1 Overall

- worked on a wide range of natural language processing problems in understanding, generation, and grounding
- published papers and served as Program Committee members on top tier AI and NLP conferences including ACL, NAACL, EMNLP, AAAI, CoNLL.
- have both constructed impactful datasets to advance the research progress and proposed effective models to solve the tasks.
- published 16 papers by successfully collaborating with peer Ph.D. and undergraduate students in the lab, professors from different universities (UMichigan, Columbia, UMaryland), and researchers from industry labs (IBM Thomas J. Watson Research Center, Grammarly Research, Salesforce Research).
- In addition to research, also dedicated to teaching and mentoring through my academic career.

2 Research

2.1 Dialog System and Conversational AI(Sapphire)

- participated in the Sapphire project from 2015 - 2017 as a collaboration between UMich and IBM which aims to improve student advising by building conversational systems; published two papers.
- first paper: Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs. In AAAI 2018. study the problem of addressee and response selection in multi-party conversations: given a responding speaker and a dialog history, select an addressee and a response from a set of candidates for the responding speaker. This task requires modeling multi-party conversations and can be directly used to build retrieval-based dialog systems. Propose the Speaker Interaction Recurrent Neural Network (SI-RNN) by updating speaker embeddings in a role-sensitive way. On the public Ubuntu IRC benchmark data set, SI-RNN significantly improves the addressee and response selection accuracy (10% accuracy improvement from the previous state-of-the-art), particularly in complex conversations with many speakers and responses to distant messages many turns in the past.
- second paper: Catherine Finegan-Dollak*, Jonathan K. Kummerfeld*, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. Improving Text-to-SQL Evaluation Methodology. In ACL 2018. run experiments to benchmark current state-of-the-art semantic parsing systems on our newly constructed text-to-sql dataset for student information database.

2.2 Cross-lingual Information Retrieval for Low-Resource Languages (MATERIAL)

- participated in the MATERIAL Project from November 2017 which aims to build a computer system that can process, translate and summarize documents from different low-resource languages so we can greatly improve the efficiency of information access for millions of people speaking different languages around the world.

- as part of IR team, responsible for the system combination and cutoff selection which is critical to optimize the final CLIR evaluation metric. designed a simple and effective solution based on the Sum-to-One score normalization to select the best systems from around 20 CLIR systems and combine their results, and executed a large number of experiments on datasets in four different languages to determine the number of documents to return for each query. In the last evaluation in October 2019, this technique was used in all submissions for our English-Lithuanian and English-Bulgarian CLIR systems which achieved the best performance among three competing teams.

- published an ACL 2019 paper "Improving Low-Resource Cross-lingual Document Retrieval by Reranking with Deep Bilingual Representations". led a team with two graduate and five undergraduate students at Yale, and proposed a deep learning based model for low-resource CLIR. By including query likelihood scores as extra features, the model effectively learns to rerank the retrieved documents by using a small number of a few hundred relevance labels for low-resource language pairs. Furthermore, by aligning word embedding spaces for multiple languages, the model can be directly applied under a zero-shot transfer setting when no training data is available for another language pair. Experimental results on the MATERIAL dataset show that our model outperforms the competitive translation-based baselines by 2%-4% MAP scores on English-Swahili, English-Tagalog, and English-Somali cross-lingual information retrieval tasks.

- helped with preparing the proposal for our Cross-Language Search and Summarization over Text and Speech Workshop at LREC 2020 and volunteer to serve as a Program Committee member.

2.3 Cross-Domain Multi-turn Semantic Parsing (Text-to-SQL)

- built three text-to-sql datasets (co-lead with Tao Yu): (1) SParC published in ACL 2019 for cross-domain Semantic Parsing in Context. It contains 4,298 unique question sequences with 12k+ questions annotated with SQL queries. (2) CoSQL published in EMNLP 2019 for building database querying dialogue systems. It consists of 30k+ turns plus 10k+ annotated SQL queries, obtained from a Wizard-of-Oz collection of 3k dialogues. (3) Both of them are built on top of our Spider dataset published in EMNLP 2018 the largest cross-domain context-independent text-to-SQL dataset available in the field, and thus span 200 complex databases over 138 domains.

- published an EMNLP 2019 paper "Editing-based sql query generation for cross-domain context-dependent questions". proposed an editing-based approach for our cross-domain multi-turn text-to-SQL generation task. Based on the observation that adjacent natural language questions are often linguistically dependent and their corresponding SQL queries tend to overlap, I utilized the interaction history by editing the previous predicted query to improve the generation quality. Experiment results on SParC showed that by generating from the previous query, the model delivered an improvement of 7% question match accuracy and 11% interaction match accuracy over the previous state-of-the-art.

- served as an Organizing Committee member for the Interactive and Executable Semantic Parsing Workshop at EMNLP 2020. focus on mapping natural language utterances to meaning representations that can be executed in a particular context such as databases, knowledge graphs, robotic environment, and software applications. allows users to seek information and control computer systems naturally and flexibly via interactive exchanges in natural language.

3 Mentoring and Teaching

- have supervised five undergraduate students at Yale University spanning various topics including text summarization, cross-lingual information retrieval, and semantic parsing, all of which make significant contributions to the publications, e.g., Graph-based Neural Multi-Document Summarization. Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, Dragomir Radev. CoNLL, 2017.
- staring from my senior undergraduate year, I have served as Teaching Assistants for five courses (Introduction to Programming, Artificial Intelligence, Natural Language Processing) at both undergraduate and graduate levels at the University of Michigan and Yale University.
- served as the sole teaching assistant for Natural Language Processing, Fall 2016, University of Michigan with 90+ students enrolled.
- designed the Neural Machine Translation programming assignments at Natural Language Processing, Spring 2018, Yale University.
- excellent feedback and evaluation from students: 4.4/5.0 for Natural Language Processing in Spring 2018 and 4.8/5.0 for Artificial Intelligence in Fall 2017.
- assisted with a Natural Language Processing Massive Open Online Course on Coursera by building the programming assignments and maintaining the online auto-graders.

4 More Reference

- Cover Letter (`https://ryanzhumich.github.io/cover_letter_rui_zhang.pdf`)
- CV (`https://ryanzhumich.github.io/cv_rui_zhang.pdf`)
- Research Statement (`https://ryanzhumich.github.io/research_statement_rui_zhang.pdf`)
- Teaching Statement (`https://ryanzhumich.github.io/teaching_statement_rui_zhang.pdf`)
- Diversity Statement (`https://ryanzhumich.github.io/diversity_statement_rui_zhang.pdf`)