Title: Interpreting and Controlling Linguistic Features in Multilingual Language Models

Author: Tomasz Limisiewicz

Department: Institute of Formal and Applied Linguistics

Supervisor: David Mareček,
Institute of Formal and Applied Linguistics

Abstract:

Language models based on neural networks have become the foundation for solving diverse tasks, yet their inner workings remain opaque. This dissertation investigates which components of language models are crucial for representing and processing information from texts. We mainly focus on multilingual models that can reuse representation for processing tasks in different languages. We hypothesize that understanding how models represent linguistic phenomena is necessary to control their function and alleviate issues hindering models' performance. To this end, we propose novel interpretability techniques to analyze specific components of language models. Our approaches enable the localization of the representation of distinct types of signals within the language models. The localization allowed us to contain our analysis to specific modules and apply precise interventions to mitigate gender bias or improve cross-lingual transfer.

Keywords: natural language processing, language models, multilinguality, interpretable machine learning