Xiang Yue

Homepage: xiangyue9607.github.io Email: yue.149@osu.edu

Short Bio: Xiang is now a final-year CS Ph.D. student at The Ohio State University. His research interests lie in Natural Language Processing (NLP) with an emphasis on *Question Answering* and *Privacy-preserving NLP*. He has published more than **20** peer-reviewed research papers, among which **7 are first-author on the top conferences or journals**, and has ∼**800** citations according to Google Scholar. He won the **Best Paper Award** at the IEEE BIBM 2021 conference and **Third Place (**$50**K)** in the First Alexa Prize TaskBot Challenge. He interned at *Microsoft Research* (Redmond) and *Tencent AI Lab* (Seattle), where he gained hands-on experience in building machine learning and NLP models for large-scale real-world data.

Education

Ph.D Student	Computer Science & Engineering, The Ohio State University (OSU), USA GPA: 3.92/4.0 Advisor: Prof. Huan Sun Research Topics: Question Answering, Dialog Systems, Text Generation, Privacy-preserving NLP	*2018-2023*
B.Eng.	Computer Science, Wuhan University (WHU), China GPA: 3.79/4.0, Rank: top 1%, Outstanding Graduates Advisor: Prof. Wen Zhang Research Topics: Data Mining, Graph Neural Network and Graph Embedding, Bioinformatics	*2014-2018*

Professional Experience

The Ohio State University Columbus, OH

- *Graduate Research Associate, OSU NLP Group* *2018.8-Present*
  Advised by: Prof. Huan Sun
  - I mostly worked on building **Question Answering (QA)** and **Task-oriented Dialog Systems** for real use.
    - We first identified that a well-trained QA model achieved much lower performance on an out-of-domain testing set compared with its in-domain testing set (i.e., generalizability issue) (**ACL 2020**, [9]).
    - We then proposed to synthesize diverse QA pairs using the question generation (QG) technique to boost the source QA model's performance on the target domain (**IEEE BIBM 2021 Best Paper**, [6]).
    - To improve the quality of the synthetic QA pairs, we further proposed a Question Value Estimator to select the most useful synthetic QA pairs. (**ACL 2022**, [1])
    - We also built a task-oriented dialog system *Tacobot* that can assist users in accomplishing cooking and DIY tasks. *Tacobot* won the **3rd-place (out of 125 teams)** in the first *Alexa Prize TaskBot Challenge*.
  - I also actively collaborated with Security and Privacy researchers to explore **Privacy-preserving AI/NLP**.
    - We developed formal privacy-preserving mechanisms with Differential Privacy (DP) to sanitize private textual data and explore how to pre-train and fine-tune language models on it. (**ACL 2021** Findings, [5])
    - We proposed DP-Forward, a differentially-private training algorithm that adds noise in the forward pass when fine-tuning large language models (e.g., BERT) (Under Review)
    - We also developed empirical protection methods, e.g., removing Personal Identifiable Information (EMNLP 2020 Workshop, [7]), and converting text corpora to term-term co-occurrence graphs (**KDD 2019**, [12]).
  - Besides, I also did ML applications on large-scale biomedical data, e.g., building Graph Neural Networks (GNNs) for representation learning on real-world medical graphs (**Bioinformatics ESI Highly Cited Paper**, [11])

Microsoft Research (MSR) Redmond, Washington

- *NLP Research Intern, Language Learning and Privacy Group* *2022.5-2022.8*
  Advised By: Dr. Huseyin A. Inan, Dr. Robert Sim, Dr. Julia McAnallen, Dr. David Levitan
  - **Project**: Synthetic Text Generation with Privacy Guarantee (Manuscript Preparation)
  - **Abstract**: Customer data provides clear benefits and valuable insights, but it comes with many compliance requirements (e.g., GDPR) due to privacy concerns. In this project, we aim to generate a synthetic version of the original dataset, which is expected to eliminate privacy concerns while preserving the statistical and semantic properties of the original dataset. To this end, we fine-tune a generative language model, GPT-2, with differential privacy and then synthesize text based on the differentially-private model. Our utility metrics demonstrate the high quality of the synthetic dataset and our privacy metrics indicate the good preservation of privacy.

Tencent AI Lab Bellevue, Washington

- *NLP Research Intern, Tencent NLP Group* *2021.5-2021.8*
  Advised By: Dr. Xiaoman Pan, Dr. Jianshu Chen, Dr. Wenlin Yao, Dr. Dian Yu, Dr. Dong Yu
  - **Project**: Pretraining to Answer Open-Domain Questions by Consulting Millions of References (**ACL 2022**, [2])
  - **Abstract**: We consider the problem of pre-training a two-stage open-domain question answering (QA) system (retriever + reader) with strong transfer capabilities. We propose to automatically construct high-quality pre-training question-answer-context triplets by consulting millions of references cited within Wikipedia. The well-aligned pre-training signals benefit both the retriever and the reader significantly. Our pretrained retriever leads to 2%-10% absolute gains in top-20 accuracy. And with our pretrained reader, the entire system improves by up to 4% in exact match.

