SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

LPT: Language-agnostic Prompt Tuning of Pre-Trained Cross-Lingual Language Model

Yixin Ji[1], Yue Wang[1], Zecheng Tang[1], Lijun Wu[2], Juntao Li[1*] & Min Zhang[1,3]

[1]*Soochow University, Suzhou* 215021*, China;*
[2]*Microsoft Research Asia, Beijing* 100080*, China;*
[3]*Harbin Institute of Technology, Shenzhen* 518055*, China*

Abstract Recently, prompt-based learning has achieved remarkable and stable performance on various downstream tasks, especially in the few-shot settings. In prompt-based learning, prompt engineering is a critical factor that primarily affects model performance. It requires a lot of human effort to design prompts manually or training data to search prompts automatically. Thus, it is expensive to manually create language-specific prompts for each language, while automatically obtaining prompts is only applicable for high-resource languages. Both of which are unrealistic for thousands of low-resource languages. As a result, how to automatically obtain high-quality prompts for target languages has become a central issue but is still under-explored. To fill this blank, we propose a novel **L**anguage-Agnostic **P**rompt **T**uning (**LPT**) method, including head prompt and multilingual verbalizers, to induce language-agnostic representation in multilingual pre-trained language models and obtain general prompts for various target languages. Extensive experiments conducted on the Multilingual Amazon dataset demonstrate that our proposed method achieves universal improvement on three different target languages under different settings, i.e., 16-shot, 32-shot, and 64-shot. Our code is available at https://github.com/Dereck0602/LPT.

Keywords Prompt learning, cross-lingual, few-shot setting, head prompt, multilingual verbalizer

Citation Yixin Ji, Yue Wang, Zecheng Tang, et al. LPT: Language-agnostic Prompt Tuning. Sci China Inf Sci, for review

1 Introduction

In the past few years, pre-trained language models (PLMs) [11, 19, 22, 26, 28] have achieved great success on various Natural Language Processing (NLP) tasks through fine-tuning on downstream data. However, in the conventional fine-tuning paradigm of using PLMs, there is a considerable gap between pre-training and fine-tuning in data size, training objectives, etc., which may cause poor and unstable performance in low-resource settings. To address this problem, prompt-based learning [1, 28–30, 32, 33, 36, 37], which reformulates the downstream tasks to be closer to the format of pre-training tasks, has attracted the attention of many researchers. Due to the reduced gap between pre-training and fine-tuning, prompt-based learning can achieve comparable performance to fully fine-tuned PLM in the few-shot setting.

In prompt-based learning, prompt engineering is a crucial step that primarily affects model performance. It requires a lot of human effort to design prompts manually or training data to search prompts automatically [31–33, 47, 48]. Currently, researches on prompt engineering are mainly focused on high-resource languages, such as English [31–33, 47, 48] and Chinese [4, 5]. Considering that humans have thousands of languages, many low-resource languages still fail to benefit from the promising prompt-based learning. For multilingual or cross-lingual tasks, high-quality language-specific handcrafted prompts are expensive to obtain because they need experts familiar with both prompt engineering and different languages. Besides, there is no sufficient training data to access in the target language in cross-lingual tasks, which leads to automatically optimized prompts for target language-specific unavailable. Due to this limitation, prompt-based learning is under-explored in the multilingual and cross-lingual settings [25].

* Corresponding author (email: ljt@suda.edu.cn)

Though prompt engineering is challenging in cross-lingual tasks, Zhao et al. [25] demonstrate the potential transferability of prompts for the source language. They also find that discrete prompts have better transferability than soft prompts. Therefore, *How to find prompt with language transferability for various target languages* is a crucial problem to be explored in cross-lingual tasks.

Considering that most of the cross-lingual tasks rely on the generalization capability of multi-lingual pre-trained language models, it is natural and straightforward to reuse the prompts from high-resource languages or translate these prompts into the target language. For reusing the source languages, we simply reuse the source language prompt and verbalizers to perform inference for the target language. As for the translation-based method, we translate both the source language manual prompts and verbalizers into the target language for inference. Note that we only utilize a few supervised samples in the source language to fine-tune the multi-lingual pre-trained language models in the paradigm of few-shot prompt learning. Through experiments, we find that reusing manual prompts in the source language is feasible, but it suffers from severe instability. For the translation-based method, it is worth trying in languages close to the source language, e.g., translating prompts and verbalizers from English into French or German, for better and more stable performance than reusing. In contrast, the translation-based method will cause a negative effect on languages remote from the source language in terms of linguistic typology. Though translated prompts reduce the gap between prompts and target language, they widen the gap between the training and inference stages because the fine-tuning is launched on the source language prompt and training pairs, while the inference stage utilizes the translated prompts to compute the target language outputs. Besides, the translated prompts also ignore the transferable language-agnostic information.

To address the above-mentioned few-shot cross-lingual transfer task, we propose a novel prompt-based cross-lingual method, which incorporates a language-agnostic prompting strategy and multilingual verbalizers. For the language-agnostic prompting, we use tokens that do not contain any specific linguistic meaning as prompts, such as null prompt [46]. To further enhance the transferability of the language-agnostic prompt, we propose the head prompt, which no longer needs to insert any tokens and reuse the [CLS] for prediction. Thus, the language-agnostic prompt does not require any human experts or large amounts of training data to search good prompt template. To balance the language-specific and language-agnostic information, we use multilingual verbalizers to provide weak target language supervision signals. Through the multilingual verbalizers, the model can be fine-tuned to fit the label space of the source language and the label space of the target languages simultaneously. We conduct experiments on the Multilingual Amazon dataset in 16-shot, 32-shot, and 64-shot settings. We choose English as the source language and German, French, and Japanese as target languages. Experiment results on all target languages and all settings confirm the effectiveness of our proposed method.

In a nutshell, this study has the following contributions.

- We propose a novel language-agnostic prompt-based learning method **LPT**, consisting of the head prompt and the multilingual verbalizers. These strategies are simple but effective in cross-lingual transfer tasks under the few-shot settings.
- Experimental results indicate that our proposed head prompt along with multilingual verbalizers achieves a promising and even the best performance on both 16-shot, 32-shot and 64-shot settings.
- We intuitively show that our head prompt and multilingual verbalizers can induce language-agnostic features from the multilingual pre-trained model.

