CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text

The recent success of natural language understanding (NLU) systems has been troubled by results highlighting the failure of these models to generalize in a systematic and robust way. In this talk, I will introduce a diagnostic benchmark suite, named CLUTRR, to clarify some key issues related to the robustness and systematicity of NLU systems. Motivated by classic work on inductive logic programming, CLUTRR requires that an NLU system infer kinship relations between characters in short stories. Successful performance on this task requires both extracting relationships between entities, as well as inferring the logical rules governing these relationships. CLUTRR allows us to precisely measure a model's ability for systematic generalization by evaluating on held-out combinations of logical rules, and it allows us to evaluate a model's robustness by adding curated noise facts. Our empirical results highlight a substantial performance gap between state-of-the-art NLU models (e.g., BERT and MAC) and a graph neural network model that works directly with symbolic inputs---with the graph-based model exhibiting both stronger generalization and greater robustness.