Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Aleksandr Rubashevskii

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Efficient active learning and fact-checking via predictive uncertainty quantification for natural language processing

Supervisor: Assistant Professor Maxim Panov

Name of the Reviewer: Assistant Professor Alexander Korotin

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 18.11.2024**

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:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications
- The summary of issues to be addressed before/during the thesis defense

Review of PhD Thesis by Aleksandr Rubashevskii

Brief overview

This thesis addresses the important problem of reliability of language models in modern applications by means of uncertainty quantification (UQ). A significant feature of uncertainty quantification is the wide class of problems to be solved and the lack of unified common standards for the methodology used.

The thesis presents efficient uncertainty estimation methods for 2 scenarios which widely appear in language models, namely active learning and fact-checking. The thesis is based on 4 papers, two of which are related to the active learning task and two of which are related to the fact-checking task. All papers have been published at well-reputed conferences: one of them at the CORE A (SIAM SDM) conference and three of them at the CORE A* conferences (ACL, EMNLP).

The content of the work consists of an introduction (section 1), related works (section 2), 3 chapters on selected published works (sections 3-5), and a conclusion (section 6). Section 3 presents a new UQ approach for active learning, while sections 4 and 5 are related to the fast checking by means of UQ. In particular, section 5 proposes a benchmark for the task. Overall, the structure of the thesis is reasonable and the actual content of the thesis is fully reflected in its title. The dissertation is logically connected and represents a coherent single work. The novelty, quality and level of the work is also good: the author carefully analyzes the designated tasks, provides justification of the methodology used, and demonstrates its effectiveness in comparison with state-of-the-art approaches. The obtained results are relevant to the application, which is confirmed by the provided experimental study.

Detailed overview

The main idea of **section 3** is to propose a new batch active learning algorithm which alleviates the limitations of previously known algorithm BatchBALD such as the computationally hard estimation of criterion. The proposed Large BatchBALD (LBB) performs an approximation of the BatchBALD method, using the BALD acquisition function and pairwise mutual information of model output components. Formally speaking, the author proposes to consider a Taylor series-like decomposition of the true criterion into many tuple-wise mutual information terms “I(yi,yj,yk,...)” and then gets rid of high-order terms leaving only the second order terms “I(yi,yj)” depending on two variables. This approach makes it possible to approximate the BatchBALD criterion up to the second order terms. The approximation is very smart and, as the author demonstrates, good performance in practice compared to baselines.

While the idea presented in Section 3 is interesting and significant, the clarity of the presentation of the idea and of the results of the experimental evaluation may be improved. This part of the thesis is the most mathematically heavy but lacks sufficient mathematical rigor. Specifically, it looks like in most of the cases the probability, entropy & information terms of label “y” should be conditioned on the given input “x”. While this is present in some cases, e.g., equation 3.8 with the entropy, this notation disappears later without any comments from the authors (and is not present in expressions after 3.5). It is still possible for a reader to understand the message and the proof, but some places are a little bit vague and may require additional clarification. In equation 3.5, KL is introduced between continuous random variables using the integral notation, but label y is always discrete, so this notation is slightly misleading. In the experimental subsection, there are many plots and tables with the results on various

setups, however, it is hard to parse them for a reader as there is no clear emphasis on the proposed method's results (something like adding "ours" to results of the proposed method could have been considered, especially taking into account that it is).

In **Section 4**, the author presented a novel approach to fact-checking and hallucination detection based on token-level uncertainty quantification. The main design is to design a formula (CCP: claim conditioned probability, equation 4.5) for uncertainty based on the meaning and claim type of the particular token in the claim and on the previous tokens in the sentence. Overall, the idea of the section is very intuitive, reasonable and clear, although some aspects could have been clarified more formally, e.g., what the "meaning" is and what it the "claim type", as well as why assumption that the meaning uniquely defines the claim type is valid. In practice, the implementation of CCP requires using a NLI (Natural language inference) model which may make the results dependent on its properties. However, the experimental evaluation is very thorough and, in particular, studies different NLI models as a part of the ablation studies. The results demonstrate that the proposed CCP exhibits minimal dependence on the NLI model (several NLI models from Hugging Faces are tested). To conclude, the experimental section demonstrates that the results of the proposed CCP are superior to existing approaches.

Section 5 is devoted to designing Fact Check-Bench: a Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. I think this section plays the role of a nice addition to the thesis as designing new benchmarks is of high importance for the transparency and reproducibility of machine learning research. At the same time, it would be nice to further establish a link between the benchmark and the previous methodological sections 3, 4. For example, adding an explanation which steps could be used to exploit this benchmark to test the method from section 4 could be beneficial.

Questions

1) While the approach to estimation of the multivariate information in LargeBatchBALD uses the approximation of the true value by using only pairwise information terms. How well does this allow us to approximate the actual value of interest? Could some theoretical bound be established or an empirical study of this approximation error conducted?
2) One of the assumptions in Section 4.1.2 that the meaning of the sentence deterministically defines the claim type. Could it be clarified how reasonable this assumption is, i.e., what are the cases/counterexamples when this does not hold and how frequent are they in practice?
3) It is claimed in Section 4.1.1 that existing methods for UQ based on deep ensembles provide a significant computational overhead. How does the proposed method compare to them in terms of computational time?
4) The approach of Section 3 is based on the (Batch)BALD which, in turn, exploits the entropy of the distribution. Are there any analogs of BALD with other notions of dispersity of the distribution instead of the entropy (e.g., something like generalized entropy or variance of the probabilities), and could the proposed LargeBatchBALD generalize to such cases?

Minor comments

1) Some of the contributions (section 1.3) could probably be merged. For example, point 1 and 2 look like two parts of the same contribution (new method + evaluation).
2) In Section 3.1.2 it would be nice to give an additional comment about the intuition of I_BALD and explain why it is always non-negative.

3) Subsection “BatchBALD” could benefit from adding explanation how this existing criterion is evaluated/approximated in practice by its authors (or just adding the reference to 3.1.4).
4) For completeness, it would be nice if the authors include the formal definition of the entropy as it appears many times throughout the thesis (for example, section 3 and 4).

Summary

The thesis is a solid piece of work addressing several important problems in language models using uncertainty quantification (UQ). The ideas presented here are novel, generally well described and supported by an extensive experimental evaluation and comparisons with existing alternatives. The results of the thesis are published in the high-ranking relevant venues. It is worth noting that some UQ-related ideas presented here are not just limited to language domains and may be used for general machine learning problems (e.g., see Section 3.2.2) which adds additional value to the results.

The comments and questions posted above serve as recommendations for the thesis improvement rather than the critical issues and they do not detract from the overall positive assessment of the work. Therefore, I think the submitted work satisfies the requirements for the thesis defense and Aleksandr Rubashevskii deserves the PhD degree.

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*