Berk Atil, CSE584 HW1

PALS: Personalized Active Learning for Subjective Tasks in NLP (Kanclerz et al., 2023)

Data quality is essential for machine learning models, and we usually need labeled data. Deep Learning approaches require large annotated datasets for supervised tasks, but data annotation is expensive. There are task-focused active learning methods to reduce the amount of supervision required for a good performance. However, this is not enough for subjective tasks such as hate speech detection, that allow different decisions for different users. Therefore, the authors propose a new approach to active learning in NLP for subjective problems.

They propose a personalized active learning technique that takes both the individual's perspective and task into account. They propose 5 metrics to select a text for a given user and task to be annotated. Those metrics consider the ambiguity of the text, subjective preferences of the users, the uncertainty of the label of a document, etc. They select the texts with higher metric values since they are supposed to be more ambiguous, controversial, and confusing. They use each metric separately, so they can be considered as different strategies. Based on their experiments, the "controversy" metric which is entropy-based is the best-performing one.

They point out an important problem about data annotation processes for subjective tasks to reduce the cost and time. To address this, they propose five metrics that can be used within the active learning framework.

One of the limitations of this paper is that they did not finetune the language model with new annotations for any downstream tasks. They use the language model only as a feature extractor and use a linear classifier that uses textual features and learns the biases of annotators and words (This model was proposed by others). This approach does not align with the current techniques, where we directly use the language model as a classifier.

Causal-Guided Active Learning for Debiasing Large Language Models (Du et al., 2024)

Despite the impressive performance of LLMs, they have some biases such as positional or stereotyping. This is because, they learn the correlations in the training data, which has biases. These biases might result in poor generalization or the possibility of harmful content generation. It is important to debias models to have better LLMs. Some previous methods rely on prior knowledge to recognize biases and eliminate them through aligning or prompt-based regularization. However, because of the diversity of biases, it is not feasible to find biases manually. Therefore, automated identification methods are needed. Previous automated methods rely on finetuning and are designed mostly for discriminative models, not generative ones.

LLMs predict subsequent text Y given text X. Ideally, they should only take the semantic relationship between X and Y into account. However, because of the biases in the training corpus, they also model the biased relationship such as negation patterns which are correlative, not causal. In this work, authors propose a causal invariance-based biased instance identification, most informative biased example selection, and bias pattern induction which is an active learning method.

To find biased instances, they assume that if the hidden representations of two texts and the similarity of subsequent text Ys are similar, then the model captures the predictive casual relationship, not the biased information. However, if the hidden representations are similar but not the subsequent text Ys, then the model might have captured biased information. Based on this idea, they find biased examples by applying thresholds for similarities. Afterward, to select more informative samples, they have two criteria: the first is having very similar subsequent text Ys which indicates a typical bias, and the second is having less similar Ys but having a low probability of producing the gold text. The idea behind the second one is that biased information hinders the LLM. To induce bias patterns, they cluster selected biased examples and prompt GPT4 to summarize bias patterns for each cluster.

To debias LLMS, they propose prompt-based approaches. For the 0-shot case, they provide the induced bias information to the model so that it should not use that information. For the few-shot case, they provide counterfactual examples where biased information would lead to wrong predictions. To find those examples, they use informative examples for which producing the gold text is unlikely.

I think the problem statement here is important and being worked on a lot in ML literature. They use active learning to find biased examples and suggest prompt-based techniques to reduce the bias of the LLM.

One concern that I have is that although they show that their method improves performance, they do not debias the LLM, they just hide those biases via prompting, but the biases still exist in the model. Second, they have a few threshold values that are hyperparameters and they do not mention exact values. So there is a doubt if we need different thresholds for different datasets. If this is the case, this reduces the generality of the approach. The last limitation of the work is it is needed to access the hidden representations and probability of producing gold text, which may not be always possible.

Active Learning Principles for In-Context Learning with Large Language Models (Margatina et al., 2023)

In-context learning does not have the drawbacks of supervised learning, such as need for large labelled data or finetuning for each new task. However, it also requires few labelled data. This has two issues: the first one is what data to sample which would be the most appropriate, and if there is unlabeled data which ones to label. Although there are some works around this, authors explore the potential of active learning, which helps identify the most informative instances from unlabeled data to annotate.

They apply different active learning techniques for a single iteration to choose the most appropriate labeled examples for in-context learning. They experiment with uncertainty, diversity and similarity sampling. In diversity sampling, they cluster data points into k clusters and choose 1 example from each cluster. For the uncertainty sampling, they calculator the perplexity of each candidate in-context example. The idea is that a high perplexity set of examples should yield better performance than random sampling. Last but not least, similarity based sampling uses k-nearest neighbors algorithm to find the most similar in-context examples for a given input.

They basically test 3 different active learning approach for in-context example selection extensively, including 15 models on 24 tasks. They conclude that similarity-based

sampling is the most efficient one, although uncertainty sampling is one of the best active learning approaches in supervised learning. They also do some analysis on the effect of model size and the importance of the trustfulness of in-context example labels.

I think the work is good in the sense that they do an extensive analysis on the effects of few-shot example selection. However, they do not propose any new active learning method and as far as I know there are similar works that suggest using similarity-based sampling strategies in the literature. But, it is still good to have these kinds of through analysis papers which can help a lot to other researchers.

References

- Kanclerz, K., Karanowski, K., Bielaniewicz, J., Gruza, M., Miłkowski, P., Kocoń, J., & Kazienko, P. (2023, December). PALS: Personalized Active Learning for Subjective Tasks in NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13326-13341).
- Zhouhao Sun, Li Du, Xiao Ding, Yixuan Ma, Yang Zhao, Kaitao Qiu, Ting Liu, and Bing Qin. 2024. Causal-Guided Active Learning for Debiasing Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14455–14469, Bangkok, Thailand. Association for Computational Linguistics.
- Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active Learning Principles for In-Context Learning with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.