Report on the Doctoral Thesis

“Interpreting and Controlling Linguistic Features in Multilingual Language Models”

by Ing. Tomasz Limisiewicz

At a high level, the thesis contributes to a deeper understanding of the current foundation models: their internal workings, types of encoded information, ways to measure this, and methods to control and alter the encoded information. Since such models are receiving increasingly wide adoption in high-stakes settings, working towards better understanding and control of the types of encoded information is vital from the safety, reliability and trustworthiness perspectives.

The thesis contributes to (i) methods to probe model representations for encoded information, and (ii) better understanding of model components: feed-forward layers, attention layers, token embeddings. Starting from methods, the thesis presents a novel alternative to standard probing for linguistic structure: orthogonal probing. This method is grounded theoretically and allows, among other things, separating representations of different kinds of information inside the model. Experimentally, this is illustrated by showing that (i) lexical and syntactic information are encoded in different substapes in the model, (ii) unwanted information can be filtered from the representation space, and (iii) analysis of multilingual models can reveal how information is shared between languages from close and distant families. As a result, I see the proposed orthogonal probing an important contribution to the NLP field since it can potentially be used as an analysis tool when developing new models and in practice as part of the safety pipelines (e.g., by removing unwanted information).

Coming to model components, the thesis covers all the main components of the current state-of-the-art models: feed-forward, attention, and embedding layers. For all of them, it makes novel observations that are not only relevant by themselves but also can lead to much broader understanding of the ways information can be encoded in a network. For example, when talking about attention layers, the thesis explains that some types of information, e.g. syntactic, can be encoded in collaboration by several attention heads. Firstly, this is interesting because previously, there was rather weak evidence of the presence of syntactic heads in language models (in contrast to machine translation models having individual high-accuracy syntactic heads). Secondly, presented work (i) suggests a novel way of finding various kinds of information, and (ii) can mean that other kinds of information are likely to manifest themselves in specialised model components if looking at them in collaboration and not individually. For other considered model components, the thesis also provides novel observations with a potential to be useful in practice. Some examples of these include locating bias in feed-forward layers,

identifying the influence of different aspects of tokenization on quality of multilingual models, etc.

Overall, this thesis highlights the author's deep understanding of their research field, ability to identify important research questions, formulate hypotheses and conduct the experiments to answer the posed questions. I can confidently conclude that Tomasz Limisiewicz has proved their ability for creative scientific work and can successfully complete their doctoral program.

21.08.2024
Dr. Elena Voita