Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages

Ercong Nie[⋆ 1,2] **Sheng Liang**[⋆ 1,2] **Helmut Schmid**[1] **Hinrich Schütze**[1,2]
[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Munich, Germany
{nie, shengliang}@cis.lmu.de

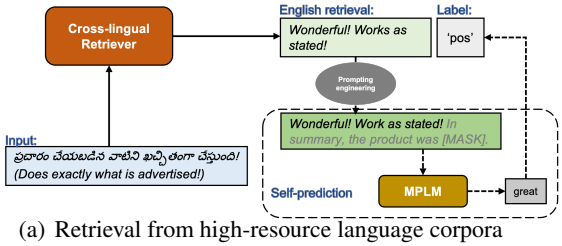
Abstract

Multilingual Pretrained Language Models (MPLMs) perform strongly in cross-lingual transfer. We propose **P**rompts **A**ugmented by **R**etrieval **C**rosslingually (**PARC**) to improve zero-shot performance on low-resource languages (LRLs) by augmenting the context with prompts consisting of semantically similar sentences retrieved from a high-resource language (HRL). PARC improves zero-shot performance on three downstream tasks (sentiment classification, topic categorization, natural language inference) with multilingual parallel test sets across 10 LRLs covering 6 language families in unlabeled (+5.1%) and labeled settings (+16.3%). PARC also outperforms finetuning by 3.7%. We find a significant positive correlation between cross-lingual transfer performance on one side, and the similarity between high- and low-resource languages as well as the amount of low-resource pretraining data on the other side. A robustness analysis suggests that PARC has the potential to achieve even stronger performance with more powerful MPLMs.

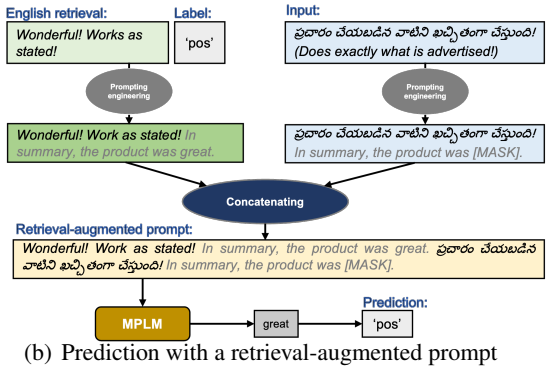
1 Introduction

Multilingual pretrained language models (MPLMs) (Devlin et al., 2019; Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021; Shliazhko et al., 2022), pretrained on multilingual corpora with >100 languages, exhibit strong multilinguality on downstream tasks (Hu et al., 2020).

Low-resource languages, for which little text data is available for pretraining monolingual pretrained language models (PLMs), benefit from MPLMs. However, the lack of LRL data leads to an imbalanced language distribution in the pretraining corpora of MPLMs (Wu and Dredze, 2020). LRLs are therefore under-represented in pretraining, resulting in bad performance. Furthermore, the scarcity of domain- or task-specific annotated data of LRLs makes it difficult to apply the pretraining-finetuning paradigm to LRLs (Lauscher et al., 2020). Given that the pretraining-finetuning paradigm always has a high demand for domain-specific labeled data, another line of research – prompt-based learning – emerges, focusing on exploiting large pretrained language models by reformulating the input. The prompt is designed to



(a) Retrieval from high-resource language corpora



(b) Prediction with a retrieval-augmented prompt

Figure 1: Main idea of PARC: we enhance zero-shot learning for low-resource languages (LRLs) by cross-lingual retrieval from **labeled/unlabeled** high-resource languages (HRLs). (a) An LRL input sample is taken as query by the cross-lingual retriever to retrieve the semantically most similar HRL sample from the HRL corpus. The label of the retrieved HRL sample is obtained either from the corpus (**labeled** setting) or by self-prediction (**unlabeled** setting). (b) The retrieved HRL sample together with its label and the input sample are reformulated as prompts. The cross-lingual retrieval-augmented prompt is created by concatenation and taken by the MPLM for prediction. Our experiments show that PARC outperforms other zero-shot methods and even finetuning.

⋆ Equal Contribution.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8320–8340
July 9-14, 2023

help PLMs "understand" the task better and "recall" what has been learned during the pretraining. In particular, Brown et al. (2020) propose a simple in-context learning approach without any finetuning, which adds training examples as additional context to test examples. Instead of using random examples as context, KATE (Liu et al., 2022a) and SOUP (Liu et al., 2022b) retrieve semantically similar examples as prompt for monolingual in-context learning. The above mentioned prompt-based learning techniques require no parameter updating, while there is also work employing sampled similar examples for prompt-based funetuning (Gao et al., 2021). Unlike Brown et al. (2020) who created prompts with manually selected examples, these approaches augment the context by retrieving related information from external corpora, allowing the PLMs to capture more domain- or task-specific knowledge. The prompt-based method offers a new form of zero-shot or few-shot learning in multilingual NLP studies. It involves performing a specific task using prompts, without labeled data in the target language and has the potential of being an effective method for LRLs lacking annotated data.

Our work improves the zero-shot transfer learning performance of LRLs on three different classification tasks by taking advantage of cross-lingual information retrieval and the multilinguality of MPLMs. Specifically, we retrieve semantically similar cross-lingual sentences as prompts and use the cross-lingual retrieval information to benefit the LRLs from the multilinguality of MPLMs and achieve better performance in the zero-shot setting[1]. Our main contributions are: (1) We propose **P**rompts **A**ugmented by **R**etrieval **C**rosslingually (**PARC**), a pipeline for integrating retrieved cross-lingual information into prompt engineering for zero-shot learning (Figure 1). (2) We conduct experiments on several multilingual tasks, showing that PARC improves the zero-shot performance on LRLs by retrieving examples from both labeled and unlabeled HRL corpora. (3) To find an optimal configuration of our PARC pipeline, we conduct a comprehensive study on the variables that affect the zero-shot performance: the number of prompts, the choice of HRL, and the robustness w.r.t. other retrieval methods and MPLMs.

[1]Different from the zero-shot cross-lingual transfer learning where MPLMs are finetuned on HRLs (Hu et al., 2020), our zero-shot setting does not involve finetuning. Details in §6.4

2 Related Work

Retrieval methods External knowledge extracted by information retrieval is often leveraged to solve NLP tasks. Two types of representations have been used for retrieval: (1) sparse bag-of-words representations (Chen et al., 2017; Wang et al., 2018), and (2) dense representation learned by neural networks (Qu et al., 2020). Dense representations come either from contextual token embeddings (May et al., 2019; Zhang et al., 2020) or from sentence encoders (Conneau et al., 2017; Cer et al., 2018). Reimers and Gurevych (2019) propose sentence transformers to create semantically meaningful sentence embeddings by applying siamese and triplet network structures to transformer-based pretrained language models. By using knowledge distillation, sentence transformers can be expanded to support various languages as multilingual sentence transformers (Reimers and Gurevych, 2020), allowing for cross-lingual retrieval.

Retrieval augmented prompt Brown et al. (2020) show that large-scale pretrained language models such as GPT-3 can learn to perform a task by putting examples of input-output pairs into the input as context. The in-context learning method simply concatenates the input with examples randomly extracted from the training set. Recent studies (Gao et al., 2021; Liu et al., 2022a,b) augment the prompts for pre-trained models by sampling semantically similar examples. They apply the retrieval augmented method to discrete prompts, which are represented by tokens instead of vectors in a continuous space. They use them either for finetuning in few-shot settings or for zero-shot learning. Chowdhury et al. (2022) use a similar kNN-based retrieval method for tuning the soft prompts in a continuous space with a standard supervised training setup. Previous work focused on monolingual retrieval-augmented prompts. Our work applies cross-lingual retrieval to discrete prompts in a scenario without parameter updating. To the best of our knowledge, our work is the first to investigate prompt learning augmented by cross-lingual retrieval.

Multilingual prompt learning Despite the success of prompting in English, prompting in multilingual tasks has not been extensively studied. Winata et al. (2021) show the multilingual skills of LMs mainly trained on English data in prompt learning

by giving them a few English examples as context but testing them on non-English data. Some recent works investigate the prompt learning with multilingual PLMs (Zhao and Schütze, 2021; Huang et al., 2022). Unlike our work, they focus on finetuning or prompt tuning requiring parameter updating. We apply our method to LRLs in a zero-shot setting without adjusting the model parameters.

3 Methodology

This work aims to improve the performance of MPLMs on LRLs in the zero-shot setting by leveraging retrieved cross-lingual contents from HRLs. For that, we design the PARC pipeline that can be applied to labeled and unlabeled scenarios, i.e., the HRL information can be retrieved from either labeled or unlabeled corpora.

As Figure 1 shows, the PARC pipeline consists of two steps: (a) Cross-lingual retrieval from high-resource language corpora, and (b) prediction with a retrieval-augmented prompt. Figure 1 shows an example: A Telugu input sentence from a sentiment classification task is firstly fed into the cross-lingual retriever to fetch the semantically closest sample from the HRL corpus, i.e. English in this case. In the second step, the retrieved HRL sample together with its label and the LRL input sentence are transformed into a prompt. For prompt-based classification, we need (i) a *pattern* which converts the input sentence into a cloze-style question with a mask token, and (ii) a representative word (called *verbalizer*) for each possible class. Converting the classification task into a cloze-style question aligns seamlessly with the framework of our proposed PARC method, because it not only performs zero-shot learning well but, more significantly, facilitates better integration of the retrieved cross-lingual contexts.

In our example, we use the pattern $P(X) = X \circ$ "`In summary, the product was [MASK].`" to convert the retrieved English sentence into "`Wonderful! Works as stated! In summary, the product was [MASK].`", where $\circ$ is the string concatenation operator. A verbalizer such as {`pos` $\rightarrow$ "great", `neg` $\rightarrow$ "terrible"}, which maps the original labels {`pos`, `neg`} onto words in the vocabulary, is then used to replace the `[MASK]` token with the verbalized label word "great", standing for the correct label `pos` of this sentence. We call the resulting English sentence (in our example: "`Wonderful! Works as stated! In summary, the product was great.`") the "cross-lingual context". At last, we fill the same pattern with the input Telugu sentence and append it to the cross-lingual context. We feed this cross-lingual retrieval augmented input to the MPLM. The MPLM returns for each of the verbalizers its probability of being the masked token.

More formally, let $X_i^L \in D^L$ be the input sample from the LRL test set, $(X_j^H, y_j) \in D_{lb}^H$ and $X_j^H \in D_{un}^H$ denote the HRL data from the *labeled* and *unlabeled* corpora, respectively, where X_j is the text sample and y_j its class label from a label set Y. As Eq. (1) shows, the cross-lingual retriever CLR takes the HRL corpora D^H and a given LRL input sentence X_i^L. It returns an ordered list of HRL sentences D^{R_i} according to the semantic similarity. We then have $(X_k^{R_i}, y_k^{R_i}) \in D_{lb}^{R_i}$ and $X_k^{R_i} \in D_{un}^{R_i}$ for labeled and unlabeled scenarios, respectively, where $X_k^{R_i}$ is the k-th most similar HRL sentence to the LRL input X_i^L.

$$D^{R_i} = CLR(X_i^L, D^H) \tag{1}$$

The prompt pattern $P(.)$ converts an HRL input sentence $X_k^{R_i}$ into a cloze-style form with a mask token. The verbalizer $v(.)$ is a bijective mapping from the set of class labels Y to a set of verbalized words V from the HRL vocabulary. We use the verbalized label word to fill in the mask token in the prompt pattern, and construct the cross-lingual context C_k^i for the input X_i^L with the k-th most similar HRL sample $X_k^{R_i}$:

$$C_k^i = P(X_k^{R_i}, v(y_k^{R_i})) \tag{2}$$

The cross-lingual context C_k^i is then concatenated with the prompted LRL input as the input I to the MPLM:

$$I_i = C_k^i \circ P(X_i^L) \tag{3}$$

The MPLM M performs masked token prediction and returns the probabilities $p = M(I_i)$ of all candidate words for the masked token in I_i. We predict the class $\hat{y}$ whose verbalizer $v(\hat{y})$ received the highest probability from model M:

$$\hat{y} = \arg\max_{y \in Y} p(v(y)) \tag{4}$$

In the labeled scenario, we obtain the correct label $y_k^{R_i}$ of the HRL sentence from $D_{lb}^{R_i}$. In the unlabeled scenario, we predict the label using the same prompt-based classification method without

cross-lingual context, similar to Eq. (4). We call this the *self-prediction* method:

$$\hat{y}_k^{R_i} = \arg\max_{y \in Y} M(P(X_k^{R_i}), v(y)) \quad (5)$$

In order to use more cross-lingual information, we retrieve the K most similar HRL samples. With each sample, we obtain verbalizer probabilities as described above and retrieve the class whose verbalizer has the largest sum of probabilities. We call this method the Bag-of-Retrieval (BoR) strategy. We also tried concatenating the different cross-lingual contexts (CONC method), but the resulting performance has been worse (see Table 15 in the Appendix).

4 Experimental Setup

4.1 Datasets

Base Datasets Three representative classification tasks are selected for evaluation in this work: binary sentiment analysis on Amazon product reviews (Keung et al., 2020), topic classification on AG News texts (Zhang et al., 2015), and natural language inference on XNLI (Conneau et al., 2018).

Amazon Reviews dataset categorizes the shopping reviews into 5 star ratings from 1 to 5. In order to satisfy a binary classification setting, we select the reviews with rating 1 as `negative` (0) and 5 as `positive` (1) for our experiments. The following pattern $P(X)$ and verbalizer v are defined for an input review text X:

- $P(X) = X \circ$ "All in all, it was [MASK]."
- $v(0) =$ "terrible", $v(1) =$ "great"

AG News is a collection of more than 1 million news articles for topic classification. The news topic categories contained in the dataset are `World` (0), `Sports` (1) , `Business` (2), and `Tech` (3). The pattern and verbalizers are as follows:

- $P(X) =$ "[MASK] News: " $\circ X$
- $v(0) =$ "World", $v(1) =$ "Sports", $v(2) =$ "Business", $v(3) =$ "Tech"

XNLI is a multilingual version of the MultiNLI dataset (Williams et al., 2018). We use a subset of the original XNLI dataset in our experiment. The text in each data item consists of two parts. Sentence A is the premise and sentence B is the hypothesis. The NLI task is to predict the type of inference between the given premise and hypothesis among the three types: `entailment` (0), `neutral` (1) and `contradiction` (2). For a given sentence pair X_1 and X_2, we design the pattern and verbalizer as:

- $P(X_1, X_2) = X_1 \circ$ "? [MASK]," $\circ X_2$
- $v(0) =$ "Yes" , $v(1) =$ "Maybe" , $v(2) =$ "No"

Construction of Multilingual Parallel Test Sets Parallel test datasets for evaluating cross-lingual transfer performance on LRLs are rare. However, recent research conducted by Hu et al. (2020); Liu et al. (2022c) shows that automatically translated test sets are useful for measuring cross-lingual performance. Hence, we adopt their methodology and construct datasets for different tasks by automatically translating English test sets to targeted LRLs. We use the Python API of the Google Translate System to implement the construction of multilingual parallel test sets in our experiment. We also validate the translation effectiveness and quality. The original XNLI datasets include two low-resource languages that are used in our experiments (Swahili and Urdu), so we use them as the "gold" standard for our translation validation. We compare the cross-lingual transfer performance on translation test sets and original test sets of XNLI. We also measure the translation quality by using the original sets as gold standard. Through the validation conducted on these two languages within the XNLI task, we infer that the translation method is effective and could be generalized to other languages and tasks. Detailed results are shown in Appendix §A.

Following Wu and Dredze (2020), we regard languages with a WikiSize[2] of less than 7 as LRLs. We construct a test set consisting of 10 LRLs in 6 language families: Indo-European (Afrikaans - af, Urdu - ur), Austronesian (Javanese - jv, Tagalog - ta), Altaic (Mongolian - mn, Uzbek - uz), Dravidian (Tamil - tl and Telugu - te), Sino-Tibetan (Burmese - my), and Niger-Congo (Swahili - sw). Table 18 in the Appendix shows more information on the test sets.

HRL Corpora To retrieve rich and diverse information, a large-scale general corpus or knowledge base in the different HRLs would be the ideal

[2]WikiSize less than 7 means that the Wikipedia corpus of the language is smaller than 0.177 GB.

sentence retrieval pool. In practice, however, a trade-off is necessary in order to save computational resources. Following Wang et al. (2022), we therefore use the task-specific labeled training set of English as the sentence pool in our experiments. The selection of the HRL will be discussed in §6.2.

4.2 Baseline

We compare PARC with the following baselines in both labeled and unlabeled settings:

MAJ The majority baseline. Since we construct the test sets to be balanced, MAJ is equivalent to random guess.

Random We randomly retrieve a cross-lingual sentence as prompt, similarly to the simple in-context learning using examples without semantic similarity to the input (Brown et al., 2020).

Direct The pattern filled with the input sample is directly fed to the MPLM for prediction, without adding cross-lingual context to the prompts.

Finetune The MPLM is first finetuned with the retrieved high resource sentences. Then the low-resource test input is predicted by the finetuned MPLM. We use the Cross Entropy Loss as the objective function for finetuning and AdamW for optimization with a learning rate of 1e-5. Since the finetuning data is very limited, we only train for a single epoch to avoid overfitting.

Our test sets are constructed by machine translation. Therefore we cannot apply a translation baseline, where we translate the input sample into the high resource language before feeding it to the MPLM. The Appendix presents a different experiment where we compare with a translation baseline.

4.3 Models

Cross-Lingual Retriever The retrieval methods used in monolingual NLP are either based on sparse or dense representations. Sparse representations such as BM25 (Manning et al., 2008) which is based on term frequency, cannot be used for cross-lingual retrieval as the shared words across different languages are normally scarce. Therefore dense representations from deep learning methods such as LASER (Artetxe and Schwenk, 2019) and sentence-BERT (Reimers and Gurevych, 2019) are more suitable for our pipeline.

We choose the multilingual sentence transformer (Reimers and Gurevych, 2020) version "*paraphrase-multilingual-mpnet-base-v2*" as the retriever in our experiments. This multilingual retriever is based on XLM-R (Conneau et al., 2020) and trained on parallel data from 50+ languages by employing knowledge distillation. Through the multilingual sentence transformer, sentences are represented by embeddings. We use the sentence embeddings to calculate the cosine similarity between the LRL inputs and HRL sentences and rank the most similar ones for retrieval. Robustness with respect to other cross-lingual retrievers will be discussed in §6.3.

	Amazon	AGNews	XNLI	Avg.
MAJ	50.0	25.0	33.3	36.1
Random	48.2	25.6	32.4	35.4
Direct	53.8	36.3	33.1	41.1
Finetune	68.6	57.9	34.5	53.7
PARC-unlabeled	58.4	46.7	33.5	46.2
PARC-labeled	**68.9**	**67.6**	**35.8**	**57.4**

Table 1: Overview of results on three classification tasks. The reported numbers are averaged across 10 evaluation LRLs. The number of prompts k=1 in relevant baselines and our methods for all three tasks.

Multilingual Pretrained Language Model In order to solve cloze-style classification tasks, we use the pretrained multilingual BERT model "*bert-base-multilingual-cased*" (Devlin et al., 2019). It contains 178M parameters and was trained on Wikipedia corpora in 104 languages. In §6.3, we will also explore XLM-R (Conneau et al., 2020), another multilingual pretrained language model.

All the models mentioned above were implemented using the Huggingface Transformers library (Wolf et al., 2020).

5 Results

Table 1 presents an overview of the results on the three tasks[3]. PARC outperforms the *MAJ*, *Direct* and *Random* baseline on all three tasks, in both labeled and unlabeled settings: When retrieving from unlabeled high-resource language corpora, the performance is improved by **4.6%**, **10.4%** and **0.4%** compared to *Direct* on Amazon Review, AG News, and XNLI respectively. When retrieving from labeled HRL corpora, the performance is improved by **15.1%**, **31.3%** and **2.7%**. The *Finetune* baseline uses the label of retrieved examples for prompt-based finetuning. Hence it is fair to compare it with *PARC* in the labeled setup rather than the unlabeled one. *PARC-labeled* outperforms *Finetune* by **0.3%**, **9.7%** and **1.3%** on the three tasks respectively.

Although our proposed methods perform better

[3] $k = 1$ unless otherwise specified.

than the baselines on all three tasks, the degree of improvement differs. A large improvement is found on AG News, the topic categorization task, while XNLI only shows a slight improvement. An explanation for this could be that the natural language inference task is more difficult than topic categorization, especially in a zero-shot setup. Also, prior work has shown that designing cloze-style patterns and searching the answer space for NLI tasks (Schick and Schütze, 2021; Webson and Pavlick, 2022) is difficult.

