PROMPTDA : Label-guided Data Augmentation for Prompt-based Few Shot Learners

Anonymous ACL submission

Abstract

Recent advances on large pre-trained language models (PLMs) lead impressive gains on many natural language understanding (NLU) tasks with task-specific fine-tuning. However, direct fine-tuning PLMs heavily rely on large amount of labeled instances, which are expensive and time-consuming to obtain. Prompt-based tuning on PLMs has proven valuable for few shot tasks. Existing works studying prompt-based tuning for few-shot NLU mainly focus on deriving proper label words with a verbalizer or generating prompt templates for eliciting semantics from PLMs. In addition, conventional data augmentation methods can enrich training data for improving few-shot learning, while ignoring the label semantics. It is promising to leverage the rich label semantics in label words for data augmentation to facilitate prompt-based tuning for the downstream NLU tasks. However, the work on this is rather limited. Therefore, we study a new problem of data augmentation for prompt-based few shot learners. We propose a novel label-guided data augmentation method PROMPTDA which exploits the enriched label semantic information for data augmentation. Experimental results on several few shot text classification tasks show that our proposed framework achieves superior performance by effectively leveraging label semantics and data augmentation in language understanding. We will open our code later for reproduction.

1 Introduction

Pre-trained language models (PLMs) have shown promising performance in various applications such as text classification (Yang et al., 2019), document summarization (Zhang et al., 2020a), question answering (Mirzaee et al., 2021). Directly *fine-tuning* PLMs such as BERT or RoBERTa is a common approach to adapt them in downstream natural language understanding (NLU) (Devlin et al., 2018; Liu et al., 2019). For example, PLMs can be appended with additional classification layers that are trained with labeled examples for text classification. However, sufficient labeled data can be expensive to acquire, which hinders the deployment of PLMs effectively in many NLU tasks (Mukherjee et al., 2021; Shu et al., 2020b). Therefore, it becomes increasingly important for developing effective PLMs with few labeled data.

The recent advancement of *prompt-based tuning* has shown significant improvement over the normal fine-tuning paradigm on tasks with few labeled data (or few-shot tasks) (Brown et al., 2020). Typically, a prompt-based tuning paradigm transforms a NLU task into a masked language modeling (MLM) problem. For example, in sentiment analysis, an original sentence "`nice movie to watch.`" can be augmented with a *template* "`It is [MASK]`" as the input x. Each class (e.g., POSITIVE) is usually represented by a *label word* (e.g., `good`) selected by a *verbalizer* from the vocabulary (Schick and Schütze, 2020). The prediction of a positive label is based on the probability of the `[MASK]` token being filled with `good`.

However, there are several challenges for applying prompt-based tuning methods for few shot scenarios. Conventional prompt-based tuning methods do not fully utilize the label semantics for each class in classification. Most of the verbalizers only choose one label word that is the most relevant to each class, namely ***one-to-one verbalizer*** (Schick and Schütze, 2020; Gao et al., 2021). Using a single label word to represent each class only utilize the limited semantic information in the token. For instance, the prediction for POSITIVE can only be inferred based on the probability score of the selected label word such as `good`, while other possible relevant tokens such as `great` and `best` are ignored. In addition, selecting the suitable label word manually (Schick and Schütze, 2020) or via automatic search (Gao et al., 2021) can have large variance and lead to unstable prediction results.

To address the aforementioned challenges, one natural solution is to design a verbalizer mapping from multiple label words to each class, namely ***multiple-to-one verbalizer***, which can leverage rich label semantics and reduce the variance of selecting a single label word. For instance, with the verbalizer mapping from set of label words {`good`, `great`, `best`} to the class POSITIVE, the prediction for POSITIVE can be inferred based on the probability score of each token in the set.

For few shot tasks, data augmentation (DA) methods are widely used to enrich training data by modifying existing data through transformations on token-level or sentence-level (Chen et al., 2021). For example, a common approach is to fetch synonyms tokens for substitutions based on pre-defined dictionaries such as WordNet (Miao et al., 2020). However, previous DA methods focus on transforming the input while not fully utilizing the label semantics, which have great potentials for few shot tasks (Luo et al., 2021). However, the work on exploring label semantics for data augmentation is rather limited. Therefore, we propose to incorporate the rich label semantic information from label words derived from the aforementioned multiple-to-one verbalizer into a data augmentation, by carefully constructing ***instance-label pairs***. For example, we aim to generate synthetic data points $\{(x, \texttt{good}), (x, \texttt{great}), (x, \texttt{best})\}$ from the original instance x and leverage them to improve the prompt-based tuning.

