Xiang Deng

Ph.D. candidate, The Ohio State University • deng.595@osu.edu • +1 (614) 254-0270 • xiang-deng.github.io

My research interests lie in NLP and data mining, with emphasis on knowledge discovery and utilization from heterogeneous sources. The aim is to ***build AI-powered data systems that can assist with information acquisition and decision making for regular users as well as domain experts*** in Digital Era. Specifically, my recent research focuses on: (1) **Large-scale pretraining and representation learning** for data from heterogeneous sources, such as plain text, web documents, knowledge graphs, and databases; both for general and domain-specific applications. (2) **Building natural language interfaces** (e.g., question answering, semantic parsing, and dialog systems) with varied data and services as backend.

EDUCATION

Ph.D. Candidate in Computer Science *2018 - 2023 (expected)*

The Ohio State University, *Columbus, OH, USA*

- Advisor: Prof. *Huan Sun*
- Major: Artificial Intelligence; Minor: Database, Graphics

B.Eng. in Computer Science *2014 - 2018*

University of Science and Technology of China, *Hefei, China*

- Advisor: Prof. *Qi Liu*
- The Talent Program in Computer and Information Science, School of The Gifted Young

PROFESSIONAL EXPERIENCE

The Ohio State University *Aug 2018 - present*

Graduate Research Associate *Columbus, OH*

Supervisor: Prof. *Huan Sun*

- Prompting and reasoning with Large Language Models for solving complex tasks. (**EMNLP**'22.)
- Building dialog system that can assist users in accomplishing tasks. Focusing on bootstrapping the system with few in-domain training data, and accommodating noisy real user input. (OSU Tacobot team, ranked **3rd place in the inaugural Alexa Prize TaskBot Challenge**.)
- Textual and tabular data understanding via pre-training and representation learning. (**VLDB**'21, **EMNLP**'21, **SIGMOD Research Highlight**'22. Collaboration with Google Research under **Google Faculty Research Award**.)
- Relation Extraction with extra signals from Web Tables. (**EMNLP**'19.)
- Question answering and tabular query resolution with Knowledge Base.

Google Research *May 2022 - Oct 2022*

Research Intern *New York City, NY*

Supervisor: *Vasilisa Bashlovkina**, *Feng Han*, *Simon Baumgartner*

- Financial sentiment analysis on social media content. Obtaining supervised data for tasks that require domain knowledge is often challenging. We propose to leverage the in-context learning ability of large language models, and inject domain knowledge via weak supervision. The resulting model obtains competitive performance on public datasets and a significant improvement on the internal benchmark.

Amazon *May 2021 - Aug 2021*

Applied Scientist Intern *Remote*

Supervisor: *Prashant Shiralkar**, *Colin Lockard*, *Binxuan Huang*

- Learning robust and generalizable representation for semi-structured web pages. The resulting model brings significant improvement under zero-shot and few-shot settings, which greatly reduces human annotation efforts.

Microsoft Research *May 2020 - Aug 2020*

Research Intern *Remote*

Supervisor: *Matthew Richardson**, *Ahmed Awadallah*, *Christopher Meek*, *Oleksandr Polozov*

- Natural language to SQL, with a focus on generalization ability in real-world applications. By weakly supervised pre-training using existing text-table parallel data on the web, we enhance the model's performance on value prediction and column selection, especially when access to actual database content is limited at runtime. (**NAACL**'21)

Microsoft Research Asia *Dec 2017 - May 2018*
Research Intern *Beijing, China*
Supervisor: *Lei Cui*

- News recommendation and summarization leveraging title, content, and trending topics information.

PUBLICATIONS

[1] 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", *Alexa Prize Proceedings*, 2022, **3rd place in the Alexa Prize TaskBot Challenge**

[2] Boshi Wang, **Xiang Deng**, Huan Sun, "Shepherd Pre-trained Language Models to Develop a Train of Thought: An Iterative Prompting Approach", *Conference on Empirical Methods in Natural Language Processing*, (**EMNLP**), 2022

[3] **Xiang Deng**, Prashant Shiralkar, Colin Lockard, Binxuan Huang, Huan Sun, "DOM-LM: Learning Generalizable Representations for HTML Documents", *arXiv preprint*, 2022

[4] **Xiang Deng**, Yu Su, Alyssa Lees, You Wu, Cong Yu, and Huan Sun, "ReasonBERT: Pre-trained to Reason with Distant Supervision", *Conference on Empirical Methods in Natural Language Processing*, (**EMNLP**), 2021

[5] **Xiang Deng**, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson, "Structure-Grounded Pretraining for Text-to-SQL", *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (**NAACL**), 2021

[6] **Xiang Deng**, Huan Sun, Alyssa Lees, You Wu, and Cong Yu, "TURL: Table Understanding through Representation Learning", *International Conference on Very Large Data Bases*, (**VLDB**), 2021, **SIGMOD Research Highlight**, 2022

[7] **Xiang Deng**, Huan Sun, "Leveraging 2-hop Distant Supervision from Table Entity Pairs for Relation Extraction", *Conference on Empirical Methods in Natural Language Processing*, (**EMNLP**), 2019

[8] Jie Zhao, **Xiang Deng**, Huan Sun, "Easy-to-Hard: Leveraging Simple Questions for Complex Question Generation", *arXiv preprint*, 2019

[9] Bortik Bandyopadhyay, **Xiang Deng**, Goonmeet Bajaj, Huan Sun, and Srinivasan Parthasarathy, "Automatic Table completion using Knowledge Base", *arXiv preprint*, 2019

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), *Amazon* *2022*
- **SIGMOD Research Highlight**, *SIGMOD* *2022*
- Student Travel Award, *KDD 2019* *2019*
- Student Scholarship, *USTC* *2015 - 2017*
- Freshman Scholarship, *USTC* *2014*

PROFESSIONAL SERVICE

Program Committee/Reviewer: ACL ARR, SUKI 2022, NLP4Prog 2021; NLPCC 2020, 2021, 2022; AAAI 2022, 2023
Secondary/External Reviewer: KDD 2020; NAACL 2019; KDD 2019

TEACHING EXPERIENCE

Syllabus of Digital Logic Lab, Teaching Assistant, *USTC* Fall, 2016

SKILLS

Python, PyTorch, Tensorflow, Spark, Ray, C++, Java, SQL, Cloud and Distributed Environments