Lingyu Gao

Toyota Technological Institute at Chicago (TTIC)
6045 S. Kenwood Ave., Chicago, IL 60637
lygao@ttic.edu | |

Summary

Final-year Ph.D. student specializing in natural language processing (NLP). My primary focus is on text classification and generation, with an aim to *identify and enhance the capabilities of the generative components of pretrained language models*. I'm proficient in Python, PyTorch, LaTeX, pandas, and more. Here are some highlights of my experience:

- **Broad Scope of Research:** My work spans a wide range of projects in areas such as text classification and generation, model structure modifications, ranking, information retrieval, etc. I remain open to exploring new directions and actively collaborate with students from diverse backgrounds, like linguistics and computer vision.
- **Sustained Collaborations:** I pride myself on my ability to foster and maintain long-term professional relationships. My ongoing collaborations with Debanjan and Xiaomeng, both initiated in 2021, attest to my adaptability and easygoing nature in team environments.
- **Independence in Research:** Several of my projects, which were published in AACL and TACL, were conducted autonomously without mentorship, showcasing my passion and capability in research.
- **Experience with Large Language Models:** During my internships at Google (2023), TikTok (2022), and Educational Testing Service (2021), I've worked on selecting better in-context learning demonstrations with Flan-PaLM 2 (M and L) and question generation after fine-tuning models like T5, mT5, ByT5, and BART.
- **Real-world Impact:** My project on distractor selection was integrated into the codebase of Cooori Japan, highlighting its practical utility. Additionally, some of my other projects are under patent application.

Areas of Expertise

- Natural Language Processing
- R&D (Research and Development)
- Project Management
- Deep Learning & Machine Learning
- Programming
- Collaboration & Teamwork
- Data Analysis & Visualization
- Problem Solving
- Adaptability

Selected Technical Skills

Python, PyTorch, TensorFlow, LaTeX, NumPy, Pandas

Education

Toyota Technological Institute at Chicago (TTIC), Chicago, IL, USA
Ph.D. candidate, Computer Science, CGPA: 3.86/4.0, Advisor: *Prof. Kevin Gimpel* — *Sep '17-Present*
M.S. within Ph.D., Computer Science, Advisor: *Prof. Kevin Gimpel* — *Sep '17-Sep '19*

Tsinghua University (THU), Beijing, China
M.E., Electrical Engineering, CGPA: 3.63/4.0, Advisor: *Prof. Xiaohua Jiang* — *Sep '14-Jun '17*
B.E., Electrical Engineering and Automation, CGPA: 3.79/4.0 — *Aug '10-Jul '14*

Internship

Research Intern, Google LLC., Mountain View, CA, USA — *May-Aug, '23*
Target: Selecting Better In-Context Learning Demonstrations for Text Classification
Key Skills: TensorFlow, Pandas, Python, NumPy, LaTeX
Models: Flan-PaLM 2 (M & L), off-the-shelf retriever (fine-tuned on mT5-base)
- Completed over 100 pages of documentation and 4,000+ lines of code. Prepared a **paper** for submission to ARR.
- Achieved a +2.6% gain on F1 macro scores over an already high baseline that matches or exceeds current benchmarks.
- Proposed constraints for demonstration selection are potentially adaptable to other applications, including ranking.

Research Intern (Remote), TikTok Inc., Chicago, IL, USA — *May-Aug, '22*
Target: Generating Questions of Different Styles Controlled with Keywords
Key Skills: PyTorch, PyTorch Lightning, Python, NumPy
Models: T5, mT5, ByT5 (all base versions)
- Authored over 3,600 lines of code.
- Demonstrated that an enhanced T5 model with additional tokens, such as emojis, excels in generating keywords together with topics over other models, surpassing spaCy on keyword extraction by an F1 score of 0.21.
- Generated questions controlled with keywords, topics, and specified length. Determined that using distinct models yields better results for generating questions with different styles.

Intern (Remote), Educational Testing Service, Princeton, NJ, USA *Jun-Aug, '21*
Target: Generating and Ranking Inquisitive Questions Controlled with Question Types
Key Skills: PyTorch, Fairseq, Pandas, Python, NumPy, LaTeX
Models: RoBERTa, BART (all large versions)
- Code is publicly available on **GitHub** (5000+ lines). Our **paper** was accepted for presentation at *SEM 2022.
- Produced diverse questions tailored to specific question types.
- Leveraged a pairwise ranker to select generated questions that matched the quality of human-crafted queries in terms of syntax, semantics, relevancy, and inquisitiveness, as validated by human assessment.

Other Industry Collaborations

Label-Description Training for Zero-Shot Text Classification *'22*
Target: Improving Zero-Shot Text Classification by Finetuning on Curated Small Datasets
Key Skills: PyTorch, Python, NumPy, LaTeX
Models: RoBERTa-large
- Code is publicly available on **GitHub**. Our **paper** was accepted by EMNLP 2023 (main conference).
- Boosted zero-shot accuracy by 17-19% across a range of sentiment and topic classification datasets.
- Enhanced model robustness, sometimes even outperforming a supervised approach with out-of-domain data.

