Enhancing Cross-lingual Prompting with Dual Prompt Augmentation*

Meng Zhou[♡,♢], **Xin Li**[♢], **Yue Jiang**[♡], **Lidong Bing**[♢]
[♡]Carnegie Mellon University
[♢]DAMO Academy, Alibaba Group
{mengzhou, yuejian2}@andrew.cmu.edu
{xinting.lx, l.bing}@alibaba-inc.com

Abstract

Prompting shows promising results in few-shot scenarios. However, its strength for multilingual/cross-lingual problems has not been fully exploited. Zhao and Schütze (2021) made initial explorations in this direction by presenting that cross-lingual prompting outperforms cross-lingual finetuning. In this paper, we conduct an empirical exploration on the effect of each component in cross-lingual prompting and derive language-agnostic Universal Prompting, which helps alleviate the discrepancies between source-language training and target-language inference. Based on this, we propose DPA, a dual prompt augmentation framework, aiming at relieving the data scarcity issue in few-shot cross-lingual prompting. Notably, for XNLI, our method achieves 46.54% with only 16 English training examples per class, significantly better than 34.99% of finetuning. Our code is available at https://github.com/DAMO-NLP-SG/DPA.

1 Introduction

Although adapting Pre-trained Language Models (PLMs) (Devlin et al., 2019) to downstream NLP tasks via *finetuning* is the de facto mainstream paradigm under fully supervised settings (Wang et al., 2018), *prompting*[1] (Gao et al., 2021; Radford et al., 2019; Brown et al., 2020; Schick and Schütze, 2021a,b) has demonstrated its superiority over *finetuning* in low-resource scenarios. Typically, *prompting* reformulates the classification task as a language modeling problem over manually-designed natural language prompts.

Despite the effectiveness of *prompting* on English tasks, its potential for cross-lingual problems, which assume the availability of the training data in high-resource languages (e.g., *English*) only, is still under-explored. Zhao and Schütze (2021) is the pioneering work to apply *prompting* to cross-lingual NLP. However, their major efforts are spent on comparing different training strategies for cross-lingual prompting such as discrete prompting and soft prompting. They do not fully investigate the design choice of key components in *prompting*, i.e., prompt template and verbalizer.

To provide a practical guide for designing cross-lingual prompting, we first conduct an empirical analysis to explore the effects of each *prompting* component on the performance of cross-lingual transfer. Our preliminary study shows that template-free prompting combined with English-only inference, dubbed as language-agnostic "Universal Prompting" (UP) in this paper, generally performs well across different few-shot settings. Intuitively, UP avoids the discrepancies between the source-language training and the target-language inference, which intrinsically better fits cross-lingual tasks.

The derived UP is a concise solution with reasonable performance but does not take advantage of other available resources in the context of multilingual problems, e.g., the translation of verbalizers in target languages. Motivated by this fact, we propose a Dual Prompt Augmentation (DPA) framework to alleviate the data scarcity issue in few-shot scenarios. **Firstly**, we introduce multilingual verbalizers as answer augmentation for prompting, where the translated label tokens are treated as additional target-language supervision. **Secondly**, we propose prompt mixup as prompt input augmentation, which mixes the prompt representations in each batch. Intuitively, given two prompt representations on real data, we can generate a virtual representation based on their interpolation, which encodes the semantics in between. Our DPA framework is not task-dependent and does not require either external unlabeled data (Xie et al., 2020) or

*This work was supported by Alibaba Research Intern Program. It was done when Meng Zhou was an intern at Alibaba. Xin Li is the corresponding author.

[1]In this work, the term "prompting" refers to prompt-based finetuning, where the parameters of PLMs are finetuned.

massive text manipulation efforts (Wei and Zou, 2019) compared with other data augmentation approaches.

In summary, our contributions are as follows:

- We develop language-agnostic **Universal Prompting**, a concise prompting baseline with competitive performance for cross-lingual transfer.
- To overcome the data scarcity issue, we propose **Dual Prompt Augmentation** for cross-lingual prompting to perform data augmentation from the views of prompt answers and prompt inputs.

2 Language-Agnostic Universal Prompting

In this section, we first empirically investigate the importance of essential elements, i.e., template and verbalizer design, in cross-lingual prompting (Zhao and Schütze, 2021). Based on our investigation, we derive a more competitive baseline called Universal Prompting. It is language-agnostic because it does not make assumptions about the input language in template design, and the verbalizer during training is taken for all other languages. Note that, since soft prompting (SP) and mixed prompting (MP) rely on an external bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to create soft prompts and do not outperform discrete prompting (DP) significantly, we mainly discuss DP in this work for a clear comparison.

As illustrated in Table 1, Zhao and Schütze (2021) directly utilize the translated templates and verbalizers for target-language inference, making templates and verbalizers language-dependent. However, the translated templates are not seen and the translated verbalizers are never modeled by the PLM during training. This leads to discrepancies between the source-language training and the target-language inference.

To alleviate such discrepancies, we consider three possible variants. Specifically, these three variants are derived by avoiding translation on the template and verbalizer tokens or removing the template words, see Table 1 for concrete examples.

We follow the experimental setup (refer to Section 4 for details) in Zhao and Schütze (2021) to evaluate the impact of the above designs [2]. In Table 2, we observe that W/O TEMPLATE TRANSLATION achieves slight but stable improvements under different shots. W/O TEMPLATE WORDS simply removes the template words and achieves more obvious improvements. W/O VERBALIZER TRANSLATION [3] avoids using translation at the verbalizer end and brings in the most significant improvements. Therefore, by alleviating discrepancies either in the aspect of verbalizer or template, the performance of cross-lingual prompting can be further improved. By combining the advances of these variants, the Universal Prompting (UP) is derived to treat various languages in a unified fashion. Specifically, UP alleviates the discrepancy of prompt templates and verbalizers simultaneously, which is a much stronger baseline than Zhao and Schütze (2021) in multilingual tasks.

Note that the idea of removing template words in UP is distinct to "null prompt" (IV et al., 2021) from the perspective of motivation. "Null prompt" is proposed to simplify the manual prompt design on monolingual tasks. Compared with "null prompt", the primary goal of UP is to alleviate the source-target discrepancies in cross-lingual transfer. Moreover, besides removing template words, our UP also involves the design choice for target-language inference (W/O VERBALIZER TRANSLATION), which proves to be a larger contribution according to the empirical results shown in Table 2. The effectiveness of using the verbalizer in the source language is also found in (Lin et al., 2022).