Specifically, our framework PROMPTDA contains three coherent modules including a label augmentation, an augmented prompt-based tuning, and a prediction transformation. First, we use a PLM to build the multiple-to-one verbalizer and derive the semantically similar tokens for each label class. Second, we construct the instance-label pairs from the original data for each label word, and training the language model with masked language modeling. Third, for inference, we utilize the trained language model to predict the label by aggregating the probability scores on label words.

The contributions of this paper are summarized as follows: (1) We study a new problem of data augmentation in prompt-based tuning for few shot learning; (2) We propose a novel label-guided data augmentation framework PROMPTDA that can derive multiple label words and exploit the rich semantic information of the label words for prompt-tuning; and (3) We conduct extensive experiments on several real-world few shot tasks and demonstrate the effectiveness of the proposed framework.

2 Related Work

Prompt-based Tuning has attracted increasing attention recently for various natural language inference tasks including text classification (Gao et al., 2021), question answering (Jiang et al., 2020; Li and Liang, 2021), language generation (Dou et al., 2020), etc. The prompt-based learning framework has shown promising performances especially in zero shot or few shot classification tasks when limited or no labels are available (Liu et al., 2021). For example, Dou *et al.* propose a prompt-based fine-tuning framework that automatically generates prompt templates and incorporates demonstrations to improve few-shot classification performances (Gao et al., 2021). Shin *et al.* proposes the AutoPrompt method to automatically generate prompts for eliciting the knowledge from language models (Shin et al., 2020). Other work on improving prompt-based model performances also focus on constructing various types of prompt and answers (Brown et al., 2020; Jiang et al., 2020), or utilizing multi-modal prompt learning (Tsimpoukelli et al., 2021). etc.

Few-shot Text Classification aims to build text classification model when few labeled data is available. Existing work mainly utilize the following categories. First, semi-supervised learning where unlabeled data, alongside usually small amounts of labeled data, is used for learning (Mukherjee and Awadallah, 2020; Lee et al., 2021). For example, Subhabrata *et al.* propose to jointly learn from a small set of labeled data and a large amount of unlabeled data with uncertainty using self-training (Mukherjee and Awadallah, 2020). Second, meta-learning frameworks such as metric-based (Sui et al., 2020) and optimization-based approaches (Bansal et al., 2019) are developed. Third, weakly supervised learning to derive weak labels (Shu et al., 2020a; Meng et al., 2020) or incorporating constraints (Stewart and Ermon, 2017) to in addition to the limited clean labels to improve text classification. Other approaches user transfer learning to learn to adapt transferable information from the source domain to the target domain (Gupta et al., 2020), or leverage auxiliary tasks to improve the target tasks (Xia et al., 2021; Yin, 2020).

Data Augmentation is to construct synthetic data from an available dataset to enlarge the data size, which can help supervised training with enriched training data (Shu et al., 2018; Guo, 2020), or self-supervised learning for constructing samples in pretext tasks (Zhang et al., 2017; Yoon et al., 2020), etc. Data augmentation techniques for natural language generally fall into data space and feature space (Bayer et al., 2021). Augmentation in data space transforms the data in the original form in the character-level, word-level, phrase-level and document-level. In the feature space, representations in the latent space is manipulated by adding noise or interpolation (Schwartz et al., 2018; Verma et al., 2019). In general, conventional data augmentation brings marginal improvements with prompt-based tuning paradigm in few shot tasks (Zhou et al., 2021; Chen et al., 2021). It is under exploring about how to design effective data augmentation methods for prompt-based few-shot scenarios. Therefore, we propose a novel label-guided data augmentation mechanism in prompt-based tuning for few shot learning tasks.

3 Problem Definition

Few-shot classification is a task where a classifier is learned to predict unseen classes with limited labeled examples during the training. Follow the widely-used few-shot setting (Gao et al., 2021; Liu et al., 2021), we assume to have a large pre-trained language model $\mathcal{M}$ that is utilized to fine-tune on a task with data $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X}$ denotes the examples and $\mathcal{Y}$ indicates the corresponding labels. For each task, the number of training instances in each class is K, which is usually small (e.g., 8 or 16). The goal is to train a prompt-learning strategy that generalize well on unseen examples in the test set $\mathcal{D}_{\text{test}}$ with few labeled training data in $\mathcal{D}_{\text{train}}$. To ensure a fair parameter setting, we assume that a validation set $\mathcal{D}_{val}$ is available, and $|\mathcal{D}_{\text{val}}| = |\mathcal{D}_{\text{train}}|$. The test set $\mathcal{D}_{\text{test}}$ is the same as full-data training setting.

4 Label-guided Data Augmentation for Prompt-based Tuning

In this section, we detail the proposed framework of PROMPTDA, illustrated in Figure 1. It mainly consists of three modules: (1) a label augmentation module to derive multiple label words from labels for enriching the label space; (2) a augmented prompt-based tuning module for augmenting training data guided by label words; and (3) a prediction transformation module to transform the prediction from the label words to original labels.

4.1 Label Augmentation

Due to the limited available labels in few-shot learning, recent work are generating label words to help prediction (Schick and Schütze, 2020; Gao et al., 2021). The goal is to extend the label space by incorporating the rich context of vocabulary. While existing work mainly focus on selecting one label word for each label category manually or automatically in prompt-tuning, the resultant label words are often inconsistent and the semantics in other candidate label words are ignored. Therefore, we propose to automatically search multiple label words for each class to better enrich the label space. Let $\mathcal{F} : \mathcal{V}_\mathcal{Y} \rightarrow \mathcal{Y}$ denotes the multiple-to-one verbalizer that maps a set of label words $\mathcal{V}_y = \{v_y^1, v_y^2, ..., v_y^{k_y}\} \subset \mathcal{V}$ to each label category $y \in \mathcal{Y}$, where $k_y = |\mathcal{V}_y|$ denotes the number of selected label words.

Firstly, we aim to search a candidate set of label word $\tilde{\mathcal{V}}_y \subset \mathcal{V}$ that is semantically similar to each class $y \in \mathcal{Y}$. Let $\mathcal{D}_{\text{train}}^y$ denote the subset of training data with class y. $\mathcal{T}(x)$ denotes the input x with a fixed template $\mathcal{T}$. `Po([mask])` denotes the position of `[mask]` in input x. We propose to select the Top- m label words from vocabulary as $\tilde{\mathcal{V}}_y$ based on the conditional likelihood over $\mathcal{D}_{\text{train}}^y$ for each class y:

$$\tilde{\mathcal{V}}_y = \underset{v \in \mathcal{V}}{\text{Top-}m} \left\{ \sum_{(x,y) \in \mathcal{D}_{\text{train}}^y} \Pr(v, \mathcal{T}(x)) \right\} \quad (1)$$

where $\Pr(v, \mathcal{T}(x))$ denotes the corresponding probability score of each token in the vocabulary filling in `Po([mask])` in PLM inference as:

$$\Pr(v, \mathcal{T}(x)) = \Pr(\texttt{Po([mask])} = v \mid \mathcal{T}(x)) \quad (2)$$

Secondly, we construct a candidate set F for the whole dataset. It is a combinatorial problem to select k_y label words from $\tilde{\mathcal{V}}_y$ to construct $\mathcal{V}_y$ for each class y. The number of possible candidates of $\mathcal{V}_y$ is $\binom{|\tilde{\mathcal{V}}_y|}{k_y}$. Then the element number of candidate set F is $|F| = \binom{|\tilde{\mathcal{V}}_y|}{k_y}^{|\mathcal{Y}|}$. We utilize each multiple-to-one verbalizer candidate of F to infer and calculate the prediction accuracy on $\mathcal{D}_{\text{train}}$ via the same *prediction transformation* method in § 4.3. Then we select the Top- n candidates from F based on the prediction accuracy. If there exist

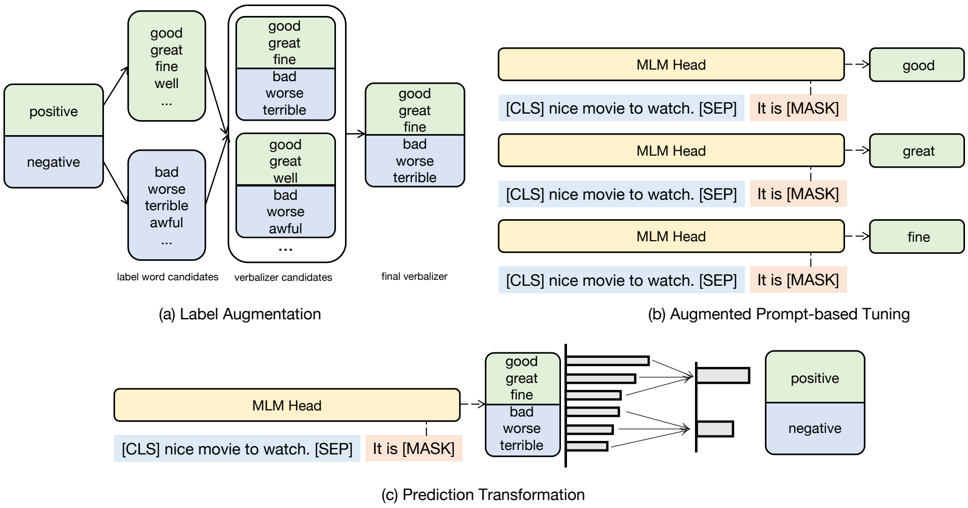


Figure 1: The proposed framework PROMPTDA for few-shot learning (with a sentiment classification task as an example): (a) the label augmentation derive multiple label words to enrich the label semantic space; (b) the augmented prompt-based tuning is trained with the augmented data using masked language modeling; and (c) the prediction transformation predicts the target labels with inferred label words.

multiple candidates have the same accuracy prediction score, we randomly select one as the final multiple-to-one verbalizer. Otherwise, we select the candidate with highest accuracy score. Note that m and n are both hyperparameters and can be adjusted according to different datasets.

4.2 Augmented Prompt-based Tuning

To enrich the training data for few-shot text classification, it is natural to utilize data augmentation methods such as token-level or sentence-level augmentations for fine-tuning (Chen et al., 2021). Most of existing data augmentation methods are focusing on enlarging training data conditioned on the original label space. Orthogonal to previous augmentation methods, our method incorporates label semantic information into prompt-tuning via augmenting sample-label pairs rather than only augmenting samples. For $(x, y) \in \mathcal{D}_{\text{train}}$, we have obtained the corresponding label word set $\mathcal{V}_y = \{v_y^1, v_y^2, ..., v_y^{k_y}\}$. Then we can include $\{(x, v_y^1), (x, v_y^2), ..., (x, v_y^{k_y})\}$ for augmentation. Let $\tilde{\mathcal{D}}_{\text{train}}$ denotes the augmented dataset. The resultant dataset can be denoted as follows,

$$\tilde{\mathcal{D}}_{\text{train}} = \cup_{(x,y)\in\mathcal{D}_{\text{train}}} \{(x, v_y^1), (x, v_y^2), ..., (x, v_y^{k_y})\} \tag{3}$$

In the tuning process, we follow the MLM training paradigm and minimize the negative log-likelihood on the whole training set $\tilde{\mathcal{D}}_{\text{train}}$. The optimization objective is:

$$\mathcal{L} = \sum_{(x,v)\in\tilde{\mathcal{D}}_{\text{train}}} -\log \Pr(v \mid x) \tag{4}$$

For $(x, v) \in \tilde{\mathcal{D}}_{\text{train}}$, the conditional probability of filling the position of `[mask]` with v is calculated as:

$$\begin{aligned} \Pr(v \mid x) &= \Pr(\texttt{Po([mask])} = v \mid x) \\ &= \frac{\exp\left(\mathbf{w}_v \cdot \mathbf{h}_{\texttt{[MASK]}}\right)}{\sum_{v'\in\mathcal{V}} \exp\left(\mathbf{w}_{v'} \cdot \mathbf{h}_{\texttt{[MASK]}}\right)} \end{aligned} \tag{5}$$

where $\mathbf{w}_v$ denotes the pre-softmax output vector for each token v in the vocabulary, and $\mathbf{h}_{\texttt{[MASK]}}$ denotes the corresponding hidden state of the `[MASK]` position. Note that we completely reuse the PLM and do not introduce new parameters in the training process, which is important for prompt-based tuning paradigm to be effective in few-shot scenarios.

4.3 Prediction Transformation

We have demonstrated the process of training the MLM classifier head with the augmented data in prompt-based tuning paradigm. Next, we describe how to perform the inference of the target label. Let h denotes the function transforms the probability scores on the label word set $\mathcal{V}_y = \{v_y^1, v_y^2, ..., v_y^{k_y}\}$ into the probability score of each category y . For instance, h can be $max()$ or $average()$. We use

$h = max()$ in our framework. The probability score of each class y can be calculated as:

$$\Pr(y \mid x) = h(\mathrm{P}(v_y^1, x), \mathrm{P}(v_y^2, x), ..., \mathrm{P}(v_y^{k_y}, x)) \tag{6}$$

where for (x, v_y^i) that satisfies $(x, v_y^i) \in \tilde{\mathcal{D}}_{\text{train}}$ and $v_y^i \in \mathcal{V}_y, (i = 1, 2, ..., k_y)$, $\mathrm{P}(v_y^i, x)$ is denoted as the conditional probability of filling the position of `[mask]` with v_y^i:

$$\mathrm{P}(v_y^i, x) = \Pr(\texttt{Po([mask])} = v_y^i \mid x) \tag{7}$$

After we obtain the probability score over each category, the final predicted label $\hat{y}$ is calculated as:

$$\hat{y} = \mathrm{argmax}_{y \in \mathcal{Y}} \Pr(y \mid x) \tag{8}$$

5 Experiments

In this section, we present the experiments to evaluate the effectiveness of the proposed PROMPTDA. Specifically, we aim to answer the following research questions:

- **RQ1** Can PROMPTDA improve the performance of few-shot prompt-based tuning?
- **RQ2** Can the proposed data augmentation strategy in PROMPTDA help for the target label prediction?
- **RQ3** Can the PROMPTDA make the prompt-based tuning method more stable?

