TrustNLP

TrustNLP: First Workshop on Trustworthy Natural Language Processing

Proceedings of the Workshop

June 10, 2021

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-33-6

Introduction

Recent progress in Artificial Intelligence (AI) and Natural Language Processing (NLP) has greatly increased their presence in everyday consumer products in the last decade. Common examples include virtual assistants, recommendation systems, and personal healthcare management systems, among others. Advancements in these fields have historically been driven by the goal of improving model performance as measured by accuracy, but recently the NLP research community has started incorporating additional constraints to make sure models are fair and privacy-preserving. However, these constraints are not often considered together, which is important since there are critical questions at the intersection of these constraints such as the tension between simultaneously meeting privacy objectives and fairness objectives, which requires knowledge about the demographics a user belongs to. In this workshop, we aim to bring together these distinct yet closely related topics.

We invited papers which focus on developing models that are "explainable, fair, privacy-preserving, causal, and robust" (Trustworthy ML Initiative). Topics of interest include:

- Differential Privacy
- Fairness and Bias: Evaluation and Treatments
- Model Explainability and Interpretability
- Accountability
- Ethics
- Industry applications of Trustworthy NLP
- Causal Inference
- Secure and trustworthy data generation

In total, we accepted 11 papers, including 2 non-archival papers. We hope all the attendants enjoy this workshop.

Organizing Committee

- Yada Pruksachatkun - Alexa AI
- Anil Ramakrishna - Alexa AI
- Kai-Wei Chang - UCLA, Amazon Visiting Academic
- Satyapriya Krishna - Alexa AI
- Jwala Dhamala - Alexa AI
- Tanaya Guha - University of Warwick
- Xiang Ren - USC

Speakers

- Mandy Korpusik - Assistant professor, Loyola Marymount University
- Richard Zemel - Industrial Research Chair in Machine Learning, University of Toronto
- Robert Monarch - Author, Human-in-the-Loop Machine Learning

Program committee

- Rahul Gupta - Alexa AI
- Willie Boag - Massachusetts Institute of Technology
- Naveen Kumar - Disney Research
- Nikita Nangia - New York University
- He He - New York University
- Jieyu Zhao - University of California Los Angeles
- Nanyun Peng - University of California Los Angeles
- Spandana Gella - Alexa AI
- Moin Nadeem - Massachusetts Institute of Technology
- Maarten Sap - University of Washington
- Tianlu Wang - University of Virginia
- William Wang - University of Santa Barbara
- Joe Near - University of Vermont
- David Darais - Galois
- Pratik Gajane - Department of Computer Science, Montanuniversitat Leoben, Austria
- Paul Pu Liang - Carnegie Mellon University

- Hila Gonen - Bar-Ilan University
- Patricia Thaine - University of Toronto
- Jamie Hayes - Google DeepMind, University College London, UK
- Emily Sheng - University of California Los Angeles
- Isar Nejadgholi - National Research Council Canada
- Anthony Rios - University of Texas at San Antonio

Table of Contents

Conference Program

June 10, 2021

9:00–9:10 *Opening*
Organizers

9:10–10:00 *Keynote 1*
Richard Zemel

10:00–11:00 Paper Presentations

Interpretability Rules: Jointly Bootstrapping a Neural Relation Extractorwith an Explanation Decoder
Zheng Tang and Mihai Surdeanu

Measuring Biases of Word Embeddings: What Similarity Measures and Descriptive Statistics to Use?
Hossein Azarpanah and Mohsen Farhadloo

Private Release of Text Embedding Vectors
Oluwaseyi Feyisetan and Shiva Kasiviswanathan

Accountable Error Characterization
Amita Misra, Zhe Liu and Jalal Mahmud

11:00–11:15 *Break*

June 10, 2021 (continued)

11:15–12:15 **Paper Presentations**

xER: An Explainable Model for Entity Resolution using an Efficient Solution for the Clique Partitioning Problem
Samhita Vadrevu, Rakesh Nagi, JinJun Xiong and Wen-mei Hwu

Gender Bias in Natural Language Processing Across Human Languages
Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie and Jeanna Matthews

Interpreting Text Classifiers by Learning Context-sensitive Influence of Words
Sawan Kumar, Kalpit Dixit and Kashif Shah

Towards Benchmarking the Utility of Explanations for Model Debugging
Maximilian Idahl, Lijun Lyu, Ujwal Gadiraju and Avishek Anand

12:15–1:30 ***Lunch Break***

13:00–14:00 ***Mentorship Meeting***

14:00–14:50 *Keynote 2*
Mandy Korpusik

14:50–15:00 ***Break***

15:00–16:00 ***Poster Session***

16:15–17:05 *Keynote 3*
Robert Munro

17:05–17:15 ***Closing Address***