3 Dual Prompt Augmentation

In prompting, the mask token is directly used for making predictions. In this section, we formalize a Dual Prompt Augmentation (DPA) framework based on this crucial element of prompting.

3.1 Prompt Answer Augmentation

In Section 2, we show that directly translating the verbalizers to the target language for inference is not helpful. In this subsection, we explore the usage of verbalizer translation at the training stage. Intuitively, their rich semantics could serve as high-quality paraphrases (Jiang et al., 2021) of the English verbalizer and provide additional supervision to train multilingual models. Motivated by this, we define a multilingual verbalizer for the English

[2] As we employ a different evaluation method, the reproduced results of Zhao and Schütze (2021) are slightly different from the original ones. More details can be found in Section 4.

[3] Note that W/O VERBALIZER TRANSLATION refers to not applying translated verbalizers during *inference*. In Section 3 we will show how to exploit the translated verbalizers as answer augmentation during *training*.

		Prompt Templates	Verbalizers
EN (source)	Zhao and Schütze (2021)	A . Question: B ? Answer: <mask> .	Entailment: yes; Contradict: no; Neutral: maybe
	Universal Prompting	A . B ? <mask> .	Entailment: yes; Contradict: no; Neutral: maybe
TR (target)	Zhao and Schütze (2021)	A . Soru: B ? Cevap: <mask> .	Entailment: Evet; Contradict: hiçbir; Neutral: belki
	w/o Template Translation	A . Question:B ? Answer: <mask> .	Entailment: Evet; Contradict: hiçbir; Neutral: belki
	w/o Template Words	A . B ? <mask> .	Entailment: Evet; Contradict: hiçbir; Neutral: belki
	w/o Verbalizer Translation	A . Soru: B ? Cevap: <mask> .	Entailment: yes; Contradict: no; Neutral: maybe
	Universal Prompting	A . B ? <mask> .	Entailment: yes; Contradict: no; Neutral: maybe

Table 1: Prompt templates and verbalizers in English (EN) and Turkish (TR). A and B indicate two sentences of a sentence pair. For XNLI, A is the premise and B is the hypothesis. With the proposed language-agnostic Universal Prompting, we could treat source-language training and target-language inference in a unified fashion.

Shots	Method	Accuracy
16	Zhao and Schütze (2021)	$38.81_{1.61}$
	w/o Template Translation	$39.15_{1.73}$
	w/o Template Words	$39.87_{2.94}$
	w/o Verbalizer Translation	$42.32_{1.81}$
	Universal Prompting	$\mathbf{43.18}_{2.77}$
32	Zhao and Schütze (2021)	$41.42_{1.66}$
	w/o Template Translation	$41.72_{1.89}$
	w/o Template Words	$43.66_{0.96}$
	w/o Verbalizer Translation	$46.50_{1.54}$
	Universal Prompting	$\mathbf{48.26}_{1.34}$
64	Zhao and Schütze (2021)	$46.42_{0.65}$
	w/o Template Translation	$46.75_{0.61}$
	w/o Template Words	$47.60_{1.09}$
	w/o Verbalizer Translation	$\mathbf{53.07}_{1.33}$
	Universal Prompting	$52.19_{1.53}$

Table 2: Comparison results between Zhao and Schütze (2021) and its variants on XNLI. We calculate the average accuracy over 15 languages. The standard deviation over 5 runs is reported as the subscript.

training data, which can be regarded as answer augmentation for the mask token. Formally, given the pre-built prompt $\boldsymbol{x}$ filled with input sentences, the training objective is to maximize the likelihood of verbalized label tokens in multiple languages:

$$\arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x}} \frac{1}{|\mathcal{L}|} \sum_{\boldsymbol{\ell} \in \mathcal{L}} \log P\big(\langle \text{mask} \rangle = V_\ell(\boldsymbol{y}) | \boldsymbol{x}; \boldsymbol{\theta}\big) \quad (1)$$

where θ denotes the parameters of the PLM. V_ℓ is the verbalizer in a certain language $\ell \in \mathcal{L}$, and it maps from the gold label to a specific word in language ℓ. [4] In comparison, UP only takes $\mathcal{L} = \{\text{EN}\}$, which is a monolingual verbalizer.

3.2 Input Augmentation with Prompt Mixup

Previous mixup methods for NLP perform the whole-sequence interpolation at the input embedding level (Zhang and Vaidya, 2021; Guo et al., 2019) or hidden representation level (Jindal et al., 2020; Chen et al., 2020). However, directly applying previous methods to prompting has been shown to even lead to a significant performance drop in Zhou et al. (2021). In prompting-based methods, the most important hidden space representation for classification is encoded at the position of mask tokens. Different training data may have different sequence lengths and their mask tokens are at different positions. The interpolation between the representation of a mask token and a normal verbal token would be meaningless in prompting. Therefore, we propose to interpolate between the top-most mask token representations to augment prompt inputs. Then the interpolated representation is fed into the masked language modeling head.

Formally, let $\boldsymbol{m}_i = h(\boldsymbol{x}_i)$ and $\boldsymbol{m}_j = h(\boldsymbol{x}_j)$ be the top-most hidden representations corresponding to the mask tokens of two prompts $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively. Then we perform linear interpolation to produce a virtual representation:

$$\hat{\boldsymbol{m}}_{ij} = \lambda h(\boldsymbol{x}_i) + (1-\lambda) h(\boldsymbol{x}_j) \quad (2)$$

where λ follows a Beta distribution, i.e., $\lambda \sim \beta(\alpha, \alpha)$. The corresponding answer labels are linearly interpolated accordingly:

$$\hat{\boldsymbol{y}}_{ij} = \lambda \boldsymbol{y}_i + (1-\lambda) \boldsymbol{y}_j \quad (3)$$

Considering an augmented multilingual verbalizer as in Section 3.1, the training objective of this particular virtual example would be:

$$\begin{aligned} \arg\max_{\boldsymbol{\theta}} \frac{1}{|\mathcal{L}|} \sum_{\boldsymbol{\ell} \in \mathcal{L}} \big\{ & \lambda \log P\big(\langle \text{mask} \rangle = V_\ell(\boldsymbol{y}_i) | \hat{\boldsymbol{m}}_{ij}; \boldsymbol{\theta}\big) \\ & + (1-\lambda) \log P\big(\langle \text{mask} \rangle = V_\ell(\boldsymbol{y}_j) | \hat{\boldsymbol{m}}_{ij}; \boldsymbol{\theta}\big) \big\} \end{aligned} \quad (4)$$

The interpolation is performed in a dynamic in-batch fashion. For a mini-batch drawn from the training set, we will split it into pairs and generate a virtual prompt representation based on each pair.

4 Experiments

4.1 Setup

Datasets We conduct experiments on two sentence-pair classification tasks: XNLI (Conneau

[4] Please refer to Appx. A for the language set we use

et al., 2018; Williams et al., 2018) for cross-lingual natural language inference and PAWS-X (Yang et al., 2019) for multilingual paraphrase identification. For these two datasets, while the evaluation data is human-translated, the golden training data is only available in English.

Evaluation We conduct our experiments by training the XLM-R base model (Conneau et al., 2020) on English. Then the model will be directly applied to other target languages, without using any training examples of the target language. To make a reasonable comparison between finetuning and prompting, we ensure finetuning to be better than a random guess on each language. Therefore, we randomly sample without replacement $K \in \{16, 32, 64, 128, 256\}$ per class for XNLI and $K \in \{256, 512\}$ per class for PAWS-X to construct the training set. Then we use the same number of shots from the validation split to select the best model (Perez et al., 2021).

The evaluation of few-shot cross-lingual transfer can be with large variance and depend on data selection (Zhang et al., 2021a; Zhao et al., 2021; Keung et al., 2020). In our work, to faithfully reflect the few-shot performance, separate training/validation sets are sampled for different runs.

4.2 Results

UP v.s. Finetuning/PCT On the XNLI dataset, even the simplest prompting method for cross-lingual transfer, namely UP, consistently outperforms the finetuning (FT) method by a large margin. Besides, our language-agnostic UP also surpasses FT on the majority of languages on the more challenging PAWS-X. These observations suggest that prompting is indeed a better solution for few-shot learning in cross-language scenarios and our UP can serve as a strong baseline for cross-lingual prompting. We also reproduce PCT (Qi et al., 2022), another recent cross-lingual prompting method based on data augmentation and consistency training, with our evaluation method. Table 3 shows that UP outperforms PCT consistently without any data augmentation approach or introducing additional loss terms.

Dual Prompt Augmentation With the proposed DPA framework, our prompting method achieves consistent improvement over UP, indicating that multilingual verbalizers from the answer view and prompt mixup from the input view are both effective ways to enhance cross-lingual prompting. The comparison results in Table 3 and Table 4 also exhibit clear superiority of our method over cross-lingual finetuning. Even in the most resource-rich settings, compared to FT, our method still obtains 7.1% (256 shots) and 4.9% (512 shots) absolute gains on XNLI and PAWS-X.

Ablation Study The performance of our prompting method will become worse when removing either prompt mixup or multilingual verbalizer, showing that both prompt input and prompt answer augmentation contribute positively to the improvement. We also notice that the negative effects brought by DPA W/O MV are generally larger, showing the necessity of target-language guidance for cross-lingual prompting.

4.3 Inference Strategy

A natural extension for the DPA framework is to leverage the multilingual verbalizer in some way for target-language inference as well. For comparisons, we heuristically devise the following inference strategies :

(1) English Verbalizer The English verbalizer is still used when transferring to target languages. This strategy is used to produce results in Table 3 and 4. To formalize:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P\big(\langle \text{mask} \rangle = V_{EN}(\boldsymbol{y}) | \boldsymbol{x}; \boldsymbol{\theta}\big) \quad (5)$$

(2) Target Language Verbalizer The verbalizer in the corresponding target language is used, which is the practice of Zhao and Schütze (2021) during inference time. However, in this case, our DPA framework has already modeled these words during the training time. To formalize:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P\big(\langle \text{mask} \rangle = V_{target}(\boldsymbol{y}) | \boldsymbol{x}; \boldsymbol{\theta}\big) \quad (6)$$

(3) Taking Maximum over the Multilingual Verbalizer In this strategy, we will take the maximum probability over the whole multilingual verbalizer. To formalize:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y},\boldsymbol{\ell}} P\big(\langle \text{mask} \rangle = V_{\ell}(\boldsymbol{y}) | \boldsymbol{x}; \boldsymbol{\theta}\big) \quad (7)$$

(4) Taking Sum over the Multilingual Verbalizer In this strategy, we will take the sum of probability over the whole multilingual verbalizer. To formalize:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \sum_{\boldsymbol{\ell} \in \mathcal{L}} P\big(\langle \text{mask} \rangle = V_{\ell}(\boldsymbol{y}) | \boldsymbol{x}; \boldsymbol{\theta}\big) \quad (8)$$

