

Jian Zhu

COMPUTATIONAL LINGUIST

☎ (+1) 734-709-9552 | ✉ lingjzhu@umich.edu | 🌐 <https://lingjzhu.github.io/> | 📄 <https://github.com/lingjzhu>

Summary

I am a Ph.D. student in computational linguistics with training in both linguistics and natural language processing. I have 5+ years of experience in language research. I am familiar with the pipelines of speech recognition/synthesis and general tasks of natural language processing. My research works are published in top-tier NLP conferences including EMNLP and NAACL.

Education

University of Michigan-Ann Arbor

PH.D. IN LINGUISTICS AND SCIENTIFIC COMPUTING (JOINT PHD PROGRAM)

Ann Arbor, MI, USA

2017 - 2023 (Expected)

Beijing Foreign Studies University

M.A. IN LINGUISTICS & B.A. IN TRANSLATION AND INTERPRETING

Beijing, China

2010 - 2017

Skills

Programming Languages Python, R, SQL, Bash shell, Julia, C++(basic)

Machine Learning PyTorch, Keras, Sklearn, Deep Graph Library, Tensorflow

Natural Language Processing Transformers, Spacy, NLTK, Stanza

Speech Processing Kaldi, ESPNet, Praat, Librosa

Data Processing Regex, Pandas, Matplotlib, Seaborn, Networkx

Natural Languages Chinese (*Native*), Cantonese (*Native*); English (*Professional*); Japanese (*Intermediate*)

Project Experience

Lexical change in Reddit [Github] (NAACL 2021)

PYTHON, NETWORKX, NETWORKIT, RAY

University of Michigan

April. 2020 - Nov. 2020

- Preprocessed 2TB of raw Reddit posts using parallel processing framework **Ray**;
- Constructed 270k social networks for subreddits at different times and extracted network features with parallelization;
- Used deep survival analysis and other machine learning methods to analyze lexical change.

Large-scale authorship verification [Github] (EMNLP 2021)

PYTHON, TRANSFORMERS, NLTK, SPACY

University of Michigan

Aug. 2020 - Nov. 2020

- Trained Siamese RoBERTa/BERT to perform authorship verification on 60k authors of Amazon reviews;
- Achieved 10% increase in accuracy over previous LSTM models;
- Analyzed the contribution of tokenization methods and the linguistic regularity in the learned style embedding.

Text-independent phone-to-audio alignment [Github] (Submitted to ICASSP)

PYTHON, PYTORCH, TRANSFORMERS

University of Michigan

June. 2021 - Now

- Developed a semi-supervised transformer model to align phones to audio;
- Implemented two text independent phone segmentation models using Wav2Vec2;
- Trained speech recognition model on over 2000 hours of speech;
- Deployed the model to speed up the otherwise time-consuming process of manual phone segmentation.

BERT-enhanced Mandarin TTS [Github] (Speech Prosody 2020)

PYTHON, PYTORCH, LIBROSA

University of Michigan

Sep. 2019 - Dec. 2019

- Implemented Chinese text frontend for Mandarin text-to-speech (TTS) models;
- Performed forced alignment with Kaldi to extract time-aligned features;
- Combined BERT embeddings with Tacotron 2 to improve the naturalness of synthesized prosody;
- Trained neural vocoders such as WaveGlow and WaveNet for Mandarin.

Coursework

GPA: 4.0

Machine Learning Linear Models, Computational Data Science, Numerical Linear Algebra, Machine Learning, Deep Learning, Optimization

Natural Language Processing Natural Language Processing, Computational Sociolinguistics, Situated Natural Language Processing, Natural Language Generation

Linguistics & Cognitive Science Phonetics, Phonology, Syntax, Semantics, Neurolinguistics, Perception