Bocconi

28 JUNE 2023
12:00 PM

Bocconi University
Room 3-E4-SR03
Via Röntgen 1, Milano (3° floor)

Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages

Speaker

Prof. Dr. Hinrich Schütze

Chair for Computational Linguistics

LMU, Ludwig-Maximilians-Universität

Abstract: Large language models (LLMs) are currently the most active area of research in NLP. Most work has focused on what we call "vertical" scaling: making LLMs even better for a relatively small number of high-resource languages. We address "horizontal" scaling instead: extending LLMs to a large subset of the world's languages, focusing on low-resource languages. Our Glot500-m model is trained on more than 500 languages, many of which are not covered by any other language model. I will talk about the major challenges we faced: (i) finding, validating and cleaning training data for low-resource languages; (ii) evaluating performance of Glot500-m on low-resource languages for which native speakers and labeled datasets were not available to us; and (iii) determining the factors that ultimately make training on a low-resource language successful. We find that trying to reduce such factors to the so-called curse of multilinguality is naive and there is in fact also a "boon of multilinguality". We have released Glot500-m and are in the process of making Glot500-c, our training corpus covering 500 languages, publicly available.

Università
Bocconi
DEPARTMENT
OF COMPUTING
SCIENCES