Selected Publications

The Full List of my 20+ published papers are available at: **(Google Scholar) (Semantic Scholar)**
* indicates equal contributions

[1] **Xiang Yue**, Ziyu Yao and Huan Sun, "Synthetic Question Value Estimation for Domain Adaptation of Question Answering", **ACL 2022**

[2] **Xiang Yue**, Xiaoman Pan, Wenlin Yao, Dian Yu, Dong Yu, Jianshu Chen, "C-MORE: Pretraining to Answer Open-Domain Questions by Consulting Millions of References", **ACL 2022**

[3] Shijie Chen, Ziru Chen, Xiang Deng, Ashley Lewis, Lingbo Mo, Samuel Stevens, Zhen Wang, **Xiang Yue**, Tianshu Zhang, Yu Su, Huan Sun, "Bootstrapping a User-Centered Task-Oriented Dialogue System", **1st Proceedings of Alexa Prize TaskBot**

[4] Frederick Zhang, Heming Sun, **Xiang Yue**, Simon Lin and Huan Sun, "COUGH: A Challenge Dataset and Models for COVID-19 FAQ Retrieval", **EMNLP 2021**

[5] **Xiang Yue***, Minxin Du*, Tianhao Wang, Yaliang Li, Huan Sun and Sherman S. M. Chow, "Differential Privacy for Text Analytics via Natural Text Sanitization", **ACL-IJCNLP 2021, Findings**

[6] **Xiang Yue***, Xinliang (Frederick) Zhang*, Ziyu Yao, Simon Lin and Huan Sun, "CliniQG4QA: Generating Diverse Questions to Improve Clinical Reading Comprehension on New Contexts", **IEEE BIBM 2021, Best Paper Award**

[7] **Xiang Yue** and Shuang Zhou, "PHICON: Improving Generalization of Clinical Text De-identification Models via Data Augmentation", **EMNLP 2020 Clinical NLP Workshop**

[8] Kaushik Mani*, **Xiang Yue***, Bernal Jimenez Gutierrez, Yungui Huang, Simon Lin and Huan Sun, "Clinical Phrase Mining with Language Models", **IEEE BIBM 2020**

[9] **Xiang Yue**, Bernal Jimenez Gutierrez and Huan Sun, "Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset", **ACL 2020**

[10] Feng Huang*, **Xiang Yue***, Zhankun Xiong, Zhouxin Yu, Shichao Liu, Wen Zhang, "Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations", **Briefings in Bioinformatics**, 2020

[11] **Xiang Yue**, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon Lin, Wen Zhang, Ping Zhang and Huan Sun, "Graph Embedding on Biomedical Networks: Methods, Applications, and Evaluations", **Bioformatics**, 2020, **(ESI Highly Cited Paper: top 1% cited paper of its academic field)**

[12] Zhen Wang, **Xiang Yue**, Soheil Moosavinasab, Yungui Huang, Simon Lin and Huan Sun, "SurfCon: Synonym Discovery on Privacy-Aware Clinical Data", **KDD 2019**

Academic Services

Program Committee/Reviewer:

- Conferences: AAAI'23, EMNLP'21,22, ACL'22, NLPCC'21, NAACL'20,21,22, KDD'19,20,21, AMIA'20, BIBM'18,20
- Journals: IEEE Transactions on Knowledge and Data Engineering (TKDE), IEEE Transactions on Neural Networks and Learning Systems, Bioinformatics, Nature Scientific Reports, ACM Transactions on Computing for Healthcare, BMC Medical Informatics and Decision Making

Honors and Awards

- **Third Place ($50K) in the First Alexa Prize TaskBot Challenge** (10 participant teams selected worldwide out of 125 initiated applications; 5 teams selected into finals; The only US team in top-3) *June 2022*
- **Best Paper Award, IEEE BIBM 2021** *Nov 2021*
- Best Student Research Poster 2021, OSU CSE *April 2021*
- KDD 2020, 2019 Student Travel Award *Aug 2019,2020*
- Excellent Graduation Thesis Award of WHU (Scale: 5%) *June 2018*
- Outstanding Graduates of WHU (Scale: 10%) *May 2018*
- **LEI JUN Scholarship** (Top 1 Winner of National Scholarship, the highest prize for students in WHU) *2016-2017*
- First Class Scholarship (Scale: 5%), Three Times, WHU *2014-2017*
- Excellent Student (Scale: 5%), Three Times, WHU *2014-2017*
- **National Scholarship (Scale: 1%), China** *2014-2015*

Skills Summary

- **Languages**: Python, MATLAB, JAVA, C/C++, HTML/CSS/JS
- **Frameworks**: Pytorch, Huggingface Transformers, Scikit-learn, Numpy, Tensorflow, NLTK, SpaCy
- **Tools**: Git, MySQL, MongoDB, SQLite
- **Platforms**: Linux, Web, Windows, AWS