Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Irina Nikishina

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Learning linguistic tree structures with text and graph methods

Supervisor: Assistant Professor Alexander Panchenko

Name of the Reviewer:

I confirm the absence of any conflict of interest Ivan Oseledets	**Date: 15-09-2022**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense. *If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.*
Reviewer's Report

Reviewers report should contain the following items:

The thesis of Irina Nikishina has 159 pages, has an introduction ,related work section, 6 main chapters and conclusion. It is devoted to a very hot topic in NLP, namely - combining text and graph structure of the dataset.

Comments:

General: Linguistic terms are used without introduction. The abstract immediately starts from a «hypernym», but before reading the thesis I did not know what this word means. Googling it gives a good explanation of that as a general concept for a word (i.e. color is a hypernym for red). Would be nice to have such description in the text to help the reader.

Page 39: I would not agree that transformers are better parallelizable than RNN, they also process data sequentially (especially, GPT-type models).

Page 45: word2vec is mentioned after transformer-based models, however it is historically older (as well as FastText). Not clear, why in the thesis it is mentioned after transformer-based models.

Page 51: the review misses Ivanov & Burnaev paper on anonymous walk embeddings (ICML 2018).

Overall, Chapter 2 is a solid review work.

Chapter 3 covers the dataset creating but from reading it it is not clear to me what is the actual challenge in creating such dataset. Once the definition of a hypernym is clear, the procedure is quite straightforward. A separate «algorithm» environment for the collection of the dataset will be useful. The Chapter also mentions another word «synset» which is also not easy to understand and also not clear why we can be interested in it. To sum up, the chapter is quite short, it is important but it is not clear what is the main challenge solved in this chapter (and should it be even be a chapter?). The next chapter describes the baseline, but again, it could be a single part of the thesis, it might be more logical.

Chapter 4:

I have never seen the description of the competition and their solutions in a thesis. Maybe it is ok for a field, but the thesis should contain original contributions of the authors. Maybe the right place for this is the Appendix.

Then, Chapter 5 has the method contribution, the DWRank. The text says that it builds upon the baseline, so it would be better to put this in one place, rather than saying it is a baseline and reporting many solution of other people. The beginning of Chapter 5 just gives the solution: these are the features. However, for me the intuition «why» these features are selected. Obviously, it is not the first attempt, and ideally ablation studies are required in order to see if a particular innovation in the feature engineering really help with the improvement of the scores.

Actually, one of the nice ideas from the viewpoint of the methods (dealing with unseen words by averaging fastText embeddings), is hidden without highlighting in Section 5.2.2. Many questions: do you use fastText embeddings? How you put fastText and node2vec into the same space?

Not clear, and highlighting your ideas is as equally important as carefully reviewing work of the others. Section 5.2.4. seems to repeat the material in the Related Work section, as well as 5.2.5, you don't need to do it! Just tell your results.

Again, I am missing a formal algorithm description (there is a nice algorithm environment in Latex) for the two proposed methods.

Chapter 6 shows that the methods proposed in the thesis outperform others. But since the feature engineering is rather complicated, ideally, ablation study is needed.

Chapter 7 again has related work (no need! Even in the papers it is better not to start with related work).

Figure 7-2 makes an attempt to describe the algorithm, but still it is not easy to decipher the algorithm from it. In the end, it seems that there is a mapping from Graph-Bert embeddings to Bert embeddings, used as an MLP (probably?). If you look next, the details come at Section 7.4.2. where it is more algorithmic and mathematical.

Table 7.2 and 7.3 are really amazing (beats baselines by a large margin).

To summarize, algorithmically, the most impactful contribution is the joint Graph/NLP model which beats the baselines, and also the development of DWRank models.

There is obviously a big work put by the author into the current thesis. Irina really tried to write down very accurately, in the book way, the description of the state of the art and different models. What could have been improved, is highlighting of the main results and the contributions. In the DWRank models, the origin of the feature engineering are not very clear, and some insights on «why» would be very helpful.

The comments above do not influence the final evaluation of the thesis, and I think it could be accepted.

Provisional Recommendation

I recommend that the candidate should defend the thesis by means of a formal thesis defense

I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report

The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense