Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Michael Hanna
Název práce	Investigating Large Language Models' Representations Of Plurality Through Probing Interventions
Rok odevzdání	2022
Studijní program	Informatika **Studijní obor** Language Technologies and Computational Linguistics

Autor posudku	David Mareček **Role** vedoucí
Pracoviště	ÚFAL MFF UK

Text posudku:

The goal of the thesis was to combine several interpretation methods to investigate how Transformers really process the subject-verb agreement in English sentences. The main finding is that even though the probing methods detect the subject plurality information in all words of the sentence, the intervention methods show that it is really used only in words that agree with the subject (article, subject, verb), and when the task is to predict the masked verb, the plurality information is being gradually transferred from the subject to the verb during the layers.

After the Introduction, the second chapter describes the background theory - neural networks in NLP, Transformers, and current post-hoc and behavioral interpretation techniques. The third chapter describes the intervention techniques and outlines the methods that are used in the experiments. The fourth chapter is the most extensive one and describes individual experiments and results. This is followed by the Discussion and Conclusion.

Hodnocení práce:

Michael worked very independently, he studied a lot of related literature and proved to be very well versed in this field. All the experiments and methods were proposed by him, I only suggested making some additional tests and visualizations. The methods used (as probing intervention and swap intervention) are not novel, however, the novelty of the thesis is in their combination and usage in investigating the plurality and subject-verb agreement.

The thesis consists of 42 pages without references and it is written in very good English. I appreciate its structure, the story of the thesis is very well written and all procedures are clear and well reasoned. I like the experiments–predictions–results partitioning in the experimental section. A bit weaker point is the graphs presented, for example, one would need to compare the graphs in Figures 4.3 and 4.4, but they are on different pages.

I suggest publishing this work as a conference paper. It demonstrates well the fact that even if a piece linguistic information can be well probed from the representations, it may not be used at all. I recommend the thesis for defense.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 20. 8. 2022

Podpis: