Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Paul Mouret
Název práce	Evaluation of gender bias of Large Language Models in Natural Contexts
Rok odevzdání	2024
Studijní program	Computer Science **Studijní obor** Artificial Intelligence

Autor posudku	David Mareček **Role** vedoucí
Pracoviště	Institute of Formal and Applied Linguistics

Text posudku:

The goal of this thesis is to build a new evaluation dataset for detecting gender bias. Unlike many other datasets that are usually generated from templates and covers only biases of a specific group of words (e.g. professions), this is based on real texts so the amount of bias can be evaluated for every word. Further, this dataset is used for causal tracing to identify which parts of the Transformer are responsible for storing unwanted stereotypical bias. Finally, existing techniques for updating Large Language Models (LoRA and DAMA) are used to create models mitigating gender bias.

Structure of the thesis:

The thesis is divided into three chapters, Introduction and Conclusion.

The first Chapter describes the background theory (transformer architecture, selected Large Language Models, evaluation of gender bias), and all the tools and techniques that were needed in the experiments (syntactic parsing, coreference resolution, causal tracing, Low-rank adaptation, DAMA).

The second chapter describes the methodology of the experiments:

- how the dataset was created: the choice of the source data, selection of relevant contexts, and filtering.
- definitions of bias measures: the "Bias score" and the "Relevance score" that are very useful and important for the evaluation
- description of techniques and choice of parameters for analysis of stereotypical gender bias

The third chapter presents the results: the statistics of the generated datasets, statistics on gender biases in the data, lists of the most biased words, transformer layers that are responsible for gender bias and the results of the Large Language Models finetuning and adaptation.

Evaluation:

The thesis is written in English, the main body covers 76 pages and it is very well structured.

I appreciate very detailed analysis on the stereotypical biases of individual words. Although I thought this would be rather simple, it turned out to be a very complex problem that would deserve much more time. Paul has shown that he is capable of working independently and designing and evaluating partial experiments on his own. He proposes the bias and the relevance scores for contexts measuring the relative difference and sum of the probabilities of predicting the words *he* and *she*. Then he shows how noising of individual words in these contexts affects these scores. (By noising, we mean adding a random noise to the word embedding, which practically removes the meaning of the word but preserve the flow of the rest of information of the context at this position in the Transformer). With this method, he was able to extract the most biased words, some of them confirming our expectations (*nurse*, *blonde* vs. *farmer, merchant*), some of them showing very interesting biases (e.g. *deciding, grabbing, twisting* are female biased). He also showed, that the word's bias strongly depends on whether the word belongs to the subject or not.

Unfortunately, Paul did not have much time for the final experiments on fine-tuning or adapting existing Large Language Models to mitigate unwanted bias in them. This was very time-consuming, since it required significant amount of computational capacity, and it wasn't possible to fully explore all the options and perform better evaluations. Anyway, Paul demonstrated that the resulting models reduce stereotypical gender bias and are at least comparable to other models using simpler datasets.

Conclusion:

I am confident that Paul has done a significant amount of work, which meets the requirements for a master thesis. I therefore fully recommend it to be defended.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 3. 9. 2024

Podpis: