Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers

Findings of ACL 2022

Paper and Code
github.com/webis-de/ACL-22


Christopher Schröder

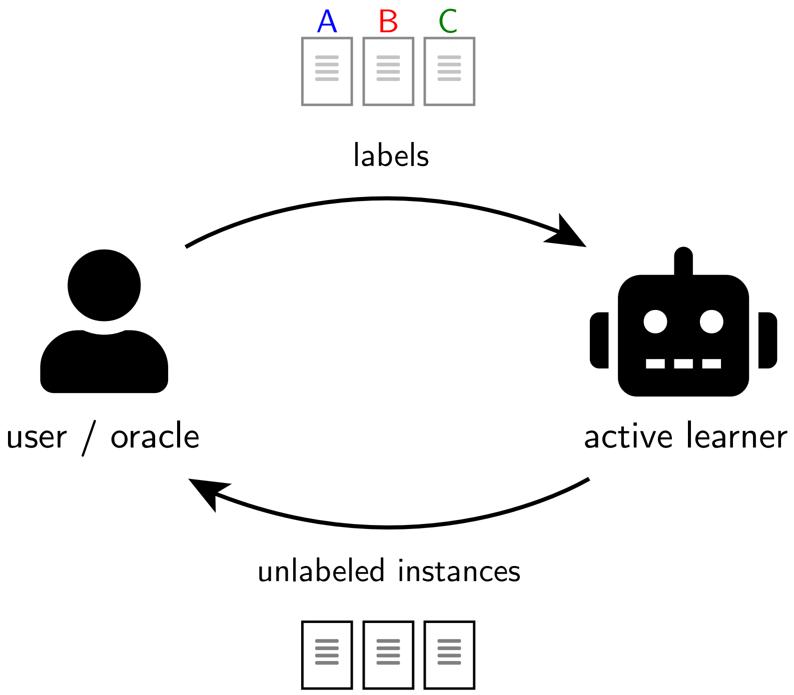

Andreas Niekler


Martin Potthast

WEBIS.DE
UNIVERSITÄT LEIPZIG

Introduction

Active Learning: minimize the labeling costs of training data acquisition while maximizing a model's performance (increase) with each newly labeled problem instance



This Paper

Motivation

- Research has started to investigate transformer models (“transformers”) for active learning but previous findings may not generalize to transformer models.
- Query strategies targeted at neural networks or text classification are computationally expensive.
- Uncertainty-based query strategies are (computationally inexpensive but) usually considered only as a baseline.

Contributions

- Systematic investigation of uncertainty-based query strategies paired with transformers.
- Evaluation on a five well-known lately neglected text classification benchmarks.
- We investigate the effectiveness of using a transformer model with fewer parameters, DistiRoBERTa, for active learning.

Experiment

Models: BERT [Devlin et al. 2019], DistilRoBERTA [Sanh et al. 2019] (and KimCNN [Kim 2014], SVM)

Query Strategies:

Prediction Entropy
[Roy and McCallum 2001; Schohn and Cohn 2000]

$$\underset{x_i}{\operatorname{argmax}} \left[-\sum_{j=1}^{c} P(y_i = j|x_i) \log P(y_i = j|x_i) \right]$$

Breaking Ties
[Scheffer et al. 2001; Luo et al. 2005]

$$\underset{x_i}{\operatorname{argmin}} \left[P(y_i = k_1^*|x_i) - P(y_i = k_2^*|x_i) \right]$$

Least Confidence
[Culotta and McCallum 2005]

$$\underset{x_i}{\operatorname{argmax}} \left[1 - P(y_i = k_1^*|x_i) \right]$$

Contrastive Active Learning
[Margatina et al. 2021]

$$\underset{x_i}{\operatorname{argmax}} \left[\frac{1}{m} \sum_{j=1}^{m} \mathsf{KL}(P(y_j|x_j^{knn}) \parallel P(y_i|x_i)) \right]$$

