Junghyun Min

jm3743@georgetown.edu | Google Scholar | Personal website | Washington, DC | (443) 414-4405

WORK EXPERIENCE

University of Toronto, Toronto, ON
Visiting Researcher, Computer Science — May 2025 – Aug 2025
- Investigate effects of transfer learning for low-resource NLU; open-source produced models, recipes, and benchmarks.
- Supervise student researchers on topics in low-resource languages, machine translation, and cross-lingual transfer.

NCSOFT, Seongnam, Korea
NLP Engineer, Financial Language Understanding — Jan 2021 – Apr 2024
- Implemented and trained tokenizers for the Language Model Task Force, precursor to the VARCO LLM family.
- Developed and served asynchronous Stanza-like NLP API, parsing 10k requests per sec on 4GB VRAM at 96% acc.
- Introduced punctuation restoration as pre-training objective, format loss and forced decoding for rel, entity extraction.
- Extract financial / biochemical entities and relations for downstream tasks like market sensing, drug discovery.

Harford Community College, Bel Air, MD
Data Analyst, Analytics & Planning — Apr 2018 – Jul 2019
- Distilled expert insight in student retention and success in machine learning models with ~80% accuracy in Wolfram.
- Automated recurring data validation and reports in SAS, SQL, leading to 20% increase in request processing volume.

EDUCATION

Georgetown University, Washington, DC
Ph.D. Computational Linguistics. Advisor: Ethan Wilcox. — Exp. 2029

Johns Hopkins University, Baltimore, MD
M.A. Cognitive Science. Advisor: Tal Linzen. — Dec 2020

Johns Hopkins University, Baltimore, MD
B.S. Physics, secondary major in Mathematics. — Dec 2017

SELECTED PUBLICATIONS

- **Junghyun Min**, Minho Lee, Woochul Lee, Yeonsoo Lee. RepL4NLP at NAACL 2025. Punctuation Restoration Improves Structure Understanding without Supervision. Tech blog (Korean).
- **Junghyun Min**, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. ACL 2020. Syntactic data augmentation increases robustness to inference heuristics.

SELECTED TECHNICAL PROJECTS

Lead, Information theoretic approach to the syntax-prosody interface.
- Develop multi-modal architecture that also receives extracted phonological features as input to best predict syntax.
- Measure the reduction of entropy in predicting syntax with additional modal feature (e.g. text, duration, pause, pitch).

Lead, Visually grounded prepositions.
- Design vision-language model pipeline, explore how visually encoded spatial relations affect lexical prob distribution.

Co-lead, LLM legal interpretation.
- Implement LLM legal interpretation judgment extraction pipeline via vLLM. Oral presentation at NLLPW at EMNLP.
- Show unreliability of LLM legal interp. due to sensitivity to surface form, unstable correlation to human judgment.

Team lead, CantoNLU.
- Train open-source, open-recipe Cantonese encoder-only LMs from scratch and via transfer from Mandarin Chinese.
- Collect training data, compile evaluation benchmark for Cantonese NLU. Evaluate baseline and commercial LLMs.

Lead, Punctuation restoration as pre-training objective.
- Proposed and discovered unsupervised, structure-related representation learning objective in punctuation restoration.
- Presented improvements in 7 structure-related tasks with PR, improving OpenIE, NER, SRL systems by 11%p on avg.

Other projects
- **Lead engineer** for **ai.ly**, a GPT-based, personalized AI lyricist. 50k visits over 3 months of service. Hip-hop sample.
- **Technical lead** for wecommit's prototype **genDOC**, an LLM-powered document automation solution for startups.

SKILLS

Programming Languages: Python (proficient), Java, JavaScript, C++, R, Unix shell, SAS, SQL, Wolfram.
Tools and Libraries: PyTorch, TensorFlow, transformers. Flask, FastAPI, async., GCP, Docker, Hydra.
LLM Use: OpenAI, LangChain, prompt engineering. vLLM, quantization, distributed training.
Natural Languages: Korean (Standard, Busan), English, German, Chinese (Mandarin)