Review of Ph.D. Thesis Proposal

Reviewer: Pavel Ircing (KKY FAV ZČU)

Candidate: Ondřej Plátek (UFAL MFF UK)

Thesis title: *Evaluation Metrics for NLG and TTS in Task-Oriented Dialog*

Supervisor: Ondřej Dušek

The presented proposal focuses on the automatic methods for evaluation of Task-oriented Dialogue (ToD) and Text-to-Speech Synthesis (TTS). The aim of the author is to design and test evaluation methods that would – most importantly - be highly correlated with human evaluation of the mentioned NLP tasks and also consistent, reproducible and general (i.e. readily usable across different domains). The fact that those method will be faster and cheaper than human evaluators comes without saying but author's ambition is to design methods that would also be as fast and computationally efficient as possible.

The author is aware of the fact that non-trainable evaluation metrics like BLEU typically do not satisfy his most important condition – high correlation with human judgement – and therefore concentrates solely on trainable metrics, more concretely those based in neural networks. However, as he also mentions immediately, this choice raises the issue of problematic interpretability of such NN-based metric that needs to be addressed as well.

I see the selected topic as extremely important and, at the same time, rather underrepresented in the research work, as various sophisticated chatbots (i.e. the systems engaging with users in an open-ended dialogue, as the author of this proposal puts it) are essentially the flagships of current NLP research and yet the evaluation of their output is often mostly anecdotal. I am therefore quite sure that the subject definitely deserves a Ph.D. dissertation thesis.

The submitted proposal with extensive list of references shows that the author is sufficiently acquainted with state-of-the-art in the field and his choice of the methods and approaches that he plans to implement and evaluate is well-grounded. In fact, he and his colleagues has already designed several trainable metric that were successfully tested in the VoiceMOS challenge (evaluation of TTS) and the DSTC11 Track4 (evaluation of open-ended dialogue).

The proposal is well-structured and clearly written, even slightly more extensive that is (I believe) usual for this type of report at the institute. The fact that the goals of the dissertation project are divided into two sections named "Immediate Plans" and "Long-term Projects" further strengthens my belief that the Ph.D. candidate has already outlined a clear roadmap his final thesis.

The author also has a decent publication record that is, however, rather sparse between 2017 and 2023. I suppose that this can be attributed to the shift of his research focus from developing spoken dialogue systems to the evaluation of their outputs and I believe that a substantial surge of

publication activity that occurred this year also indicates that the direction of the research that the candidate is pursuing is very promising.

To conclude, I declare that the submitted proposal presents a clear research plan that is very likely to lead to a successfully completed Ph.D. thesis over the horizon of the next couple of years.

I have one final question/comment:

In the *Abstract*, the author states that "*Additionally, we address the broader concern of the lack of interpretability in neural network metrics.*". However, I did not find this concern really addressed in the proposal. Could you please provide a more detailed comment on this issue?

Štěnovice, 3.9.2023 Pavel Ircing