# Tao Yu

| | | |
|---|---|---|
| <small>CONTACT INFORMATION</small> | https://taoyds.github.io/ <br> Rm 204E, Chow Yei Ching Building, Pokfulam, Hong Kong | tyu@cs.hku.hk <br> +852-9768-6811 |

**RESEARCH INTERESTS**

Natural Language Processing. The development and design of the next generation of natural language interfaces employing large language models to facilitate human interaction with data analysis, web applications, and robotic instruction through conversation. It involves Executable Language Grounding, Semantic Parsing/Code Generation, Efficient and Generalizable Language Models, and Dialog/Interactive Systems.

**ACADEMIC POSITION**

**University of Hong Kong**, Hong Kong — 08/2021 - *Present*
Assistant Professor in Computer Science, Co-director of the HKU NLP Group

**University of Washington**, Seattle, WA — 09/2021 - 08/2022
Postdoctoral Fellow, the UW NLP Group
Working with Noah Smith, Luke Zettlemoyer, and Mari Ostendorf

**EDUCATION**

**Yale University**, New Haven, CT — 08/2017 - 05/2021
Ph.D. Computer Science
Advisors: Dragomir Radev

**Columbia University**, New York, NY — 09/2015 - 05/2017
M.S. Data Science   GPA: 3.82/4.0
Advisors: Owen Rambow and Kathleen McKeown

**University of Utah**, Salt Lake City, UT — 01/2012 - 05/2015
B.S. Mathematics and Economics   GPA: 3.86/4.0

**AWARDS**

| | |
|---|---|
| **Google Research Scholar Award**, aka **Google Faculty Award** | 2023 |
| **Amazon Research Award**, aka **Amazon Faculty Award** | 2022 |

**TALKS & PRESENTATIONS**

| | |
|---|---|
| **Apple KP Tech Talks** | 06/2023 |
| **Cornell DB Seminar** | 05/2023 |
| **Microsoft Research Asia** | 05/2023 |
| **Columbia NLP Seminar** | 04/2023 |
| **Amazon AWS** | 11/2022 |
| **AllState Tech Talks** | 06/2022 |
| **ElementAI/ServiceNow Research** | 02/2022 |
| **Google Research** | 04/2021 |
| **CS, The University of Hong Kong** | 03/2021 |
| **CS, The National University of Singapore** | 03/2021 |
| **CS, The University of Wisconsin-Madison** | 03/2021 |
| **CS, Simon Fraser University** | 03/ 2021 |
| **CSE, The University of Minnesota Twin Cities** | 03/2021 |

PUBLICATIONS &
PAPERS IN
PREPARATION

1. Hongjin Su\*, Weijia Shi\*, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer,**Tao Yu**
One Embedder, Any Task: Instruction-Finetuned Text Embeddings
**ACL Findings 2023**

2. Yuhang Lai\*, Chengxi Li\*, Yiming Wang\*, Tianyi Zhang\*, Ruiqi Zhong\*, Luke Zettlemoyer, Scott Wen-tau Yih, Daniel Fried, Sida Wang, **Tao Yu**
DS-1000: A Natural and Reliable Benchmark for Data Science Code Generation
**ICML 2023**

3. Tianyi Zhang, **Tao Yu**, Tatsunori B Hashimoto, Mike Lewis, Wen-tau Yih, Daniel Fried, Sida I Wang
Coder Reviewer Reranking for Code Generation
**ICML 2023**

4. Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, **Tao Yu**, Lingpeng Kong
Compositional Exemplars for In-context Learning
**ICML 2023**

5. Zhoujun Cheng\*, Tianbao Xie\*, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, **Tao Yu**
Binding Language Models in Symbolic Languages
**ICLR 2023**

6. Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, **Tao Yu**
Selective Annotation Makes Language Models Better Few-Shot Learners
**ICLR 2023**

7. Tianbao Xie\*, Chen Henry Wu\*, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, **Tao Yu**
UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models
**EMNLP 2022**

8. Jiacheng Ye\*, Jiahui Gao\*, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, **Tao Yu**, Lingpeng Kong
ZeroGen: Efficient Zero-shot Learning via Dataset Generation
**EMNLP 2022**

9. Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, **Tao Yu**, Noah A. Smith, Mari Ostendorf
In-Context Learning for Few-Shot Dialogue State Tracking
**EMNLP Findings 2022**

10. Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, **Tao Yu**, Lingpeng Kong
ProGen: Progressive Zero-shot Dataset Generation via In-context Feedback
**EMNLP Findings 2022**

11. Qi Liu, Zihuiwen Ye, **Tao Yu**, Phil Blunsom, Linfeng Song
Augmenting Multi-Turn Text-to-SQL Datasets with Self-Play
**EMNLP Findings 2022**

