Towards Scaling Large Language Models to 1000 Languages – Challenges and Advances

Professor Lei Li, Carnegie Mellon University

Location: **Campus Center, 2nd Floor, Room 2551**

Time 4-5:15 pm, Nov. 4, 2024

Zoom: https://umassboston.zoom.us/j/94125081677

Abstract: Large Language Models (LLMs) demonstrate general NLP capabilities, but they are limited to a few high-resource languages. How can we efficiently scale LLM's capabilities to massively many languages? In this talk, we present our analysis on the fundamental challenges of extending LLMs to more languages, including improper linguistic token design, insufficient multilingual raw text, and insufficient cross-lingual training. We introduce two approaches towards scaling LLM to massive languages: i) LLaMAX, a continual pre-training approach to enable translation support across more than 100 languages; and ii) LingoLLM, a training-free method to incorporate linguistic description to enhance language capabilities for endangered languages. Finally, I will share my vision for advancing LLM and machine translation for 1000 languages.

Bio: Lei Li is an Assistant Professor in Language Technologies Institute at Carnegie Mellon University. His research focuses on machine translation, trustworthy generative AI, agentic LLM, and AI-powered drug discovery. He received Ph.D. from CMU School of Computer Science in 2011. He is a recipient of the ACL 2021 Best Paper Award, the CCF Young Elite Award in 2019, the CCF Distinguished Speaker in 2017, the Wu Wen-tsün AI prize in 2017, and the 2012 ACM SIGKDD dissertation award (runner-up), and is recognized as Notable Area Chair of the ICLR 2023. Previously, he was an associate professor (tenured) at UC Santa Barbara. Before that, he was the Founding Director of ByteDance AI Lab, a principal scientist at Baidu, and a postdoc researcher at UC Berkeley. He led and developed ByteDance's machine translation system VolcTrans and AI writing system Xiaomingbot, and many of his algorithms have been deployed in products (Toutiao, Douyin, Tiktok, Lark), serving over one billion users.

UMass Boston

Web: https://www.cs.cmu.edu/~leili/