We also find that PARC-labeled noticeably outperforms PARC-unlabeled, indicating that the performance of self-prediction is limited by the capabilities of mBERT. More powerful MPLMs and better pattern designs might further improve the performance.

To analyze the performance for every language in detail, we present the complete experimental results for the topic categorization task on AG News in Table 2. Here, we use the BoR method to take advantage of multiple retrieved HRL sentences. As expected, PARC outperforms the *Direct* baseline on all languages in both labeled and unlabeled settings.

However, it is worth noting that the sensitivity to cross-lingual retrieval differs from language to language. For some LRLs, e.g. Urdu (Ur) and Uzbek (Uz), PARC's improvement from cross-lingual retrieval is smaller. Others gain more, e.g. Javanese (Jv). Retrieving more samples increases the performance up to k=30 except for Telugu (Te) and Swahili (Sw) where the max is reached for k=20.

We now turn to the following two questions: 1) How does k affect the performance on other tasks than topic categorization? 2) Which LRLs profit most from our PARC method and which HRLs are best suited to retrieve prompts?

6 Analysis

6.1 Effect of k

We investigated how the performance changes as the number of retrieved HRL samples k increases. As shown in Figure 2, an abrupt accuracy increase can be seen in both labeled and unlabeled scenarios by concatenating the most similar cross-lingual sample. In labeled scenarios, the performance tends to increase up to k=20 and then levels off. This can be explained by the fact that later retrieved samples are less similar to the input sample, so their contribution as prompts decreases. In unlabeled

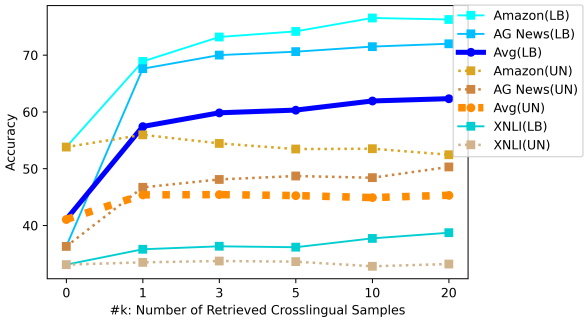


Figure 2: Accuracy on three tasks with different k in the labeled (LB) and unlabeled (UN) setup.

scenarios, there is no clear improvement beyond k=1 except for AGNews(UN), where the accuracy increases monotonically except for k=10. The performance of XNLI is less obviously influenced by the value of k than binary sentiment analysis and topic categorization. We assume that this could be attributed to the difficulty of the inference task. Unlike the other two single sentence classification tasks, XNLI identifies the relationship between a pair of sentences. Transferring knowledge about sentence relationships is more complicated and requires more samples to learn, in contrast to the other two tasks where semantic information from similar cross-lingual sentences can be transferred directly.

6.2 Effect of Languages

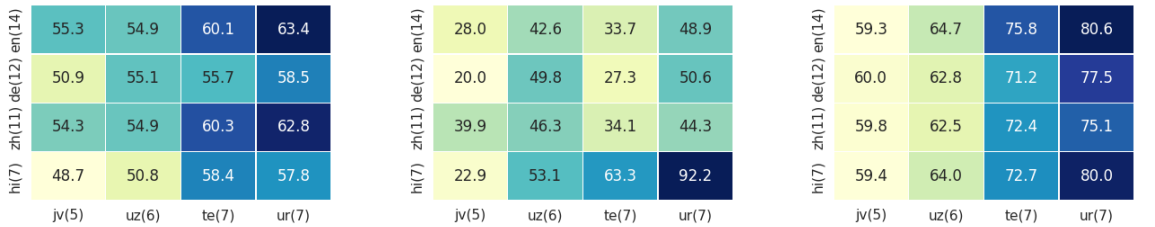
Lauscher et al. (2020) pointed out that two linguistic factors exert crucial effects on cross-lingual transfer performance: (1) the size of the pretraining corpus for the target language and (2) the similarity between the source and target language. In our study, we also consider a third factor: (3) the size of the pretraining corpus for the source language. In this section, we conduct a correlation analysis between PARC's cross-lingual transfer performance and the three language-related factors mentioned above. To achieve that, we have to measure these factors in a proper way at first. The size of the pretraining corpus can be easily measured by the log_2 value of the Wikipedia size in MB, as we mentioned in §4. Thus the remaining problem is how to properly represent language similarity.

6.2.1 Measurement of Language Similarity

Malaviya et al. (2017) and Littell et al. (2017) propose LANG2VEC from linguistic, typological, and phylogenetic perspectives. LANG2VEC employs different vectors to represent various types of lin-

		En	Af	Jv	Mn	My	Sw	Ta	Te	Tl	Ur	Uz	Avg
MAJ		25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Direct		52.5	41.8	27.4	42.5	32.2	31.3	31.5	33.0	31.6	46.9	44.8	36.3
UN	k=1	53.7	52.8	46.2	46.5	46.1	42.8	43.3	44.3	45.0	51.0	49.7	46.7
	k=3	55.8	53.6	46.2	47.1	48.2	44.9	44.5	46.3	47.1	52.6	51.0	48.1
	k=5	57.1	54.4	47.0	47.0	48.0	46.6	44.8	45.8	48.5	53.1	52.3	48.7
	k=10	57.5	55.3	46.3	46.4	47.6	45.6	44.1	46.7	47.7	53.0	51.4	48.4
	k=20	59.7	57.2	48.1	46.7	50.0	47.9	46.0	**48.9**	49.6	55.4	53.2	50.3
	k=30	**60.1**	**57.4**	**49.0**	**47.4**	**51.1**	**49.2**	**47.1**	48.7	**50.1**	**56.5**	**54.4**	**51.1**
LB	k=1	74.9	75.4	68.1	63.5	68.2	64.0	62.8	65.6	64.8	72.5	71.4	67.6
	k=3	77.1	77.1	69.6	65.6	71.1	67.6	65.6	68.4	65.9	74.6	74.4	70.0
	k=5	78.1	78.6	69.0	64.4	72.9	68.8	65.9	69.3	66.4	75.8	75.4	70.6
	k=10	78.7	79.4	70.5	67.0	72.9	68.3	66.6	70.7	67.2	76.6	75.9	71.5
	k=20	**79.0**	79.7	70.7	67.5	72.5	**70.0**	67.5	70.7	68.1	**77.4**	76.3	72.0
	k=30	79.0	**79.7**	**71.3**	**67.6**	72.8	69.9	**68.1**	**71.1**	**69.4**	77.2	**76.7**	**72.4**

Table 2: Results of topic categorization task on AG News dataset. k is the number of retrieved cross-lingual sample. MAJ is the majority baseline. Avg is the average accuracy across 10 LRLs. En is the HRL for retrieval. BoR strategy is adopted.



(a) Zero-Shot Performance (Unlabeled) (b) Language Similarity (c) Zero-Shot Performance (labeled)

Figure 3: Visualization of the correlation between zero-shot performance and language similarity, pretraining data size of source and target language. On the X(Y)-axis are target(source) languages with an increasing order of pretraining data size from left(bottom) to right(top). (a) and (c) show the zero-shot performance with PARC-unlabeled and PARC-labeled on Amazon review task respectively. (b) shows the language similarity of each pair.

guistic features for different languages. Each language is encoded with 5 vectors corresponding to different linguistic features including three typological features (syntax, phonology and phonetic inventory), phylogenetic and geographical features. In typological vectors, each dimension represents a linguistic property. For example, one dimension of the syntax vector represents the word order feature SVO. If a language has a SVO order, then its syntax vector would have the value 1 on this dimension. Missing values in the typological vectors could have detrimental effects. Therefore we replace them with values predicted from the k most similar typological vectors (Malaviya et al., 2017). The phylogenetic vector embodies the position of a language in the world language family tree (Harald et al., 2015), while the geographical vector contains the position information of languages w.r.t. their speakers.

Following prior work (Rama et al., 2020), we consider all 5 linguistic features when measuring the language similarity: syntax (SYN), phonology (PHO), phonological inventory (INV), language family (FAM), and geography (GEO). Given these different types of vectors, we calculate 5 cosine similarities for each pair of source language (i) and target language (j) and average them to get the final language similarity $sim(i, j)$:

$$sim(i,j) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} s(\mathbf{v}_f(i), \mathbf{v}_f(j)) \quad (6)$$

where $\mathcal{F}$ is the set of features, $\mathbf{v}_f(i)$ and $\mathbf{v}_f(j)$ stand for the language vectors representing the feature f for i and j, and $s(\cdot)$ computes the min-max normalized cosine similarity of the two vectors. The detailed cosine similarities between English and 10 LRLs evaluated in our experiment are shown in Table 9 in Appendix §B.

6.2.2 Correlation Analysis

We conduct a correlation analysis between cross-lingual performance and the three language factors mentioned above: language similarity between the

Unlabeled	Sim.		source size		target size	
	corr	p	corr	p	corr	p
Spearman	0.28	0.05	0.20	0.16*	0.31	0.03
Pearson	0.27	0.06*	0.22	0.12*	0.38	6e-03
labeled	Sim.		source size		target size	
	corr	p	corr	p	corr	p
Spearman	0.42	2e-03	0.08	0.54*	0.44	1e-03
Pearson	0.41	3e-03	-3e-4	1.00*	0.46	8e-4

Table 3: Correlations between Amazon review performance and three features. Sim.: language similarity between an LRL and an HRL; source (target) size: the log of the data size (MB) of source (target). *: insignificant result with a p value larger than 0.05.