Random Sampling

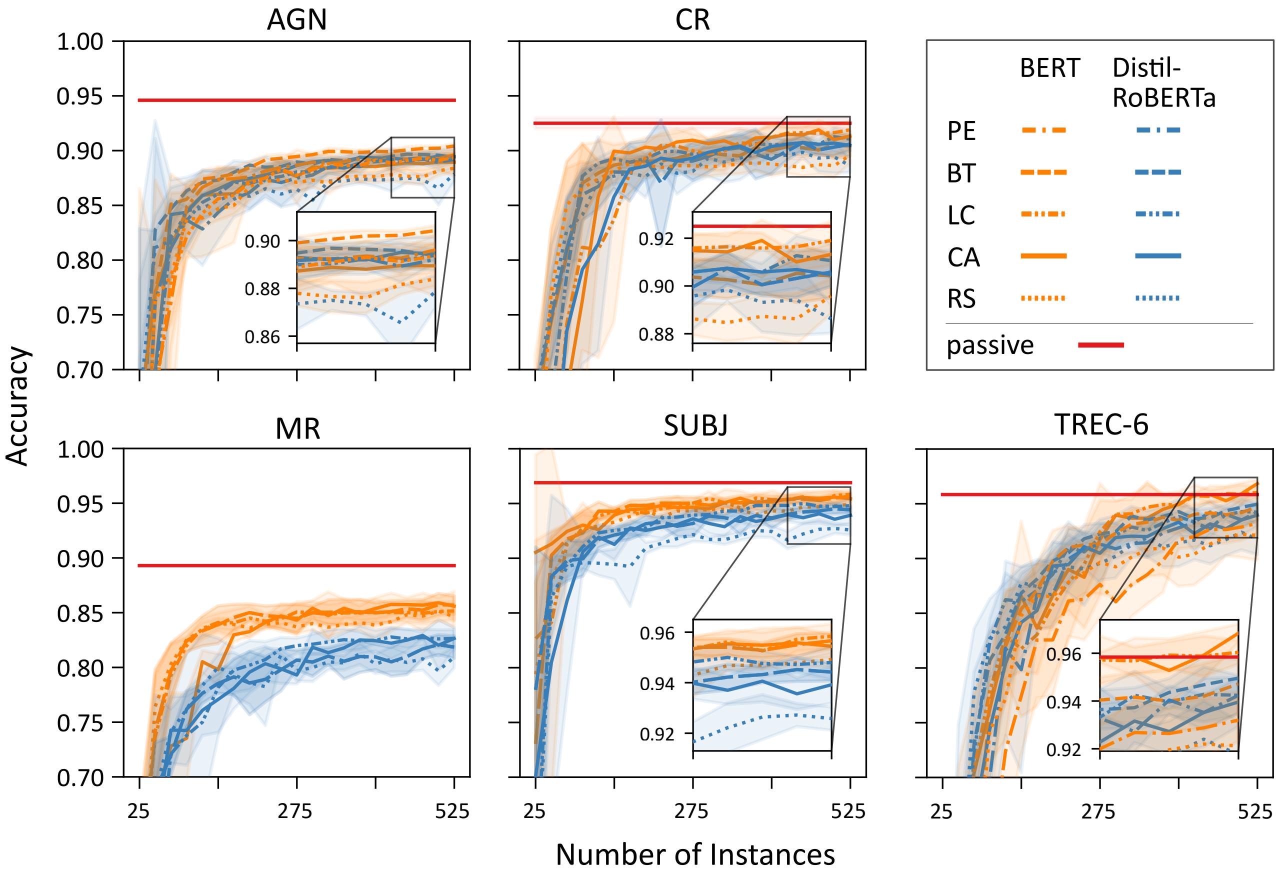
Sample i.i.d. from the unlabeled pool.

Experiment: Datasets

Dataset Name (ID)	**Type**	**Classes**	**Training**	**Test**
AG's News (AGN) [Zhang et al. 2015]	News	4	120,000	(*) 7,600
Customer Reviews (CR) [Hu and liu 2004]	Sentiment	2	3,397	378
Movie Reviews (MR) [Pang and Lee 2005]	Sentiment	2	9,596	1,066
Subjectivity (SUBJ) [Pang and Lee 2004]	Sentiment	2	9,000	1,000
TREC-6 (TREC-6) [Li and Roth 2002]	Questions	6	5,500	(*) 500

(*): Predefined test sets were available and adopted.