2 Related Work

2.1 Prompt-Based Learning

With the development of prompt-based learning [28–30], PLMs have recently achieved remarkable performance in the few-shot setting. We take a simple sentiment classification case, for example, to illustrate prompt-based learning. Given a sentence *"This movie is great!"* as input, prompt-based learning first appends a prompt to the sentence, which changes the input to *"This movie is great! The sentiment of this sentence is [MASK]."* Then, the expected classification outputs are extracted from the word prediction probability of the MLM head at the position of [MASK]. More concretely, the output score of each label is obtained from the prediction probability of the corresponded label words (e.g., 'positive', 'negative'), where the label words can be manually designed or automatically extracted from data.

Previous works find that different prompts have a significant influence on the prediction accuracy

of models [29, 30]. To construct better prompts for improving the performance of pre-trained models, many researchers [27–30] design prompts manually to deal with different downstream tasks, including knowledge probing, question answering, text classification, etc. Since the handcrafted prompts need lots of human efforts and may fail to be optimal [47], many works search prompts automatically, which can be divided into two categories, i.e., *discrete prompts* [31–33, 47, 48] and *continuous prompts* [1, 34–37, 49–51]. Besides the above two types of prompts, Logan IV et al. [46] further propose the null prompt, which only consists of a mask token without any task-specific prompt. Surprisingly, the simple null prompt can achieve comparable performance to handcrafted prompts. Unlike previous works that mainly focus on monolingual settings, we propose to explore prompt-based tuning methods for cross-lingual transfer.

2.2 Cross-Lingual Transfer Learning

With the significant progress in multilingual pre-trained language models (mPLMs), e.g., mBERT [11], mT5 [6], XLM [17], XLM-R [12], and XLM-E [23], few-shot and even zero-shot cross-lingual transfer achieve surprising performance in various NLP tasks [13–16, 24, 39]. Through exploring the potential and limitations of pre-trained language models, a series of effective and efficient adaptation methods for cross-lingual transfer are proposed. Ye et al. [18] adapt the features of pre-trained models to new domains or languages without fine-tuning. Pfeiffer et al. [40] propose an adapter-based [21] framework that enables parameter-efficient transfer to arbitrary tasks and languages by learning modular language and task representations. Recently, Zhao et al. [25] demonstrate the capability of discrete and soft prompting in cross-lingual transfer scenarios. Since mPLMs are trained without using parallel data, language representations are not correctly aligned in mPLMs [41]. To further improve the transferability of mPLMs, researchers have proposed many methods to align representations and induce language-agnostic representations [20, 42–45]. Unlike previous works, we are the first to explore language-agnostic prompts in cross-lingual transfer to our best knowledge.

3 Method

3.1 Prompting

Manually Designed Prompt. Following Schick et al. [29], we convert the sentiment classification task into a cloze task using manually designed prompts. Specifically, we use four manual templates for each dataset which are either introduced by Schick et al. [29] or tailored to fit the dataset [58] in Table 1.

ID	Template
$T_1(\mathbf{x})$	It was [MASK] . $\mathbf{x}$
$T_2(\mathbf{x})$	Just [MASK] ! $\mathbf{x}$
$T_3(\mathbf{x})$	$\mathbf{x}$ All in all, it was [MASK].
$T_4(\mathbf{x})$	$\mathbf{x}$ In summary, it was [MASK].

Table 1 The templates of manually designed prompts.

In these templates, $\mathbf{x}$ is a sentence sampled from the dataset, and the [MASK] is a special token to request the pretrained model to fill with words in the preset vocabulary $\mathcal{V}$, called verbalizer. Although recent studies [59,60] reveal that the model performance is dependent on the choice of verbalizers and can benefit from the ensemble of verbalizers, we only utilize a fixed verbalizer of source language under the zero-shot cross-lingual transfer setting to simplify experiments and demonstrate the effectiveness of our proposed method. Concretely, for sentiment classification task, $\mathcal{V} = \{`negative', `positive'\}$, we define a map $v : \mathcal{Y} \mapsto \mathcal{V}$, which maps label '0' to 'negative' and '1' to 'positive'. The optimization objective is to minimize the cross-entropy loss between the predicted and the golden label words, formulated as

$$-\underset{\theta,\phi}{argmin} \sum_{(x,y)} logp([MASK] = v(y)|g(Encoder(T(x))); \theta, \phi) \tag{1}$$

where the *Encoder* is a pre-trained multilingual encoder with parameters θ, and $g(\cdot)$ is the masked language model (MLM) head with parameters ϕ. If the probability of 'positive' is larger than 'negative', we classify the instance into label '1'.

Null Prompt. To avoid prompt engineering in few-shot prompt-based finetuning, Logan IV et al. [46] propose the null prompt method, which simply concatenates inputs and the [MASK] token. In our view, different from a manually designed prompt, the null prompt is language-agnostic because it does not contain any tokens with specific linguistic meaning. In our experiments, we insert the [mask] token in the tail of a sentence:

$$T(\mathbf{x}) = \mathbf{x}\ \text{[MASK]}$$

The effectiveness of the null prompt in monolingual settings indicates that prompt templates may not be as important as expected in prompt-based tuning. Meanwhile, it also breaks the prevailing view that the success of few-shot learning is due to inductive bias in prompt. Thus, in this paper, we pay more attention to the language-agnostic prompt and explore the transferability potential of null prompt in zero-shot cross-lingual transfer settings.

3.2 Language-Specific Prompting

For manually designed prompts, if we use the same prompt in both the training and inference stages in multilingual tasks, a linguistic gap will exist between prompts and sentences. To mitigate the gap, we translate the prompt and verbalizer of the source language into the target language, called language-specific prompts. Taking French as an example, we translate manually designed prompts of English into French, e.g., *"It was [MASK] .***x** $\rightarrow$ *C'était [MASK] .* **x***"*. Accordingly, the verbalizer should also be translated, i.e., *"negative* $\rightarrow$ *négatif"*, *"positive* $\rightarrow$ *positif"*. Table 2 presents more examples of translated language-specific prompts and verbalizers. As demonstrated by Zhao et al. [25], language-specific prompts and verbalizers are beneficial to in-language results for each language, where the model is fine-tuned on in-language labeled data and prompts translated from high-resource languages. In contrast, we mainly explore the effectiveness of language-specific prompts for the unsupervised cross-lingual transfer setting, i.e., the model is trained on the source language prompts and data but performs inference in the target language with translated language-specific prompts.

Language	Template	Verbalizer
English	$T_1(\mathbf{x})$ = It was [MASK] . **x**	positive
	$T_2(\mathbf{x})$ = Just [MASK] ! **x**	
	$T_3(\mathbf{x})$ = **x** All in all, it was [MASK].	negative
	$T_4(\mathbf{x})$ = **x** In summary, it was [MASK].	
German	$T_1(\mathbf{x})$ = Es war [MASK] . **x**	positiv
	$T_2(\mathbf{x})$ = Einfach [MASK] ! **x**	
	$T_3(\mathbf{x})$ = **x** Alles in allem war es [MASK].	negativ
	$T_4(\mathbf{x})$ = **x** Zusammenfassend war es [MASK].	
French	$T_1(\mathbf{x})$ = C'était [MASK] . **x**	positif
	$T_2(\mathbf{x})$ = Juste [MASK] ! **x**	
	$T_3(\mathbf{x})$ = **x** Dans l'ensemble, c'était [MASK].	négatif
	$T_4(\mathbf{x})$ = **x** En résumé, c'était [MASK].	
Japanese	$T_1(\mathbf{x})$ = [MASK] でした . **x**	ポジティブ
	$T_2(\mathbf{x})$ = ただ [MASK] ! **x**	
	$T_3(\mathbf{x})$ = **x** 全体として、それは [MASK] でした.	ネガティブ
	$T_4(\mathbf{x})$ = **x** 要約すると、それは [MASK] でした.	

Table 2 Language-specific Prompts and verbalizers in English, German, French and Japanese.

3.3 Head Prompt

During the development of prompt-based tuning [3], inserting [MASK] tokens into sentences and using the representation of [MASK] token to predict seems to be a kind of standard practice. Therefore, in practice, even with as simple as the null prompt, for users, it is still inevitable to decide where to place the [MASK] token. Recently, some methods [51] have tried to break this routine. For prompt-based methods [3], e.g., discrete prompts and the null prompt, inserting [MASK] tokens into sentences and using the representation of [MASK] token to predict are prevalent in practice. Thus, it is inevitable to decide where to place the [MASK] token(s). However, in the cross-lingual setting, it is difficult to find an optimal place for inserting the [MASK] token(s) due to the diversity of different languages in syntax,

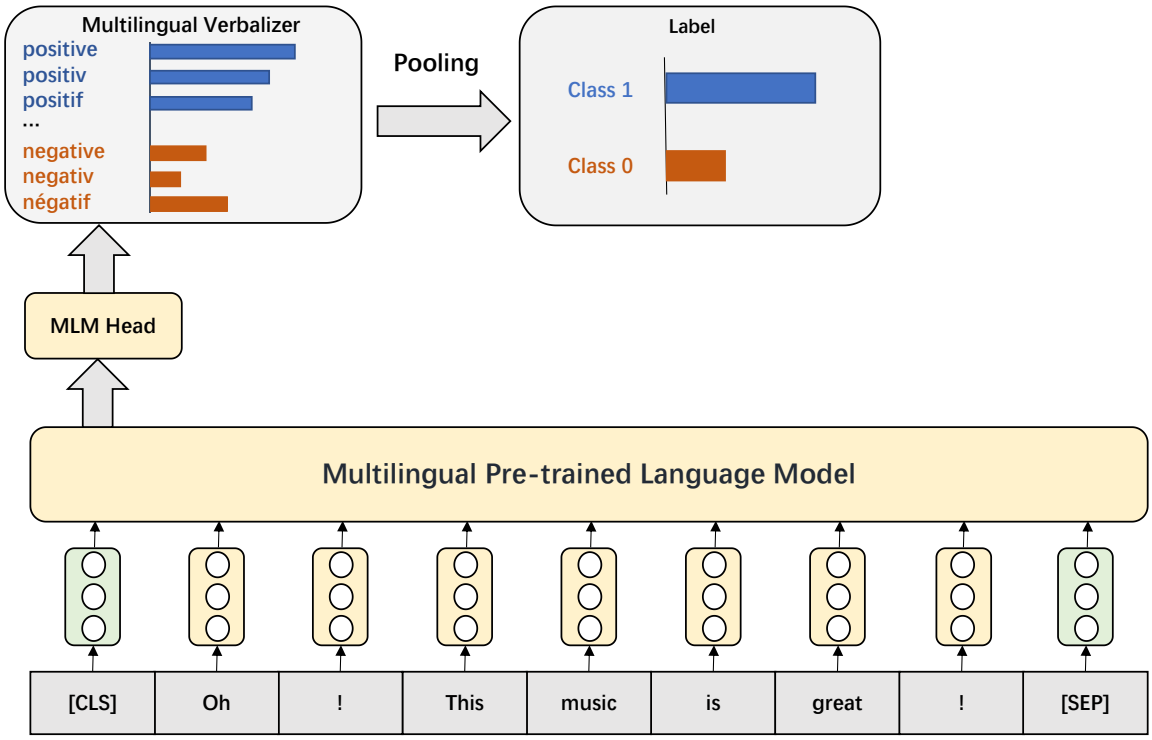


Figure 1 The overall framework of our proposed head prompting and multilingual verbalizers.

morphology, etc. Furthermore, we hope the representations used for prediction can obtain more language-independent global semantic information. Based on the above considerations, we propose a simple and effective head prompt strategy. Similar to the null prompt, the head prompt also discards language-specific prompt templates, which can be considered language-independent. As shown in Figure 1, unlike all other prompt-based tuning methods, the head prompt predicts label words upon the representation of the [CLS] token other than [MASK] tokens, which can be formulated as:

$$-\underset{\theta,\phi}{argmin}\sum_{(x,y)} logp([CLS]=v(y)|g(Encoder(x));\theta,\phi) \tag{2}$$

3.4 Multilingual Verbalizers

The language-specific prompt only considers the translated verbalizers in the evaluation stage, leading to inconsistency between the training and testing phases. We propose to exploit the multilingual verbalizers in the training stage to narrow such inconsistency.

For instance, the source language verbalizer is $\mathcal{V}_{en}$={'negative', 'positive'}, and the correlated target language verbalizers are $\mathcal{V}_{de}$={'negativ', 'positiv'}, $\mathcal{V}_{fr}$={'négatif', 'positif'}, and $\mathcal{V}_{ja}$={'ネガティブ', 'ポジティブ'}. We define the multilingual verbalizers as a set of these target language verbalizers, i.e., $\mathcal{V}$={$\mathcal{V}_{en}$,$\mathcal{V}_{de}$,$\mathcal{V}_{fr}$,$\mathcal{V}_{ja}$}. We expect the multilingual verbalizers can provide additional target language supervision to induce better multilingual representation. Formally, in the training stage, the optimization objective is to minimize the loss function between the predicted and the golden multilingual verbalizers.

$$-\underset{\theta,\phi}{argmin}\sum_{(x,y)}\frac{1}{|\mathcal{V}|}\sum_{\ell\in\mathcal{V}} logp(< x >= v_\ell(y)|g(Encoder(x));\theta,\phi) \tag{3}$$

where $< x >$ is the token for prediction. In manually designed prompts and the null prompt, $< x >$ refers to the [MASK] token, while in the head prompt, $< x >$ represents the [CLS] token.

Dataset	English			Each Other Language		
	Book	Dvd	Music	Book	Dvd	Music
#Class	2	2	2	2	2	2
shot	K	K	K	Full	Full	Full
Usage	Train/Dev	Train/Dev	Train/Dev	Test	Test	Test

Table 3 The statistical results of datasets used in experiments.

4 Experimental Setup

4.1 Dataset and Setup

Dataset. We conduct our experiments on the multilingual sentiment analysis task. This multilingual dataset [52] includes reviews in English, German, French, and Japanese, and each language contains three domains, i.e., books, DVDs, and music. For each domain in each language, there are 1,000 positive reviews and 1,000 negative reviews in the training set, validation set, and test set.

In order to simulate a realistic few-shot scenario [53,54], we randomly sample K shots per class without repetition from the training set as the few-shot training set for each domain in each language. For the few-shot validation set, we also perform sampling according to the construction strategy of the training set. For example, a 32-shot experiment uses 32 training, 32 validation shots per class, and the whole test set. The statistics of datasets are given in Table 3. We use the widely acknowledged metric for evaluation, that is, accuracy on the test set to show cross-lingual transfer performance.

Setup. All our experiments are conducted on the pre-trained XLM-RoBERTa-base model [12] and rely on the Pytorch-based [55] HuggingFace Transformers repository [10] [1)].

We choose English as the source language and others as the target languages. Our experiments explore the 16-shot, 32-shot, and 64-shot settings. Following [25, 56, 57], we train the model with 50 epochs in the source language and select the checkpoint which performs best on the English validation set. Given the instability problem of few-shot learning, we repeat the experiments with 5 random seeds and report the average of 5 runs as the final results, along with the standard deviation of 5 runs.

4.2 Training Details

The pre-trained XLM-RoBERTa-base model contains 270M parameters, which is trained on 2.5 TB CommonCrawl data in 100 languages. Its vocabulary contains 250k tokens. The architecture of XLM-RoBERTa-base is the same as RoBERTa-base. We use the batch size of 32, a learning rate of 5e-5 for all experiments, and leverage a single RTX 3090 GPU for training. For all experiments, the random seeds we used are ({5,10,15,20,42}). More hyperparameters values of the training stage are listed in Table 4.

Hyperparameters	
Number of epochs	50
Batch Size	32
Max Sentence Length	512
Optimizer	Adam [8]
Adam setting	(β_1=0.9, β_2=0.98)
Learning rate	5e-5
Warmup step	200
Dropout ratio	0.1

Table 4 Hyper-parameter values of the training stage.

5 Results and Analysis

Table 5 and 6 summarize the overall experiment results of baseline models and our proposed models under 16-shot, 32-shot, and 64-shot settings in the source language and target languages. Although the

1) https://github.com/huggingface/transformers

Method	16-shot			Ave	32-shot			Ave	64-shot			Ave
	Book	Dvd	Music		Book	Dvd	Music		Book	Dvd	Music	
Manual Prompt	$87.41_{2.5}$	$83.10_{1.5}$	$85.01_{3.0}$	85.17	$87.63_{1.6}$	$86.05_{1.3}$	$87.65_{1.0}$	**87.11**	$90.25_{1.3}$	$88.38_{1.0}$	$88.89_{1.2}$	**89.17**
language-specific prompt	$87.41_{2.5}$	$83.10_{1.5}$	$85.01_{3.0}$	85.17	$87.63_{1.6}$	$86.05_{1.3}$	$87.65_{1.0}$	**87.11**	$90.25_{1.3}$	$88.38_{1.0}$	$88.89_{1.2}$	**89.17**
null prompt	$86.19_{0.5}$	$79.13_{1.2}$	$82.36_{4.0}$	82.56	$85.43_{1.3}$	$84.60_{1.6}$	$84.96_{1.3}$	85.00	$90.19_{0.8}$	$86.10_{1.1}$	$87.11_{1.0}$	87.80
+ multilingual verbalizers	$85.70_{0.7}$	$81.89_{2.0}$	$83.08_{1.7}$	83.56	$84.86_{1.2}$	$85.01_{0.3}$	$85.61_{1.0}$	85.16	$90.70_{0.9}$	$84.56_{0.8}$	$86.84_{0.7}$	87.37
head prompt	$87.42_{0.8}$	$83.99_{1.8}$	$84.67_{3.1}$	85.36	$87.46_{0.7}$	$84.72_{0.9}$	$86.16_{1.0}$	86.11	$91.21_{0.6}$	$87.98_{1.2}$	$87.77_{0.4}$	88.99
+ multilingual verbalizers	$86.66_{0.9}$	$85.23_{1.6}$	$85.43_{3.3}$	**85.77**	$87.15_{2.0}$	$85.95_{1.3}$	$87.06_{1.4}$	86.72	$90.25_{1.2}$	$89.13_{0.8}$	$87.78_{1.6}$	89.05

Table 5 The few-shot classification accuracy (%) results on the source language (English). Models are evaluated by 5 random runs and we report mean and standard deviation.

shot	**Method**	German			Ave	French			Ave	Japanese			Ave
		Book	Dvd	Music		Book	Dvd	Music		Book	Dvd	Music	
16	Manual Prompt	$83.63_{3.6}$	$77.79_{3.6}$	$81.94_{4.8}$	81.12	$75.33_{12.1}$	$69.73_{11.8}$	$68.66_{11.2}$	71.24	$75.95_{3.9}$	$69.78_{6.3}$	$75.18_{4.7}$	73.64
	language-specific prompt	$81.21_{6.8}$	$80.75_{2.8}$	$82.48_{4.0}$	81.48	$79.85_{8.3}$	$76.16_{2.4}$	$76.93_{7.1}$	77.64	$62.78_{6.7}$	$68.94_{12.1}$	$67.42_{13.2}$	66.38
	null prompt	$88.53_{0.5}$	$80.76_{1.2}$	$83.09_{2.4}$	84.13	$81.30_{0.9}$	$66.69_{8.9}$	$65.91_{13.7}$	71.30	$79.24_{0.6}$	$73.40_{2.6}$	$76.18_{5.8}$	76.27
	+ multilingual verbalizers	$88.30_{0.6}$	$82.61_{2.5}$	$84.50_{1.8}$	85.14	$84.93_{0.9}$	$80.05_{3.8}$	$78.52_{2.9}$	81.17	$80.39_{1.2}$	$77.47_{3.0}$	$80.41_{1.9}$	79.42
	head prompt	$87.52_{1.0}$	$85.13_{0.9}$	$80.63_{5.2}$	84.43	$87.00_{1.2}$	$79.84_{1.7}$	$72.74_{7.2}$	79.86	$83.17_{1.0}$	$73.43_{1.2}$	$78.20_{4.6}$	78.27
	+ multilingual verbalizers	$85.68_{0.7}$	$85.65_{0.3}$	$84.89_{3.4}$	**85.41**	$85.04_{1.1}$	$84.92_{0.9}$	$80.60_{5.1}$	**83.52**	$82.30_{1.4}$	$79.47_{0.9}$	$82.49_{3.8}$	**81.42**
32	Manual Prompt	$84.88_{1.4}$	$84.14_{1.7}$	$87.23_{1.6}$	85.41	$77.29_{14.6}$	$79.90_{12.9}$	$80.55_{9.1}$	79.25	$77.97_{4.1}$	$79.45_{7.0}$	$80.63_{3.6}$	79.35
	language-specific prompt	$85.66_{2.2}$	$85.55_{1.6}$	$87.08_{1.9}$	86.09	$85.63_{2.6}$	$86.86_{0.8}$	$83.79_{3.1}$	85.43	$60.28_{7.0}$	$61.03_{7.6}$	$60.40_{9.1}$	60.57
	null prompt	$86.51_{0.6}$	$85.78_{0.4}$	$87.59_{1.0}$	**86.63**	$78.96_{9.2}$	$85.84_{3.1}$	$71.37_{9.6}$	78.72	$82.43_{1.1}$	$85.17_{1.2}$	$79.90_{2.2}$	82.50
	+ multilingual verbalizers	$85.05_{0.7}$	$85.10_{0.4}$	$87.59_{0.7}$	85.91	$86.65_{0.5}$	$87.97_{0.4}$	$83.69_{1.1}$	86.10	$81.31_{0.6}$	$85.70_{0.5}$	$81.77_{0.8}$	82.93
	head prompt	$87.08_{0.7}$	$85.14_{0.9}$	$85.95_{1.1}$	86.06	$87.85_{0.7}$	$85.61_{1.5}$	$79.59_{12.6}$	84.35	$84.98_{0.8}$	$83.30_{1.7}$	$79.35_{8.7}$	82.54
	+ multilingual verbalizers	$86.20_{1.7}$	$84.89_{0.6}$	$86.91_{0.9}$	86.00	$86.42_{1.9}$	$87.29_{0.7}$	$85.39_{1.9}$	**86.37**	$83.91_{1.3}$	$84.71_{1.4}$	$83.21_{1.7}$	**83.94**
64	Manual Prompt	$85.41_{4.9}$	$85.06_{1.8}$	$89.16_{1.5}$	86.54	$81.72_{9.6}$	$81.67_{10.9}$	$80.58_{11.5}$	81.32	$79.16_{3.9}$	$81.00_{4.7}$	$84.25_{3.2}$	81.47
	language-specific prompt	$87.80_{2.2}$	$86.21_{1.1}$	$89.72_{0.8}$	87.91	$86.48_{2.3}$	$87.52_{1.3}$	$86.21_{1.2}$	86.74	$73.71_{8.9}$	$72.87_{10.8}$	$75.09_{9.4}$	73.89
	null prompt	$89.56_{1.4}$	$85.30_{1.6}$	$88.96_{0.7}$	87.94	$83.55_{2.8}$	$86.72_{1.2}$	$86.29_{1.5}$	85.52	$83.11_{2.2}$	$83.79_{1.1}$	$85.50_{0.4}$	84.13
	+ multilingual verbalizers	$91.04_{1.0}$	$83.68_{2.1}$	$88.49_{0.8}$	87.74	$87.61_{2.7}$	$85.06_{2.5}$	$86.43_{1.3}$	86.37	$84.99_{1.5}$	$81.78_{1.6}$	$84.59_{0.8}$	83.79
	head prompt	$89.95_{2.5}$	$86.39_{0.5}$	$89.46_{0.6}$	**88.60**	$86.08_{3.7}$	$87.92_{1.3}$	$86.20_{1.2}$	86.73	$84.57_{2.1}$	$85.41_{1.0}$	$86.94_{0.6}$	**85.64**
	+ multilingual verbalizers	$89.49_{1.6}$	$87.19_{0.8}$	$89.12_{1.0}$	**88.60**	$85.95_{2.8}$	$88.90_{0.7}$	$87.26_{0.7}$	**87.37**	$83.73_{1.6}$	$85.21_{0.8}$	$86.61_{0.4}$	85.18

Table 6 The few-shot classification accuracy (%) results on MultiAmazon dataset. Models are evaluated by 5 random runs and we report mean and standard deviation.

results vary under different settings, the combination of language-agnostic head prompt and multilingual verbalizers achieves good in-language performance and the best cross-lingual performance. Manually designed prompts consistently achieve the best in-language performance. It can be seen from the standard deviation of the experimental results that the performance of manual prompts and the language-specific prompt are more unstable than the null prompt and our proposed head prompt since the optimal manually designed prompts for different languages vary dramatically. And the usage of our introduced multilingual verbalizers is also helpful for the stability of zero-shot cross-lingual transfer.

5.1 The Ablation Analysis of Language-Specific Prompts

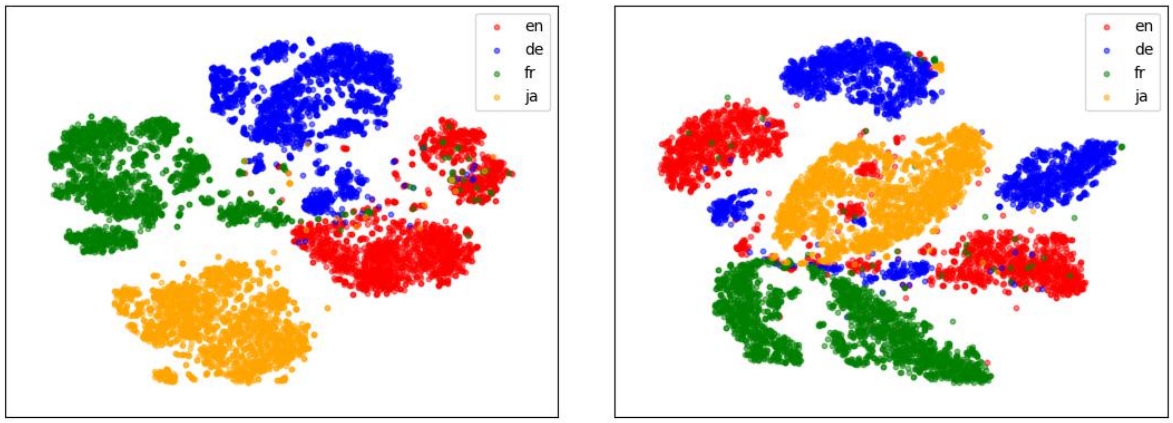
As Table 6 shows, though language-specific prompts do not perform best in our experiments, we observe that they can effectively boost the performance of languages close to the source language but meanwhile negatively affect languages distinct from the source language. To investigate the puzzling phenomenon, we launch the ablation analysis of language-specific prompts in 32-shot settings and report results in Table 7. In general, the language-specific templates under-perform the original template on average, especially in Japanese. Comparing the results without using the translated templates and verbalizers, we can find that using the translated verbalizers in the evaluation stage has considerable side effects regardless of the typological differences between the target language and source language. We attribute this to the optimization gap between training and evaluation. In the training stage, we optimize the representation to fit the source language verbalizer. In contrast, in the evaluation stage, we require the model to predict in the target language verbalizer space. Based on this analysis, we can further understand the effectiveness of our multilingual verbalizers from the perspective of optimization.

5.2 The Language-Agnostic Representation of Null Prompt and Head Prompt

To intuitively explore why the cross-lingual performance of our head prompt works better than the null prompt, we give the t-SNE plots [2] of the null prompt with multilingual verbalizers and our head prompt with multilingual verbalizers. Specifically, we fine-tune the XLM-R model with two different prompts under 16-shot settings in the DVD domain. Then, we use the fine-tuned model to extract features of two

Method	English	German	French	Japanese	Average
Manual $T_1(\mathbf{x})$	88.05	86.23	**85.37**	**82.20**	85.46
language-specific $T_1(\mathbf{x})$	88.05	85.96	84.72	68.78	81.88
- w/o translate templates	88.05	**86.57**	72.80	72.71	80.03
- w/o translate verbalizers	88.05	86.25	77.00	73.06	81.09
Manual $T_2(\mathbf{x})$	88.21	86.27	63.21	73.85	77.89
language-specific $T_2(\mathbf{x})$	88.21	86.66	**86.13**	50.75	77.94
- w/o translate templates	88.21	85.33	67.54	73.41	78.62
- w/o translate verbalizers	88.21	**87.63**	73.78	**78.36**	82.00
Manual $T_3(\mathbf{x})$	86.02	84.79	**85.30**	**81.11**	84.31
language-specific $T_3(\mathbf{x})$	86.02	85.48	85.08	57.97	78.64
- w/o translate templates	86.02	84.65	81.08	76.86	82.15
- w/o translate verbalizers	86.02	**85.72**	83.76	80.83	84.08
Manual $T_4(\mathbf{x})$	86.16	84.37	83.05	80.24	83.46
language-specific $T_4(\mathbf{x})$	86.16	**86.28**	**85.78**	64.77	80.75
- w/o translate templates	86.16	84.47	73.84	77.24	80.43
- w/o translate verbalizers	86.16	**86.28**	84.30	**82.93**	84.92

Table 7 The ablation study results of language-specific prompts in the 32-shot settings. Models are evaluated by 5 random runs and we report the mean values. We show the best results with bold numbers.



(a) null prompt with multilingual verbalizers (b) head prompt with multilingual verbalizers

Figure 2 T-SNE plots, where the left figure refers to features from the null prompt, and the right figure corresponds to features from the head prompt.

thousand Amazon reviews from each language in the test sets. We extract the feature from the [MASK] token for the null prompt, while for our head prompt, we extract the feature from the [CLS] token.

As presented in Figure 2, each data point in the plots represents an input text. We can observe from the left plot that the features of different languages in the null prompt still have a clear border that can distinguish, and the features of Japanese are almost isolated from others, which suggests the weak capability of the null prompt in inducing language-agnostic representations. The right plot in Figure 2 further demonstrates that the features of different languages from the head prompt on XLM-R are partly entangled, including Japanese, which is distinct from English. It can indicate that our head prompt can induce better language-agnostic representation. Through the qualitative analysis, we attribute the successful cross-language transferability of the head prompt to better language-agnostic representation.

5.3 The Effectiveness of Multilingual Verbalizers

Yang et al. [9] demonstrate that the cross-lingual transfer performance is highly related to the cross-lingual representation discrepancy. The smaller cross-lingual representation discrepancy correlates to better cross-lingual transfer performance. Following Conneau et al. [38], we utilize the linear centered

CKA	German	French	Danish	Russian	Japanese	Chinese
null prompt	0.65	0.67	0.66	0.70	0.40	0.47
+multilingual verbalizers	0.67	0.70	0.68	0.71	0.44	0.51
Δ	+0.02	+0.03	+0.02	+0.01	+0.04	+0.04
head prompt	0.29	0.31	0.30	0.27	0.11	0.10
+multilingual verbalizers	0.32	0.38	0.29	0.30	0.14	0.12
Δ	+0.03	+0.07	-0.01	+0.03	+0.03	+0.02

Table 8 CKA scores on WikiMatrix dataset, where Δ denotes the score difference between models with or without multilingual verbalizers.

Method	English	German	French	Japanese	Average
head prompt	86.11	**86.06**	84.35	82.54	84.76
+multi-verbalizer	**86.72**	86.00	86.37	83.94	**85.76**
w/o en	86.04	85.30	86.35	84.02	85.43
w/o de	86.43	85.71	**86.39**	**84.15**	85.67
w/o fr	86.68	85.94	85.92	84.04	85.65
w/o ja	84.87	84.52	85.16	82.34	84.22

Table 9 The ablation studies about language selection for multilingual verbalizers in 32-shot settings. Models are evaluated by 5 random runs and we report mean. We show the best results with bold numbers.

kernel alignment (CKA) [27] score to indicate the cross-lingual representation discrepancy, written by

$$\mathrm{CKA}(X,Y) = \frac{\left\|Y^\top X\right\|_{\mathrm{F}}^2}{\left\|X^\top X\right\|_{\mathrm{F}}^2 \left\|Y^\top Y\right\|_{\mathrm{F}}^2} \tag{4}$$

where X and Y are features of parallel sequences from the source and target languages, respectively. Specifically, in the null prompt, X and Y are the last hidden state of the [MASK] token, while in the head prompt, they are from the representation of the [CLS] token. A higher CKA score means a smaller cross-lingual representation discrepancy.

To verify whether our introduced head prompt can lead to smaller cross-lingual representation discrepancy with the enhancement of the multilingual verbalizers, we randomly select 2000 parallel sequences for each English-German, English-French, English-Danish, English-Russian, English-Japanese, and English-Chinese pair from WikiMatrix [7]. In table 8, we report CKA scores on these datasets with models fine-tuned under 16-shot settings. The result shows that models with multilingual verbalizers consistently achieve higher CKA scores. Thus, we claim that the multilingual verbalizers help induce better aligned cross-lingual representation.

5.4 Language Selection for Multilingual Verbalizers

In this section, we investigate the effect of language selection for multilingual verbalizers, which can guide the design of better multilingual verbalizers. Specifically, we fine-tune the XLM-R model with the head prompt and different configurations of multilingual verbalizers under the 32-shot setting and report the averaged results over three domains in Table 9.

As shown in Table 9, not every configuration of multilingual verbalizers will improve cross-lingual performance. The verbalizers that only contains English, German and French achieve the worst result, even underperforming the method with only English verbalizer. It indicates that the key to the success of the multilingual verbalizers is likely to be those languages that are pretty different from the source language. Besides, we can observe that the verbalizers without German and those without French have comparable results. It indicates that similar languages will not improve performance consistently.

6 Conclusion

In this paper, we propose a simple yet effective prompt-based tuning method in cross-lingual transfer tasks under few-shot settings. Through introducing the head prompt and multilingual verbalizers, our proposed model can induce language-agnostic features from the multilingual pre-trained language model with only a small labeled dataset in the source language as the training data. Experimental results

indicate that, with the enhancement of our proposed model, the discrete prompt-based tuning, which has been confirmed success in cross-lingual transfer tasks, achieves continuous gains, leading to state-of-the-art performance on the Multilingual Amazon review benchmark dataset under few-shot settings. In the near future, we will explore the effect of our proposed method on more challenging cross-lingual transfer tasks and datasets, such as POS tagging, named entity recognition, and multilingual retrieval.

Acknowledgements This work is supported by the National Science Foundation of China (NSFC No. 62036004).

