PINGAN Omini-Sinitic at SemEval-2022 Task 4: Multi-prompt Training for Patronizing and Condescending Language Detection

Ye Wang **Yanmeng Wang** **Baishun Ling** **Zexiang Liao**
Shaojun Wang **Jing Xiao**
Ping An Technology, Beijing 100191, China
{wangye430,wangyanmeng219,lingbaishun271,liaozexiang901,
wangshaojun851,xiaojing661}@pingan.com.cn

Abstract

This paper describes the second-placed system for subtask 2 and the ninth-placed system for subtask 1 in SemEval 2022 Task 4: Patronizing and Condescending Language Detection. We propose an ensemble of prompt training and label attention mechanism for multi-label classification tasks. Transfer learning is introduced to transfer the knowledge from binary classification to multi-label classification. The experimental results proved the effectiveness of our proposed method. The ablation study is also conducted to show the validity of each technique.

1 Introduction

Patronizing and Condescending Language (PCL) is proposed by Pérez-Almendros et al. (2020), which builds a dataset for PCL detection. The Patronizing and Condescending Language Detection (Pérez-Almendros et al., 2022) contains two text classification tasks. Subtask 1 is a binary classification task, which requires a system to predict whether the paragraph contains any form of PCL. Subtask 2 is a multi-label classification task, which must identify which PCL categories express the condescension.

PCL Detection is a sentiment analysis task and we treated subtask 2 as an Aspect-Based Sentiment Analysis (ABSA) (Jo and Oh, 2011) task, which is also a multi-label classification to classify the sentence sentiment on different aspects.

Early works on ABSA focus on feature engineering (Wagner et al., 2014) and subsequent neural network-based methods (Wang et al., 2017). Recently, Pre-trained Language Models (LMs) such as BERT (Devlin et al., 2019), ROBERTA (Liu et al., 2019), ALBERT (Lan et al., 2020) have been proposed and brought significant improvement in various NLP applications. As there is not much improvement in ABSA task through direct implementation of those pre-trained LMs, Bert-pair (Sun et al., 2019) has been proposed to help them adapt to ABSA effectively. Furthermore, Bu et al. (2021) proposed an attention between sentence embeddings and label embeddings to focus on the crucial tokens which are related to the label. Recently, CapsNet-Bert (Jiang et al., 2019) used the attention between the label and the input with capsule network to improve the performance.

Nowadays, the Pattern-Exploiting Training (PET) (Schick and Schütze, 2021) and P-tuning (Liu et al., 2021) has been proposed to utilize the pre-trained LMs more effectively, which trains fine-tune model by the pre-training tasks such as Mask Language Modeling (MLM) by adding a prompt sequence to the input.

Considering the similarity between PCL and ABSA, we applied PET and Bert-pair in PCL Detection through the following steps: Firstly, we treat the subtask 1 as a prompt training task (as described in PET) by using a PCL description as the prompt. Secondly, we transform the subtask 2 from a multi-label classification task to multiple binary classification tasks by using the label names as the prompts. Thirdly, we conduct transfer learning from subtask 1 to subtask 2 to further improve the performance. Last but not least, we also proposed a label attention mechanism based on the multi-prompt training for multi-label classification.

Our contribution can be summarized as follows: we apply prompt training on PCL detection subtask 1 and 2, and prove its effectiveness on both tasks; in subtask 2, we propose a label attention mechanism with multi-prompt training; we apply transfer learning from subtask 1 to subtask 2, where the transfer learning is proved to be effective in ABSA.

2 Background

2.1 Task Description

The tasks intend to detect whether the input paragraphs have any forms of PCL, and which PCL categories are contained. The PCL categories defined

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 313 - 318

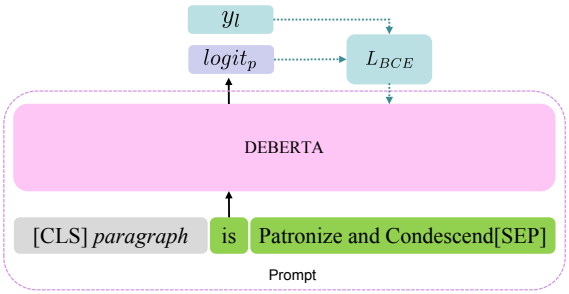

Figure 1: Subtask 1 overview, $logit_p$ is the corresponding logits of the word "is" in prompt.

in the task paper Pérez-Almendros et al. (2022) are 'Unbalanced power relations', 'Shallow solution', 'Presupposition', 'Authority voice', 'Metaphors', 'Compassion','The poorer, the merrier'. The tasks can be formalized as follows:

The training data of subtask 1 consists of tuples (q, l_{binary}), where q is a paragraph extracted from articles, and l_{binary} is the binary classification label with values 0, 1. The training data of subtask 2 consists of tuples (q, $l_{category}$), where q is a paragraph extracted from articles and $l_{category}$ is the label for multi-label multi-classification task. $l_{category}$ is a 7-digit binary vector. Each digit represents whether the paragraph contains the corresponding PCL category and it is possible for one paragraph to contain multiple PCL categories. We also defined a set A, which contains the names of categories. $a \in A$ represents a category name from all PCL categories.

2.2 DEBERTA

Our comparison of present large pre-trained Language Models(LM) showed that DEBERTA (He et al., 2021b) seems to be the most effective model. Different from other works, DEBERTA implemented a disentangled attention mechanism which utilizes the input contents and relative positions. We conduct most experiments based on DEBERTA-v3 (He et al., 2021a), which trained the DEBERTA model with replaced token detection (RTD) pre-training task, proposed by ELECTRA (Clark et al., 2020). RTD is a more sample-efficient pre-training task than replacing Mask Language Modeling (MLM). The experiments in He et al. (2021a) shows that DeBerta-v3-large model achieves better performance on GLUE benchmark even compared with larger pre-trained LMs.

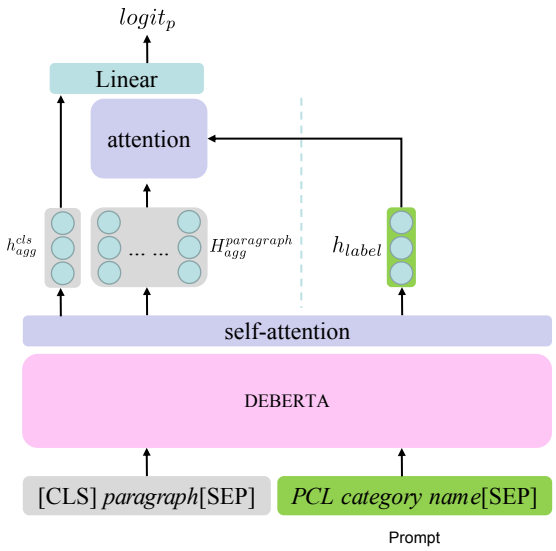

Figure 2: Subtask 2 overview, we conduct a attention mechanism between the logits of paragraph and the logits of label name.

3 System overview

3.1 Prompt Classification

Inspired by PET, we adopt a prompt design on traditional classification tasks based on pre-trained LMs. As shown in Figure 1, we adopt DEBERTA-v3 as the pre-trained LM. The input sequence consists of the original paragraph q and the prompt sequence, which is "is patronize and condescend" in subtask 1, denoted as r. The total input sequence can be described as "$[CLS]\, q\, r\, [SEP]$" for subtask1. Since the DEBERTA-v3 trained by RTD as same as ELECTRA, we don't use the hidden states of [CLS] for classification. Instead we use the hidden state of word "is" (as described in Figure 1) for a binary classification, and the label is 1 when the paragraph contains PCL, 0 when the paragraph has no PCL. The input sequence is forward to DEBERTA and the hidden states are calculated by Equation 1