		Amazon	**AGNews**	**XNLI**	**Avg.**
Direct		53.8	36.2	33.1	41.0
UN	mBERT+pooling	53.1	36.9	33.6	41.2
	mBERT+distiluse	54.7	38.4	34.0	42.3
	mBERT+paraphrase	59.6	46.7	33.7	46.7
	XLM-R+paraphrase	**70.1**	**57.4**	34.7	**54.1**
	mBERT+LaBSE	59.4	43.8	**35.1**	46.1
LB	mBERT+pooling	53.6	58.0	33.8	48.5
	mBERT+distiluse	62.8	63.8	34.6	53.7
	mBERT+paraphrase	72.9	67.6	36.8	59.1
	XLM-R+paraphrase	**73.0**	76.0	35.7	61.6
	mBERT+LaBSE	72.2	**80.0**	**37.5**	**63.2**

Table 4: Accuracy with different models used in our approach. pooling: cosine similarity of the last hidden states from the MPLM; distiluse: *distiluse-base-multilingual-cased-v2*, sentence transformer of multilingual distilBERT; paraphrase: *paraphrase-multilingual-mpnet-base-v2*, sentence transformer of XLM-R. UN: unlabeled setup; LB: labeled setup.

source (retrieved) and *target* (input) language, pretraining data size of the source language and of the target language. We use the log value of Wikipedia size to represent the size of pretraining corpus for target and source languages and $sim(i, j)$ computed by Eq. (6) to represent the similarity between the source and target language. Four other HRLs – Chinese, German, Hindi, Cebuano – are selected as source languages in addition to English. We measure the cross-lingual performance of PARC on the Amazon product review task in both the labeled and the unlabeled settings. Full results can be found in Appendix §D.2.

Table 3 shows the outcome of the correlation analysis. We observe a significant positive correlation between cross-lingual performance and language similarity as well as target language pretraining data size, in both the labeled and the unlabeled setting. The correlation between performance and source language size is not significant. Figure 3 visualizes the correlations and further clarifies the findings by selecting 4 source languages and 4 target languages and showing the cross-lingual performance and similarity between them.

		Ig	**Sn**	**Mt**	**Co**	**Sm**
Direct		30.3	32.1	29.8	32.6	30.4
LB	k=1	56.5	59.7	63.9	75.0	52.0
	k=3	58.1	61.4	65.2	78.2	54.1
	k=5	**58.8**	**61.6**	**65.9**	**79.8**	**55.4**
UN	k=1	36.6	37.3	39.1	42.6	34.4
	k=3	34.8	36.2	37.6	40.6	33.9
	k=5	34.8	35.3	37.2	40.4	34.1
		St	**Haw**	**Zu**	**Ny**	**Avg.**
Direct		30.4	27.1	34.4	29.8	30.8
LB	k=1	53.5	49.9	58.0	54.9	58.1
	k=3	55.5	49.7	58.5	57.0	59.7
	k=5	**56.8**	**51.4**	**58.8**	**58.0**	**60.7**
UN	k=1	36.3	31.6	35.6	35.3	36.5
	k=3	33.7	31.0	34.3	32.9	35.0
	k=5	34.2	30.6	34.0	32.0	34.7

Table 5: Results of several unseen languages on a topic categorization task (AG News dataset). Ig - Igbo, Sn - Shona, Mt - Maltese, Co - Corsican, Sm - Samoan, St - Sesotho, Haw - Hawaiian, Zu - Zulu, Ny - Chiechewa.

6.3 Robustness

In this section, we test the robustness of the PARC method w.r.t. other cross-lingual retrievers and MPLMs as well as unseen languages.

6.3.1 Retriever and MPLM

Apart from the multilingual sentence transformer based on XLM-R ("paraphrase") used in our previous experiments, we explore several other types of cross-lingual retrievers: a "pooling" retriever which averages the last hidden states of the MPLM and computes the cosine similarity between these pooled sentence representations; "distiluse" retriever, a sentence transformer based on multilingual distilBERT (Sanh et al., 2019); and the "LaBSE" retriever (Feng et al., 2020), a BERT-based model trained for sentence embedding for 109 languages. As an alternative to mBERT, we also investigate the performance of XLM-R, which has the same architecture as mBERT but is more powerful. We follow the setup described in §4.

Results are shown in Table 4. We can find that even the worst combination – *mBERT+pooling* – outperforms the *Direct* baseline on average under both labeled and unlabeled settings. If the retriever is replaced by a slightly more powerful one, such as the combination *mBERT+distiluse*, higher accuracies in the unlabeled and labeled setting are achieved, suggesting that our proposed method PARC is robust w.r.t. other cross-lingual retrievers. In the result of *XLM-R+paraphrase*, the obviously better performance of XLM-R in the unlabeled setup shows that a stronger MPLM can

		p1		p2		p3		p4		Avg	
		en	te	en	te	en	te	en	te	en	te
Finetune	Direct	84	76	83	70	86	67	85	73	85	74
	PARC-UN	84 –	65↓	85↑	62↓	83↓	60↓	82↓	64↓	84↓	67↓
	PARC-LB	83↓	64↓	83 –	64↓	83↓	64↓	82↓	70↓	83↓	69↓
w/o Finetune	Direct	54	53	59	54	54	50	53	51	55	52
	PARC-UN	59↑	55↑	55↓	58↑	52↓	52↑	53 –	52↑	55 –	54↑
	PARC-LB	**90**↑	**82**↑	**90**↑	**82**↑	**90**↑	**82**↑	**90**↑	**82**↑	**90**↑	**82**↑

Table 6: Result of English and Telugu on Amazon review task using MPLMs with and without finetuning on English train set. UN: Unlabeled, LB: labeled. p_i represents different prompt patterns.

noticeably improve the self-prediction. We expect that an even better performance could be obtained by applying our proposed PARC approach to larger and/or more powerful MPLMs such as InfoXLM (Chi et al., 2021).

6.3.2 Unseen Languages

Our previous experiments show that the LRLs pretrained by MPLMs can benefit well from PARC. However, popular MPLMs are pretrained only on approx. 100 languages, accounting for a tiny part of all languages in the world (∼100/7000). We wonder if our proposed method could potentially benefit a wider range of LRLs, so we apply PARC to several unseen LRLs, i.e. languages not included in the pretrained corpora of the MPLM. We conduct experiments on a topic categorization task for nine unseen languages. The results in Table 5 show that PARC is also effective for unseen LRLs. It can be observed from the result that PARC is also effective for unseen LRL languages.

6.4 Zero-shot Setting

Different from the cross-lingual transfer paradigm where a MPLM is first finetuned on annotated training data of one language, and then directly applied to the test data of other languages for inference, our proposed approach is employed in the zero-shot setting for LRLs, i.e., the model parameters are not adjusted by finetuning with HRL data. Table 6 shows results from a preliminary experiment where our PARC method combined with a finetuned MPLM fails to outperform the Direct baseline. When using finetuned MPLM to evaluate with PARC, we do not see sufficient performance improvement. However, without finetuning, PARC performs better in both unlabeled and labeled setup, and PARC-LB without finetuning also outperforms it with finetuning.

6.5 Qualitative Analysis

Table 7 shows results of the PARC pipeline for an example from the Amazon review task. The review

Amazon Review

Case #963
Input:
အဝတ်လျှော်အများအပြားဝန်နှင့်အတူအသုံးပြုခဲ့ကြသည်။ထည်အတွက်နူးညံ့သိမ်မွေ့ခြင်းနှင့်ငါ့အသားအရေကိုနူးညံ့သိမ်မွေ့ပုံရသည်။
(Used with several loads of laundry. Gentle on the fabric and gentle on my skin.) pos
Retrieved:
R1: Hard to wash. The fur on top gets all over the sides in the wash. :/ pos
R2: Very nice and thick high quality towels. pos
R3: Smelled really bad mold! I had to wash them before use. neg
Predictions: No retrieval - neg, **k=1 -** neg, **k=3 -** pos

Table 7: A PARC pipeline example for Amazon review task in the labeled setting.

in Telugu is positive, but the class predicted without cross-lingual context is negative. The prediction stays the same when a single positive English sample is added as prompt context. When two more English samples are added, the prediction becomes correct.

This case indicates that the retrieved cross-lingual samples help the MPLM make a correct decision. Furthermore, more similar HRL samples could rectify the deviation. More cases are shown in Table 10 and Table 11 in Appendix §C.

7 Conclusion

We propose PARC, a pipeline that augments prompts for zero-shot learning on low resource languages by retrieving semantically similar cross-lingual sentences from HRL corpora. We test PARC on three classification tasks with parallel test sets across 10 LRLs, and it performs better than the baselines in both unlabeled and labeled settings. Increasing the number of retrieved prompts improves performance at first, but deteriorates it after a certain point. A robustness study shows that PARC also performs well with other cross-lingual retrievers or MPLMs, suggesting potential applications of PARC to a wider scope of scenarios.

Limitations

The PARC pipeline proposed in this work is designed to improve the cross-lingual transfer performance for low-resource languages in a zero-shot setting. We tested our method on different LRLs contained in MPLMs and also investigate its effectiveness for several unseen languages. These are not included in pretraining corpora of the MPLM but use a seen script and share some subwords with the seen languages. However, our proposed method is not applicable for unseen languages with new scripts, which restricts its extension towards a wider range of languages. Besides, PARC is a retrieval-based method. More time and computational resources are required in the cross-lingual retrieval phase. Therefore, it is computationally less efficient to use PARC for inference.

Acknowledgements

This work was supported by European Research Council (# 740516), Munich Center for Machine Learning (MCML) and China Scholarship Council (CSC).

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Jishnu Ray Chowdhury, Yong Zhuang, and Shuyi Wang. 2022. Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning. In *AAAI*.

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, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Matthew Cer, N. Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. In *Annual Meeting of the 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.

Hammarström Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. glottolog-data: Glottolog database 2.6.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. *ArXiv*, abs/2202.11451.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yanchen Liu, Timo Schick, and Hinrich Schütze. 2022b. Semantic-oriented unlabeled priming for large-scale language models. *arXiv preprint arXiv:2202.06133*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2022c. MulZDG: Multilingual code-switching framework for zero-shot dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 648–659, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *ArXiv*, abs/1903.10561.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *North American Chapter of the Association for Computational Linguistics*.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual bert for genetic and typological signals. In *International Conference on Computational Linguistics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179, Dublin, Ireland. Association for Computational Linguistics.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2018. R3: Reinforced ranker-reader for open-domain question answering. In *AAAI*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. 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. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *ArXiv*, abs/1509.01626.

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.

A Effect of Translations

In our experiment, we use multilingual parallel test sets created by machine translation from English to target low-resource languages. To explore the effect of machine translation-created test sets, we compare the cross-lingual transfer performance on translation test sets and original test sets of XNLI. The original XNLI datasets include two low-resource languages that we used in our experiments, i.e., Swahili (sw) and Urdu (ur). We also measure the translation quality by using the original sets as gold standard. The analysis results (Table 8) suggests that machine translated test sets are useful as a proxy for evaluating cross-lingual performance on LRLs.

