Computer Science
Seminar

Positive AI with Social Commonsense Models

Maarten Sap
University of Washington

Abstract: To effectively understand language and safely communicate with humans, machines must not only grasp the surface meanings of texts, but also their underlying social meaning. This requires understanding interpersonal social commonsense, such as knowing to thank someone for giving you a present, as well as understanding harmful social biases and stereotypes. Failure to account for these social and power dynamics could cause models to produce redundant, rude, or even harmful outputs.¡br¿ ¡br¿
In this talk, I will describe my research on enabling machines to reason about social dynamics and social biases in text. I will first discuss ATOMIC, the first large-scale knowledge graph of social and interpersonal commonsense knowledge, with which machines can be taught to reason about the causes and effects of everyday events. Then, I will show how we can make machines understand and mitigate social biases in language, using Social Bias Frames, a new structured formalism for distilling biased implications of language, and PowerTransformer, a new unsupervised model for controllable debiasing of text.¡br¿ ¡br¿
I will conclude with future research directions on making NLP systems more socially-aware and equitable, and how to use language technologies for positive societal impact.

Friday, February 5, 2021, 1:00 pm
https://emory.zoom.us/j/92294085195

Computer Science
Emory University