Shiyue Zhang

https://www.cs.unc.edu/shiyue

June 30, 2024

Email:shiyue@cs.unc.edu

Work Experience

Aug. 2023–Present **LLM Research Engineer**
Bloomberg AI, NYC, USA

- Benchmarking open-sourced large language models (LLMs) on finance-related benchmarks
- Finetuning LLMs on specific finance-related tasks
- Continued pre-training LLMs to inject financial domain knowledge
- Building retrieval systems and leveraging retrieval in model training and evaluation
- Supporting production teams to use LLMs in products

Education

Aug. 2018–Aug. 2023 **Ph.D. in Computer Science**
University of North Carolina at Chapel Hill, Chapel Hill, USA
Advised by Mohit Bansal

- Research topic was natural language processing (NLP)
- Mainly focused on language generation, including question generation, machine translation, summarization, and language modeling
- Thesis: Towards Reliable and Inclusive Natural Language Generation

Sep. 2011–Mar. 2018 **M.E. and B.E. in Communication Engineering**
Beijing University of Posts and Telecommunications, Beijing, China

Internship Experience

May 2022–Aug. 2022 **Software Engineer Intern**
Bloomberg AI, NYC, USA
Supervised by Ozan Irsoy, Steven Lu, Shijie Wu, David Rosenberg, and Mark Dredze

- Worked on language modeling
- To overcome the over-generalization problem of MLE-trained LMs, we propose a novel training objective, MixCE, that combines forward and reverse cross-entropies

May 2021–Aug. 2021 **Research Intern**
Facebook AI, remote from Chapel Hill, USA
Supervised by Vishrav Chaudhary and Francisco (Paco) Guzmán

- Worked on multilingual tokenization
- Analyzed how downstream translation performance is affected by the language imbalance in the data used to train a multilingual tokenizer

June 2020–Aug. 2020 **Research Intern**
Microsoft Research, remote from Chapel Hill, USA
Supervised by Asli Celikyilmaz

- Worked on email thread summarization
- Collected a dataset with email threads and human-written summaries, and benchmarked multiple generation models on this dataset

Sep. 2016–June 2018 **Research Intern**
CSLT at Tsinghua University, Beijing, China
Supervised by Dong Wang and Yang Feng

- Worked on machine translation
- Augmented neural machine translation models with a memory component that stores discrete dictionary information

Selected Awards

2023 EECS Rising Star
2021 Bloomberg Data Science PhD Fellowship
2015 Excellent Graduate
2014, 2013 National Scholarship

Publications

2023 David Wan, **Shiyue Zhang**, and Mohit Bansal. *HistAlign: Improving Context Dependency in Language Generation by Aligning with History* EMNLP 2023

2023 **Shiyue Zhang**, Shijie Wu, Ozan Irsoy, Steven Lu, Mohit Bansal, Mark Dredze and David Rosenberg. *MixCE: Training Autoregressive Language Models by Mixing Forward and Reverse Cross-Entropies* ACL 2023

2023 Derek Tam, Anisha Mascarenhas, **Shiyue Zhang**, Sarah Kwan, Mohit Bansal, Colin Raffel. *Evaluating the Factual Consistency of Large Language Models Through Summarization* Findings of ACL 2023

2023 **Shiyue Zhang***, David Wan*, and Mohit Bansal. *Extractive is not Faithful: An Investigation of Broad Unfaithfulness Problems in Extractive Summarization* ACL 2023

2023 Swarnadeep Saha, **Shiyue Zhang**, Peter Hase, Mohit Bansal. *Summarization Programs: Interpretable Abstractive Summarization with Neural Modular Trees* ICLR 2023

2022 Xiang Zhou, **Shiyue Zhang**, and Mohit Bansal. *Masked Part-Of-Speech Model: Does modeling long context help unsupervised POS-tagging?* NAACL 2022

2022 Yinuo Hu*, **Shiyue Zhang***, Viji Sathy, A. T. Panter, and Mohit Bansal. *SETSum: Summarization and Visualization of Student Evaluations of Teaching* NAACL Demo 2022

2022 **Shiyue Zhang**, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. *How Robust is Neural Machine Translation to Language Imbalance in Multilingual Tokenizer Training?* AMTA 2022