Shots	Method	EN	AR	BG	DE	EL	ES	FR	HI	RU	SW	TH	TR	UR	VI	ZH	Avg.
16	FT	35.62	35.11	34.85	35.07	35.08	35.21	34.95	34.89	34.52	35.07	34.92	34.79	35.02	35.02	34.71	34.99±1.84
	PCT	42.43	35.80	37.48	36.02	40.23	36.14	38.79	39.79	37.96	36.32	39.01	37.41	35.46	38.84	38.90	38.04±3.52
	UP	47.68	42.01	45.50	44.51	46.68	36.61	46.81	40.29	45.43	42.06	44.21	41.04	40.61	45.79	38.42	43.18±2.77
	DPA	48.55	**46.24**	**47.95**	**48.00**	47.41	**47.47**	**48.61**	**44.36**	46.76	44.35	45.95	**45.83**	**44.80**	**47.31**	**44.55**	**46.54±1.83**
	W/O MV	**49.54**	41.55	46.84	45.53	47.59	34.63	48.55	42.39	**47.18**	43.95	**46.37**	43.82	43.32	46.52	40.09	44.52±2.15
	W/O MIXUP	48.38	45.59	47.74	47.72	**47.60**	44.38	47.83	42.44	46.69	**44.38**	44.65	45.52	43.48	46.65	40.83	45.59±1.91
32	FT	37.62	36.82	36.61	37.03	37.07	37.39	37.53	37.35	36.83	36.42	36.40	36.40	36.71	36.84	36.96	36.93±1.96
	PCT	46.63	41.33	44.30	43.35	45.31	45.61	46.79	43.32	44.13	40.88	42.86	43.19	38.94	44.85	43.81	43.69±2.11
	UP	53.33	47.70	50.87	49.74	51.41	41.48	51.09	44.97	50.11	46.76	49.50	45.92	45.64	51.00	44.33	48.26±1.34
	DPA	52.79	**49.37**	**51.48**	**50.84**	**51.78**	50.05	**51.77**	**48.08**	**50.46**	47.30	49.35	**50.14**	**47.44**	**50.84**	**48.25**	**49.99±2.21**
	W/O MV	**53.75**	48.42	50.71	50.57	51.76	41.98	51.54	45.64	**50.46**	45.84	**49.65**	47.42	45.58	50.56	47.54	48.76±1.56
	W/O MIXUP	52.38	49.29	51.39	50.76	51.60	**50.21**	51.54	47.57	50.35	**47.56**	49.07	49.56	47.02	50.65	46.24	49.68±1.46
64	FT	42.97	40.70	41.29	41.68	42.09	42.46	42.23	40.59	40.38	39.96	40.65	40.84	40.24	42.09	40.53	41.25±3.60
	PCT	52.26	46.39	48.73	48.39	49.64	49.46	50.46	47.48	48.52	45.27	48.28	48.55	44.76	49.81	49.12	48.47±2.82
	UP	57.76	51.67	54.85	54.99	54.69	51.63	54.96	47.97	53.32	48.12	51.91	49.89	47.86	54.14	49.13	52.19±1.54
	DPA	**59.97**	53.18	56.51	**56.67**	55.63	**56.79**	**56.97**	51.77	55.46	50.71	53.35	**54.21**	50.76	56.05	53.09	54.74±0.93
	W/O MV	59.17	**53.79**	**56.95**	56.53	**56.18**	55.35	56.48	**52.17**	**55.72**	**50.89**	**54.55**	53.35	**51.62**	**56.43**	**54.42**	**54.91±1.18**
	W/O MIXUP	59.56	53.06	55.98	55.65	55.16	56.67	56.66	51.44	55.18	49.99	52.90	53.76	49.80	55.43	53.70	54.33±0.98
128	FT	47.24	43.91	44.13	43.96	44.38	45.25	44.48	42.38	42.81	42.87	42.87	42.93	42.36	44.60	42.87	43.80±2.58
	PCT	55.31	48.55	52.09	50.75	52.92	52.69	52.79	50.43	51.60	47.86	50.88	50.37	48.04	52.20	51.79	51.22±2.58
	UP	60.08	51.31	56.60	55.10	56.17	51.25	56.97	49.62	55.18	48.71	53.87	50.42	49.20	55.03	53.15	53.51±3.51
	DPA	**62.57**	54.91	58.72	**58.81**	**58.25**	**59.47**	58.76	**52.93**	**57.35**	50.95	54.30	**54.94**	51.47	57.80	54.99	**56.42±1.37**
	W/O MV	61.51	**55.31**	58.67	58.15	58.12	58.10	58.42	52.31	56.99	50.80	**55.40**	53.88	**51.74**	**57.96**	**56.12**	56.23±0.90
	W/O MIXUP	61.84	54.59	**58.77**	58.57	57.77	59.13	**58.89**	52.70	56.99	**52.05**	54.15	54.69	51.31	57.27	55.59	56.29±1.46
256	FT	59.49	52.87	55.92	55.51	55.07	57.44	56.32	51.75	54.19	49.88	52.38	53.68	50.38	55.37	53.95	54.28±2.15
	PCT	60.09	53.51	57.21	56.60	57.63	58.78	58.42	54.07	56.35	51.80	54.57	54.62	50.56	56.36	56.14	55.78±1.63
	UP	65.08	56.57	61.03	60.65	60.74	59.21	61.01	55.18	59.41	53.73	57.66	57.62	54.08	60.58	58.71	58.75±1.92
	DPA	**67.97**	**59.54**	**63.59**	**63.26**	**62.34**	**64.80**	**63.93**	**58.39**	**61.87**	**55.83**	**59.19**	**60.32**	**56.00**	**62.41**	**61.29**	**61.38±0.92**
	W/O MV	65.80	58.07	62.04	61.33	61.05	63.03	62.36	56.16	60.14	54.17	58.23	57.62	54.12	60.52	59.81	59.63±0.92
	W/O MIXUP	67.40	58.02	62.33	62.18	61.35	63.61	62.93	56.89	60.75	54.68	58.06	59.00	54.74	61.17	59.33	60.16±0.97

Table 3: Zero-shot cross-lingual transfer accuracy on XNLI. FT: finetuning; MV: Multilingual Verbalizer. Reported results are averaged with 5 random seeds.

Shots	Method	EN	DE	ES	FR	JA	KO	ZH	Avg.
256	FT	63.18	60.81	60.95	61.39	58.60	58.48	59.78	60.46±4.23
	UP	65.50	62.21	63.24	62.82	54.11	54.30	55.99	59.74±4.12
	DPA	**71.87**	**68.59**	**69.10**	**69.02**	**60.41**	**60.88**	**62.75**	**66.09±3.62**
	W/O MV	69.06	66.26	66.47	65.79	59.28	58.34	60.77	63.71±4.37
	W/O MIXUP	70.95	67.14	67.58	67.63	59.01	60.44	61.16	64.84±2.91
512	FT	77.64	73.41	73.19	74.33	65.55	65.19	68.25	71.08±5.81
	UP	83.31	76.18	77.63	77.42	63.41	65.03	68.06	73.01±1.52
	DPA	**84.97**	**78.63**	79.60	**80.48**	**67.86**	68.13	**72.34**	**76.00±1.04**
	W/O MV	84.81	78.56	**79.67**	79.64	67.04	**68.34**	71.50	75.65±0.64
	W/O MIXUP	84.84	77.85	79.36	79.69	66.76	68.03	71.03	75.37±2.00

Table 4: Zero-shot cross-lingual transfer accuracy on PAWS-X. FT:finetuning; MV: Multilingual Verbalizer. Reported results are averaged with 5 random seeds.

