Computer Science Colloquium

You Are What You Train On: Creating Robust Natural Language Interfaces

Speaker: Jonathan K. Kummerfeld

Host: Dragomir Radev

Monday, February 15, 2021
4:00 p.m.

Zoom Presentation

ABSTRACT:
Natural Language Interfaces like Siri and Alexa help people do things more efficiently, but they are brittle, unable to handle the full range of ways people naturally express themselves. Each of their actions is manually defined by developers, with limited ability to compose actions to make more sophisticated ones. The choice of action is made by a statistical model that is limited by the range of data seen in training. Despite steady progress in the accuracy of these systems, the true scope of remaining challenges has been obscured by the way researchers collect and prepare data.

In this talk, I will describe two of my projects that have revealed previously unknown limitations of natural language interfaces and ways to address them. First, I will show that systems for converting questions to SQL queries have limited generalizability beyond examples seen in training (ACL 2018). I propose a new model and a new way to split data into training and test sets that explore this challenge. Second, I will show that standard crowd-worker data collection processes miss the long and heavy tail of ways people speak (ACL 2017). I propose an outlier-based data collection workflow (NAACL 2019), and a complementary taboo list workflow (EMNLP 2020), that improve data diversity and reduce the cost of data cleaning. I will conclude by outlining a research agenda for fundamentally changing the capabilities of these systems. Today we use these systems to do simple tasks, e.g. “start a 5 minute timer”. My work will enable systems to do complex tasks as part of applications, e.g. “Plot population over the last 2000 years with a trend line only and a log scale on the y-axis”.

BIO:
Jonathan K. Kummerfeld is a Postdoctoral Research Fellow in Computer Science and Engineering at the University of Michigan. He completed his Ph.D. at the University of California, Berkeley, advised by Prof. Dan Klein. Jonathan’s research has revealed new challenges in syntactic parsing, coreference resolution, and dialogue. He has proposed models and algorithms to address these challenges, improving the speed and accuracy of natural language processing systems. He has been on the program committee for 55 conferences and workshops, including Area Chair at ACL and Shared Task Coordinator for the DSTC workshops. He currently serves as a standing reviewer for the Computational Linguistics journal and the Transactions of the Association for Computational Linguistics journal. For more details, see his website: https://www.jkk.name

Yale University