Evaluation: Learning Curves



Evaluation: Summary

Model	Strategy	Mean Rank Acc.	Mean Rank AUC	Mean Result Acc.	Mean Result AUC
SVM	PE	1.80	2.60	0.764	0.663
	BT	**1.60**	**1.60**	**0.767**	**0.697**
	LC	3.00	2.60	0.751	0.672
	CA	5.00	5.00	0.667	0.593
	RS	3.00	2.60	0.757	0.686
KimCNN	PE	1.60	2.40	**0.818**	0.742
	BT	**1.60**	**2.00**	**0.818**	**0.750**
	LC	3.80	2.80	0.810	0.732
	CA	3.80	4.80	0.793	0.711
	RS	3.60	2.40	0.804	0.749
D.RoBERTa	PE	2.60	3.00	0.901	0.856
	BT	2.20	**1.80**	0.902	**0.864**
	LC	**1.40**	2.00	**0.904**	0.860
	CA	3.00	3.40	0.901	0.852
	RS	5.00	4.20	0.884	0.853
BERT	PE	2.40	2.40	0.909	0.859
	BT	**2.00**	**1.60**	0.914	**0.873**
	LC	2.20	3.80	**0.917**	0.866
	CA	2.80	2.60	0.916	0.872
	RS	5.00	4.00	0.899	0.861

- Surprisingly: prediction entropy is outperformed by breaking ties.
- For DistilRoBERTa: least confidence also outperforms prediction entropy.
- DistilRoBERTa performs only slightly worse than BERT

Evaluation: Further Results

- Using transformer models we reach considerably higher AUC scores compared to Zhang et al. (2017).

- Active learning is very close (and even surpasses) previous state-of-the-art results, and our own passive classification, in terms of final accuracy (using a fraction of the data).

- Detailed results and runtimes per configuration are reported in the paper's appendix.

Conclusion

Experiment: Active Learning for Text Classification

- BERT, DistilRoBERTa
- Several sentence classification datasets
- Four query strategies and a baseline

Findings

- The supposedly strongest baseline, prediction entropy, “is not so strong”.
- Breaking ties consistently outperforms prediction entropy in multi-class scenarios.
- DistilRoBERTa achieves results close to BERT while using only about 25% of the parameters.

Conclusion

Experiment: Active Learning for Text Classification

- BERT, DistilRoBERTa
- Several sentence classification datasets
- Four query strategies and a baseline

Findings

- The supposedly strongest baseline, prediction entropy, “is not so strong”.
- Breaking ties consistently outperforms prediction entropy in multi-class scenarios.
- DistilRoBERTa achieves results close to BERT while using only about 25% of the parameters.

Thank you!

Bibliography

- Aron Culotta and Andrew McCallum. 2005.
  Reducing labeling effort for structured prediction tasks.
  In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*, pages 746–751.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019.
  BERT: Pre-training of deep bidirectional transformers for language understanding.
  In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186.

- Minqing Hu and Bing Liu. 2004.
  Mining and summarizing customer reviews.
  In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177.

Yoon Kim. 2014.
Convolutional neural networks for sentence classification.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Tong Luo, Kurt Kramer, Dmitry B. Goldgof, Lawrence O. Hall, Scott Samson, Andrew Remsen, and Thomas Hopkins. 2005.
Active learning to recognize multiple types of plankton.
Journal of Machine Learning Research (JMLR), 6:589–613.

Xin Li and Dan Roth. 2002.
Learning question classifiers.
In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021.
Active learning by acquiring contrastive examples.
In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 650–663.

Bo Pang and Lillian Lee. 2004.
A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.
In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 271–278.

Bo Pang and Lillian Lee. 2005.
Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.
In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 115–124.

Nicholas Roy and Andrew McCallum. 2001.
Toward optimal active learning through sampling estimation of error reduction.
In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 441–448.

Greg Schohn and David Cohn. 2000.
Less is more: Active learning with support vector machines.
In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 839–846.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019.
Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
arXiv preprint arXiv:1910.01108.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001.
Active hidden markov models for information extraction.
In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis (IDA)*, pages 309–318.

Ye Zhang, Matthew Lease, and Byron C. Wallace. 2017
Active discriminative text representation learning.
In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 3386–3392.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015.
Character-level convolutional networks for text classification.
In *Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS)*, pages 649–657.