Strategy Num.	Accuracy
1	$\mathbf{56.42_{1.37}}$
2	$56.31_{1.15}$
3	$56.23_{1.09}$
4	$56.33_{1.11}$
5	$56.39_{1.21}$

Table 5: Accuracy of different inference strategies, averaged over 15 testing languages of XNLI and 5 random seeds.

(5) Bilingual Verbalizer In this strategy, we will take the sum of probability over the target language verbalizer and the English verbalizer. To formalize, the predicted label $\hat{y}$ is given by:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \{ P\big(\langle \text{mask} \rangle = V_{EN}(\boldsymbol{y}) | \boldsymbol{x}; \boldsymbol{\theta}\big) + P\big(\langle \text{mask} \rangle = V_{target}(\boldsymbol{y}) | \boldsymbol{x}; \boldsymbol{\theta}\big) \} \quad (9)$$

We use the checkpoint of XLM-R trained by 128 shots on the XNLI dataset and make inference with different strategies. Table 5 shows the accuracy by employing different inference strategies. We show that with our DPA framework, the inference is quite robust to the utilization of the verbalizer. This can probably be attributed to answer augmentation via multilingual verbalizers, which help to model label tokens in multiple languages. We choose to simply employ English-only inference due to its simplicity and slightly better performance to produce results in Tables 3 and 4.

5 Conclusion

In this paper, we first derive language-agnostic Universal Prompting, a concise but competitive baseline for cross-lingual prompting. The proposed DPA framework can further enhance cross-lingual prompting as shown on two sentence-pair classification tasks. In the future, we will consider verifying the effectiveness of prompting and the DPA framework in cross-lingual sequence tagging or question-answering tasks (Xu et al., 2023).

6 Limitations

Our work mainly focuses on cross-lingual sentence-pair classification tasks. While it is directly applicable to single-sentence classification tasks (Li et al., 2020; Ye et al., 2020) but may require additional efforts to adapt our DPA framework to more complex cross-lingual tasks such as sequence tagging (Liu et al., 2021; Zhou et al., 2022, 2023; Zhang et al., 2021b) or question answering (Xu et al., 2022, 2023). Another limitation is that the proposed multilingual verbalizer in the DPA framework requires an external machine translator to produce the translated verbalizers. Finally, we limit the language set of the multilingual verbalizer to the set of target languages in a multilingual dataset. Extending this language set might give us greater improvement for cross-lingual tasks.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *ArXiv*, abs/1905.08941.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Robert L Logan IV, Ivana Balavzevi'c, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2021. Cutting down on prompts and parameters: Simple few-shot learning with language models. *ArXiv*, abs/2106.13353.

Zhengbao Jiang, J. Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Amit Jindal, Arijit Ghosh Chowdhury, Aniket Didolkar, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Augmenting NLP models using latent feature interpolations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6931–6936, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Juntao Li, Ruidan He, Hai Ye, Hwee Tou Ng, Lidong Bing, and Rui Yan. 2020. Unsupervised domain adaptation of a pretrained cross-lingual language model. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3672–3678. ijcai.org.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Ves Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Conference on Empirical Methods in Natural Language Processing*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *ArXiv*, abs/2105.11447.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Timo Schick and Hinrich Schütze. 2021a. 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*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Weiwen Xu, Xin Li, Wai Lam, and Lidong Bing. 2023. mpmr: A multilingual pre-trained machine reader at scale. In *The 61th Annual Meeting of the Association for Computational Linguistics.*

Weiwen Xu, Xin Li, Wenxuan Zhang, Meng Zhou, Lidong Bing, Wai Lam, and Luo Si. 2022. From clozing to comprehending: Retrofitting pre-trained language model to pre-trained machine reader. *arXiv preprint arXiv:2212.04755*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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)*, pages 3687–3692, Hong

Kong, China. Association for Computational Linguistics.

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7386–7399, Online. Association for Computational Linguistics.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021a. Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*.

Wancong Zhang and I. Vaidya. 2021. Mixup training leads to reduced overfitting and improved calibration for the transformer architecture. *ArXiv*, abs/2102.11402.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021b. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

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

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, and Chunyan Miao. 2023. Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning. In *The 61th Annual Meeting of the Association for Computational Linguistics.*

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. ConNER: Consistency training for cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8438–8449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Additional Implementation Details

Implementation Package Our implementation is based on PyTorch (Paszke et al., 2019) and Huggingface Transformer (Wolf et al., 2019) framework.

Model Details XLM-R base model, containing 270M parameters, is pretrained on 2.5TB of filtered CommonCrawl on 100 languages. It contains 12 Transformer layers with hidden space dimensions of 768 and 12 attention heads in each layer.

Computing Infrastructure All of our experiments are conducted on a single *Tesla V100-SXM2 32G*. Gradient accumulation steps of 4 is used for prompting to overcome resource limitations.

Hyperparameter Settings Our major hyperparameter settings follow Zhao and Schütze (2021). A fixed learning rate (1e-5) is used for all of our experiments without any learning rate schedule to compare finetuning with prompting (Le Scao and Rush, 2021). We use a smaller batch size of 8 for finetuning and prompting because it achieves slightly better performance. We use the max sequence length of 256. The model is trained for 50 epochs and we select the checkpoint by validation accuracy for testing as suggested in Mosbach et al. (2021); Zhang et al. (2021a). The α value for β distribution in prompt mixup is set to 1.2 for all of the experiments.

Prompting The language sets $\mathcal{L}$ used for multilingual verbalizers are determined by the language availability of the dataset. Specifically, for XNLI, $\mathcal{L}$ = {EN, AR, BG, DE, EL, ES, FR, HI, RU, SW, TH, TR, UR, VI, ZH}. For PAWS-X, $\mathcal{L}$ = {EN, DE, ES, FR, JA, KO, ZH}

For simplicity, the verbalizers of target languages are translated by Google Translate. Similar with XNLI, we use "paraphrase → yes" and "non-paraphrase → no" as the verbalizer of PAWS-X in English. Table 6 presents the full multilingual verbalizer we use for the PAWS-X dataset.

We discuss Universal Prompting across languages for multilingual sentence-pair classification tasks in Section 2. Moreover, we believe the same notion of alleviating source-target discrepancies in terms of prompt template and verbalizer is generally applicable for cross-lingual tasks, which is left for future work.

