Prompt Tuning for Condescending Detection

Yuchen Niu
Imperial College London
yn621@ic.ac.uk

Ziwei Lin
Imperial College London
zl5721@ic.ac.uk

Yuzhao Peng
Imperial College London
yp220@ic.ac.uk

1 Introduction

Pre-trained language models (PLMs) are widely used for natural language processing (NLP), under the first pre-train, then fine-tune paradigm (Liu et al., 2019). Recently, a new paradigm called prompt-based methods has received attention from the community (Schick and Schütze, 2021). By reformulating NLP downstream tasks into language modeling analogous tasks (*e.g.,* cloze-style question), prompt-based methods, powered by PLMs, are particularly useful in text classification tasks (Chen et al., 2022). Most existing studies focus on tasks where classification objects are explicitly expressed by the text semantic, such as news topic classification (Gao et al., 2021) and sentiment classification (Li et al., 2021).

A natural question is whether prompt-based methods are still useful for tasks where the classification objects are more implicitly expressed on the pragmatic level, *e.g.,* patronizing and condescending language (PCL) detection (Perez Almendros et al., 2020). In this project, we explore this question by comparing two paradigms, namely, prompt-tuning (P-T) and fine-tuning (F-T), for the PCL detection. Given a paragraph, the task is to identify whether it contains PCL or not. Experiments on the Don't Patronize Me! (DPM) dataset show that two tuning methods achieve similar performance under full data setting. With half of the data, P-T with demonstration achieves competitive performance compared with full data setting[1].

2 Don't Patronize Me! Dataset Analysis

Each data in the DPM dataset are assigned a score from 0 to 4, indicating their PCL level. For this task, the data with level 0 and 1 are `negative` and others are `positive`. Table 1 shows the data distribution. We find that the DPM dataset is highly biased, where only 1% data are `positive`. Table 2 shows the input length distributions of two classes. Given the location and scale of the two distributions, we find that input lengths and labels have a low correlation.

Dataset	Totoal	Positive	Negative
Train	8375	794 (+2068)	7581
Val	2094	199	1895
Test	3832	-	-

Table 1: Statistics of DMP dataset. The number of augmented data is in the bracket. The test set is hidden.

Class	Official		+Augment	
	Mean	Std.	Mean	Std.
PCL	53.52	33.65	41.53	26.32
Non-PCL	48.16	29.18		

Table 2: Distribution of input lengths in training set.

Following Li et al. (2020), we measure the point-wise mutual information between the words of a paragraph and its keyword given `positive` or `negative` class to find the top-100 most related words for each keyword (Table 3). Then, we calculate word similarities of top-100 words between two classes given a keyword using word2vec (Mikolov et al., 2013). As shown in Table 3, the word similarities are above 90%, indicating that semantic features of different labels are very close at the word level. This suggests that this task requires understanding of pragmatic functions of languages, which can be very subjective.

3 Data Augmentation

Highly biased data can lead to consistent prediction for one major label. To balance the data, we augment positive samples using Parrot paraphraser (Damodaran, 2021). To obtain high-quality data, we set the paraphrase space to 10, and the adequacy and fluency thresholds to 0.9,

[1]Code is available at https://github.com/NiuJ1ao/PromptTuningPCL

Keyword	Positive (PCL)	Negative (Non-PCL)	Averaged Word Similarity
poor-families	anybody; absent; roadblocks; schools; answers; davao; ...	urchins; osme; dswd; federal-state; nagpur; convergence; ...	0.94
vulnerable	gorman; peaceful; emission; inquest; reduction; coexistence; ...	rendle; neutralise; lieut; adikpo; handset; dispensing; ...	0.95
...	...	...	...

Table 3: Top-100 PMI words and their similarities given keywords and classes in the training data.