Languages		sw	ur
Performance	MT Acc.	34.00	33.92
	OV Acc.	34.07	33.87
	Diff	0.07	-0.05
	P-Value	0.85	0.92
Translation Quality	BLEU	56.39	64.96
	chrF	49.58	59.89
	Sim.	81.82	81.19

Table 8: Comparison of performance on machine translation-created XNLI test sets (MT) and the original version of XNLI test sets (OV) in sw and ur languages. BLEU & chrF scores and semantic similarities (Sim.) are computed to measure the translation quality of machine translation-created test sets.

B Language Features

Table 9 shows the language features of all 10 LRLs evaluated in our experiments. Language similarity refers to the similarity between each LRL and English. SIM score is computed by Eq. (6). WikiSize is the log value of the Wikipedia size in MB.

C Case Study

Table 10 shows two examples from the Amazon Review task. We compare the predictions for three scenarios: no retrieval information (i.e., Direct baseline, see §4.2), one retrieved sample, and three retrieved samples. Similarly, Table 11 shows the same comparison on the AG News task.

D Detailed Results

D.1 Results for each task

We show the detailed experimental results for all tasks in Table 12 (Amazon reviews), Table 13 (AG News) and Table 14 (XNLI), respectively.

Lang	Language Similarity						Wiki Size
	SYN	PHO	INV	FAM	GEO	SIM	
Af	84.9	60.3	38.4	50.4	33.1	53.4	6
Jv	48.0	39.2	52.7	0.0	0.0	28.0	5
Mn	31.0	100.0	39.4	0.0	56.8	45.4	5
My	17.4	80.3	100.0	0.0	37.6	47.1	5
Ta	28.9	60.3	51.5	0.0	72.7	42.7	7
Te	36.0	56.2	31.3	0.0	45.2	33.7	7
Tl	35.0	70.5	26.7	0.0	38.8	34.2	6
Sw	27.0	87.0	62.1	0.0	57.2	46.6	5
Ur	50.2	72.0	47.1	12.6	62.5	48.9	7
Uz	39.8	75.6	24.1	0.0	73.7	42.6	6

Table 9: List of language features of the 10 LRLs that we evaluate.

Amazon Review

Case 1 #37
Input:
ငါ့ဆံပင်ပေါ်မှာအလွန်ခြောက်သွေ့။
(Very dry on my hair.) neg
Retrieved:
R1: It's a little bit too greasy in my opinion. Doesn't really seem to soak into the hair very well. pos
R2: The tiniest amount leaves my hair stringy and oily. neg
R3: could smell this stuff all day but I don't feel like it moisturizes my skin enough, and my skin isn't overly dry to begin with. pos
Predictions: No retrieval - pos, **k=1 -** neg, **k=3 -** neg

Case 2 #963
Input:
အဝတ်လျှော်အများအပြားဝန်နှင့်အတူအသုံးပြုခဲ့ကြသည်။ထည်အတွက်နူးညံ့သိမ်မွေ့ခြင်းနှင့်ငါ့အသားအရေကိုနူးညံ့သိမ်မွေ့ပုံရသည်။
(Used with several loads of laundry. Gentle on the fabric and gentle on my skin.) pos
Retrieved:
R1: Hard to wash. The fur on top gets all over the sides in the wash. :/ pos
R2: Very nice and thick high quality towels. pos
R3: Smelled really bad mold! I had to wash them before use. neg
Predictions: No retrieval - neg, **k=1 -** neg, **k=3 -** pos

Table 10: PARC examples for Amazon Review task.

D.2 Detailed data for Correlation Analysis

Table 16 shows the detailed data used for correlation analysis of language similarity, high- and low-resource language pretraining data size with cross-lingual performance in the unlabeled setting as well as labeled setting.

D.3 Complete Results for Robustness Analysis

Table 17 shows the results of each language using different combinations of retriever and MPLM for validating the robustness on three tasks.

AG News

Case 1 #1939
Input:
ပန်းပွင့်ပါဝါသည်ပန်းများကိုအသံချဲ့စက်များသို့လှည့်လာသည့်နည်းလမ်းဖြင့်ဂျပန်ကုမ္ပဏီ တစ်ခုပေါ်လာသည်။

(Flower Power A Japanese company has come up with a way to turn flowers into amplifiers.) Tech
Retrieved:
R1: Japanese firms step up spending Japanese firms continue to spend on new equipment and production plants, a survey finds, underlining a continuing recovery in the world's second-largest economy. Business
R2: IBM, Honda deliver in-car speech-recognition navigation system IBM and Honda have jointly developed a hands-free and natural sounding in-vehicle speech-recognition system that will be offered as standard equipment on the 2005 Acura RL Tech
R3: Scientists Make Phone That Turns Into a Sunflower (Reuters) Reuters - Scientists said on Monday they have come up with a cell phone cover that will grow into a sunflower when thrown away. Tech
Predictions: No retrieval - World, **k=1 -** Tech, **k=3 -** Tech

Case 2 #1302
Input:
လျှပ်တပြက်အတွင်း ရုပ်ရှင်များ- Netflix နှင့် TiVo တို့သည် ဒေါင်းလုဒ်များကို Bee Staff Writer ဆွေးနွေးကြသည်။ TiVo Inc ၏ပိုင်ရှင်များကိုခွင့်ပြုမည့် Silicon Valley မဟာမိတ်အသစ်၏အသံများကြားတွင်နည်းပညာမြေပြင်မြေပြင်သည်ခြေလျင်အောက်သို့ ပြောင်းလာသည်။

(Movies in a Snap: Netflix and TiVo Discuss Downloads Bee Staff Writer. The high-tech terrain is shifting underfoot amid rumblings of a new Silicon Valley alliance that would allow the owners of TiVo Inc.) Business
Retrieved:
R1: NETFLIX, TIVO HOOKUP CLOSE Netflix and TiVo are in late-stage talks on a partnership that would let subscribers use the Internet to download Netflix movies directly into their TiVo box, The Post has learned. Business
R2: TiVo and NetFlix: Picture-Perfect Duo? With TiVo (TIVO) and NetFlix (NFLX) finally announcing a long-rumored partnership to launch a video-on-demand service sometime next year, investors smiled on the deal that will keep the two popular, but under-fire, innovators ahead of competitors. Tech
R3: New Treo and more unveiled at CTIA CTIA stands for the Cellular Telecommunications and Internet Association. Each year they host two shows for the industry. This week is their fall Wireless IT and Entertainment expo in San Francisco. Business
Predictions: No retrieval - World, **k=1 -** Tech, **k=3 -** Business

Table 11: PARC examples for AG News task

pattern 0	[X] [MASK]
pattern 1	It was [MASK]. [X]
pattern 2	[X] All in all, it was [MASK].
pattern 3	Just [MASK]! [X]
pattern 4	[X] In summary, the product is [MASK].

		en					**af**					**ur**				
		p0	p1	p2	p3	p4	p0	p1	p2	p3	p4	p0	p1	p2	p3	p4
MAJ		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Direct		50.5	54.3	58.9	53.7	52.6	53.3	50.7	50.4	49.8	51.5	49.9	51.7	54.6	49.9	50.3
Unlabeled	k=1	**50.9**	**55.4**	**59.1**	**51.9**	**52.6**	**51.0**	**54.9**	**57.9**	**52.9**	**52.8**	**51.6**	**56.7**	**60.0**	**52.2**	**52.2**
	k=3	50.7	53.7	57.7	50.8	50.4	50.4	52.5	56.2	50.7	51.0	51.3	52.9	57.1	50.8	50.9
	k=5	50.8	52.2	56.0	50.3	50.9	50.8	52.2	55.0	50.2	50.6	51.2	52.5	56.4	50.3	50.7
	k=10	50.7	51.9	56.0	50.0	50.6	50.7	52.0	55.8	50.2	50.7	51.4	52.4	55.5	50.0	50.3
	k=20	50.5	50.8	53.6	49.9	50.1	50.5	51.1	53.5	50.0	50.2	51.1	51.2	54.0	49.8	50.0
labeled	k=1	**60.0**	82.4	82.4	82.3	82.4	66.0	79.0	79.2	79.2	79.2	**57.0**	80.4	80.6	80.6	80.6
	k=3	58.5	86.2	86.2	86.2	86.2	65.0	80.7	81.1	81.1	81.0	56.4	83.8	84.3	84.3	84.3
	k=5	57.3	87.2	87.2	87.2	87.2	65.4	82.7	82.9	82.9	82.8	56.2	84.6	85.0	85.0	85.0
	k=10	57.7	88.9	88.9	88.9	88.9	**66.5**	85.2	85.4	85.4	85.4	56.6	87.0	87.3	87.3	87.3
	k=20	56.4	**89.5**	**89.5**	**89.5**	**89.5**	64.3	85.3	**85.7**	**85.7**	**85.6**	55.4	**87.6**	**87.9**	**87.9**	**88.0**
	k=30	56.3	88.9	88.9	88.9	88.9	63.6	85.4	85.6	85.6	85.6	55.7	87.4	87.6	87.6	87.6

		sw					**te**					**ta**				
		p0	p1	p2	p3	p4	p0	p1	p2	p3	p4	p0	p1	p2	p3	p4
MAJ		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Direct		47.3	50.2	51.9	49.9	50.3	50.8	52.5	53.9	49.9	51.4	54.1	59.0	56.2	50.5	51.9
Unlabeled	k=1	**51.4**	**50.4**	**50.5**	**50.5**	**50.1**	**51.6**	**54.8**	**57.5**	**52.3**	**52.1**	**57.1**	**55.3**	**57.2**	**52.6**	**51.6**
	k=3	50.5	50.3	50.3	50.1	50.1	51.3	52.8	55.3	50.6	51.3	55.7	52.5	55.0	50.5	50.6
	k=5	50.6	50.1	50.0	50.1	50.1	51.6	51.7	54.0	50.4	50.3	56.1	51.4	54.0	50.1	50.1
	k=10	50.8	50.1	50.0	50.1	50.1	51.8	52.1	53.5	50.4	50.3	57.3	51.5	53.9	50.0	50.1
	k=20	50.5	50.1	50.0	50.1	50.1	51.4	50.6	52.9	50.0	50.0	56.9	50.5	52.9	50.0	50.0
labeled	k=1	50.5	50.0	49.9	49.9	49.9	**58.2**	75.9	75.8	75.8	75.8	68.1	75.3	75.4	75.4	75.4
	k=3	51.0	54.1	54.1	54.1	54.1	58.0	78.4	78.4	78.4	78.4	70.2	79.1	79.3	79.3	79.2
	k=5	50.7	54.4	54.4	54.4	54.4	56.8	79.1	79.0	79.0	79.1	70.7	80.5	80.5	80.5	80.5
	k=10	**51.3**	**55.5**	**55.5**	**55.5**	**55.5**	57.2	81.3	81.6	81.6	81.6	**70.9**	**83.7**	**83.9**	**83.9**	**83.9**
	k=20	50.9	54.3	54.4	54.4	54.4	56.9	**82.0**	**82.1**	**82.1**	**82.1**	70.8	82.8	83.1	83.1	83.1
	k=30	50.7	54.3	54.3	54.3	54.3	56.8	82.0	82.0	82.0	82.0	70.5	83.3	83.5	83.4	83.4

