KUL@SMM4H'22: Template Augmented Adaptive Pre-training for Tweet Classification

Sumam Francis and **Marie-Francine Moens**
KU Leuven, Belgium

Abstract

This paper describes models developed for the Social Media Mining for Health 2022 shared tasks. Our team participated in the first subtask that classifies tweets with Adverse Drug Effect mentions. Our best-performing model comprises of a template augmented task adaptive pre-training and further fine-tuning on target task data. Augmentation with random prompt templates increases the amount of task-specific data to generalize the pretrained language model to the target task domain. We explore 2 pre-training strategies: masked language modeling and simple contrastive pre-training and the impact of adding template augmentations with these pre-training strategies. Our system achieves an F1 score of 0.433 on the test set without using supplementary resources and medical dictionaries.

1 Introduction

The goal of the shared task as part of the Social Media Mining for Health (SMM4H) - 2022 (Weissenbacher et al., 2022) involves detecting tweets that have Adverse Drug Effect (ADE) mentions. Organizers of SMM4H Task 1 provided datasets of English tweets with annotations ADE and noADE indicating the presence or absence of ADE mentions in the tweet. Recent work (Lee et al., 2020; Gururangan et al., 2020; Beltagy et al., 2019) showed that downstream performance can be improved by further adapting a general pre-trained model by continued pre-training on more relevant set of downstream tasks. The representations learned in the task adaptive pre-training (TAPT) (Gururangan et al., 2020) involves pre-training using an unsupervised objective on not just the similar domain but the actual end-task and dataset itself. It has shown to improve the performance of the model for the target task and is less computationally demanding. We explore task adaptive pre-training strategies together with fine-tuning for the classification problem. We further explore the use templates as augmentations to improve the task adaptive pre-training.

2 Data

The dataset (Magge et al., 2021) comprises of a training set (18,000 tweets), validation set (915 tweets), and test set (10,000 tweets). The dataset is very imbalanced with only around 7% of the tweets consisting of ADE mentions. We use oversampling to deal with this class imbalance.

As part of the pre-processing on the dataset, we removed URL, retweets, mentions, extra space, non-ascii words and characters. Further we lowercased and striped off white spaces at both ends. We inserted space between punctuation marks.

3 Model

We first apply TAPT (Gururangan et al., 2020) on the pre-trained model to expose it to the task domain sentences and increase the overlap between the language model domain and the target task domain.

The two pre-training strategies we explored in this system setup are: 1) **Simple Contrastive pre-training (SimSCE)** (Gao et al., 2021): The idea is to encode the same sentence twice. Due to the dropout used in the transformer models, both sentence embeddings will be at slightly different positions in the vector space. The distance between these two embeddings will be minimized, while the distance to other embeddings of the other sentences in the same batch will be maximized which serve as negative examples.

2) **Masked language modelling (MLM) (Devlin et al., 2019)** : MLM is an efficient pre-training strategy for learning sentence embeddings. It is trained with a masked language modeling objective (i.e., cross-entropy loss on predicting randomly masked tokens).

Proceedings of the 29th International Conference on Computational Linguistic, pages 153–155
October 12–17, 2022.

Given that amount of task data is insufficient for TAPT,we further augment the task data and generate multiple sentences with various prompt templates appended at the end of each sentence. The templates are randomly assigned from a set of predefined templates based on label information. Examples of templates include: "It contains an Adverse Drug Effect.","It contains an ADE","ADE is present","It mentions an ADE". We thus expose the pre-trained language model (LM) to syntactically diverse template prompts to either cluster the similar representations (SimSCE_aug) or probe the pre-trained LM using masked tokens (mlm_aug). We further fine-tune the continually task pre-trained model on the task data. The supervised TAPT with augmentations phase helps the pre-trained LM encapsulate a broader range of task distribution.

We use BERTweet (Nguyen et al., 2020) as the base encoder to compute the sentence representations. To tackle class imbalance, we experiment with oversampling. For oversampling approach, we randomly sampled positive examples with replacement until ADE class contained 10,000 tweets.

4 Experiments

For the classification task, each model is fine-tuned for 10 epochs with a learning rate of $5e-5$ using Adam optimizer (Loshchilov and Hutter, 2019). We set the batch size to 32 and the maximum sequence length to 128. We utilize PyTorch (Paszke et al., 2017) and HuggingFace library [1] for training BERT using cross-entropy loss. We save model_checkpoint every 200 steps against the validation set using the F1-score of the ADE class for evaluation. For adaptive pre-training with MLM, we train for 10 epochs with masking probability set as 0.15. For adaptive contrastive pre-training, we train for 10 epochs with batch size of 64 and the maximum sequence length 100.

5 Discussion

It is evident from Table 1 that template augmented TAPT yields improved classification accuracy in detecting ADE mentions in tweets. It demonstrates that augmenting with prompt templates can better generalize LM to target task distribution compared to conventional adaptive pre-training method. MLM pre-training with template augmentations (mlm_aug) outperforms MLM only pre-training quite significantly on the validation set. This difference can be attributed to the lack of sufficient task-related data in MLM only pre-training compared to template augmented pre-training. The results from Table 1 indicate that oversampling the ADE class (BERTweet) improved the class imbalance compared to not oversampling (BERTweet_n) and is clearly indicative in the F1-scores.

Table 1: Micro-average Precision (P), Recall (R) and F1 scores (F1) on the validation set of the SMM4H 2022 Task 1a.

TAPT	FT	P	R	F1
-	BERTweet_n	0.6912	0.7231	0.7068
-	BERTweet	0.7065	0.7223	0.7143
mlm	BERTweet	0.7042	0.7692	0.7353
mlm_aug	BERTweet	**0.7605**	**0.8307**	**0.7941**
simsce	BERTweet	0.7344	0.7231	0.7287
simsce_aug	BERTweet	0.7385	0.75	0.7441

Table 2: Micro-average Precision (P), Recall (R) and F1 scores (F1) on the test set of the SMM4H 2022 Task 1a.

model (BERTweet)	P	R	F1
mlm_aug	0.614	0.334	0.433
mlm_aug+ (post-eval)	**0.776**	0.4680	**0.584**
Mean_results	0.646	0.497	0.562

Table 2 shows the performance of the best performing model on the test set of SMM4H 2022 Task1a. Our model's performance is relatively less on the test set compared to the validation set, which can be attributed to over-fitting. Further the performance was improved in post-eval by augmenting data with more templates and achieved better F1 scores than mean_results. To further improve the performance, back translations can be used in addition to template augmentations.

6 Conclusion

In this work, we explore and propose the use of template based TAPT for improving the downstream task performance of detection ADE entity mentions in English tweets. We demonstrate the significance of the template augmentation in comparison to traditional adaptive pre-training strategies. Experiments have shown that our model with template augmentations with TAPT has achieved an F1-score of 0.43, precision of 0.61, and recall of 0.33 on the test set. The future directions would be to use supplementary data and back-translation together with template augmentations.

[1] https://huggingface.co/models

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

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, NAACL-HLT*. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasovic, 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, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Deepademiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on twitter. *J. Am. Medical Informatics Assoc.*, 28(10):2184–2192.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Davy Weissenbacher, Ari Z. Kleinand Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications smm4h shared tasks at coling 2022. In *In Proceedings of the Seventh Social Media Mining for Health (SMM4H) Workshop and Shared Task.*