hate speech detection datasets with an out-of-domain challenge set to reach positive conclusions on its use for reducing hate speech bias. **8.61%** F-1 score improvement indicates the effectiveness of our method. This work has been submitted to ***ACL 2023***.

Shanghai Jiao Tong University, **John Hopcroft Center for Computer Science** May 2021 – Jun 2022

Research Assistant, Advisor: Prof. Shuai Li, Co-mentor: Handong Zhao and Tong Yu Shanghai, China

- Introduced context and label space information for prompt initialization in the setting of few-shot continual learning with state-of-the-art performance (**20.95% accuracy improvement** in 4-shot setting) in commonly-used text classification benchmarks. Successfully addressed catastrophic forgetting and fast adaptation simultaneously and submitted the work to ***ACL 2023***.
- Analyzed CLIP (Contrastive Language–Image Pre-training) model based on the embedding's isotropy, ensuring its zero-shot effectiveness by confirming its uni-variance and uncorrelatedness of its dimensions using histogram and heatmaps visualization.

Singapore University of Technology and Design, **Stat NLP group**, Mar 2021 – Mar 2022

Research Assistant, Advisor: Prof. Wei Lu Singapore

- Systematically analyzed the architecture of Transformers and its variants (e.g., Linear Transformers) to implement a proof of concept model from scratch for machine translation on WMT14 datasets, with comparative performance (bleu and perplexity) to the original.
- Studied and presented the explainability of neural networks, especially from Neural Tangent Kernel's perspective (NTK), through 4 comprehensive English oral presentations to Stat group meetings of over 30 members with strong positive feedback.

Industry Experiences

Microsoft Research Asia, **Data, Knowledge, and Intelligence Lab** December 2022 – Present

Research Intern, Advisor: Dr. Yan Gao Beijing, China

- Explored Large Language Models'(*e.g.,* GPT-3, Instruct-GPT, etc.) reasoning ability on structured data. Working on constructing a large-scale Table-based implicit Question Answering dataset which requires the model to have multi-step complex reasoning capability.

Selected Projects

Chinese Medical Named Entity Recognition (NER) Apr 2022 – Jun 2022

- Located and classified medical-related entities (e.g., symptoms, organs, etc.) on a large-scale Chinese biomedical dataset (CBLUE). Implement BERT and Roberta with Conditional random field (or Long short-term memory) baseline model for both vanilla and nested settings. The F-1 score of 63.392 indicates the effectiveness of this baseline model.
- Further introduce the Flat-Lattice Transformer model to incorporate word information (besides character information) and other techniques, including adversarial training and layer-wise learning rate decay, improving 3.18 F-1 scores over the baseline model.

Skills

Programming Languages: C/C++, Python, MATLAB
Tools and Frameworks: Git, GitHub,LaTeX, PyTorch, Huggingface transformers, Numpy, Scikit-learn, OpenCV, pandas
Spoken Language: English (IELTS overall band score 7.5), Mandarin