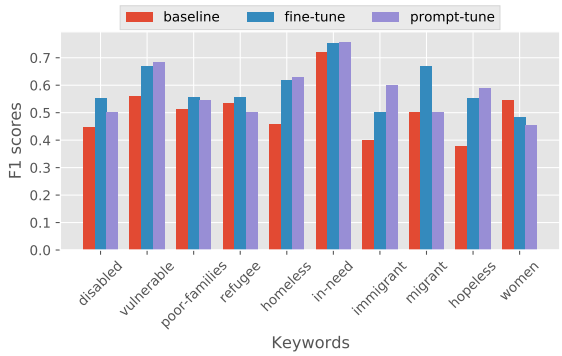

Figure 1: F-1 scores of models given keywords.

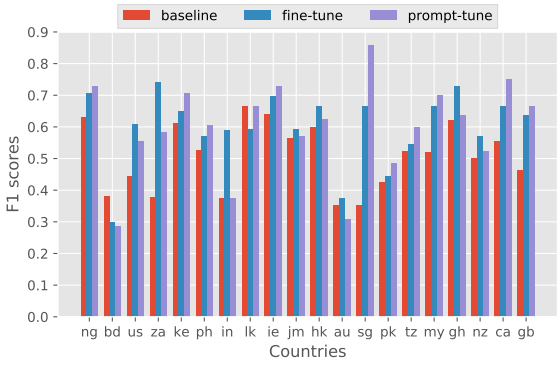

Figure 2: F-1 scores of models given countries.

following Damodaran (2021) and set the maximum length to 90 percentile of the maximum input length based on empirical consideration. In this way, we obtain 2068 augmented positive data.

4 Prompt-based Text Classification

Prompt-based methods for text classification consist of patterns, which convert an input into a cloze-style question, and verbalizers, which map predicted words into labels (Schick and Schütze, 2021). In this project, we define a pattern, P:

$$P(S) = S.\ It\ was\ <mask>. \quad (1)$$

where S is the input, and the verbalizer, V:

$$\begin{aligned} V(\text{“}patronizing\text{”}) &= \texttt{positive} \\ V(\text{“}respectful\text{”}) &= \texttt{negative} \end{aligned} \quad (2)$$

Using such pattern-verbalizer pairs, we ask a PLM to perform mask filling for $<mask>$ that can directly give scores for each label.

In addition, for few-shot setting, following Gao et al. (2021), we sample one positive example and one negative example as the demonstration, and attach it to the input pattern. The mask tokens are replaced with target words in examples. Multiple models are trained by different subsets of data and the logits are ensembled uniformly in inference.

5 Experiments

5.1 Settings

We use BART-Large (Lewis et al., 2019) as our foundation model. For F-T, we train a classifier on the top of BART, taking sentences representation as input. For P-T, we use the regular prompt tuning method (Schick and Schütze, 2021). For both F-T and P-T, we train models using augmented data.

The hyper-parameter choice is based on prior work (Schick and Schütze, 2021). For full data learning, we use a linear learning rate scheduler with the learning rate $1e{-}5$. The batch size and maximum sequence size are 32 and 128, respectively. The weight decay is 0.01 and early stop is used. The best checkpoint is picked using F-1 on the val set. All models are trained with 20 epochs and cross-entropy loss on an Nvidia RTX 3090.

For few-shot learning, we conduct experiments with 500 samples and 2000 samples per class, respectively. The maximum input length extends to 512 since demonstrations are appended and the batch size reduces to 4 to fit into our GPU and the number of epochs decreases to 10 accordingly.

5.2 Full-data learning

Table 4 shows the model performance under F-T and P-T paradigms. We observe that both F-T and P-T models outperform the baseline by 10 F-1 score on the val set. On the test set, P-T shows

	Model	Val Set			Test Set		
		F-1	Precision	Recall	F-1	Precision	Recall
Full-data	Baseline	0.5151	0.4134	0.6834	0.5065	0.6183	0.5568
	F-T	0.6025	0.5161	**0.7236**	0.5462	0.4600	**0.6719**
	P-T	**0.6028**	**0.5536**	0.6615	**0.5605**	**0.5263**	0.5994
Few-shot	P-T (1000 data)	0.5610	0.4816	0.6717	0.5556	0.4784	0.6625
	P-T (4000 data)	0.5991	0.5341	0.6821	0.5568	0.5065	0.6183

Table 4: Performance of different tuning paradigms. The baseline is a fine-tuned RoBERTa-base model.