		mn					**uz**					**my**				
		p0	p1	p2	p3	p4	p0	p1	p2	p3	p4	p0	p1	p2	p3	p4
MAJ		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Direct		49.1	49.7	51.4	49.7	50.0	48.5	50.2	52.4	49.7	**51.2**	**54.4**	**56.1**	**56.1**	50.5	**52.6**
Unlabeled	k=1	**51.1**	**54.7**	**58.6**	**52.6**	**52.8**	50.4	**53.1**	**53.6**	**51.8**	50.9	53.0	53.9	56.0	**52.3**	52.0
	k=3	50.2	53.2	56.4	51.0	51.1	50.5	51.9	52.1	50.2	50.3	53.0	51.5	55.0	51.2	50.7
	k=5	50.2	52.0	55.3	50.4	50.5	50.5	50.3	50.7	50.0	50.2	52.9	51.1	53.6	50.5	50.3
	k=10	50.4	52.2	56.3	**50.6**	50.5	50.6	50.3	50.6	50.1	50.0	53.4	51.1	54.2	50.2	50.1
	k=20	50.4	51.1	54.5	50.0	50.0	50.5	50.0	50.7	50.0	50.0	53.2	50.5	52.8	50.0	50.0
labeled	k=1	60.8	74.9	74.9	74.9	74.9	**56.0**	65.0	64.7	64.7	64.7	65.3	73.9	73.8	73.8	73.8
	k=3	60.3	79.5	79.7	79.7	79.7	55.2	65.3	65.2	65.2	65.2	66.6	77.5	77.7	77.7	77.7
	k=5	59.7	80.6	80.6	80.6	80.6	55.5	66.1	66.0	66.0	65.8	65.8	78.6	78.9	78.9	78.9
	k=10	**62.2**	**83.9**	**84.3**	**84.3**	**84.3**	55.9	**68.1**	**68.2**	**68.2**	**68.3**	**67.8**	80.9	81.1	81.1	81.1
	k=20	60.3	82.5	83.2	83.2	83.2	53.8	67.0	67.1	67.1	67.1	67.4	**81.8**	**81.8**	**81.8**	**81.8**
	k=30	59.7	83.3	83.8	83.8	83.8	54.4	67.5	67.7	67.7	67.7	67.6	81.7	81.8	81.8	81.8

		jv					**tl**					**Avg.**				
		p0	p1	p2	p3	p4	p0	p1	p2	p3	p4	p0	p1	p2	p3	p4
MAJ		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Direct		**50.9**	52.3	54.1	50.1	**52.3**	49.6	50.4	**51.9**	50.0	**51.2**	50.8	52.5	53.8	50.3	51.4
Unlabeled	k=1	50.6	**53.0**	54.2	**50.9**	50.5	**50.4**	**50.6**	50.9	50.1	50.2	**51.7**	**53.9**	**56.0**	**51.8**	**51.6**
	k=3	50.2	51.7	**53.5**	50.4	50.3	50.0	50.3	50.3	**50.2**	50.0	51.2	52.1	54.4	50.6	50.6
	k=5	50.2	50.9	52.9	50.1	50.2	50.1	50.2	50.1	50.0	50.1	51.4	51.3	53.5	50.2	50.4
	k=10	50.1	50.7	52.5	49.9	50.0	50.2	50.0	50.3	50.0	50.0	51.6	51.3	53.5	50.1	50.2
	k=20	50.5	50.1	51.7	50.0	50.0	50.2	50.0	50.4	50.0	50.0	51.4	50.5	52.5	50.0	50.0
labeled	k=1	**54.1**	59.3	59.3	59.3	59.3	52.4	55.4	55.4	55.4	55.4	58.9	70.1	68.9	70.1	70.1
	k=3	52.7	61.6	61.6	61.6	61.6	52.1	57.7	57.7	57.7	57.7	58.7	73.1	73.2	73.2	73.2
	k=5	52.8	61.5	61.5	61.5	61.5	51.6	60.2	60.2	60.2	60.1	58.4	74.1	74.2	74.2	74.2
	k=10	51.6	**62.6**	**62.6**	**62.6**	**62.6**	**52.4**	**63.2**	**63.3**	**63.3**	**63.3**	**59.1**	**76.4**	**76.5**	**76.5**	**76.5**
	k=20	51.6	61.5	61.5	61.5	61.5	51.5	62.8	62.9	62.9	62.9	58.1	76.1	76.3	76.3	76.3
	k=30	51.6	60.9	61.0	61.0	61.0	51.5	62.3	62.4	62.4	62.4	58.0	76.1	76.2	76.2	76.2

Table 12: Results on Amazon reviews dataset.

pattern 0	[X] [MASK]
pattern 1	[MASK]: [X]
pattern 2	[MASK] News: [X]
pattern 3	[X] Category: [MASK]

		en				**af**				**ur**			
		p0	p1	p2	p3	p0	p1	p2	p3	p0	p1	p2	p3
MAJ		25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Direct		52.5	47.8	**47.3**	53.0	41.8	41.3	40.2	**57.8**	27.4	32.4	33.0	**53.5**
Unlabeled	k=1	53.7	47.6	45.6	53.2	52.8	**46.8**	**46.2**	53.2	46.2	**41.8**	**41.0**	49.7
	k=3	55.8	47.6	43.4	54.3	53.6	46.5	44.3	54.3	46.2	40.5	38.2	49.9
	k=5	57.1	**48.3**	41.7	**55.6**	54.4	46.9	43.7	55.1	47.0	40.9	37.2	51.4
	k=10	57.5	45.7	41.9	55.3	55.3	44.6	42.3	55.6	46.3	38.3	35.3	51.9
	k=20	**59.7**	46.7	41.5	55.3	**57.2**	45.9	42.2	56.1	**48.1**	39.7	35.5	51.6
labeled	k=1	74.9	83.5	83.8	83.8	75.4	81.2	82.9	82.7	68.1	76.9	78.8	78.7
	k=3	77.1	86.5	86.8	86.7	77.1	84.3	85.4	85.2	69.6	79.4	81.7	81.8
	k=5	78.1	87.7	88.0	87.9	78.6	86.8	87.1	87.1	69.0	79.9	82.7	82.7
	k=10	78.7	88.2	88.5	88.5	79.4	87.2	87.7	87.5	70.5	81.5	**83.6**	**83.4**
	k=20	**79.0**	**89.1**	**89.4**	**89.4**	**79.7**	**87.4**	**87.8**	**87.5**	**70.7**	**81.6**	83.3	83.2

		sw				**te**				**ta**			
		p0	p1	p2	p3	p0	p1	p2	p3	p0	p1	p2	p3
MAJ		25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Direct		42.5	37.6	33.3	**56.6**	32.2	37.2	32.5	**55.4**	31.3	37.2	28.6	55.1
Unlabeled	k=1	46.5	**42.1**	**42.0**	46.4	46.1	**41.5**	**43.3**	48.6	42.8	**41.6**	**39.2**	47.6
	k=3	**47.1**	41.2	39.9	47.9	**48.2**	40.0	42.4	50.3	44.9	41.0	36.9	50.1
	k=5	47.0	41.5	39.3	48.6	48.0	40.4	41.0	52.4	46.6	39.8	36.0	50.9
	k=10	46.4	38.5	37.0	50.0	47.6	39.0	39.3	51.8	45.6	37.8	33.9	51.5
	k=20	46.7	39.1	36.9	49.9	50.0	40.1	39.7	51.6	**47.9**	38.8	34.7	**52.5**
labeled	k=1	63.5	68.4	70.3	70.3	68.2	73.9	75.0	75.0	64.0	69.7	71.5	71.5
	k=3	65.6	70.8	72.3	72.4	71.1	77.6	78.2	78.2	67.6	74.4	75.7	75.7
	k=5	64.4	72.2	73.5	73.4	**72.9**	79.7	79.9	79.8	68.8	75.8	76.6	76.5
	k=10	67.0	72.5	**74.1**	**73.9**	72.9	79.9	80.0	80.0	68.3	76.5	77.2	77.1
	k=20	**67.5**	**72.7**	73.6	73.6	72.5	**80.2**	**80.6**	**80.6**	**70.0**	**77.5**	**78.1**	**78.2**

		mn				**uz**				**my**			
		p0	p1	p2	p3	p0	p1	p2	p3	p0	p1	p2	p3
MAJ		25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Direct		31.5	30.9	32.0	47.3	33.0	37.5	33.8	50.7	31.6	37.4	33.7	51.9
Unlabeled	k=1	43.3	**42.5**	**41.5**	48.2	44.3	**44.4**	**42.3**	49.0	45.0	43.9	**43.6**	50.0
	k=3	44.5	41.2	40.5	51.1	46.3	42.2	40.7	50.9	47.1	**44.5**	41.7	53.7
	k=5	44.8	41.5	39.6	51.8	45.8	41.7	39.2	52.3	48.5	43.8	41.4	54.2
	k=10	44.1	39.7	38.0	**53.3**	46.7	39.7	37.9	**53.4**	47.7	41.4	40.0	**54.4**
	k=20	**46.0**	39.7	37.9	52.8	**48.9**	41.2	36.9	53.1	**49.6**	42.2	40.3	53.6
labeled	k=1	62.8	70.9	72.7	72.8	65.6	71.5	73.2	73.3	64.8	76.2	77.4	77.2
	k=3	65.6	75.4	77.3	77.2	68.4	73.6	75.7	75.7	65.9	79.5	80.1	79.8
	k=5	65.9	75.8	78.0	77.9	69.3	76.1	77.9	77.8	66.4	81.4	82.5	81.8
	k=10	66.6	77.0	**78.7**	**78.6**	70.7	76.4	78.3	78.2	67.2	82.4	82.9	82.3
	k=20	**67.5**	**77.4**	78.2	78.0	**70.7**	**77.3**	**78.8**	**78.7**	**68.1**	**83.1**	**83.6**	**83.3**

