Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

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: Asst. Prof. Maxim Panov

The thesis document includes the following changes in answer to the external review process.

0)Modified a proceedings section

Ekaterina Artemova
1) A "Limitations" section has been added after the "Future Work" section (Chapter 6).
2) Added a contribution note at the end of section 1.4. Thesis outline.
3) Both papers on active learning have received some citations. One of the citations of the Large BatchBALD method is made by a laboratory from Oxford University, which independently obtained similar approximation results for the BatchBALD method. This further confirms the relevance and significance of our results. The ALToolBox paper received several citations with suggestions for extending and improving the proposed framework, which also shows the interest in the published work.

Mikhail Burtsev
1) The captions to Figure 3-1 have been modified accordingly, and information about the error bars in Section 3.2.1 has been added. Results are presented as filled error bars representing the average of 5 runs of the same experiment with different seeds with the corresponding standard deviation (tabular results are obtained in the same way). In active learning, random acquisition function is often considered as a competitive baseline method. From Figure 3-1, we conclude that our PLBB method is the leader in the presented figure, while LBB is comparable to random sampling and outperforms two other baselines, namely BALD and BatchBALD.
2) The corresponding information about table results is added in Section 3.2.1.
3) In the limitations section at the end of the thesis, we have added information that for active learning, training on each data volume is done from scratch for both image and text data. As for fine-tuning the model at each step, there are a number of works that address this task (active fine-tuning), but it is beyond the scope of this work.
4) The parts of the pipeline for extracting and matching claims for all data and all methods are performed identically using GPT-4, and the inference time is highly dependent on the amount of text generated, the number of claims extracted, and the size of the claims. A full evaluation of these parts also requires regeneration of all data. But if we roughly estimate the time on a few generated biographies, the operation of splitting a single sentence into atomic claims and mapping a single atomic claim back to sentence requires on average a 3-4 seconds of inference time each.

Alexander Korotin
Comments from the text
1) " … in most of the cases the probability, entropy & information terms of label "y" should be conditioned on the given input "x"…"

In the text, we have omitted the conditioning of mutual information I on x and D_train for convenience of notation and similarly to Kirsch et al. BatchBALD. 2019. We have added this clarification in Section 3.1.2 BALD.
2) "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)."
The tag "ours" has been added to our methods in tables and figure captions.
3) "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."
Active learning methodology (Section 3) is not applicable to our benchmark since the dataset is built and labeled for an already trained ChatGPT model and the issue of effective training is not relevant for this model. The white-box uncertainty quantification methodology (Section 4) can be applied to this benchmark to detect annotated hallucinations, but this would require a) regenerating all ChatGPT responses to obtain the corresponding generated logits, b) since the model responses are subject to randomness, it would require a complete reannotation of the generated responses.

Questions
1) Approximation by mutual information triples has a cubic complexity with respect to the size of the pool of unlabeled data. Even with a more accurate approximation, such complexity is not feasible for a large amount of unlabeled data, so it was decided to focus on approximation using pairwise mutual information.
2) The assumption that a sentence unambiguously defines a claim type follows from the assumption that in a word distribution, each word refers to a single claim type. Otherwise, the word in a phrase must have more than one meaning, that the statement formed with it refers to more than one type of claim. Since the same word cannot have opposite meanings to each other, sentences with this exact property do not exist. It seems possible to construct a sentence with multiple meanings where one meaning is neutral to another because of the polysemous nature of the keyword, but I assume that in the case of biographies this is a negligible number of cases.
3) For ensembles, model inference must be done 5-10 times (once for each model in the ensemble), which gives a large overhead multiple of the number of runs.
4) Related to this topic is the recent work of Kotelevskii & Panov. Predictive Uncertainty Quantification via Risk Decompositions for Strictly Proper Scoring Rules. (2024). The authors showed that BALD is just a special Bayesian approximation of excess risk (epistemic uncertainty) with a special loss substitution (proper scoring rule with logscore). You can use any other proper scoring rule and get different metrics.

Minor comments
1) Dividing the active learning contribution into two points we would like to emphasize that have been 2 separate significant contributions related to the same topic, namely: intuition and theoretical derivation of the new method, as well as extensive experimental study for several domains.
2) Added in Section 3.1.2 the relevant intuition underlying BALD and an explanation of why BALD scores are non-negative.
3) Corresponding citation is added to Section 3.1.4 BatchBALD.
4) The expression for the entropy acquisition function is presented in equation 3.4

Ivan Oseledets
1) Added a contribution note at the end of 1.4. Thesis Outline.
2) The conclusions in each section have been modified accordingly.
3) Added a description of the NLI term in the Implementation Section of 4.1.2.
4) The results presented in Chapter 4 refer to different models such as Mistral 7b, Vicuna 13b, Jais 13b, GPT-3.5-turbo. All these models differ in size, performance, and tendency to hallucinate. We showed that for different models, our CCP uncertainty estimation method is a quality leader. We believe that for newer models, the conclusions will remain similar because they are based on the nature of the uncertainty quantification rather than the architectural features of the model. More detailed conclusions require a separate experimental study, which we leave as future work.
5) All annotations of ChatGPT responses were done manually. Moreover, all annotation steps were performed according to an internal protocol, and each annotation fragment was checked by at least two people associated with the NLP domain. The manual annotation results obtained at each stage are considered as the gold standard.