2022 **Shiyue Zhang**, Benjamin Frey, and Mohit Bansal. *How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language* ACL 2022 Theme Track

2021 **Shiyue Zhang** and Mohit Bansal. *Finding a Balanced Degree of Automation for Summary Evaluation* EMNLP 2021

2021 **Shiyue Zhang**, Benjamin Frey, and Mohit Bansal. *Cherokee-English Machine Translation Demo with Quality Estimation and Corrective Feedback* ACL Demo 2021, [News: The sanctity of Cherokee]

2021 **Shiyue Zhang**, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. *EmailSum: Abstractive Email Thread Summarization* ACL 2021

2021 Zineng Tang, **Shiyue Zhang**, Hyounghun Kim, and Mohit Bansal. *Continuous Language Generative Flow* ACL 2021

2020 **Shiyue Zhang**, Benjamin Frey, and Mohit Bansal. *ChrEn: Cherokee-English Machine Translation for Endangered Language Revitalization* EMNLP 2020

2020 Peter Hase, **Shiyue Zhang**, Harry Xie, and Mohit Bansal. *Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?* Findings of EMNLP 2020

2019 **Shiyue Zhang** and Mohit Bansal. *Addressing Semantic Drift in Question Generation for Semi-Supervised Question Answering* EMNLP 2019

2018 Jiyuan Zhang, Zheling Zhang, **Shiyue Zhang**, and Dong Wang. *VV-Couplet: An open source Chinese couplet generation system* APSIPA ASC 2018

2017 Lantian Li, Zhiyuan Tang, Dong Wang, Andrew Abel, Yang Feng, and **Shiyue Zhang**. *Collaborative learning for language and speaker recognition* NCMMSC 2017

2017 Yang Feng, **Shiyue Zhang**, Andi Zhang, Dong Wang, and Andrew Abel. *Memory-augmented Neural Machine Translation* EMNLP 2017

2017 **Shiyue Zhang**, Gulnigar Mahmut, Dong Wang, and Askar Hamdulla. *Memory-augmented Chinese-Uyghur Neural Machine Translation* APSIPA ASC 2017

2017 Aodong Li, **Shiyue Zhang**, Dong Wang, and Thomas Fang Zheng. *Enhanced Neural Machine Translation by Learning from Draft* APSIPA ASC 2017

2017 Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, **Shiyue Zhang**, and Andi Zhang. *Flexible and Creative Chinese Poetry Generation Using Neural Memory* ACL 2017

2017 Dong Wang, Thomas Fang Zheng, Zhiyuan Tang, Ying Shi, Lantian Li, **Shiyue Zhang**, Hongzhi Yu, Guanyu Li, Shipeng Xu, Askar Hamdulla, Mijit Ablimit, and Gulnigar Mahmut. *M2ASR: Ambitions and first year progress* O-COCOSDA 2017

2017 Zhiyuan Tang, Ying Shi, Dong Wang, Yang Feng, **Shiyue Zhang**. *Memory visualization for gated recurrent neural networks in speech recognition* ICASSP 2017

Technical Reports

2017 **Shiyue Zhang**, Pengtao Xie, Dong Wang, and Eric P. Xing. *Medical Diagnosis From Laboratory Tests by Combining Generative and Discriminative Learning* on arXiv

2017 Yang Wang, Dong Wang, **Shiyue Zhang**, Yang Feng, Shiyao Li, and Qiang Zhou. *Deep Q-trading*

2016 Zhiyuan Tang, Ying Shi, Dong Wang, Yang Feng, and **Shiyue Zhang**. *Visualization analysis for recurrent networks*

Professional Services

Area Chair	EACL 2024, NAACL 2024, EMNLP 2024
Session Chair	AMTA 2022
Reviewer	ACL Rolling Review, NSF Proposal
Conference Reviewer	ACl 2023, EMNLP 2021/2022, COLING 2022, AAAI 2020/2022, AKBC 2020/2022
Journal Reviewer	Natural Language Engineering (JNLE)
Workshop Reviewer	SRW@ACL 2020/2021, Eval4NLP 2020/2021, SRW@EACL 2023

Skills

Deep Learning Frameworks:	PyTorch, TensorFlow
Programming Languages:	Python
Speaking Languages:	Chinese (Native), English (Full Professional Proficiency)