COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Shane Storks

Coherent Physical Commonsense Reasoning in Foundational Language Models

Wednesday, September 4, 2024
2:00pm – 4:00pm
2725 Beyster
Hybrid – Zoom

ABSTRACT: Recent years in natural language processing (NLP) research have seen a paradigm shift toward foundational language models (LMs), which are trained on large amounts of text data from the web and serve as flexible foundations that can be easily applied to downstream tasks. While *commonsense reasoning*, i.e., the ability to incorporate implicit background knowledge into natural language understanding (NLU), is a long-standing grand challenge in NLP research with decades of effort spent, these foundational LMs exhibit an apparent human-level proficiency on traditional NLU benchmarks requiring it. However, given limitations of these LMs, including their lack of transparency, tendency to exploit statistical bias in language data, and capability to hallucinate factual information, we argue that traditional benchmarking practices are no longer appropriate to evaluate the commonsense reasoning capabilities of foundational LMs.

In this thesis, we develop a new evaluation paradigm targeting *coherent commonsense reasoning*. While traditional benchmarks boil down NLU into high-level text classification tasks targeting various semantic phenomena, we propose a concept of *consistency* of LMs' decisions by requiring them to localize these semantic phenomena within long language contexts, serving as evidence for decisions. Further, we propose a concept of *verifiability* specific to physical commonsense reasoning (PCR), which requires LMs to explicitly model the physical states associated with actions occurring in text. We densely annotate several benchmark datasets to capture these concepts. While we find that traditional fine-tuning and in-context learning strategies to apply LMs to downstream tasks severely lack coherence in their commonsense reasoning, we develop new strategies inspired by dual process theory in human cognition, significantly improving LMs' coherence and faithfulness of attention.

Lastly, we adapt this evaluation paradigm to the challenging multimodal task of procedural mistake detection in video frames. We develop automated, reference-free metrics for the relevance and informativeness of LM-generated explanations in this problem, using them to create a novel, multi-tiered coherence evaluation. We then systematically investigate the impact of various interventions in LMs' textual and visual inputs on performance, and show how our evaluation framework can reveal a wealth of insights into the strengths and weaknesses of LMs, enabling auditing and possible future improvement.

CHAIR: Prof. Joyce Chai