Model	High PCL (4)	Low PCL (2)
Baseline	0.8848	0.3636
F-T	**0.9425**	**0.6154**
P-T	0.9000	0.5000

Table 5: F-1 scores of models given two levels of PCL.

Model	Short (F-1)	Long (F-1)
Baseline	0.4771	0.5342
F-T	0.6203	0.5677
P-T	**0.6352**	**0.5714**

Table 6: Model performance under different input lengths, where long inputs are those longer than 75%ile of inputs (63) and short inputs are shorter than 45%ile of inputs (41) in val set.

better performance than F-T with respect to F-1 score, indicating P-T has better generalization ability. This is consistent with the findings of (Han et al., 2021) that prompt-tuning can better stimulate the knowledge distributed in the PLM. Also, we find that compared with other text classification tasks, model performance on the PCL task is very low, which can be that PCL detection requires more than semantic information and the task is too different from language modeling tasks. Similar findings are also observed by Min et al. (2021) on the TREC dataset (Voorhees and Tice, 2000).

5.3 Few-shot learning

Table 4 shows the model performance using prompt-tuning under few-shot settings. We find that few-shot P-T models achieve similar results compared with full data training. This proves that PLMs can be few-shot learners even the classification object is relatively implicitly expressed.

6 Analysis and Discussion

We further compare the model performance influenced by different PCL levels, input lengths and categories, under full-data learning.

PCL levels. As shown in Table 5, the F-T model outperforms the P-T model at predicting two levels of patronising content. The great improvement at predicting a lower level of PCL indicates the latent representation of PCL is reinforced in the knowledge distribution during training. All models have a higher F-1 score on higher-level PCL than on lower-level PCL, because their semantic information is more distinguishable.

Input length. Table 6 shows that baseline model is better at predicting long inputs. The baseline is trained with the official data that have long-tail input length distribution. Since the paraphraser tends to generate shorter sentences (Table 2), F-T and P-T are better at predicting shorter inputs. In opposition, since the dependency relations in the long inputs is more complicated, the improvement at predicting long inputs is limited. Overall, F-T and P-T outperform the baseline.

Categorical data. Figure 1 & 2 present model performance under different categories, *i.e.,* different keywords and countries. Despite that the F-1 scores increase for most categories, we observe that models under-fit some categories, *e.g.,* "women", which can be because data are imbalanced between categories in quantity and quality, and models are lack fine-granularity control at categorical prediction. For categories that show dramatic improvements, such as "sg", the number of positive data in val set is comparably less, where the model only needs to predict fewer positive samples during inference to get high F-1 scores. In a few categories like "migrant", P-T underperforms F-T. This can be we use one fixed pattern, which can be not suitable for all data, and can generalize poorly on specific categories.

7 Conclusion

In this project, we compare two PLM tuning paradigms, namely fine-tuning and prompt-tuning, for the PCL detection task. We find that, there is a marginal distinction between F-T and P-T since PLM does not have prior information on condescending. Besides, PLMs are few-shot learners,

where careful design of prompts and demonstrations can be necessary.

References

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. Adaprompt: Adaptive model training for prompt-based nlp. *arXiv preprint arXiv:2202.04824*.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Chengxi Li, Feiyu Gao, Jiajun Bu, Lu Xu, Xiang Chen, Yu Gu, Zirui Shao, Qi Zheng, Ningyu Zhang, Yongpan Wang, et al. 2021. Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis. *arXiv preprint arXiv:2109.08306*.

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. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Noisy channel language model prompting for few-shot text classification. *arXiv preprint arXiv:2108.04106*.

Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze questions for few shot text classification and natural language inference.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 200–207, New York, NY, USA. Association for Computing Machinery.