5.1 Experimental Settings

Datasets. We evaluate the proposed framework on various few shot text classification datasets from the widely-used NLU benchmark GLUE (Wang et al., 2018) including SST-2, MR (Pang and Lee, 2005), CR, and Subj (Pang and Lee, 2004). These datasets covers different tasks such as sentiment analysis, topic classification, from various domains including movie reviews, news pieces, etc. The statistics of the datasets are shown in Table 4 in Appendix.

Baselines. We compare the proposed approach with the following state-of-the-art methods:

- **Majority**: The label is predicted by taking the majority of the class in the training set.
- **Fine-Tuning**: The prediction is based on the pre-trained language model that is fine-tuned with the training data.
- **LM-BFF** (Gao et al., 2021): Prompt-based Tuning with automatic generation of demonstration with templates.
- **Prompt Tuning**: The standard Prompt-based Tuning augmented by a simple template or template-free.

Evaluation setting. Evaluation is critical in few-shot scenarios because small changes of training dataset can result in a large variance in performance in test dataset. Following the few-shot setting in (Perez et al., 2021), (Zhang et al., 2020b), (Gu et al., 2021) and (Gao et al., 2021), we randomly select K-shot samples from original dataset for each class to construct the training set $\mathcal{D}_{\text{train}}$. and select another K-shot samples to construct the development set $\mathcal{D}_{\text{val}}$. For enhancing the stability of evaluation, we utilize the whole development set of original dataset as out test set $\mathcal{D}_{\text{test}}$ and change the random seed of sampling $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$ for five times. We select RoBERTa-large as our backbone model to make fair comparison with baselines like LM-BFF (Gao et al., 2021).

5.2 Experimental Results

In this section, we present our main results, and address the aforementioned research questions pertaining to our PROMPTDA approach.

In addition to comparing with baselines such as Majority, normal fine tuning and prompt-based method LM-BFF, we conduct more experiments to verify the effectiveness of our proposed method PROMPTDA as a plug-in module. Because different template choices can result in a large variance of performance (Gao et al., 2021), we design two groups of experiments, namely template-free and template-augmented, to show that our method can improve over standard prompt-based tuning method regardless of template design. For the template-augmented group of experiments, we manually choose "`It is [MASK]`" as the template, following (Wang et al., 2021). For the template-free group of experiments, we only append "`[MASK]`" in the input. We report the results of PROMPTDA in Table 3 when the size of data augmentation is $\times 3$ (i.e., $k_y = 3$). We also consider two scenarios that the label words are derived manually or automatically with our label augmentation mechanism. We choose 8 samples ($K = 8$) per class as the few-shot setting of our main experiments. For fair comparison, we choose the same

	SST-2	MR	CR	Subj
Majority	50.9	50.0	50.0	50.0
Fine-Tuning (full)	95.0	90.8	89.4	97.0
LM-BFF (*K=16*)§	92.3 (1.0)	85.5 (2.8)	89.0 (1.4)	91.2 (1.1)
Few-shot scenario with K=8				
Fine-Tuning	60.5 (3.1)	60.3 (7.5)	61.9 (5.1)	78.3 (8.2)
LM-BFF¶	79.9 (6.0)	85.4 (3.9)	88.6 (2.3)	81.6 (6.2)
Prompt Tuning‡	82.0 (3.2)	83.0 (3.7)	86.5 (3.0)	85.8 (6.4)
+ PROMPTDA(m.)‡	87.3 (4.4)	83.5 (2.1)	88.1 (2.7)	82.9 (4.2)
+ PROMPTDA(au.)‡	87.6 (4.1)	79.5 (2.3)	**89.8 (1.5)**	86.6 (3.6)
Prompt Tuning†	89.0 (2.2)	83.1 (3.2)	86.2 (3.2)	82.2 (8.8)
+ PROMPTDA(m.)†	89.8 (0.6)	83.8 (3.0)	85.9 (2.1)	**87.5 (3.4)**
+ PROMPTDA(au.)†	**89.9 (2.2)**	**84.4 (2.6)**	88.8 (2.4)	85.7 (3.9)

Table 1: The main results using RoBERTa-large on representative NLU tasks. All the results are evaluated on full dev sets and averaged across 5 different training sets. $K = 8$: 8 samples per class for the experiments; †: template augmented; ‡: template-free; m.: manual label augmentation; au.: automatic label augmentation; §: result from (Gao et al., 2021); ¶: results from (Wang et al., 2021).

random seed of training set sampling as LM-BFF. We train for 10 epochs for each dataset following (Wang et al., 2021). We report the average performance and standard variance of our result over five runs of sampling for each dataset. The main results can be seen in Table 3.

