Supervisor’s review of master thesis

Author of the review: **doc. RNDr. Pavel Pecina, Ph.D.**

Author of the thesis: **Goutham Venkatesh Karunakaran**
Title of the thesis: **Automatic Relation Extraction from Clinical Documents: A Study of Fine-Tuned Transformer Models and LLMs**

The thesis of Goutham Venkatesh Karunakaran deals with information extraction from clinical documents. It focuses on extraction of entities and their relations from clinical narratives. The task is adopted from two shared tasks organized in 2023: TestLink (Multilingual relation extraction of clinical measurements in clinical narratives) organized within IberLEF 2023 (data in Spanish and Basque) and CLinkaRT (Linking a Lab Result to its Test Event in the Clinical Domain) organized within Evalita 2023 (data in Italian). The goal of the thesis is to implement strong baselines for the task and the three languages.

The main text of the thesis spans 41 pages and includes all sections typical for a research thesis: introduction (unnumbered section), overview of related work (Section 1), description of the dataset (Section 2) and methods (Section 3), experiments and results (Section 4), discussion (Section 5), and conclusion (unnumbered section). The text is accompanied with a list of references, figures, tables, and abbreviations, plus three appendices. In total, the thesis has 69 pages. The experiments include finetuning of three LMs (mBERT, XLM-RoBERTa, and BioBERT) in monolingual and multilingual settings, their evaluation on the three datasets plus prompting experiments with ChatGPT.

The text is written in English with only a few grammatical errors, it is understandable and well-structured. It must be stated that this is a resubmission of a previously submitted diploma thesis that has been improved. Most of the issues in the first version have been addressed and solved (or at least improved) regarding both the form and content. In addition to the monolingual models fine-tuned for both the tasks, the author trained multilingual models and presented that in most of the cases the multilingual models perform better. In total, for each language and each task, the author compared 9 models (3 monolingual ones, 3 multilingual ones, and 3 from the ChatGPT family). All the experiments are presented with more detailed error analysis of the results (including the prompting experiments). Other improvements are also visible, including e.g. hyperparameter tuning, evaluation metrics definition, and minor improvements in literature overview and model architecture description. Contrary, some of the weaknesses are still present, e.g. regarding the fact that the work was done in the context of two shared tasks but there is no discussion of the results in that context, no comparison of the author's results with those achieved in the shared task etc. Also, no confidence estimation of the results was done (the experiments were probably executed only once without averaging the resulting scores) and it is difficult to interpret them.

To conclude, the goal of the thesis was achieved, the thesis was significantly improved since the previous submission and I recommend it for a defense.

Pavel Pecina
Prague, Sept 3, 2024