		jv				**tl**				**Avg**			
		p0	p1	p2	p3	p0	p1	p2	p3	p0	p1	p2	p3
MAJ		25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Direct		46.9	39.3	38.0	**59.3**	44.8	44.4	42.6	**60.4**	37.8	38.4	36.2	50.9
Unlabeled	k=1	51.0	**45.5**	**45.4**	51.6	49.7	**45.8**	**43.7**	52.2	47.4	**44.2**	**43.5**	48.9
	k=3	52.6	44.6	42.0	53.5	51.0	45.3	42.7	54.0	48.8	43.6	41.9	50.3
	k=5	53.1	44.5	41.3	53.6	52.3	45.2	41.8	54.2	49.5	43.7	41.2	51.0
	k=10	53.0	42.4	39.9	54.0	51.4	44.0	39.8	54.9	49.2	41.7	39.7	51.2
	k=20	**55.4**	42.8	40.1	54.2	**53.2**	44.4	38.9	55.3	**51.1**	42.6	39.9	**51.4**
labeled	k=1	72.5	77.8	79.1	79.1	71.4	76.6	78.9	79.0	68.3	74.6	75.9	75.9
	k=3	74.6	80.5	82.3	82.3	74.4	80.7	82.1	82.2	70.6	77.8	78.9	78.9
	k=5	75.8	81.3	82.8	82.8	75.4	81.2	83.4	83.5	71.3	79.1	80.2	80.1
	k=10	76.6	82.0	84.0	84.2	75.9	82.4	**84.5**	**84.6**	72.1	79.8	80.9	80.8
	k=20	**77.4**	**82.8**	**84.6**	**84.8**	**76.3**	**82.8**	84.0	84.0	**72.6**	**80.4**	**81.1**	**81.1**

Table 13: Results on AG News dataset.

pattern 0	$[X_1]$ [MASK] $[X_2]$
pattern 1	$[X_1]$? [MASK], $[X_2]$ (Yes - No)
pattern 2	$[X_1]$? [MASK], $[X_2]$ (Right - Wrong)

		en			**af**			**ur**			**sw**		
		p0	p1	p2	p0	p1	p2	p0	p1	p2	p0	p1	p2
MAJ		33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
Direct		33.3	**34.2**	34.3	33.2	33.0	33.4	**33.6**	34.0	33.2	33.2	32.2	33.1
Unlabeled	k=1	**34.1**	33.7	**34.5**	**34.0**	**34.1**	33.7	32.4	**35.3**	32.7	33.5	**33.7**	**33.7**
	k=3	33.7	**34.1**	34.3	33.0	32.9	34.1	33.3	34.0	**33.9**	**33.6**	33.0	33.5
	k=5	31.9	33.7	34.3	32.5	32.8	33.9	31.2	34.1	33.6	33.2	32.7	32.9
	k=10	31.9	33.6	33.3	31.9	33.3	32.6	32.2	34.2	33.2	33.0	32.7	32.5
	k=20	32.0	34.4	33.3	31.6	33.6	34.1	31.6	34.4	33.9	33.1	33.1	32.0
labeled	k=1	38.9	39.1	38.8	38.7	38.9	38.1	37.0	37.4	36.7	33.3	33.4	33.4
	k=3	39.2	39.1	38.6	37.9	37.9	37.4	37.0	37.8	36.8	33.7	33.5	33.7
	k=5	40.0	39.8	39.5	38.0	38.0	37.1	40.2	40.6	39.8	32.7	32.5	32.6
	k=10	41.5	41.6	40.9	41.1	41.1	40.5	42.0	42.4	41.0	33.7	33.7	34.1
	k=20	**44.5**	**44.1**	**43.5**	**42.3**	**43.0**	**41.3**	**42.4**	**43.4**	**42.2**	**35.9**	**35.7**	**35.9**

		te			**ta**			**mn**			**uz**		
		p0	p1	p2	p0	p1	p2	p0	p1	p2	p0	p1	p2
MAJ		33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
Direct		31.9	33.0	33.2	32.4	34.1	32.9	**33.0**	32.7	32.6	**33.3**	33.3	32.9
Unlabeled	k=1	**34.1**	34.1	**34.1**	**34.5**	34.3	33.3	32.8	33.6	**34.7**	33.2	33.9	32.8
	k=3	32.8	34.9	33.4	33.7	34.7	**34.2**	32.2	**34.5**	33.7	32.3	34.5	33.4
	k=5	32.9	**35.1**	33.8	32.9	34.3	33.9	31.9	33.9	34.1	33.1	**34.5**	**33.9**
	k=10	32.0	34.1	32.7	32.3	34.7	32.5	30.8	34.1	32.5	32.8	33.9	32.6
	k=20	31.5	34.6	32.7	32.5	**34.8**	32.9	32.0	34.1	33.4	32.6	33.5	32.6
labeled	k=1	37.8	38.1	37.7	37.7	38.0	37.0	36.5	36.5	36.5	35.5	34.8	35.0
	k=3	38.9	39.5	38.4	38.7	39.4	37.5	39.1	39.1	38.9	35.1	34.7	34.7
	k=5	37.5	37.1	35.9	38.3	38.7	36.3	37.1	36.9	36.9	36.0	35.9	35.9
	k=10	39.2	39.5	37.9	41.1	40.8	38.0	39.5	39.3	39.3	38.3	37.9	37.8
	k=20	**41.2**	**41.5**	**39.3**	**42.7**	**43.1**	**39.7**	**40.3**	**40.2**	**40.0**	**40.0**	**39.9**	**39.6**

		my			**jv**			**tl**			**Avg**		
		p0	p1	p2	p0	p1	p2	p0	p1	p2	p0	p1	p2
MAJ		33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
Direct		**33.7**	33.6	33.7	**33.3**	33.3	33.6	33.3	33.5	32.3	33.1	33.3	33.1
Unlabeled	k=1	33.3	33.5	**33.8**	32.4	32.0	33.3	33.8	32.7	32.8	**33.4**	33.7	33.5
	k=3	32.6	33.9	33.7	32.1	31.4	34.2	33.7	**33.9**	**33.3**	32.9	**33.7**	**33.7**
	k=5	32.5	**34.3**	33.6	32.4	31.6	34.3	**34.1**	33.5	32.1	32.7	33.6	33.6
	k=10	30.5	33.9	33.3	32.1	32.6	33.5	33.2	33.1	32.6	32.1	33.5	32.8
	k=20	30.9	33.5	32.7	30.8	**33.6**	**34.7**	32.9	32.5	33.1	32.0	33.6	33.2
labeled	k=1	36.8	36.7	36.1	34.2	33.5	33.3	34.7	34.4	34.3	36.2	36.2	35.8
	k=3	36.7	36.9	36.2	34.6	33.9	33.9	35.7	35.7	35.7	36.7	36.8	36.3
	k=5	37.7	37.7	37.3	**35.2**	**34.8**	**34.6**	35.7	35.7	35.3	36.9	36.8	36.2
	k=10	39.5	39.3	38.1	34.7	34.4	33.6	37.2	36.9	36.9	38.6	38.5	37.7
	k=20	**41.7**	**41.3**	**39.6**	32.8	32.8	32.4	**37.4**	**37.0**	**37.0**	**39.7**	**39.8**	**38.7**

Table 14: Results on XNLI dataset.

			En	**Af**	**Jv**	**Mn**	**My**	**Sw**	**Ta**	**Te**	**Tl**	**Ur**	**Uz**	**Avg**
MAJ			25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Direct			52.5	41.8	27.4	42.5	32.2	31.3	31.5	33.0	31.6	46.9	44.8	36.3
UN	k=1		53.7	52.8	46.2	46.5	46.1	42.8	43.3	44.3	45.0	51.0	49.7	46.7
	k=3	BoR	55.8	53.6	46.2	47.1	48.2	44.9	44.5	46.3	47.1	52.6	51.0	48.1
		CONC	53.5	52.4	45.9	44.9	44.8	42.9	41.7	46.6	46.0	52.0	51.6	46.9
	k=5	BoR	57.1	54.4	47.0	47.0	48.0	46.6	44.8	45.8	48.5	53.1	52.3	48.7
		CONC	53.5	48.0	38.2	41.3	36.3	36.9	39.5	41.4	42.9	50.5	49.6	42.4
	k=10	BoR	57.5	55.3	46.3	46.4	47.6	45.6	44.1	46.7	47.7	53.0	51.4	48.4
		CONC	46.4	41.1	36.2	38.3	36.6	34.9	34.6	35.8	40.7	46.3	45.0	38.9
	k=20	BoR	59.7	57.2	48.1	46.7	50.0	47.9	46.0	48.9	49.6	55.4	53.2	50.3
		CONC	50.0	48.4	42.3	41.4	43.3	43.1	39.3	44.3	48.1	47.9	48.4	44.6
	k=30	BoR	60.1	57.4	49.0	47.4	51.1	49.2	47.1	48.7	50.1	56.5	54.4	51.1
		CONC	50.7	47.6	43.9	38.2	42.9	42.5	41.8	44.5	47.7	47.1	47.3	44.3
LB	k=1		74.9	75.4	68.1	63.5	68.2	64.0	62.8	65.6	64.8	72.5	71.4	67.6
	k=3	BoR	77.1	77.1	69.6	65.6	71.1	67.6	65.6	68.4	65.9	74.6	74.4	70.0
		CONC	75.6	74.8	67.3	63.1	60.3	59.0	60.5	67.1	65.9	73.3	72.4	66.4
	k=5	BoR	78.1	78.6	69.0	64.4	72.9	68.8	65.9	69.3	66.4	75.8	75.4	70.6
		CONC	74.6	66.5	48.2	53.9	44.9	45.4	52.1	59.5	56.0	70.9	63.6	56.1
	k=10	BoR	78.7	79.4	70.5	67.0	72.9	68.3	66.6	70.7	67.2	76.6	75.9	71.5
		CONC	61.2	52.7	43.2	48.0	44.5	42.5	41.3	45.0	50.1	62.3	56.7	48.6
	k=20	BoR	79.0	79.7	70.7	67.5	72.5	70.0	67.5	70.7	68.1	77.4	76.3	72.0
		CONC	67.4	65.1	55.8	55.6	57.6	58.3	51.2	61.0	62.8	66.4	66.0	60.0
	k=30	BoR	79.0	79.7	71.3	67.6	72.8	69.9	68.1	71.1	69.4	77.2	76.7	72.4
		CONC	72.8	71.1	62.1	57.0	61.6	60.4	57.9	67.9	64.6	71.6	69.3	64.3

Table 15: Results of topic categorization task on AG News Dataset. k is the number of retrieved cross-lingual sample. MAJ is the majority baseline. Avg is the average accuracy across 10 LRLs. En is the HRL for retrieval. BoR refers to the *Bag of Retrieval* strategy, CONC refers to the *Concatenation* strategy.