$$H_q = F([q; r]) \tag{1}$$

$$logit_p = D(H^p_q) \tag{2}$$

where F is the pre-trained 24-layer transformers and D is a linear layer which classify the hidden states of prompt word from DEBERTA. Then we use the hidden states $H^p_q \in \mathbb{R}^d$ selected from H_q as the input to D, which denotes the hidden states of the prompt word "is", the d is the hidden size of DEBERTA. BCE (Binary Cross Entropy) loss,

which measures the Binary Cross Entropy between the golden label and the output, is used for binary classification as Equation 3.

$$L_{BCE} = BCE(Sigmoid(logit_p), l_{binary}) \quad (3)$$

where l_{binary} is the binary label of subtask 1.

In this way, the semantics of prompt can be jointly learned with the input paragraph. By using the same training method with pre-trained LM, the representations stored in the pre-trained LM has been maximum reserved.

3.2 Multi-label Prompt Classification

For subtask 2, we transform the multi-label classification task to multi binary classification tasks by concatenating the paragraph with each PCL category label names. As shown in Figure 2, the input paragraph q is concatenated with one PCL category name a, which can be denoted as "[CLS]q[SEP]a[SEP]". We use an self-attention layer as an aggregator to aggregate the sequence embeddings, which is more effective then using the [CLS] embedding. It can be formulated as:

$$H_q = F([q; a]) \quad (4)$$

$$H_{agg} = S(H_q) \quad (5)$$

where S is the self-attention layer proposed by BERT (Devlin et al., 2019), and H_{agg} represents the aggregated hidden states of input sequence. Then we use the [CLS] aggregated hidden states H^{cls}_{agg} to classify whether the paragraph has the corresponding PCL category, which can be formulated as:

$$logit_{cls} = D(H^{cls}_{agg}) \quad (6)$$

$$L_{BCE} = BCE(Sigmoid(logit_{cls}), l_c) \quad (7)$$

where l_c is the binary label of PCL category detection, selected from the category label $l_{category}$. The real output for subtask 2 is the combination of all PCL category detection results.

From previous works, this approach have been proved effectively for improving the multi-label classification tasks. We named this approach MPrompt in this paper.

3.3 Label Attention

As shown in Figure 2, we also propose a label attention above the MPrompt approach. Although the transformers model has self-attention between every tokens, an external attention between paragraph and label prompt is still helpful for the model to focus on more important words in the paragraph. First, We split the output hidden states H_{agg} into $H^{paragraph}_{agg}$ and H^{label}_{agg}, where $H^{paragraph}_{agg}$ is the hidden states corresponding to the sequence "[CLS]q[SEP]", and H^{label}_{agg} is the hidden states corresponding to the sequence "a[SEP]". Then, the label embedding h_{label} is computed as average pooling over H^{label}_{agg}, where $h_{label} \in \mathbb{R}^d$.

As shown in equation 8, we use an attention layer to combine the h_{label} and $H^{paragraph}_{agg}$, where h_{label} is used as query (Q), and $H^{paragraph}_{agg}$ is used as key (K) and value (V) followed the definition of Scaled Dot-Product Attention in Vaswani et al. (2017).

$$h_{LA} = softmax(\frac{QK}{\sqrt{d}})V \quad (8)$$

where h_{LA} denotes the output of attention layer. Then we concatenate h_{LA} with the aggregated hidden states of [CLS] as follows:

$$logit_p = D([H^{cls}_{agg} : h_{LA}]) \quad (9)$$

where D is the Linear layer for binary classification.

3.4 Transfer Learning

We also implemented multi-task learning, joint learning and transfer learning between subtask 1 and 2. Only the transfer learning has improved the performance of subtask 2. First we trained a subtask 1 model with prompt training, then we use the subtask 1 model as the initial checkpoint, trained a subtask 2 model with MPrompt. Experiments in section 5.1 proved the effectiveness of this approach. Unfortunately, transfer learning from subtask 2 to subtask 1 or other approach has no improvement.

3.5 Other Tricks

The DEBERTA which has large amount of parameters tends to over-fit on small training dataset. We utilize RecAdam (Chen et al., 2020) to fine-tune the pre-trained model to address the over-fitting problem. RecAdam optimizer is proposed to address the catastrophic forgetting problem of sequential transfer learning paradigm by introducing

a recall and learning mechanism into Adam optimizer, which maintain the learned knowledge in pre-trained model while learning a new task.

The numbers of each category is very unbalanced in train dataset, we over-sample the positive samples to alleviate this problem. On subtask 2, we keep the proportion of each positive category unchanged during oversampling.

Data augmentation is not applied in our approach because we haven't find other proper datasets.

4 Experimental setup

4.1 Data

We use the official released dataset of SemEval2022 Task4 for experiments. The dataset contains 8375/2094/3832 samples for train, dev and test data. The subtask 1 and subtask 2 share the same input paragraphs, and has different labels respectively. The maximum, mean length of training data is 1005 and 55.28 in words perspective, and 90% of training data are shorter than 95 words. The ratio of positive samples is only 9.48% in subtask 1, even much smaller in subtask 2, since some of PCL category such as "The poorer, the merrier" is quiet few. The shortage of positive samples makes the model more difficult to distinct the PCL descriptions from negative samples, and is hard to train a model with good generalization.

4.2 Parameter settings

Our implementation is based on the Pytorch framework for transformer-based models Wolf et al. (2020). We trained our model based on the pre-trained DEBERTA-v3-large model. We use Adam/RecAdam optimizer with a learning rate of 3e-5, batch-size of 32 to train our models. The max sequence length is 256 and the epoch of training is set to 10 in subtask 1, 3 in subtask 2. To address the over-fitting problem, we apply RecAdam with sigmoid annealing function, where the annealing rate is 0.01 and the annealing time-step is 500. Specially, in subtask 2, we apply one self-attention layer as the aggregator and one attention layer to calculate the label attention. We pick the best checkpoint based on the performance on the dev set. Besides using DEBERTA, we also trained models based on ROBERTA, which is competitive with DEBERTA on subtask 2.

Since only 2 submissions are permitted in submitting phase, we trained multiple models under different settings for model ensemble. We also adopt 7-fold cross-validation training to improve the system generalization.

4.3 Ensemble

Two strategies are used for our final submissions on test data: 1) we ensemble all 7 models from 7-fold cross-validation training by averaging their outputs, which is trained on the train data of each subtask; 2) we trained multiple models on the train data with different model structures. Eight top different models are selected based on the dev accuracy for models ensemble, then average their outputs as the final output.

5 Results and Analysis

5.1 Single Model Performance

In subtask 1, the F1-score of PCL is used as the official metric and the results of dev set is shown in Table 2. We implement a baseline model for comparison, which is the traditional classification model with pre-trained LMs by using the [CLS] embedding. We trained two traditional classification models which are based on ROBERTA and DEBERTA. The results shown that DEBERTA is better than ROBERTA for subtask 1. The prompt classification improved the F1-scores by 2.80% on the same pre-trained LM, which proved that prompt classification is more suitable with the pre-trained LMs.

In subtask 2, the official metric is the average score of F1-scores of all PCL categories. As shown in Table 1, our method achieves significant improvement compared with baselines. The ROBERTA and DEBERTA denote the baseline models which uses a multi binary classification head upon pre-train LMs. The MPrompt both improved the performance on ROBERTA and DEBERTA by 8.28%, 20.29%, respectively. The information of label names is utilized by MPrompt method effectively.

For Transfer Learning(TL), as shown in Table 1, the scores is improved by 3.62% based on DEBERTA, while is dropped by 3.17% based on ROBERTA. We assume that is because the structure of ROBERTA in subtask 1 is not suitable for MPrompt in subtask 2.

