Supervisor’s review of master thesis

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

Author of the thesis: **Amrita Harikrishnan Nair**
Title of the thesis: **Unsupervised Open Information Extraction with Large Language Models**

The diploma thesis by Amrita Harikrishnan Nair tackles the problem of Open Information Extraction. It is the task of generating a structured, machine-readable representation of the information in (unstructured) text, usually in the form of triples: <subject, predicate, object>. The task is traditionally solved in two steps: 1) generating a pool of triples from a sentence, 2) ranking the triplets according to their quality. The presented work builds on such a two-step solution called DeepEx that generates the triples using the attention mechanism and ranks them using a supervised contrastive ranking model. The aim of the thesis replaces the contrastive ranking model with a lighter model that ranks the triples based on their linguistic acceptability in an unsupervised way (compared to DeepEx). The detailed specific objectives are listed in Section 3.2 and include evaluation of differ models on several benchmarks (NYT, OIE20016, PENN, WEB, and CaRB).

The main text of the thesis spans 47 pages structured into Introduction, 4 numbered chapters, and Conclusions. The text is accompanied with a rich Bibliography (6 pages), a List of Figures and List of Tables and an Appendix with additional tables and figures. Chapter 1 presents a theoretical background of the field, Chapter 2 presents the DeepEx system and proposes the idea of replacing the triple ranking step of DeepEx by a linguistic acceptability model. Chapter 3 describes the methodology of this work including the dataset and their analysis and the setup of the experiments. Chapter 4 presents the results together with their detailed analysis and thorough comparison. The work is concluded in the Conclusion chapter.

The thesis is written in English without obvious grammatical or stylistic errors. The text is well structured, very readable and understandable, with a lot of tables and nicely presented figures. The author provides a very smooth introduction to the problem, explains the idea of using linguistic acceptability models for scoring (and ranking) the triples, proposes the solution, describes the datasets with lots of details, and present the extensive experimental part of the work – again well structured, nicely planned, and very well analyzed. The results are positive. In most cases, the proposed models outperformed the DeepEx baseline. I also appreciate the two additional studies presented in Section 4.2, error analysis in Section 4.3 and limitations in Section 4.5 (the proposed approach can work with SVO languages only).

To conclude, I really enjoyed reading the thesis. It tackled the problem very well. The goal was completed in full, the baseline outperformed, and a lot of interesting results and findings were achieved. I recommend the thesis for a defense.

Pavel Pecina
Prague, Sept 3, 2024