INF 381 Report

Shih-Chieh Dai
School of Information, University of Texas at Austin
sjdai@utexas.edu

1 Introduction

In this semester, I joined the project, Detecting Political Ideological Leaning. This project aims to design a human-in-the-loop classification using active learning and rationales. The concept here is to build a framework that predicts the political ideology leans by the rationale provided by humans or extracted by the machine learning model.

We explore rationale on the benchmark **ERASER** (DeYoung et al., 2019). ERASER is a benchmark for the rationale task. The authors released seven datasets with labeled rationale sentences: Evidence inference, BoolQ, Movie Reviews, FEVER, MultiRC, CoS-E, and e-SNLI. The authors of ERASER also proposed a simple pipeline model, Bert2Bert. They train the encoder to extract the rationale and then train the decoder to make the prediction using extracted rationale sentences.

We first explore rationale with active learning on Movie Reviews dataset (Pang and Lee, 2004). Movie Reviews is a dataset for the sentiment analysis task. It contains 1000 positive movie reviews and 1000 negative movie reviews.

My parts in this project are running the ERASER benchmark Bert2Bert model, trying active learning with uncertainty-based query strategy and random sampling [1], and implementing the random sampling baseline using Hugging Face [2].

2 Backgrounds

We can divide this project into three main components: active learning, rationale, and machine learning model.

[1] https://github.com/sjdai/AL-project
[2] https://huggingface.co

2.1 Active Learning

Active learning is a machine learning strategy that starts with a certain amount of training data and adds more labeled data in each iteration. This kind of learning strategy can deal with the problem of insufficient data. The key point of active learning is how to select the added labeled data in each iteration, which calls query strategy. There are a lot of query strategies have been proposed. We currently focused on uncertainty-based, confidence-based, and random sampling in this project.

Uncertainty-based

Following by (Schröder et al., 2021b), we tried two uncertainty-based query strategies in this project. The first one is Predict Entropy (PE). PE is a query strategy that, with the goal of reducing total entropy, picks instances in the predicted label distribution with the highest entropy. The second one is Breaking Ties (BT), selected examples with the smallest difference in confidence scores between the top two predicted classes. Notice that BT and PE are equivalent when the classification is a binary task.

Confidence-based

For confidence-based query strategy, we follow the work Cartography (Swayamdipta et al., 2020). They define confidence as the mean model probability of the true label (y_i^*) across epochs:

$$\hat{\mu}_i = \frac{1}{E} \sum_{e=1}^{E} p_{\theta^{(e)}} \left(y_i^* \mid x_i \right)$$

Where $p_{\theta^{(e)}}$ denotes the model's probability with parameter θ at the end of each epoch. Note that $\hat{\mu}_i$ is with respect to the true label y_i^*, not the probability assigned to the model's highest-scoring label.

Random Sampling Random sampling (RS) is a common baseline in machine learning. The idea is just randomly select labeled data in each iteration and add the labeled data to retrain the model.

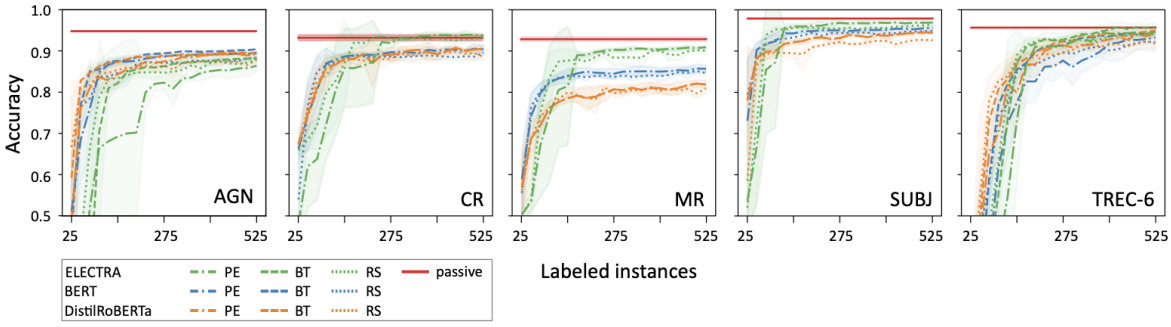

Figure 1: The result of active learning reported by (Schröder et al., 2021b)

2.2 Rationale

The rationale is the snippets that support outputs. There are two kinds of rationales: the rationale label is from humans or the rationale extracted by the machine learning model. For extractive rationale, we use BERT transformer to extract the rationale.

