DaSH 2024

Data Science with Human-in-the-Loop

Proceedings of the DaSH Workshop at NAACL 2024

June 20, 2024

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-101-8

Introduction

We are delighted to welcome to you DaSH 2024, the Fifth Workshop on Data Science with Human-in-the-loop at NAACL 2024!

The aim of this workshop is to stimulate research on the cooperation between humans and computers within the broad area of natural language processing, including but not limited to information extraction, information retrieval and text mining, machine translation, dialog systems, question answering, language generation, summarization, model interpretability, evaluation, fairness, and ethics. We invite researchers and practitioners interested in understanding how to optimize human-computer cooperation and how to minimize human effort along an NLP pipeline in a wide range of tasks and applications.

We hope to bring together interdisciplinary researchers from academia, research labs and practice to share, exchange, learn, and develop preliminary results, new concepts, ideas, principles, and methodologies on understanding and improving human-computer interaction in natural language processing. We expect the workshop to help develop and grow a strong community of researchers who are interested in this topic and to yield future collaborations and scientific exchanges across the relevant areas of computational linguistics, natural language processing, data mining, machine learning, data and knowledge management, human-machine interaction, and intelligent user interfaces. We are thankful to IBM research for sponsoring the workshop and best paper awards.

We hope you have a wonderful time at the workshop.

Cheers!

DaSH 2024 Organizers
Eduard Dragut, Temple University
Yunyao Li, Adobe
Lucian Popa, IBM Research
Shashank Srivastava, UNC Chapel Hill
Slobodan Vucetic, Temple University

Organizing Committee

Organizers

Eduard Dragut, Temple University
Yunyao Li, Adobe
Lucian Popa, IBM Research - Almaden
Shashank Srivastava, University of North Carolina at Chapel Hill
Slobodan Vucetic, Temple University

Program Committee

Chairs

Eduard Dragut, Temple University
Yunyao Li, Adobe
Lucian Popa, IBM Research - Almaden
Shashank Srivastava, University of North Carolina at Chapel Hill
Slobodan Vucetic, Temple University

Program Committee

Arjun Bhalla, Bloomberg
Eleftheria , University of Maryland
Zhijia Chen, Temple University
Aritra Dasgupta, New Jersey Institute of Technology
Rotem Dror, University of Haifa
Varun Embar, Apple
Shivali Goel, Columbia University
Sairam Gurajada, Megagon
Maeda Hanafi, IBM
Lihong He, IBM
Farnaz Jahanbakhsh, MIT
Eser Kandogan, Megagon
Edith Law, University of Waterloo
Akash Maharaj, Adobe
Yiwen Sun, Apple
Yuan Tian, Purdue University

Keynote Talk

Show It or Tell It? Text, Visualization, and their Combination

Marti Hearst
University of California, Berkeley

Abstract: In this talk, Dr. Marti Hearst will share observations about the role of language in information visualization. I will pose questions such as: how do we decide what to express via language vs via visualization? How do we choose what kind of text to use when creating visualizations, and does that choice matter? Does anyone prefer text over visuals, under what circumstances, and why?

Bio: Dr. Marti Hearst is the Interim Dean of the School of Information and a Professor at UC Berkeley in the School of Information and the Computer Science Division. Her research encompasses user interfaces with a focus on scientific document understanding, information visualization with a focus on text, and computational linguistics. She is the author of Search User Interfaces, the first academic book on that topic. She is past President of the Association of Computational Linguistics, an ACM Fellow, a member of the CHI Academy, a SIGIR Fellow, and ACL Fellow, and has received four Excellence in Teaching Awards.

Keynote Talk

Show Reasoning Myths about Language Models: What is Next?

Dan Roth
University of Pennsylvania and Amazon

Abstract: The rapid progress made over the last few years in generating linguistically coherent natural language has blurred, in the minds of many, the difference between natural language generation, understanding, and the ability to reason with respect to the world. Nevertheless, robust support of high-level decisions that depend on natural language understanding, and that require dealing with "truthfulness" are still beyond our capabilities, partly because most of these tasks are very sparse, often require grounding, and may depend on new types of supervision signals.
Dan will discuss some of the challenges underlying reasoning and argue that we should focus on LLMs as orchestrators – coordinating and managing multiple models, applications, and services, to execute complex tasks and processes. I will discuss some of the challenges and present some of our work in this space, focusing on supporting task decomposition and planning.