Performance analysis We analyze the performance from three perspectives to answer the aforementioned research questions.

To answer **RQ1**, we compare the proposed method with existing baselines. In general, we can see that the standard prompt-based tuning method with PROMPTDA consistently perform better than or is comparable with baselines like LM-BFF and normal fine tuning (results of "+ PROMPTDA (au.)" in Table 3). Compared with normal fine tuning method, standard prompt-based tuning with our proposed method PROMPTDA performs much better over all the datasets. For example, regardless of template, our method has achieved around 20% improvements on SST-2, MR and CR over normal fine tuning method. Compared with LM-BFF, our method also performs better on SST-2, CR and Subj. On MR dataset, our method perform slightly worse, which could be attributed to different choices of template. Our manually designed template may not be optimal compared to the automatically searched template in LM-BFF.

In addition, PROMPTDA consistently improve over standard prompt-based tuning method regardless of automatic label word selection or manual label word selection, tuning augmented with template or not (two group of results, ‡ and † in Table 3). Whether augmented with template or not, our method can achieve at least 1-2% gain compared to standard prompt-based tuning, which verify that our method PROMPTDA has no relation with template design and can be used as a plug-in module for improving performance. Even using the manually chosen label words for guiding data augmentation, we can still achieve better performance than standard prompt-based tuning.

To answer **RQ2**, we perform an ablation study of PROMPTDA, and compare the results of "+ PROMPTDA (m.)" and "+ PROMPTDA (au.)" in Table 3. We can see that regardless of template design, our proposed method for automatically searching label words generally perform better than manually searching label words. We analyze the reason from two perspectives. First, we hypothesize that human bias may hinder selecting optimal label words and our proposed automatic method rely on language model itself and can minimize human bias. Second, it may be easier for human to select similar words with label name "`positive, negative`" as label words for sentiment-related datasets, but it is hard to select semantically similar words as label words for tasks in other domains. For example, it is hard to manually identify semantically similar words as label words for Subj dataset with label name "`subjective, objective`", which illustrates the necessity of our proposed automatic method for searching label words.

To answer **RQ3**, we analyze the stability of performances of PROMPTDA. In general, we ob-

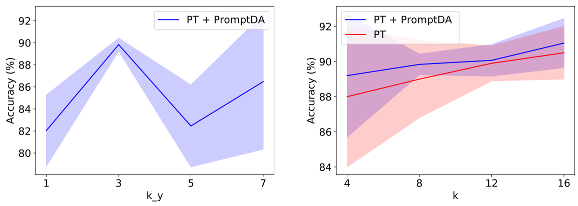
SST-2	label name	`positive | negative`
	label words (au.)	`wonderful, brilliant, fantastic | terrible done disappointing`
		`brilliant, amazing, wonderful | not, awful, terrible`
		`great, perfect, brilliant | terrible, disappointing, bad`
	label words (m.)	`positive, great, good | negative, terrible, bad`
Subj	label name	`objective | subjective`
	label words (au.)	`amazing, all, disturbing | ridiculous, important, wrong`
		`amazing, life, real | all, not, ridiculous`
		`all, life, significant | true, awesome,brilliant`
	label words (m.)	`objective, real, actual | subjective, individual, personal`

Table 2: An illustration of the label words searched automatically or manually on SST-2 and Subj datasets.

serve that PROMPTDA generally reduces the variance of prompt-tuning. (Standard variance of "+ PROMPTDA (au.)" in Table 3). The uncertainty of prompt-based tuning methods mainly comes from different distribution of small training set, different designs of template and different selections of label words for each class. Compared with LM-BFF and normal fine tuning methods, our method generally reduces the variance of prediction. For example, the standard variance of prediction over five runs for "+ PROMPTDA (au.)† " has decreased around 50% on SST-2 and Subj compared to LM-BFF and has decreased 60% on MR compared with normal fine tuning. Compared with standard prompt-based tuning method, PROMPTDA can improve the stability of tuning on most of the datasets.

