August 30, 2024

PhD Dissertation Review

Thesis title: Interpreting and Controlling Linguistic Features in Multilingual Language Models
Author: Tomasz Limisiewicz
Reviewer: Yonatan Belinkov

Summary and evaluation

The dissertation by Tomasz Limisiewicz, describes a comprehensive investigation of multilingual language models. It analyzes the main components in modern language models, namely, MLPs, attention, and the LM vocabulary. These are analyzed for different kinds of linguistic features that may be encoded in them, including, lexical, semantic, and syntactic features. In several cases the analyses lead to concrete techniques for controlling the language models, such as in cross-lingual transfer or gender debiasing. These are important and useful applications.

The thesis makes novel contributions to natural language processing in both methodology and findings. Its focus on multilingual models and datasets is appreciates, as most work in this area tends to focus on English. The methods and results are clearly presented. The dissertation demonstrates the author's ability to conduct high quality creative scientific work. Below I include specific comments and a few questions.

Details and questions

One general question that I would love to hear feedback about is how relevant the results and methods are in the fast-changing world of language models, especially given the tendency to focus on more high-level downstream tasks and capabilities and less on low-level linguistic features such as those studies in the thesis.

Chapters 2 and 3 provide comprehensive background information on language models, linguistic features, and evaluation metrics.

Yonatan Belinkov
Senior Lecturer, Faculty of Computer Science
Taub Building 733, Technion City, Haifa 3200003, Israel.
Phone: +972-4-8294958, Email: belinkov@technion.ac.il
Web: http://www.cs.technion.ac.il/~belinkov

יונתן בלינקוב
מרצה בכיר, הפקולטה למדעי המחשב
בניין טאוב 733, הטכניון, חיפה 3200003, ישראל
טלפון: +972-4-8294958, דוא"ל: belinkov@technion.ac.il
דף בית: http://www.cs.technion.ac.il/~belinkov

Chapter 4 introduces a new method for probing for linguistic features, named orthogonal probes. By introducing orthogonality to the probe, it allows disentangling different types of features in the projected space. This is shown to allow for improvements in cross-lingual transfer and mitigating gender bias. The experiments focus on multilingual models and are compared to standard methods in the probing literature. The approach is well motivated and explained clearly, with nice results.

Chapter 5 analyzes gender bias in auto-regressive language models. It first localize gender bias to certain layers and modules of the model, specifically MLP layers. Then it develops a new algorithm for removing biased associations by projecting representations on the biased subspace. The experiments on large LLaMA models demonstrate effects in debiasing with little damage to other model capabilities. This is an interesting analysis and application to an important problem. The analysis is extensive and the results appear sound. I do have two questions:

1. How does DAMA compare to linear removal methods like INLP or LEACE, both theoretically and empirically?

2. Some recent work by Hase et al. casts doubts on the connection between localization and model editing. They found no such correlation, in their case being able to do editing with ROME/MEMIT at virtually any layer. However, that might be because ROME/MEMIT are too strong. Would DAMA also work at any layer(s) or do we really need to localize the information first for it to work well?

Chapter 6 investigates how well attention weights in BERT and mBERT encode syntactic dependency relations. It proposes an algorithm for selecting sets of heads and ensembling them to get edge scores. Then it runs a MST algorithm. The results are compared with labeled trees in multiple languages, showing better than random scores. The most interesting parts are the analyses of specific cases, where the same syntactic relation is encoded in multiple heads (explaining why ensembles are needed) or when a single head encodes multiple relations. It's also interesting to see heads that encode the same relation in multiple languages. One question I have is why the LAS scores are so much lower than the UAS scores. From my memory of dependency parsing, the difference is usually much smaller, although those results where with supervised parsers.

Chapter 7 explores and understudies aspect of language models - their tokenization algorithms. It compares existing algorithms and proposes two new ones, in the setting of a multilingual model trained on data from different languages. Interestingly, the no-overlap tokenization makes models perform quite well on some tasks, but poorly on others. I also found that distinction between masked LM performance and downstream performance very instructive, and in line with some results in the literature (see the PMI masking paper). This is an important distinction. I would love to know if a similar pattern holds with auto-regressive LMs, where generally LM performance correlated better with downstream performance. This would be a good avenue for future work.

Sincerely,

Yonatan Belinkov, Ph.D.
Senior Lecturer, The Henry and Merilyn Taub Faculty of Computer Science
Technion – Israel Institute of Technology