ZHENGXIANG WANG

Computational Linguistics Ph.D. Candidate

zhengxiang.wang@stonybrook.edu | 631-739-7389 | Stony Brook, NY

EDUCATION

Ph.D. in Computational Linguistics

Stony Brook University

Sep 2022 – Present | Stony Brook, NY

Advisor: Owen Rambow

Selected Courses: CSE 548 Analysis of Algorithms, A-; CSE 538 Natural Language Processing, A; AMS 580 Machine Learning, A.

M.A. in Applied Linguistics

University of Saskatchewan

Sep 2019 – May 2021 | Saskatoon, Canada

B.A. in Chinese Language and Literature

Hunan University

Sep 2015 – June 2019 | Changsha, China

EMPLOYMENT

Ph.D. Data Science Intern

The Home Depot

May 2024 – Aug 2024 | Hybrid

- Developed a clustering-based topic modeling pipeline reducing LLM runtime and API calls by 90+%. Further eliminated redundancy in the raw identified topics by 80+%.
- Finetuned SBERT embeddings to build an efficient semantic search network (with Faiss) to select most likely customer pain points
- Fine-tuned BERT-like models and LLMs (e.g., full fine-tuning and PEFT) to predict customer pain points
- Built various LLM-based autonomous validation systems (e.g., multi-agent debating), saving up to 24% in manual annotation efforts

PUBLICATIONS

- **Wang, Z.**, Kodner, J. & Rambow, O. (2024). Exploring the Zero-Shot Capabilities of LLMs Handling Multiple Problems at once. Preprint arXiv 2406.10786. PDF | Code
- **Wang, Z.** & Rambow, O. (2024). Clustering Document Parts: Detecting and Characterizing Influence Campaigns From Documents. *Proceedings of the 6th Workshop on Natural Language Processing and Computational Social Science.* (At NAACL 2024) PDF | Code
- **Wang, Z.** (2023). Probabilistic Linguistic Knowledge and Token-level Text Augmentation. In *Practical solutions for Diverse Real-World NLP Applications*. Book: *Signals and Communication Technology*. PDF
- **Wang, Z.** (2023). Learning Transductions and Alignments with RNN Seq2seq Models. *Proceedings of the 16th International Conference on Grammatical Inference*. In *PMLR, volume 217*. PDF | Code
- Hao, H., Cui, Y., **Wang, Z.** & Kim, Y. (2022). Thirty-Two Years of IEEE VIS: Authors, Fields of Study and Citations. *IEEE Transactions on Visualization and Computer Graphics*. PDF | Code | Website
- **Wang, Z.** (2022). Linguistic Knowledge in Data Augmentation for Natural Language Processing: An Example on Chinese Question Matching. *Proceedings of the 5th International Conference on Natural Language and Speech Processing*. PDF | Code

RECOGNITIONS

- IACS Junior Researcher Award ($37,000), 2024
- NSF BIAS-NRT Research Traineeship, 2023
- SBU Distinguished Travel Award ($1,750), 2023
- Globalink Graduate Fellowship ($15,000), 2019
- Chinese Government Scholarship ($7,000), 2018
- Chinese National Student Innovation Training Program Grant (¥10,000), 2017
- Chinese National Scholarship (¥8,000), 2016

PROJECTS

Analytic Assessment Capabilities of LLMs

- Examined the capabilities of LLMs providing multi-dimensional analytic scoring and feedback jointly for graduate-level academic English writing
- Created a multi-LLM debating framework to compare the quality of human- and LLM-generated feedback on multiple evaluation criteria

Multi-problem Evaluation of LLMs

- Comprehensively evaluated the capabilities of 7 LLMs from 4 model families concurrently handling multi-problem prompts constructed based on 6/12 classification/reasoning benchmarks
- Proposed multi-problem prompting, which can save up to 82% LLM inference cost per problem

Influence Campaigns Modeling (DARPA INCAS)

- Created and deployed an end-to-end generative LLM-based clustering pipeline to detect and characterize influence campaigns from documents
- Pipeline included spaCy preprocessing, Flan-T5 fine-tuning, and SBERT/UMAP/HDBSCAN for text embedding, embedding reduction, and clustering

TUTORIAL

Notes for Stanford CS224N *Natural Language Processing with Deep Learning*

2021

- Covered the conceptual and mathematical basics of word embedding, neural networks, deep learning models. [*GitHub*] 63 stars and 32 forks.

SKILLS

- *Programming*: Python, R, LaTeX, Bash, Git, SQL
- *Tools*: OpenAI API, LangChain, Transformers, PyTorch, TensorFlow, NumPy, Pandas, scikit-learn
- *Cloud Platforms*: Google Cloud Platform, AWS
- *Languages*: English, Mandarin, Fuqing Dialect