Language	Verbalizer
EN	Paraphrase → yes Non-paraphrase → no
DE	Paraphrase → Ja Non-paraphrase → Nein
ES	Paraphrase → sí Non-paraphrase → no
FR	Paraphrase → Oui Non-paraphrase → non
JA	Paraphrase → はい Non-paraphrase → ない
ZH	Paraphrase → 是 Non-paraphrase → 否
KO	Paraphrase → 예 Non-paraphrase → 아니

Table 6: The multilingual verbalizer for PAWS-X.

Shots	Method	Accuracy
256	Universal Prompting	$59.74_{4.12}$
	w/ TEMPLATE WORDS	$57.01_{2.64}$
512	Universal Prompting	$73.01_{1.52}$
	w/ TEMPLATE WORDS	$73.39_{2.54}$

Table 7: The ablation study of the impact of removing template words on PAWS-X. We calculate the average accuracy over 7 languages. The standard deviation over 5 runs is reported as the subscript.

B Generalizability of Prompting Word Removal

In Section 2, we show that by removing template words, UP provides a more reasonable baseline for cross-lingual prompting on XNLI. To see whether such a removal generalizes to other cross-lingual sentence-pair classification task, we also investigate the impact of removing template words on PAWS-X, as shown in Table 7. We find that UP still performs reasonably well on PAWS-X without template words. It was also shown in IV et al. (2021) that hand-engineering prompt is less important when PLMs are finetuned for monolingual tasks. Our UP generalizes this in cross-lingual tasks.

C Performance with Standard Deviation

In Table 8 and 9, we show the performance with standard deviation specifically in every language.

