Data is All You Need – Adapt Language Models to Target Domains for Text Classification with Data Augmentation.

Renxi Wang
Northeastern University, China
realreasonwang@gmail.com

Abstract

It has been a paradigm for current research on text classification tasks that a transformer-based model is first pre-trained on massive unlabeled data, then finetuned on the downstream task. A large number of studies have shown that the pre-trained language models (PLM) can well transfer the learned knowledge to downstream tasks. However, the scarcity of labeled downstream data makes the model hard to adapt to the task domain, which limits the model's performance. In this paper, we utilize adaptive pre-training (APT) and pseudo-labeling (PL) to make the model more adaptable to the task domain. The experimental results show that our methods improve the performance of PLMs with a variety of model structures and pre-training objectives, indicating the effectiveness and generalization ability of APT and PL.[1]

1 Introduction

Pre-trained language models (Qiu et al., 2020) obtain knowledge and store it in their parameters from pre-training on unlabeled data. Among pre-training objectives, masked language modeling (MLM) (Devlin et al., 2019) is the most common. In MLM, some words in a sentence are masked and the model learns to predict the masked words. Although PLMs have shown superior performance on downstream tasks, the rareness of the labeled data limited the PLM to transfer from the original domain to the target domain. To alleviate this issue, we mainly focus our methods on data augmentation.

In computer vision, data augmentation is usually achieved by distortion or rotation of the images. Similarly, in natural language processing, the typical data augmentation method replaces or deletes words in a sentence or changes the word order. Also, some special augmentation methods can be adopted, such as back translation. However, with the popularity of PLMs and the emergence of some strong baselines (e.g. DeBERTa (He et al., 2021b)), these methods are not always effective in practice. Simple data augmentation adds noise to the original data, which may cause the degradation of the model's effectiveness. In this paper, we adopt adaptive training method to make the model continually be pre-trained on the text of the same domain as the downstream task. We also generate pseudo labels with a finetuned model. We then re-incorporate the pseudo labels into the training set. The two methods can improve the performance of models with different pre-training objectives and different structures.

This paper focuses on the task of text classification. Instead of categorizing the whole sentence into some classes, our task is to classify each word in the sentence. Therefore, our task is more difficult and complex than simple text classification tasks.

2 Methods

2.1 Adaptive Pre-training (APT)

In APT, the PLM is further pre-trained on texts within the same domain of the downstream task. Additional knowledge related to the downstream task is injected into the model during adaptive pre-training. There are two kinds of APT: Task Adaptive Pre-training (TAPT) and Domain Adaptive Pre-training (DAPT) (Gururangan et al., 2020). The main difference between TAPT and DAPT is the pre-trained text of TAPT comes from the training set of downstream tasks, while the text of DAPT comes from the same domain as downstream tasks. In this paper we mainly use DAPT. The process of DAPT is shown in Figure 1. First, the PLM continues to be pre-trained on the same domain texts with MLM pretraining objective. Then the trained model is finetuned on the downstream task.

[1]The original report can be found here

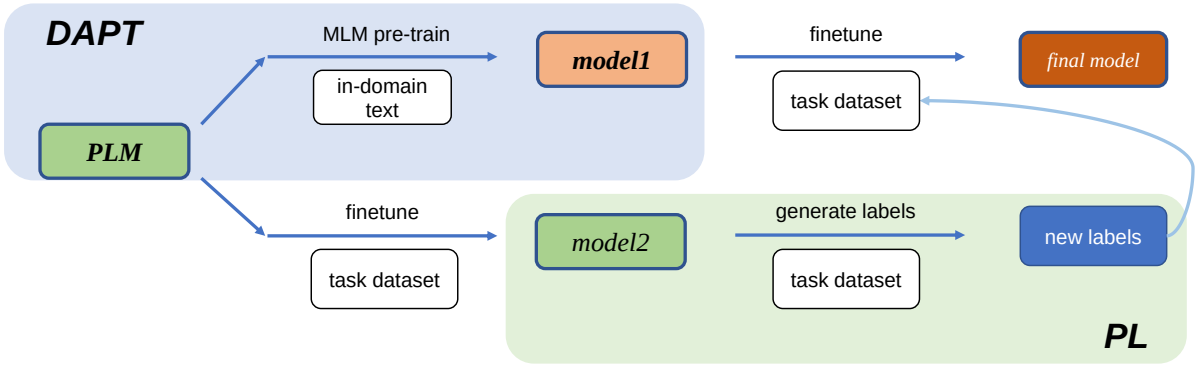

Figure 1: Domain Adaptive Pre-training and Pseudo-Labeling and the process to combine them together.

2.2 Pseudo Labeling (PL)

In PL, the PLM is first finetuned on the downstream task. Then we use the fine-tuned model to predict labels of the unlabeled texts. We take the results as labels for these texts. Finally, these labeled data are used as part of the training set. The process of PL is shown in Figure 1.

In our method, DAPT and PL are combined. First, the PLM is further pre-trained on texts within the same domain as downstream tasks. Then the model obtained from DAPT is fine-tuned on downstream tasks. Then, pseudo-labels are generated by the finetuned model and new training data are obtained. Finally, the new training set is used to fine-tune the model obtained by DAPT. The whole process is shown in Figure 1.

3 Experiments

3.1 Datasets

The dataset used in our experiment is from NBME-Score Clinical Patient Notes, a competition on the Kaggle platform. The dataset is provided by the National Board of Medical Examiners (NBME). Given a patient note and a feature text, the task is to identify the keywords corresponding to the feature text in the patient note. We transform it into a classification task for each word: a word is positive if the word is included in the feature text. Otherwise, it is negative. In addition to the annotated data, this dataset also provides 42,146 unannotated paragraphs on which our DAPT is trained.

3.2 Baselines

We choose DeBERTa as the baseline model. The DeBERTa model, proposed by Microsoft in 2021, used disentangled attention mechanisms and achieved human-level performance on the SuperGLUE benchmark. The latest DeBERTa-v3 (He et al., 2021a) is pre-trained with the ELECTRA-style training objective (Clark et al., 2020) and the performance is further improved. We also apply DAPT and PL on models with different structures and pre-training objectives.

3.3 Implementation

We use the F1 score to evaluate the performance of models as the competition. Since the classification is performed at word level, the calculation of the F1 score is also performed at word level. For DAPT, we train 10 epochs on the unlabeled 40k paragraphs, the learning rate is set to 2e-5 and linearly reduced to 0. The batch size is 32. After each epoch, we save the model and evaluate the Perplexity (PPL). We select the model with the lowest PPL and use them in finetuning. For fine-tuning, we split 20% of the training data as the validation set and train models for 5 epochs. The learning rate is set to 2e-5 and cosinely reduced to 1e-6. Finally, the model with the best performance on the validation set was selected as the final model. Since the test set of the data is not publicly available, we test all final models on the Kaggle platform and obtain the performance on the test set.

3.4 Results

The results are shown in Table 1. Most of the models are pre-trained using MLM objective. However, BART (Lewis et al., 2020) also uses token-deletion, text-infilling, sentence-permutation and document rotation, while DeBERTa-v3 uses ELECTRA-style pre-training objective. The PPL of models during DAPT is shown in Figure 2. Although BART and DeBERTa-v3 do not MLM in pre-training, DAPT can still decrease the PPL of these models. Most

Model	dev			test		
	vanilla	DAPT	PL	vanilla	DAPT	PL
RoBERTa (Liu et al., 2019)	82.07	85.16	84.15	81.73	84.26	83.71
BART (Lewis et al., 2020)	80.09	81.68	82.46	79.63	81.28	81.95
Longformer (Beltagy et al., 2020)	80.88	83.02	83.42	80.64	82.73	82.74
Big Bird RoBERTa (Zaheer et al., 2020)	77.67	82.49	80.06	76.91	82.58	79.62
DeBERTa (He et al., 2021b)	86.17	87.03	86.99	85.32	86.43	86.14
DeBERTa-v3 (He et al., 2021a)	85.70	86.24	86.68	84.96	85.87	86.09

Table 1: Performance of different models on the development set and test set. Vanilla means no DAPT or PL. All results are reported based on 5-fold cross-validation

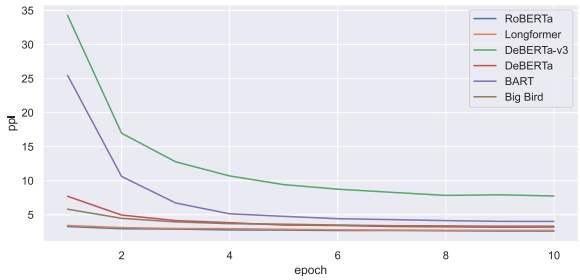

Figure 2: Perplexity of different models in each epoch with MLM pre-training.

Model	setting	dev	test
DeBERTa-v3-Large	w/ DAPT & PL	88.91	88.40
	w/ DAPT	88.67	88.08
	w/ PL	88.59	88.05
	-	88.31	87.30

Table 2: Ablation studies on development set and test set. "-" denotes no DAPT and PL.

of the models are implemented with the original Transformer structure. However, Longformer utilizes sliding window and global attention mechanisms. Big Bird employs the block sparse attention mechanism. And DeBERTa uses the disentangled attention mechanism. DAPT and PL greatly improve the performance of all models, indicating that these two methods are effective for different pre-training objectives and different model structures.

Ablation Study We use DeBERTa-v3-large as the baseline model and do ablation studies on it. The results are shown in Table 2. DAPT and PL perform similarly to each other. However, the results of DAPT and PL in Table 1 show that DAPT performs better for RoBERTa, Big Bird RoBERTa and DeBERTa, while PL performs better for BART, Longformer and DeBERTa-v3. More experiments may be needed to find which one is better.

4 Conclusion

We adopt domain adaptive pre-training and pseudo-labeling to make PLMs more adaptable to the domain of the downstream task. Experimental results show that our methods improve the performance of PLMs with different structures and pre-training objectives.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. CoRR, abs/2004.05150.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR.

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: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. CoRR, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced

bert with disentangled attention. In International Conference on Learning Representations.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63(10):1872–1897.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33.