Label Attention(LA) is also proved to be an effective approach both on ROBERTA and DEBERTA. Since the transfer learning is not useful in ROBERTA, we further experiment the ROBERTA MPrompt with LA and the DEBERTA MPrompt

Method	Unb	Sha	Pre	Aut	Met	Com	Poo	avg-F1
ROBERTA	59.28	47.05	25.00	30.76	42.85	49.73	42.85	42.50
ROBERTA MPrompt	58.95	38.09	46.15	35.00	48.71	55.23	39.99	46.02
+TL	61.53	36.36	40.00	34.61	45.83	51.74	42.10	44.56
+LA	61.42	47.05	45.87	37.38	43.03	54.63	39.99	47.05
DEBERTA	55.24	29.16	34.28	25.28	45.45	49.26	19.99	36.95
DEBERTA MPrompt	57.14	40.54	32.81	30.76	44.44	52.77	52.63	44.45
+TL	57.65	29.63	43.10	39.34	60.67	47.61	44.44	46.06
+TL+LA	60.25	30.55	43.69	45.97	53.48	49.49	47.61	**47.29**

Table 1: Single model performance of subtask 2 on dev dataset, Unb...Poo denotes the F1-scores of each categories, avg-F1 is the average F1-score. TL denotes the Transfer Learning, and LA denotes the Label Attention.

Method	Acc	Recall	F1
ROBERTA	60.56	64.82	62.62
DEBERTA	66.67	61.31	63.87
DEBERTA Prompt	65.50	65.82	**65.66**

Table 2: Single model performance of subtask 1 on dev dataset.

	Method	Subtask1	Subtask2
Dev set	7-fold ensemble	-	-
	top ensemble	72.16	52.99
Test set	7-fold ensemble	**62.73**	**43.87**
	top ensemble	58.93	43.20

Table 3: Ensemble performance on dev and test dataset, where 7-fold is the models from 7-fold cross-validation training, top ensemble means that ensembles the models with top dev accuracy.

with TL and LA. The results shown that LA can improve the performance by 2.23% on ROBERTA, and 2.67% on DEBERTA, which proved that an external attention between tokens and label names is benefit for picking more important tokens related to the label category.

5.2 Ensemble Performance

The performances of ensemble models are shown on Table 3, which is obtained from the competition leader-board. Our system got the second place in subtask 2 and the ninth place in subtask 1. Ensemble results on dev dataset are exhibited for comparison. Since the 7-fold training has trained the dev set, we don't exhibit the dev ensemble results of 7-fold training. It is obvious that the top ensemble method is much over-fit on dev set, for the scores of test set dropped much on top ensemble method. 7-fold ensemble method is an effective way to avoid over-fitting and got the best test scores of our system.

In top ensemble, we also found that integrating different pre-trained models improves the results significantly. The top ensemble method will be more useful if the distribution of dev set and test set are similar.

6 Conclusion

In this paper, we propose a multi-prompt training with label attention mechanism to improve the performance of multi-label classification task. Pre-trained models have made great performance gain compared to traditional neural network models in many natural language tasks. The above experimental results may suggest that the current pre-trained model mechanism still has room for improvement in ABSA tasks.

References

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A chinese review dataset towards aspect category sentiment analysis and rating prediction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2069–2079.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7870–7881.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

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 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining 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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. 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 6279–6284.

Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 815–824.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *CoRR*, abs/2103.10385.

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.

Carla Pérez-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.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2022. SemEval-2022 Task 4: Patronizing and Condescending Language Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 380–385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. 2014. DCU: aspect-based polarity classification for semeval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 223–229.

Bo Wang, Maria Liakata, Arkaitz Zubiaga, and Rob Procter. 2017. Tdparse: Multi-target-specific sentiment recognition on twitter. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 483–493.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.