References

1 Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.
2 Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11).
3 Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. arXiv preprint arXiv:2107.13586, 2021.
4 Xu L, Lu X, Yuan C, et al. Fewclue: A chinese few-shot learning evaluation benchmark[J]. arXiv preprint arXiv:2107.07498, 2021.
5 Xu L, Lu X, Yuan C, et al. Few-Shot Learning for Chinese NLP Tasks[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2021: 412-421.
6 Xue L, Constant N, Roberts A, et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 483-498.
7 Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong and Paco Guzman, WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia arXiv, July 11 2019.
8 Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
9 Yang H, Chen H, Zhou H, Li L. Enhancing Cross-lingual Transfer by Manifold Mixup[C]//International Conference on Learning Representations. 2022.
10 Wolf T, Chaumond J, Debut L, et al. Transformers: State-of-the-art natural language processing[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020: 38-45.
11 Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//NAACL-HLT (1). 2019.
12 Conneau A, Khandelwal K, Goyal N, et al. Unsupervised Cross-lingual Representation Learning at Scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8440-8451.
13 Hsu T Y, Liu C L, Lee H Y. Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model[C]//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: 5933-5940.
14 Artetxe M, Schwenk H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 597-610.
15 Li J, He R, Ye H, et al. Unsupervised domain adaptation of a pretrained cross-lingual language model[J]. arXiv preprint arXiv:2011.11499, 2020.
16 Lauscher A, Ravishankar V, Vulić I, et al. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4483-4499.
17 Conneau A, Lample G. Cross-lingual language model pretraining[J]. Advances in Neural Information Processing Systems, 2019, 32: 7059-7069.
18 Ye H, Tan Q, He R, et al. Feature Adaptation of Pre-Trained Language Models across Languages and Domains with Robust Self-Training[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7386-7399.
19 Lan Z, Chen M, Goodman S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[C]//International Conference on Learning Representations. 2019.
20 Ruder S, Vulić I, Søgaard A. A survey of cross-lingual word embedding models[J]. Journal of Artificial Intelligence Research, 2019, 65: 569-631.
21 Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP[C]//International Conference on Machine Learning. PMLR, 2019: 2790-2799.
22 Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
23 Chi Z, Huang S, Dong L, et al. Xlm-e: Cross-lingual language model pre-training via electra[J]. arXiv preprint arXiv:2106.16138, 2021.
24 Yu P, Fei H, Li P. Cross-lingual Language Model Pretraining for Retrieval[C]//Proceedings of the Web Conference 2021. 2021: 1029-1039.
25 Zhao M, Schütze H. Discrete and Soft Prompting for Multilingual Models[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 8547-8555.
26 Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.
27 Kornblith S, Norouzi M, Lee H, et al. Similarity of neural network representations revisited[C]//International Conference on Machine Learning. PMLR, 2019: 3519-3529.
28 Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[J]. arXiv preprint arXiv:2005.14165, 2020.
29 Schick T, Schütze H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 255-269.
30 Schick T, Schütze H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 2339-2352.
31 Haviv A, Berant J, Globerson A. BERTese: Learning to Speak to BERT[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 3618-3623.

32 Shin T, Razeghi Y, Logan IV R L, et al. Eliciting Knowledge from Language Models Using Automatically Generated Prompts[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4222-4235.

33 Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners[J]. arXiv preprint arXiv:2012.15723, 2020.

34 Zhong Z, Friedman D, Chen D. Factual Probing Is [MASK]: Learning vs. Learning to Recall[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 5017-5033.

35 Qin G, Eisner J. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 5203-5212.

36 Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

37 Hambardzumyan K, Khachatrian H, May J. Warp: Word-level adversarial reprogramming[J]. arXiv preprint arXiv:2101.00121, 2021.

38 Conneau A, Wu S, Li H, et al. Emerging Cross-lingual Structure in Pretrained Language Models[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6022-6034.

39 Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT?[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4996-5001.

40 Pfeiffer J, Vulić I, Gurevych I, et al. MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7654-7673.

41 Conneau A, Wu S, Li H, et al. Emerging Cross-lingual Structure in Pretrained Language Models[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6022-6034.

42 Cao S, Kitaev N, Klein D. Multilingual alignment of contextual word representations[J]. arXiv preprint arXiv:2002.03518, 2020.

43 Libovický J, Rosa R, Fraser A. How language-neutral is multilingual BERT?[J]. arXiv preprint arXiv:1911.03310, 2019.

44 Zhao W, Eger S, Bjerva J, et al. Inducing language-agnostic multilingual representations[J]. arXiv preprint arXiv:2008.09112, 2020.

45 Tanti M, van der Plas L, Borg C, et al. On the Language-specificity of Multilingual BERT and the Impact of Fine-tuning[J]. arXiv preprint arXiv:2109.06935, 2021.

46 Logan IV R L, Balažević I, Wallace E, et al. Cutting down on prompts and parameters: Simple few-shot learning with language models[J]. arXiv preprint arXiv:2106.13353, 2021.

47 Jiang Z, Xu F F, Araki J, et al. How Can We Know What Language Models Know?[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 423-438.

48 Davison J, Feldman J, Rush A M. Commonsense knowledge mining from pretrained models[C]//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: 1173-1178.

49 Liu X, Zheng Y, Du Z, et al. GPT Understands, Too[J]. arXiv preprint arXiv:2103.10385, 2021.

50 Zhang N, Li L, Chen X, et al. Differentiable prompt makes pre-trained language models better few-shot learners[J]. arXiv preprint arXiv:2108.13161, 2021.

51 Liu X, Ji K, Fu Y, et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks[J]. arXiv preprint arXiv:2110.07602, 2021.

52 Prettenhofer P, Stein B. Cross-language text classification using structural correspondence learning[C]//Proceedings of the 48th annual meeting of the association for computational linguistics. 2010: 1118-1127.

53 Kann K, Cho K, Bowman S. Towards Realistic Practices In Low-Resource Natural Language Processing: The Development Set[C]//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: 3342-3349.

54 Perez E, Kiela D, Cho K. True Few-Shot Learning with Language Models[J]. arXiv preprint arXiv:2105.11447, 2021.

55 Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[J]. Advances in neural information processing systems, 2019, 32: 8026-8037.

56 Zhang T, Wu F, Katiyar A, et al. Revisiting few-sample BERT fine-tuning[J]. arXiv preprint arXiv:2006.05987, 2020.

57 Mosbach M, Andriushchenko M, Klakow D. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines[C]//International Conference on Learning Representations. 2020.

58 Hu S, Ding N, Wang H, et al. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification[J]. arXiv preprint arXiv:2108.02035, 2021.

59 Holtzman A, West P, Schwartz V, et al. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right[J]. arXiv preprint arXiv:2104.08315, 2021.

60 Webson A, Pavlick E. Do Prompt-Based Models Really Understand the Meaning of their Prompts?[J]. arXiv preprint arXiv:2109.01247, 2021.