	Method	EN	AR	BG	DE	EL	ES	FR	HI	RU	SW	TH	TR	UR	VI	ZH	Avg.
16shots	FT	$35.62_{2.45}$	$35.11_{1.53}$	$34.85_{1.87}$	$35.07_{2.11}$	$35.08_{2.22}$	$35.21_{2.20}$	$34.95_{2.09}$	$34.89_{1.84}$	$34.52_{1.56}$	$35.07_{2.01}$	$34.92_{1.50}$	$34.79_{1.77}$	$35.02_{1.70}$	$35.02_{1.89}$	$34.71_{1.52}$	$34.99_{1.84}$
	UP	$47.68_{1.65}$	$42.01_{3.68}$	$45.50_{2.32}$	$44.51_{2.43}$	$46.68_{2.75}$	$36.61_{4.18}$	$46.81_{1.85}$	$40.29_{5.19}$	$45.43_{2.11}$	$42.06_{4.00}$	$44.21_{3.64}$	$41.04_{4.19}$	$40.61_{4.76}$	$45.79_{2.10}$	$38.42_{5.54}$	$43.18_{2.77}$
	OURS	$48.55_{1.43}$	$\mathbf{46.24}_{2.61}$	$\mathbf{47.95}_{1.72}$	$\mathbf{48.00}_{1.53}$	$47.41_{2.27}$	$\mathbf{47.47}_{1.83}$	$\mathbf{48.61}_{1.48}$	$\mathbf{44.36}_{2.48}$	$46.76_{1.55}$	$44.35_{1.88}$	$45.95_{2.41}$	$\mathbf{45.83}_{1.51}$	$\mathbf{44.80}_{2.20}$	$\mathbf{47.31}_{1.67}$	$\mathbf{44.55}_{3.76}$	$\mathbf{46.54}_{1.83}$
	W/O MV	$\mathbf{49.54}_{2.62}$	$41.55_{3.56}$	$46.84_{2.05}$	$45.53_{2.39}$	$47.59_{2.59}$	$34.63_{1.11}$	$48.55_{2.21}$	$42.39_{4.25}$	$\mathbf{47.18}_{1.87}$	$43.95_{2.40}$	$\mathbf{46.37}_{2.31}$	$43.82_{3.92}$	$43.32_{3.66}$	$46.52_{2.31}$	$40.09_{3.51}$	$44.52_{2.15}$
	W/O MIXUP	$48.38_{1.80}$	$45.59_{1.96}$	$47.74_{1.94}$	$47.72_{2.07}$	$\mathbf{47.60}_{2.57}$	$44.38_{5.78}$	$47.83_{1.54}$	$42.44_{3.31}$	$46.69_{1.31}$	$\mathbf{44.38}_{1.39}$	$44.65_{2.53}$	$45.52_{2.02}$	$43.48_{2.66}$	$46.65_{1.67}$	$40.83_{4.31}$	$45.59_{1.91}$
32shots	FT	$37.62_{2.66}$	$36.82_{2.09}$	$36.61_{2.28}$	$37.03_{2.56}$	$37.07_{1.90}$	$37.39_{2.17}$	$37.53_{1.71}$	$37.35_{1.55}$	$36.83_{2.37}$	$36.42_{1.93}$	$36.40_{1.72}$	$36.40_{1.59}$	$36.71_{1.50}$	$36.84_{2.32}$	$36.96_{2.17}$	$36.93_{1.96}$
	UP	$53.33_{1.60}$	$47.70_{3.58}$	$50.87_{1.41}$	$49.74_{2.66}$	$51.41_{1.98}$	$41.48_{4.77}$	$51.09_{1.49}$	$44.97_{1.40}$	$50.11_{0.76}$	$46.76_{3.00}$	$49.50_{1.36}$	$45.92_{2.53}$	$45.64_{2.07}$	$51.00_{1.54}$	$44.33_{2.51}$	$48.26_{1.34}$
	OURS	$52.79_{2.14}$	$\mathbf{49.37}_{2.85}$	$\mathbf{51.48}_{1.79}$	$\mathbf{50.84}_{1.98}$	$\mathbf{51.78}_{2.07}$	$50.05_{3.60}$	$\mathbf{51.77}_{2.13}$	$\mathbf{48.08}_{1.84}$	$\mathbf{50.46}_{1.89}$	$47.30_{2.89}$	$49.35_{2.30}$	$\mathbf{50.14}_{1.77}$	$\mathbf{47.44}_{2.30}$	$\mathbf{50.84}_{2.50}$	$\mathbf{48.25}_{4.08}$	$\mathbf{49.99}_{2.21}$
	W/O MV	$\mathbf{53.75}_{1.53}$	$48.42_{1.06}$	$50.71_{1.67}$	$50.57_{1.21}$	$51.76_{1.76}$	$41.98_{4.85}$	$51.54_{1.78}$	$45.64_{3.76}$	$\mathbf{50.46}_{1.14}$	$45.84_{3.32}$	$\mathbf{49.65}_{0.91}$	$47.42_{3.57}$	$45.58_{2.57}$	$50.56_{1.63}$	$47.54_{2.30}$	$48.76_{1.56}$
	W/O MIXUP	$52.38_{2.50}$	$49.29_{2.05}$	$51.39_{1.71}$	$50.76_{1.75}$	$51.60_{1.44}$	$\mathbf{50.21}_{3.00}$	$51.54_{1.83}$	$47.57_{0.51}$	$50.35_{1.66}$	$\mathbf{47.56}_{1.02}$	$49.07_{1.76}$	$49.56_{1.25}$	$47.02_{1.23}$	$50.65_{1.45}$	$46.24_{2.07}$	$49.68_{1.46}$
64shots	FT	$42.97_{3.88}$	$40.70_{4.06}$	$41.29_{3.93}$	$41.68_{3.66}$	$42.09_{3.81}$	$42.46_{4.09}$	$42.23_{4.38}$	$40.59_{3.39}$	$40.38_{3.58}$	$39.96_{3.10}$	$40.65_{3.48}$	$40.84_{3.26}$	$40.24_{3.01}$	$42.09_{3.68}$	$40.53_{3.46}$	$41.25_{3.60}$
	UP	$57.76_{1.49}$	$51.67_{2.31}$	$54.85_{1.52}$	$54.99_{1.71}$	$54.69_{1.00}$	$51.63_{3.98}$	$54.96_{0.87}$	$47.97_{1.74}$	$53.32_{1.73}$	$48.12_{1.21}$	$51.91_{1.27}$	$49.89_{2.80}$	$47.86_{2.10}$	$54.14_{1.21}$	$49.13_{3.38}$	$52.19_{1.54}$
	OURS	$\mathbf{59.97}_{1.25}$	$53.18_{1.12}$	$56.51_{1.14}$	$\mathbf{56.67}_{1.12}$	$55.63_{0.68}$	$\mathbf{56.79}_{1.17}$	$\mathbf{56.97}_{1.52}$	$51.77_{1.09}$	$55.46_{1.26}$	$50.71_{1.50}$	$53.35_{1.36}$	$\mathbf{54.21}_{1.06}$	$50.76_{1.39}$	$56.05_{0.97}$	$53.09_{1.20}$	$54.74_{0.93}$
	W/O MV	$59.17_{1.59}$	$\mathbf{53.79}_{1.66}$	$\mathbf{56.95}_{0.90}$	$56.53_{1.23}$	$\mathbf{56.18}_{0.99}$	$55.35_{3.39}$	$56.48_{1.96}$	$\mathbf{52.17}_{1.22}$	$\mathbf{55.72}_{1.18}$	$\mathbf{50.89}_{1.36}$	$\mathbf{54.55}_{0.91}$	$53.35_{1.27}$	$\mathbf{51.62}_{1.05}$	$\mathbf{56.43}_{1.14}$	$\mathbf{54.42}_{1.04}$	$\mathbf{54.91}_{1.18}$
	W/O MIXUP	$59.56_{1.22}$	$53.06_{1.02}$	$55.98_{1.05}$	$55.65_{1.01}$	$55.16_{0.48}$	$56.67_{1.10}$	$56.66_{1.30}$	$51.44_{1.45}$	$55.18_{1.18}$	$49.99_{1.60}$	$52.90_{0.85}$	$53.76_{0.92}$	$49.80_{1.89}$	$55.43_{0.87}$	$53.70_{1.45}$	$54.33_{0.98}$
128shots	FT	$47.24_{4.50}$	$43.91_{2.43}$	$44.13_{2.63}$	$43.96_{2.83}$	$44.38_{2.13}$	$45.25_{3.38}$	$44.48_{2.89}$	$42.38_{2.68}$	$42.81_{2.42}$	$42.87_{1.77}$	$42.87_{2.48}$	$42.93_{2.53}$	$42.36_{2.37}$	$44.60_{2.74}$	$42.87_{2.57}$	$43.80_{2.58}$
	UP	$60.08_{2.56}$	$51.31_{6.02}$	$56.60_{3.30}$	$55.10_{3.84}$	$56.17_{1.81}$	$51.25_{10.31}$	$56.97_{2.19}$	$49.62_{4.33}$	$55.18_{1.75}$	$48.71_{3.94}$	$53.87_{2.05}$	$50.42_{3.79}$	$49.20_{3.39}$	$55.03_{4.04}$	$53.15_{2.61}$	$53.51_{3.51}$
	OURS	$\mathbf{62.57}_{1.69}$	$54.91_{1.93}$	$58.72_{1.31}$	$\mathbf{58.81}_{1.49}$	$\mathbf{58.25}_{0.99}$	$\mathbf{59.47}_{1.79}$	$58.76_{1.26}$	$\mathbf{52.93}_{1.65}$	$\mathbf{57.35}_{1.38}$	$50.95_{1.34}$	$54.30_{2.39}$	$\mathbf{54.94}_{1.76}$	$51.47_{2.37}$	$57.80_{1.50}$	$54.99_{1.86}$	$\mathbf{56.42}_{1.37}$
	W/O MV	$61.51_{1.58}$	$\mathbf{55.31}_{1.26}$	$58.67_{1.38}$	$58.15_{1.83}$	$58.12_{1.02}$	$58.10_{1.74}$	$58.42_{0.50}$	$52.31_{1.03}$	$56.99_{1.09}$	$50.80_{1.30}$	$\mathbf{55.40}_{0.97}$	$53.88_{0.81}$	$\mathbf{51.74}_{0.76}$	$\mathbf{57.96}_{1.32}$	$\mathbf{56.12}_{1.50}$	$56.23_{0.90}$
	W/O MIXUP	$61.84_{1.64}$	$54.59_{1.41}$	$\mathbf{58.77}_{1.55}$	$58.57_{1.24}$	$57.77_{1.74}$	$59.13_{1.92}$	$\mathbf{58.89}_{1.34}$	$52.70_{1.86}$	$56.99_{1.83}$	$\mathbf{52.05}_{1.71}$	$54.15_{1.58}$	$54.69_{1.38}$	$51.31_{1.07}$	$57.27_{1.92}$	$55.59_{2.44}$	$56.29_{1.46}$
256shots	FT	$59.49_{2.11}$	$52.87_{2.34}$	$55.92_{2.15}$	$55.51_{2.25}$	$55.07_{2.60}$	$57.44_{1.74}$	$56.32_{2.04}$	$51.75_{2.72}$	$54.19_{2.46}$	$49.88_{2.51}$	$52.38_{2.96}$	$53.68_{1.80}$	$50.38_{2.34}$	$55.37_{2.02}$	$53.95_{2.08}$	$54.28_{2.15}$
	UP	$65.08_{1.49}$	$56.57_{2.08}$	$61.03_{1.86}$	$60.65_{1.27}$	$60.74_{1.84}$	$59.21_{2.64}$	$61.01_{2.41}$	$55.18_{2.54}$	$59.41_{2.29}$	$53.73_{2.27}$	$57.66_{2.08}$	$57.62_{2.65}$	$54.08_{2.47}$	$60.58_{1.84}$	$58.71_{2.16}$	$58.75_{1.92}$
	OURS	$\mathbf{67.97}_{1.02}$	$\mathbf{59.54}_{1.36}$	$\mathbf{63.59}_{1.23}$	$\mathbf{63.26}_{1.23}$	$\mathbf{62.34}_{0.91}$	$\mathbf{64.80}_{1.17}$	$\mathbf{63.93}_{1.06}$	$\mathbf{58.39}_{0.71}$	$\mathbf{61.87}_{1.55}$	$\mathbf{55.83}_{1.52}$	$\mathbf{59.19}_{1.06}$	$\mathbf{60.32}_{1.08}$	$\mathbf{56.00}_{0.95}$	$\mathbf{62.41}_{0.52}$	$\mathbf{61.29}_{1.14}$	$\mathbf{61.38}_{0.92}$
	W/O MV	$65.80_{1.25}$	$58.07_{1.22}$	$62.04_{1.60}$	$61.33_{1.45}$	$61.05_{1.24}$	$63.03_{1.43}$	$62.36_{1.22}$	$56.16_{0.52}$	$60.14_{1.53}$	$54.17_{1.00}$	$58.23_{1.24}$	$57.62_{1.02}$	$54.12_{0.95}$	$60.52_{1.34}$	$59.81_{1.24}$	$59.63_{0.92}$
	W/O MIXUP	$67.40_{0.65}$	$58.02_{1.41}$	$62.33_{1.47}$	$62.18_{1.00}$	$61.35_{1.24}$	$63.61_{1.24}$	$62.93_{1.25}$	$56.89_{1.06}$	$60.75_{1.67}$	$54.68_{1.60}$	$58.06_{1.51}$	$59.00_{1.44}$	$54.74_{1.10}$	$61.17_{1.19}$	$59.33_{1.08}$	$60.16_{0.97}$

