Opponent's review of PhD thesis proposal

PhD student: **Ing. Tomasz Limisiewicz**

Title: **Interpreting and Controlling Linugistc Features in Neural Networks' Representations**

Summary of the proposal content

The thesis proposal deals with the quest for interpretability in vector space representations used in the modern NLP. The first section presents four directions of motivation for this topic. The second section briefly surveys current deep learning models used in NLP. Section 3 presents a structured overview of existing approaches for searching interpretability. Section 4 focuses specifically on language biases in order to improve fairness of arising NLP technology. Section 5 summarizes the author's achievements presented in his publications along four different experimental directions. Section 6 discusses possible directions of future work. Section 7 summarizes the author's current position with respect to the four aims formulated in the introductory section.

Evaluation:

In my opinion, Tomasz has gained a really deep insight into how Neural Networks are used within the contemporary NLP (and wide knowledge of related literature), and has presented a systematized survey of where (and for what purpose) linguistic interpretability is searched for in the current NLP. Tomasz has already performed several novel experiments of his own along a few directions. I find the survey part as well as his own research overview quite mature and promising for his future dissertation.

As for the formal quality of the text, it is clearly structured, and as far as I can judge, Tomasz's English is basically flawless (I noticed only extremely small amount of language imperfections). As for the typesetting quality, it is OK too, just that more attention should be paid to bibtex entries: incomplete references as well as references with incorrectly lowercased titles should be fixed before being used in other publications. \citep should be perhaps used instead of \citet in the itemization in 2.1 (or subject-verb agreement in those descriptions should respect the number of authors).

A question for the exam:
(could you please prepare a brief reaction?)

In the context of disentangling, you speak about identifying subspaces of vector space representations. I assume that a linear subspace is meant, strictly in term of linear algebra. Would it make sense to try to isolate interpretable components of vector space representations also in some other mathematical form, e.g. as total orderings, or posets, or groups, especially in cases where multiple distinguishable values are considered in linguistic analysis?

Other comments and suggestions:

- There are quite a few other research activities aimed at merging vector space representations with linguistic intuitions carried out at UFAL (mostly very recently), such as Jan Bodnár's study on comparing embedding representations of allomorhps, Jonáš Vidra's experiments on prediction derivational relations based on embeddings, Ebrahim Ansari's experiments aimed

at making fastttext to assign more weight to morpheme-like subwords, Niyati Bafna's crosslingual embedding transfer which profits from similarities in morphematic segmentation between a high-resource and a genealogically related low-resource language, or Rudolf Rosa's thoughts on embedding-based unsupervised lemmatization and on differentiating inflection and derivation. Perhaps with the exception of the last one, the ideas and experiments are currently waiting for being published, but I believe there could be some space for interesting discussions already now.

- Multilingual studies seem to be belong to the author's priorities in the future ("6.3 Analysis beyond English"). This reminds me that recently I went across Piasecki et al.,(2018), and even if I don't trust wordnets much, the fact that wordnets are abundant nowadays suggests that wordnet-based probes could be performed for massive amounts of languages relatively easily (perhaps for even more languages than in the case of UD-based probes). Reference: *Maciej Piasecki et al., 2018: Wordnet-based evaluation of large distributional models for Polish. In Proceedings of the 9th Global Wordnet Conference, pages 229–238, Nanyang Technological University (NTU), Singapore.*

- Tomasz has gained a quite promising publication record since the beginning of his PhD study. However, at least from my viewpoint, it could be difficult to make all his started research threads sufficiently deep and to make them all meet in a single dissertation. This is just a humble suggestion, but I think that some prioritizing and a narrower focus might be more efficient in order to complete the dissertation within the recommended study length (even if the choice could be difficult: maybe I am biased, but it seems to me that the anti-biased-language topic is the most distant one, however, on the other hand, the resulting publication is the most cited one among Tomasz's publications).

Conclusion

I do recommend Tomazs's thesis proposal to be accepted.

In Jahodov, 17th September 2021

doc. Ing. Zdeněk Žabokrtský, Ph.D.