Understanding the Dynamics of Learning Linguistic Knowledge in Large Language Models

Hadi Amiri*

Abstract

Language models serve as the foundation of numerous AI applications today. However, despite their remarkable proficiency in processing textual data, questions linger regarding their linguistic capabilities. In this talk, I will empirically study the linguistic capabilities of recent large language models (LLMs) using a characterization of linguistic complexity from psycholinguistic and language acquisition research, and introduce datadriven linguistic curricula to understand the learning dynamics of these models. Through extensive analyses, we find that recent LLMs show limited efficacy in addressing linguistic queries, particularly struggling with inputs of greater linguistic complexity and controlled generation. In addition, I demonstrate that the proposed linguistic curricula can identify sets of linguistic metrics that inform the challenges and reasoning required for addressing different natural language processing tasks. The talk will conclude by offering insights to inform future endeavors in LLM development.

---

*Miner School of Computer & Information Sciences University of Massachusetts (UMass) Lowell, USA, hadi@cs.uml.edu