2.3 Machine learning model

In this project, we adopted Bert2Bert as the model for our task. Following the setting of the ERASER benchmark, We use uncased BERT in our task (Devlin et al., 2018). It is two-stage for this model. In the first phase, the BERT model would learn what sentence to select for making the prediction. Then, in the second phase, the input is the sentence chosen from the previous phase, and the model would learn how to predict the correct sentiment label.

3 Experiment

Currently, we are still in the pilot stage. We use the Movie Review dataset to explore our concept is works or not.

Implementation

We use off-shelf tool, small-text for active learning (Schröder et al., 2021a). small-text is a tool for active learning in Python [3]. The authors of small-text implement several query strategies, including uncertainty-based and random sampling. We tried two query strategies on Movie Reviews: uncertainty-based and random sampling. Notice that there are only two types of labels in Movie Reviews (i.e., Positive and Negative). As we mentioned in the previous section, when the dataset only contains two classes, both PE and BT query strategies would get the same result. Thus, we only tried PE for an uncertainty-based setting. For both uncertainty-based and random sampling, we train the model using 60 samples and then add 50 labeled samples after each iteration. We trained the model on TACC Frontera node rtx-dev. The node provides four NVIDIA Quadro RTX 5000 GPUs. It takes 1.5 hours to complete the whole training process.

[3]https://small-text.readthedocs.io/en/latest/

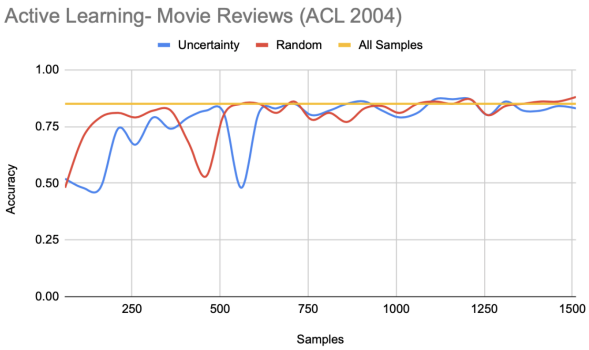

Figure 2: The result shows the performance of the classifier in each iteration. The yellow line indicates the accuracy (0.85) of the classifier trained by whole labeled data.

Result

Figure 2 shows the result of two query strategies on Movie Reviews. The uncertainty-based query strategy achieves the same accuracy score (0.85) as using whole labeled training data when using 710 labeled data. The highest accuracy score is 0.87 using 1110 labeled training data. Random sampling can achieve the same performance using 560 labeled samples, and the highest accuracy score is 0.88 using 1510 samples. From my perspective, the result is not expected. The performance trend does not have the same behavior regards to active learning.

For example, Figure 1 indicates the result of active learning from (Schröder et al., 2021b). When the labeled instances increase, the performance would improve steadily initially and then become

stable. However, our result did not have the same trend. Our result shows that the model can perform well when using 300-400 samples. However, the performance drops around 450 - 550 samples.

There are two possible explanations for the result. First, there are bugs in the toolkit we use. We implement the result using small-text. This toolkit is a new toolkit, and it may have some bugs. The solution would be implementing a random sampling baseline just using Hugging Face. The other possible explanation is the dataset is too easy. For example, (Schröder et al., 2021b) result shows that they only use 0.547% data to train an ELECTRA transformer model, and it can achieve a 0.909 accuracy score. To sum up, I am working on implementing a random sampling baseline using Hugging Face to see the result of active learning. Moreover, we are also trying to figure out what could be wrong in the package, small-text.

4 Conclusion

In this semester, we are still in the pilot stage. We are mainly exploring whether we can corporate rationale and active learning. One main point I learned from the experiment result is that when I use a new toolkit, I must check whether the result is correct. For example, I can implement the same model using other robust packages. We have tried any dataset related to political ideology this semester. I think the political lean is implicit in the context, and it makes it challenging for the machine learning model to figure it out. I believe it will have more interesting research questions to explore in the future. In my opinion, this project's key novelty is getting humans involved in the machine learning training process, which is human-in-the-loop. It is my first time working on this concept, and I think it is interesting to learn about it.

Acknowledgement

Thanks, Sooyong, for discussing issues we met in the project. Thanks to Anubrata, Venelin, and Prof. Matt's advice and suggestion this semester.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. ERASER: A benchmark to evaluate rationalized NLP models. *CoRR*, abs/1911.03429.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of ACL*, pages 271–278.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2021a. Small-text: Active learning for text classification in python. *CoRR*, abs/2107.10314.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2021b. Uncertainty-based query strategies for active learning with transformers. *CoRR*, abs/2107.05687.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *CoRR*, abs/2009.10795.