5.3 Analysis of Label Word Selection

Without loss of generosity, we take dataset SST-2 and Subj for example to analyze the quality of label augmentation (see results in Table 2). *The goal of label augmentation is to find semantically similar words to enrich the label space.* With regards to the manual way, we find the synonyms of label name from dictionary as the label words and ensure these words are in the vocabulary. And we select the same label words for different seeds. With regards to our proposed automatic method, we only rely on the training set and language model (e.g., RoBERTa-large) to find the semantically similar words from vocabulary and do not rely on label name itself. The table shows the automatically searched label words for three different sampling seeds on dataset SST-2 and Subj respectively. For sentiment related datasets like SST-2 with the label name {`positive/negative` }, the label words automatically searched are literally similar to the manually selected label words, which probably means the way language models like RoBERTa-large reason about what are similar words is close



(a) # size of augmentation (b) # samples per class

Figure 2: The impact analysis of the size of label words and training samples per class.

to the human way in sentiment domain. Nonetheless, for other datasets like Subj with the label name {`objective/subjective` }, it is interesting to observe that the label words automatically searched are not literally similar to label name or manually selected label words, which may infer that the way language models like RoBERTa-large reason about what are similar words is different from the human way in other domains. We argue that *how to define word similarity in label semantic space* needs more research in the future.

5.4 Assessment of Data Augmentation

We analyze data augmentation from three perspectives including the size of data augmentation, the size of training set and combination with conventional data augmentation method.

The size of data augmentation We choose to study the effect of the size of PROMPTDA with manual label selection on template-free prompt-based tuning. The results over five different sampling seeds for 10 epochs are presented in Figure 2 (a). We can observe that PROMPTDA can generally improve over prompt-based tuning regardless of the size of augmentation. However, larger augmentation may result in more unstable final prediction. We analyze the reason from two perspectives. First, larger data augmentation may contain larger noise.

	SST-2	MR	CR	Subj
Few-shot scenario with K=8				
Prompt Tuning	89.0 (2.2)	83.1 (3.2)	86.2 (3.2)	82.2 (8.8)
Prompt Tuning with Conventional DA	89.2 (1.3)	83.6 (3.4)	86.5 (4.5)	82.6 (5.2)
Prompt Tuning with PROMPTDA	89.9 (2.2)	84.4 (2.6)	88.8 (2.4)	85.7 (3.9)
Prompt Tuning with PROMPTDA & Conventional DA	**90.8 (2.0)**	**85.0 (1.9)**	**89.3 (3.1)**	**86.5 (3.1)**

Table 3: The main results of evaluating PROMPTDA and conventional DA method on NLU tasks. All the results are evaluated on full dev sets and averaged across 5 different training sets. $K = 8$: 8 samples per class for the experiments. Conventional DA refers to *synonym substitution*.

Considering we utilize multiple-to-one verbalizer to guide data augmentation, the size of data augmentation is equal to the number of label words per class. Because we manually choose the label words for each class here, unsuitable label word choice may worsen the performance and increase the variance of final prediction. Second, a larger number of label words per class may cause the model harder to converge on small training sets. When training for the same epochs, prompt-based tuning with more label words per class may perform more unstable.

The size of training set We study the effect of the size of training set on prompt-based tuning with PROMPTDA and standard prompt-based tuning. The size of data augmentation is $\times 3$. The results over five different sampling seeds for 10 epochs are presented in Figure 2 (b). We have several observations from the results. First, Our method PROMPTDA consistently improves over standard prompt-based tuning regardless of the size of training sets. Second, our proposed method generally decreases the variance of prompt-based tuning. Third, the improvement space of PROMPTDA over prompt-based tuning decreases as the number of samples per class increases.

Combination with conventional DA Although conventional data augmentation methods are still effective when training data is limited (Chen et al., 2021), previous works verified that they can bring marginal improvement for prompt-based tuning methods (Zhou et al., 2021). It is interesting to explore whether or not our proposed method PROMPTDA can complement with conventional DA methods for further enhancing the performance of prompt-based tuning paradigm.

We follow the same setting with the main experiments and test conventional DA, PROMPTDA and the combination of PROMPTDA and conventional DA on standard prompt-based tuning paradigm with template. With regards to conventional DA, we select *synonym substitution* method from nlpaug toolkit (Ma, 2019) and enlarge the training set by $\times 2$. With regards to our proposed PROMPTDA, we enlarge the training set by $\times 3$. The experiment results over five different sampling seeds for 10 epochs are shown in Table 3.

We can observe that the combination of PROMPTDA and Conventional DA method consistently outperform only using PROMPTDA or Conventional DA method. Conventional DA methods mostly focus on exploiting the semantic information of the instance itself. Our method proposes to utilize label semantic information to guide data augmentation and does not change instances. Therefore, our proposed method PROMPTDA can be regarded orthogonal to conventional DA methods to some extent and complement with each other.

6 Conclusion and Future Work

In this paper, we study a new problem of data augmentation in prompt-based tuning for few shot learners. To leverage the label semantic information, we propose a novel label-guided data augmentation approach PROMPTDA, which can derive multiple label words and exploit the rich semantic information of the label words into masked language modeling. We conduct extensive experiments in various benchmark datasets and demonstrate the effectiveness of PROMPTDA for few shot learning, with fine-grained analysis on the effects of different base language models, size of label words, and manual/automatic label augmentation.

There are several interesting directions for future work. First, we will extend PROMPTDA to multi-label few shot tasks and leverage multi-aspect label space. Second, we will explore prompt-based data augmentation for token-level NLU tasks such as few shot name entity recognition (NER). Third, we will explore prompt-base tuning to enhance its interpretability capacity for various NLP tasks.

References

Trapit Bansal, Rishikesh Jha, and Andrew McCallum. 2019. Learning to few-shot learn across diverse natural language classification tasks. *arXiv preprint arXiv:1911.03863*.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp. *arXiv preprint arXiv:2106.07499*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

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

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*.

Hongyu Guo. 2020. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4044–4051.

Aakriti Gupta, Kapil Thadani, and Neil O'Hare. 2020. Effective few-shot classification with transfer learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1061–1066.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Ju-Hyoung Lee, Sang-Ki Ko, and Yo-Sub Han. 2021. Salnet: Semi-supervised few-shot text classification with attention-based lexicon construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13189–13197.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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*.

Qiaoyang Luo, Lingqiao Liu, Yuhao Lin, and Wei Zhang. 2021. Don't miss the labels: Label-semantic augmented meta-learner for few-shot text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2773–2782.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. *arXiv preprint arXiv:2010.07245*.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*, pages 617–628.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. Spartqa:: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems*, 33.

Subhabrata Mukherjee, Xiaodong Liu, Guoqing Zheng, Saghar Hosseini, Hao Cheng, Greg Yang, Christopher Meek, Ahmed Hassan Awadallah, and Jianfeng Gao. 2021. Clues: Few-shot learning evaluation in natural language understanding. *arXiv preprint arXiv:2111.02570*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *arXiv preprint arXiv:2105.11447*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Rogerio Feris, Abhishek Kumar, Raja Giryes, and Alex M Bronstein. 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *arXiv preprint arXiv:1806.04734*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Kai Shu, Subhabrata Mukherjee, Guoqing Zheng, Ahmed Hassan Awadallah, Milad Shokouhi, and Susan Dumais. 2020a. Learning with weak supervision for email intent detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1051–1060.

Kai Shu, Suhang Wang, Thai Le, Dongwon Lee, and Huan Liu. 2018. Deep headline generation for clickbait detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 467–476. IEEE.

Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. 2020b. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732*.

Russell Stewart and Stefano Ermon. 2017. Label-free supervision of neural networks with physics and domain knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Dianbo Sui, Yubo Chen, Binjie Mao, Delai Qiu, Kang Liu, and Jun Zhao. 2020. Knowledge guided metric learning for few-shot text classification. *arXiv preprint arXiv:2004.01907*.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. 2021. Metaxl: Meta representation transformation for low-resource cross-lingual learning. *arXiv preprint arXiv:2104.07908*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey. *arXiv preprint arXiv:2007.09604*.

Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. 2020. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020b. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

Jing Zhou, Yanan Zheng, Jie Tang, Jian Li, and Zhilin Yang. 2021. Flipda: Effective and robust data augmentation for few-shot learning. *arXiv preprint arXiv:2108.06332*.

A Appendix

A.1 Limitations and Risks

We only explore the prompt-based paradigm for the text classification task in this work and do not consider more NLU applications like Natural Language Inference. We only conduct experiments on English data in this work. We may consider multi-lingual scenarios as our next step.

A.2 Implementation Details

We implement our model and all baselines with PyTorch and run each experiment on a single NVIDIA GPU. The hyperparameters are the same for all methods based on RoBERTa-large (the learning rate is 3e-6, the batch size is 4, the number of training epochs is 10).

A.3 Dataset Statistics

Dataset	SST-2	MR	CR	Subj
Task	sentiment	sentiment	sentiment	subjectivity
Domain	movie			news
# Classes	2	2	2	2
# Train ($\|D_{train}\|$)	67349	8662	1775	8000
# Test ($\|D_{test}\|$)	872	2000	2000	2000

Table 4: Statistics of selected few shot text classification tasks.