Bio: Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, University of Pennsylvania, a VP/Distinguished Scientist at AWS AI, and a Fellow of the AAAS, the ACM, AAAI, and the ACL. In 2017 Roth was awarded the John McCarthy Award, the highest award the AI community gives to mid-career AI researchers. Roth was recognized "for major conceptual and theoretical advances in the modeling of natural language understanding, machine learning, and reasoning." Roth has published broadly in machine learning, natural language processing, knowledge representation and reasoning, and learning theory. He was the Editor-in-Chief of the Journal of Artificial Intelligence Research (JAIR), has served as the Program Chair for AAAI, ACL and CoNLL. Prof. Roth received his B.A Summa cum laude in Mathematics from the Technion, Israel, and his Ph.D. in Computer Science from Harvard University in 1995.

Keynote Talk

Training Social Skills via Human-AI Collaboration

Diyi Yang
Stanford University

Bio: Diyi Yang is an assistant professor in the Computer Science Department at Stanford University. Her research focuses on human-centered natural language processing and computational social science. She is a recipient of the Microsoft Research Faculty Fellowship (2021), NSF CAREER Award (2022), ONR Young Investigator Award (2023), and Sloan Research Fellowship (2024). Her work has received multiple paper awards or nominations at top NLP and HCI conferences (e.g., ACL, EMNLP, SIGCHI, and CSCW).

Keynote Talk

Model-Aided Human Annotation at Scale

Hadas Kotek
Apple

Bio: Dr. Hadas Kotek is a senior data scientist on the Siri Natural Language Understanding team at Apple. She earned a PhD in Linguistics from MIT and previously held faculty positions at McGill University, New York University, and Yale University. Dr. Kotek develops methodologies for measuring the accuracy and efficiency of data annotation at scale, as well as the safety, robustness, and diversity of the resulting datasets and models, leveraging cross-functional teams to support innovative, product-centric research. Her most recent research is in the domains of model-in-the-loop annotation, ethical AI, and the efficacy of Large Language Models. In Fall 2023, she taught a full-semester seminar on Large Language Models at MIT, where she is currently a Research Affiliate.

Table of Contents

APE: Active Learning-based Tooling for Finding Informative Few-shot Examples for LLM-based Entity Matching
Kun Qian, Yisi Sang, Farima Bayat†, Anton Belyi, Xianqi Chu, Yash Govind, Samira Khorshidi, Rahul Khot, Katherine Luna, Azadeh Nikfarjam, Xiaoguang Qi, Fei Wu, Xianhan Zhang and Yunyao Li 1

Towards Optimizing and Evaluating a Retrieval Augmented QA Chatbot using LLMs with Human-in-the-Loop
Anum Afzal, Alexander Kowsik, Rajna Fani and Florian Matthes 4

Evaluation and Continual Improvement for an Enterprise AI Assistant
Akash Maharaj, Kun Qian, Uttaran Bhattacharya, Sally Fang, Horia Galatanu, Manas Garg, Rachel Hanessian, Nishant Kapoor, Ken Russell, Shivakumar Vaithyanathan and Yunyao Li 17

Mini-DA: Improving Your Model Performance through Minimal Data Augmentation using LLM
Shuangtao Yang, Xiaoyi Liu, Xiaozheng Dong and Bo Fu 25

CURATRON: Complete and Robust Preference Data for Rigorous Alignment of Large Language Models
Son The Nguyen, Niranjan Uma Naresh and Theja Tulabandhula 31

Program

Thursday, June 20, 2024

09:00 - 09:10 *Opening Remarks*

09:10 - 10:00 *Keynote Talk 1*

10:00 - 10:30 *Invited Talk 1*

10:30 - 11:00 *Coffee Break*

11:00 - 12:15 *Workshop Papers*

12:15 - 12:30 *Invited Paper (from Findings of NAACL'24)*

12:30 - 14:00 *Lunch Break*

14:00 - 14:50 *Keynote Talk 2*

14:50 - 15:20 *Invited Papers (from Findings of NAACL'24)*

15:20 - 16:00 *Coffee Break*

16:00 - 16:30 *Invited Talk 2*

16:30 - 17:15 *Panel Discussion*

17:15 - 17:30 *Conclusion of Workshop (Open Discussions)*