12. Yiru Chen, Ryan Li, Austin Mac, Tianbao Xie, **Tao Yu**, Eugene Wu
NL2INTERFACE: Interactive Visualization Interface Generation from Natural Language Queries
**IEEE Visualization Conference NLVIZ Workshop**, 2022

13. Ziming Mao*, Chen Henry Wu*, Ansong Ni, Yusen Zhang, Rui Zhang, **Tao Yu**, Budhaditya Deb, Chenguang Zhu, Ahmed H Awadallah, Dragomir Radev
DYLE: Dynamic Latent Extraction for Abstractive Long-Input Summarization
**ACL 2022**

14. Yusen Zhang, Ansong Ni, **Tao Yu**, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, Dragomir Radev
An Exploratory Study on Long Dialogue Summarization: What Works and What's Next
**EMNLP Findings 2021**

15. Ansong Ni, Zhangir Azerbayev, Mutethia Mutuma, Troy Feng, Yusen Zhang, **Tao Yu**, Ahmed Hassan Awadallah, Dragomir Radev
SummerTime: Text Summarization Toolkit for Non-experts
**EMNLP 2021**

16. Heather Lent, Semih Yavuz, **Tao Yu**, Tong Niu, Yingbo Zhou, Dragomir Radev, Xi Victoria Lin
Testing Cross-Database Semantic Parsers Using Canonical Utterances
*Evaluation & Comparison of NLP Systems Workshop at EMNLP 2021*, **Best Paper Award**

17. Chang Shu, Yusen Zhang, Xiangyu Dong, Peng Shi, **Tao Yu**, Rui Zhang
Logic-Consistency Text Generation from Semantic Parses
**ACL Findings 2021**

18. **Tao Yu**, Rui Zhang, Oleksandr Polozov, Chris Meek, Ahmed Hassan Awadallah
SCoRe: Pre-Training for Context Representation in Conversational Semantic Parsing
**ICLR 2021**

19. **Tao Yu**, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, Caiming Xiong
GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing
**ICLR 2021**

20. Ming Zhong, Da Yin, **Tao Yu**, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, Dragomir Radev
QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization
**NAACL 2021**

21. Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Nazneen Fatema Rajani, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Murori Mutuma, Yasin Tarabar, Ankit Gupta, **Tao Yu**, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, Nazneen Fatema Rajani
DART: Open-Domain Structured Data Record to Text Generation
**NAACL 2021**

22. Ruiqi Zhong, **Tao Yu**, Dan Klein
Semantic Evaluation for Text-to-SQL with Distilled Test Suites
**EMNLP 2020**

23. Yusen Zhang, Xiangyu Dong, Shuaichen Chang, **Tao Yu**, Peng Shi, Rui Zhang
Did You Ask a Good Question? A Cross-Domain Question Intention Classification Benchmark for Text-to-SQL
**Interactive and Executable Semantic Parsing Workshop at EMNLP 2020**

24. **Tao Yu**, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter Lasecki, Dragomir Radev

CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases
**EMNLP 2019**

25. Rui Zhang, **Tao Yu**, He Yang Er, Sungrok Shim, Eric Xue, Xi Victoria Lin, Tianze Shi, Caiming Xiong, Richard Socher, Dragomir Radev
Editing-based SQL Query Generation for Cross-Domain Context-Dependent Questions
**EMNLP 2019**

26. **Tao Yu**, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, Dragomir Radev
SParC: Cross-Domain Semantic Parsing in Context
**ACL 2019**

27. Richard A Plunz, Yijia Zhou, Maria Isabel Carrasco Vintimilla, Kathleen Mckeown, **Tao Yu**, Laura Uguccioni, Maria Paola Sutto
Twitter Sentiment in New York City Parks as Measure of Well-being
**Landscape and Urban Planning 2019**

28. **Tao Yu**, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, Dragomir Radev
Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task
**EMNLP 2018**

29. **Tao Yu**, Michihiro Yasunaga, Kai Yang, Rui Zhang, Dongxu Wang, Zifan Li, Dragomir Radev
SyntaxSQLNet: Syntax Tree Networks for Complex and Cross Domain Text-to-SQL Task
**EMNLP 2018**

30. **Tao Yu**, Zifan Li, Zilin Zhang, Rui Zhang, Dragomir Radev
TypeSQL: Knowledge-based Type-Aware Neural Text-to-SQL Generation
**NAACL 2018**

31. Mohammad Sadegh Rasooli, Noura Farra, Axinia Radeva, **Tao Yu**, and Kathleen McKeown
Cross-Lingual Sentiment Transfer with Limited Resources
**Machine Translation 2017**

32. Owen Rambow, **Tao Yu**, Axinia Radeva, Sardar Hamidian, Alexander R. Fabbri, Debanjan Ghosh, Christopher Hidey, Tianrui Peng, Mona Diab, Kathleen McKeown, Smaranda Muresan
The Columbia-GWU System at the 2016 TAC KBP BeSt Evaluation
**NIST TAC KBP Workshop 2016**