Table 8: Zero-shot cross-lingual transfer accuracy with standard deviation on XNLI. FT:finetuning; UP: Universal Prompting; MV: multilingual verbalizer. Reported results are averaged with 5 random seeds.

	Method	EN	DE	ES	FR	JA	KO	ZH	Avg.
256shots	FT	$63.18_{5.69}$	$60.81_{5.04}$	$60.95_{3.83}$	$61.39_{4.38}$	$58.60_{3.85}$	$58.48_{2.59}$	$59.78_{4.71}$	$60.46_{4.23}$
	UP	$65.50_{3.98}$	$62.21_{3.90}$	$63.24_{3.52}$	$62.82_{4.31}$	$54.11_{5.37}$	$54.30_{4.98}$	$55.99_{4.64}$	$59.74_{4.12}$
	OURS	$\mathbf{71.87}_{5.03}$	$\mathbf{68.59}_{4.50}$	$\mathbf{69.10}_{4.92}$	$\mathbf{69.02}_{3.97}$	$\mathbf{60.41}_{2.69}$	$\mathbf{60.88}_{1.76}$	$\mathbf{62.75}_{3.46}$	$\mathbf{66.09}_{3.62}$
	W/O MV	$69.06_{4.58}$	$66.26_{4.13}$	$66.47_{4.05}$	$65.79_{4.13}$	$59.28_{5.36}$	$58.34_{4.24}$	$60.77_{6.00}$	$63.71_{4.37}$
	W/O MIXUP	$70.95_{4.17}$	$67.14_{3.53}$	$67.58_{4.36}$	$67.63_{4.06}$	$59.01_{1.67}$	$60.44_{1.54}$	$61.16_{2.41}$	$64.84_{2.91}$
512shots	FT	$77.64_{8.38}$	$73.41_{6.30}$	$73.19_{6.84}$	$74.33_{6.47}$	$65.55_{3.89}$	$65.19_{3.81}$	$68.25_{5.28}$	$71.08_{5.81}$
	UP	$83.31_{2.43}$	$76.18_{1.93}$	$77.63_{1.73}$	$77.42_{1.64}$	$63.41_{2.39}$	$65.03_{2.12}$	$68.06_{0.78}$	$73.01_{1.52}$
	OURS	$\mathbf{84.97}_{1.60}$	$\mathbf{78.63}_{1.36}$	$79.60_{1.31}$	$\mathbf{80.48}_{1.09}$	$\mathbf{67.86}_{1.27}$	$68.13_{0.88}$	$\mathbf{72.34}_{1.47}$	$\mathbf{76.00}_{1.04}$
	W/O MV	$84.81_{1.19}$	$78.56_{0.78}$	$\mathbf{79.67}_{0.81}$	$79.64_{0.36}$	$67.04_{1.85}$	$\mathbf{68.34}_{0.88}$	$71.50_{1.76}$	$75.65_{0.64}$
	W/O MIXUP	$84.84_{1.57}$	$77.85_{1.94}$	$79.36_{2.01}$	$79.69_{1.99}$	$66.76_{3.24}$	$68.03_{2.19}$	$71.03_{2.65}$	$75.37_{2.00}$

Table 9: Zero-shot cross-lingual transfer accuracy with standard deviation on PAWS-X. FT:finetuning; UP: Universal Prompting; MV: multilingual verbalizer. Reported results are averaged with 5 random seeds.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4, Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Section 4, Appendix B, Appendix C

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.