Distractor Analysis and Selection for Multiple-Choice Cloze Questions *'19*
Target: Automatically Suggesting Distractors for Multiple-Choice Cloze Questions
Key Skills: PyTorch, Python, NumPy, LaTeX
Models: ELMo, BERT (base & large, cased versions)
- Code was incorporated into Cooori Japan codebase. Our **paper** was accepted for presentation at BEA 2020.
- Identified that models prioritize distractors morphologically similar to correct answers, yet semantically distinct.
- Optimized our models to achieve, and at times exceed, human performance.

ArXiv Preprints

- **Ambiguity-Aware In-Context Learning with Large Language Models**
  Lingyu Gao, Aditi Chaudhary, Kazuma Hashimoto, Krishna Srinivasan, Karthik Raman, Michael Bendersky
  Submitted to ARR

Publications

†: Co-senior authors

- **ToMChallenges: A Principle-Guided Dataset and Diverse Evaluation Tasks for Exploring Theory of Mind**
  Xiaomeng Ma, **Lingyu Gao**, Qihui Xu
  The SIGNLL Conference on Computational Natural Language Learning (CoNLL). 2023
- **The Benefits of Label-Description Training for Zero-Shot Text Classification**
  Lingyu Gao, Debanjan Ghosh†, Kevin Gimpel†
  The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2023
- **Evaluating Transformer Models and Human Behaviors on Chinese Character Naming**
  Xiaomeng Ma, **Lingyu Gao**
  Transactions of the Association for Computational Linguistics (TACL). 2023
- **How do we get there? Evaluating transformer neural networks on English past tense inflection**
  Xiaomeng Ma, **Lingyu Gao**
  The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP). 2022
- **"What makes a question inquisitive?"A Study on Type-Controlled Inquisitive Question Generation**
  Lingyu Gao, Debanjan Ghosh, Kevin Gimpel
  The 11th Joint Conference on Lexical and Computational Semantics (*SEM). 2022
- **Distractor Analysis and Selection for Multiple-Choice Cloze Questions for Second-Language Learners**
  Lingyu Gao, Kevin Gimpel, Arnar Jensson
  Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications (BEA). 2020
- **A Cross-Task Analysis of Text Span Representations**
  Shubham Toshniwal, Haoyue Shi, Bowen Shi, **Lingyu Gao**, Kevin Gimpel and Karen Livescu
  Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP). 2020
- **Design and Heat Leak Analysis of a HTS DC Cable System**
  Lingyu Gao, Guolin Chai, and Xiaohua Jiang
  Cryogenics and Superconductivity 45.9 (2017): 41-45. (in Chinese)
- **Closed-Loop Distribution Network by a Midvoltage Flexible HTS DC System**
  Xianglong Zhang, Lin Cheng, **Lingyu Gao**, Yiqun Zhang, Zhongxi Li, Yingyu Zeng, Zhenqian Zhang, and Xiaohua Jiang
  IEEE Transactions on Applied Superconductivity 26.4 (2016): 1-4.

Ongoing Projects

- **Reducing Object Hallucinations in Large Vision-Language Models with Decoding Algorithms**
  Collaborative Research *Sep'23-Present*
- **Enhancing Retriever-Based In-Context Learning with Large Language Models**
  Mentor: Aditi Chaudhary, Krishna Srinivasan, Kazuma Hashimoto, et al. (Google Research) *May'23-Present*
- **Exploring the Benefits of Incorporating Lookahead Heuristics into Decoding Strategies**
  Mentor: Debanjan Ghosh (ETS), Kevin Gimpel (TTIC) *Mar'23-Present*

Awards & Honors

- ETS Pre-Doctoral Fellowship *'21*
- Mitsubishi Heavy Industries Scholarship *'14*
- NARI-RELAYS Scholarship *'13*
- 1^{st} prize in Schneider Electric Programmable Logic Controller Competition *'13*
- 1^{st} grade Academic Excellence Scholarship *'11*
- 2^{nd} grade Freshman Scholarship *'10*

Teaching

Teaching Assistant | *Introduction to Machine Learning* | Instructor: *Prof. Kevin Gimpel* *Autumn '19*

Services

- **Reviewer** for NAACL-HLT 2021, BEA (2022, 2023), EMNLP (2022, 2023), ACL 2023, TALLIP 2023
- **Secondary Reviewer** for EMNLP 2019 and RepL4NLP 2020
- **Volunteer** in a non-profit organization Circle Cat *May'23-Present*
- **Student Member** of TTIC DEI committee in 2020 - 2021, PhD admissions committee in 2020
- **Teaching Volunteer** in Mabian Yi Autonomous County, Sichuan, China *Jul'11*
- **Member** of Student Association for Science and Technology, EE Department *Feb-Jun, '11*