PROFESSIONAL SERVICES

**Organizing Committee Member**

**ACL 2023**
SUKI: Workshop on Structured and Unstructured Knowledge Integration at **NAACL 2022**
The Workshop on Interactive and Executable Semantic Parsing at **EMNLP 2020**

**Program Committee Member**

**ACL Rolling Review**, **ACL**: (2020, 2021, 2022, 2023), **EMNLP**: (2019, 2020, 2021, 2022), **ICLR**: (2022), **ICML**: (2023), **NAACL**: (2019, 2021), **NeurlPS**: (2022), **COLING**: (2020), **AACL-IJCNLP**: (2020), and other various conferences and workshops.

| | | |
|---|---|---|
| STUDENTS | Tianbao Xie (HKU PhD, 2022) | |
| | Hongjin Su (HKU PhD, 2022) | |
| | Yiheng Xu (HKU PhD, 2022, co-advised with Lingpeng Kong) | |
| | Jiacheng Ye (HKU PhD, 2022, co-advised with Lingpeng Kong) | |
| | Fan Zhou (Intern 2023, SJTU BS/MS) | |
| | Leo Liu (Intern 2023, UW BS/MS → UT Austin PhD) | |
| | Zhoujun Cheng (Intern 2022, SJTU BS/MS) | |
| | Chen Henry Wu (Intern 2022, Tsinghua BS → CMU PhD) | |
| | Ryan Li (Intern 2022, UW BS → Stanford MS) | |
| | Chengzu Li (Intern 2022, Xi'an Jiaotong BS → Cambridge PhD) | |
| | Shuyang Jiang (Intern 2023, SJTU BS → Fudan PhD) | |
| | Yiming Wang (Intern 2022, PKU BS → Harvard MS) | |
| | Yuhang Lai (Intern 2022, BIT BS → Fudan MS) | |
| | Chengxi Li (Intern 2022, HIT BS) | |
| | Ming Zhong (Intern 2020, Fudan MS → UIUC PhD) | |
| | Da Yin (Intern 2020, PKU BS → UCLA PhD) | |
| | Yusen Zhang (Intern 2020, Emory MS → PSU PhD) | |
| | Naihao Deng (Intern 2020, UMich BS → UMich PhD) | |
| | Michihiro Yasunaga (Project Student, 2018-19, Yale BS → Stanford PhD) | |
| | Suyi Li (Project Student, 2019, Yale MS → HKUST PhD) | |
| | Kai Yang (Project Student, 2018, Yale MS → Google SE) | |
| | James Ma (Project Student, 2018, Yale BS → LinkedIn DS) | |

| | | |
|---|---|---|
| GRANTS | **Google Research Scholar Award** (USD$35,000) | 2023 |
| | **Amazon Research Award** (USD$40,000 + $80,000 AWS Credits) | 2022 |

| | | |
|---|---|---|
| RESEARCH & INDUSTRY EXPERIENCE | **Microsoft Research**, Redmond, WA | 06/2020 - 08/2020 |
| | NLP Research Intern<br>Mentors: Ahmed Hassan Awadallah, Oleksandr Polozov, and Chris Meek | |
| | **Salesforce Research**, Palo Alto, CA | 06/2019 - 08/2019 |
| | NLP Research Intern<br>Mentors: Victoria Lin, Caiming Xiong, and Richard Socher | |
| | **Samsung Research America**, Mountain View, CA | 06/2018 - 08/2018 |
| | NLP Research Intern | |
| | **Columbia CCLS & NLP Group**, New York, NY | 05/2016 - 10/2016 |
| | Graduate Research Assistant<br>Mentors: Owen Rambow and Kathleen McKeown | |

| | | |
|---|---|---|
| TEACHING EXPERIENCE | Introduction to Data Science with Python, University of Hong Kong | 2022 |
| | Natural Language Processing (CPSC 577), Yale University | 2019 |
| | Database Systems (CPSC 537), Yale University | 2019 |
| | Introduction to Natural Language Processing (CS 4705), Columbia University | 2017 |

| | | |
|---|---|---|
| | Natural Language Processing (Coursera) | 2016 |
| Honors | HKU-100 Scholar | 2021 |
| | Yale Nathan Hale Associates Scholarship Award | 2020 |
| | Yale Conference Travel Fellowship | 2018 |
| | Yale PhD Fellowship | 2017 |
| | Connie and Robert White Honor Roll Scholarship | 2014 |
| | Mathematics Department Scholarship | 2014 |
| | International Student and Scholar Services Scholarship | 2014 |
| | NSF Research Experience for Undergraduates (REU) Fellowship | 2014 |