	Performance		Language Similarity						WikiSize	
	Unlabeled	labeled	SYN	PHO	INV	FAM	GEO	SIM	source	target
en-af	79.2	62.0	84.9	60.3	38.4	50.4	33.1	53.4	14	6
en-ur	80.6	63.4	50.2	72.0	47.1	12.6	62.5	48.9	14	7
en-sw	49.9	51.0	27.0	87.0	62.1	0.0	57.2	46.6	14	5
en-te	75.8	60.1	36.0	56.2	31.3	0.0	45.2	33.7	14	7
en-ta	75.4	60.2	28.9	60.3	51.5	0.0	72.7	42.7	14	7
en-mn	74.9	62.9	31.0	100.0	39.4	0.0	56.8	45.4	14	5
en-uz	64.7	54.9	39.8	75.6	24.1	0.0	73.7	42.6	14	6
en-my	73.8	60.3	17.4	80.3	100.0	0.0	37.6	47.1	14	5
en-jv	59.3	55.3	48.0	39.2	52.7	0.0	0.0	28.0	14	5
en-tl	55.4	53.5	35.0	70.5	26.7	0.0	38.8	34.2	14	6
de-af	71.6	56.5	87.1	33.1	90.3	77.2	43.1	66.2	12	6
de-ur	77.5	58.5	50.7	68.3	45.8	15.4	72.6	50.6	12	7
de-sw	50.6	48.9	29.5	33.1	36.2	0.0	66.7	33.1	12	5
de-te	71.2	55.7	45.6	29.4	5.2	0.0	56.5	27.3	12	7
de-ta	76.3	57.6	43.0	56.7	48.7	0.0	81.3	45.9	12	7
de-mn	74.7	59.1	44.4	68.3	42.8	0.0	61.8	43.4	12	5
de-uz	62.8	55.1	48.3	91.9	27.8	0.0	81.1	49.8	12	6
de-my	72.0	59.3	31.3	29.9	63.9	0.0	47.5	34.5	12	5
de-jv	60.0	50.9	41.5	14.4	32.5	0.0	10.3	19.8	12	5
de-tl	54.5	52.1	48.1	42.1	0.0	0.0	50.8	28.2	12	6
zh-af	70.4	58.6	53.9	9.5	25.2	0.0	12.1	20.1	11	6
zh-ur	75.1	62.8	59.0	43.5	36.3	0.0	82.6	44.3	11	7
zh-sw	53.9	51.5	5.7	33.1	27.0	0.0	27.6	18.7	11	5
zh-te	72.4	60.3	49.9	29.4	4.5	0.0	86.7	34.1	11	7
zh-ta	73.0	61.8	19.0	56.7	16.8	0.0	40.5	26.6	11	7
zh-mn	71.6	60.4	56.5	43.5	8.7	0.0	99.0	41.5	11	5
zh-uz	62.5	54.9	49.0	69.3	26.2	0.0	87.2	46.3	11	6
zh-my	69.6	59.3	42.5	71.8	32.7	37.8	95.7	56.1	11	5
zh-jv	59.8	54.3	41.1	42.1	31.4	0.0	85.1	39.9	11	5
zh-tl	54.7	52.4	44.7	14.4	6.9	0.0	83.4	29.9	11	6
hi-af	78.2	59.0	55.4	50.1	30.8	14.3	52.3	40.6	7	6
hi-ur	80.0	57.8	100.0	88.1	73.0	100.0	99.9	92.2	7	7
hi-sw	50.7	50.5	27.4	24.6	24.9	0.0	66.9	28.8	7	5
hi-te	72.7	58.4	74.7	74.4	67.2	0.0	100.0	63.3	7	7
hi-ta	74.2	57.0	48.9	50.1	36.8	0.0	75.8	42.3	7	7
hi-mn	74.6	57.7	57.9	61.3	31.2	0.0	89.4	48.0	7	5
hi-uz	64.0	50.8	57.8	64.8	45.6	0.0	97.2	53.1	7	6
hi-my	74.3	58.7	36.7	46.7	37.5	0.0	97.6	43.7	7	5
hi-jv	59.4	48.7	21.2	0.0	13.6	0.0	79.6	22.9	7	5
hi-tl	56.6	52.9	73.1	59.8	41.3	0.0	98.2	54.5	7	6
ceb-af	63.9	58.1	42.4	44.1	52.5	0.0	8.9	29.6	11	6
ceb-ur	68.7	57.1	29.3	84.3	22.5	0.0	62.9	39.8	11	7
ceb-sw	53.4	49.2	33.0	16.1	76.3	0.0	12.0	27.5	11	5
ceb-te	69.3	59.0	4.8	98.6	17.9	0.0	75.9	39.4	11	7
ceb-ta	66.3	55.8	22.4	72.1	63.0	0.0	16.6	34.8	11	7
ceb-mn	65.9	59.7	16.5	55.0	37.6	0.0	79.3	37.7	11	5
ceb-uz	56.2	52.6	26.2	61.3	17.9	0.0	60.6	33.2	11	6
ceb-my	64.8	56.3	3.0	43.5	57.7	0.0	88.1	38.4	11	5
ceb-jv	57.1	51.2	60.2	17.1	70.0	54.8	97.6	59.9	11	5
ceb-tl	53.0	56.2	0.0	82.7	50.0	0.0	76.2	41.8	11	6

Table 16: Detailed data of 50 source-target language pairs used for correlation analysis of language similarity, source and target language pretraining data size with cross-lingual performance in unlabeled and labeled setup. Task performance is measured on Amazon review task with $k = 1$.

Amazon Review

		en	af	ur	sw	te	ta	mn	uz	my	jv	tl	Avg
UN	mBERT+pooling	57.8	54.4	54.9	52.4	53.5	54.8	51.1	49.3	52.4	56.1	52.1	53.1
	mBERT+distiluse	63.1	60.1	61.0	46.1	50.1	50.0	59.9	55.2	56.7	57.2	50.1	54.7
	mBERT+paraphrase	**69.3**	63.8	67.1	51.4	62.2	61.4	61.1	56.6	62.9	55.6	54.0	59.6
	XLM-R+paraphrase	69.2	**75.4**	**80.8**	**64.1**	**71.0**	**70.4**	**69.7**	**68.2**	**70.4**	**63.8**	**66.6**	**70.1**
LB	mBERT+pooling	65.6	56.8	57.0	51.8	53.8	53.1	52.7	51.2	52.5	53.5	53.2	53.6
	mBERT+distiluse	80.4	76.0	80.0	51.2	48.9	50.0	77.9	57.7	70.7	60.5	55.4	62.8
	mBERT+paraphrase	**87.2**	**82.9**	**85.0**	54.4	**79.0**	**80.5**	**80.6**	66.0	**78.9**	61.5	60.2	72.9
	XLM-R+paraphrase	77.6	81.7	82.2	**64.0**	74.2	73.9	75.1	**70.6**	76.4	**66.3**	**66.1**	**73.0**

AG News

		en	af	ur	sw	te	ta	mn	uz	my	jv	tl	Avg
UN	mBERT+pooling	37.9	37.3	34.8	37.7	32.9	38.0	36.0	33.7	37.4	42.0	38.8	36.9
	mBERT+distiluse	43.3	43.5	38.8	40.6	25.4	29.1	39.7	39.6	42.7	42.0	42.9	38.4
	mBERT+paraphrase	53.7	52.8	46.2	46.5	46.1	42.8	43.3	44.3	45.0	51.0	49.7	46.7
	XLM-R+paraphrase	**62.7**	**61.9**	58.9	**52.2**	58.1	**55.8**	55.6	**56.0**	**58.6**	59.2	**58.4**	**57.4**
LB	mBERT+pooling	77.4	68.2	55.4	58.5	54.7	52.1	50.7	54.6	49.0	66.7	70.2	58.0
	mBERT+distiluse	85.1	82.0	76.0	65.5	25.3	28.7	70.8	64.4	71.3	77.8	76.5	63.8
	mBERT+paraphrase	74.9	75.4	68.1	63.5	68.2	64.0	62.8	65.6	64.8	72.5	71.4	67.6
	XLM-R+paraphrase	**83.8**	**82.9**	**78.8**	**70.4**	**75.1**	**71.7**	**72.7**	**73.2**	**77.4**	**79.2**	**79.0**	**76.0**

XNLI

		en	af	ur	sw	te	ta	mn	uz	my	jv	tl	Avg
UN	mBERT+pooling	34.7	34.3	34.4	33.2	33.9	33.5	34.3	33.3	33.3	32.9	32.7	33.6
	mBERT+distiluse	32.9	32.6	33.4	**33.2**	**36.1**	36.1	33.8	34.6	31.9	**34.0**	34.1	**34.0**
	mBERT+paraphrase	34.1	32.9	**34.0**	33.0	34.9	34.7	34.5	34.5	33.9	31.4	33.9	33.7
	XLM-R+paraphrase	**35.5**	**33.7**	34.0	32.3	35.0	**36.5**	**38.1**	**34.7**	**35.1**	33.5	**34.1**	**34.7**
LB	mBERT+pooling	35.5	34.1	34.0	35.3	33.3	34.1	35.7	32.8	33.1	33.5	32.3	33.8
	mBERT+distiluse	34.5	35.6	33.6	**35.1**	31.3	31.4	38.5	35.6	34.8	35.7	34.3	34.6
	mBERT+paraphrase	**39.1**	**37.9**	**37.8**	33.5	**39.5**	**39.4**	**39.1**	34.7	36.9	**33.9**	**35.7**	**36.8**
	XLM-R+paraphrase	36.8	35.7	35.0	32.8	37.5	37.5	37.3	**36.7**	**37.5**	32.8	33.9	35.7

Table 17: Results of all languages using different combinations of retriever and MPLM for robustness analysis on Amazon review task ($k = 5$), AG News tasks ($k = 1$), and XNLI task ($k = 3$), respectively.

Task	Dataset	Size	#Label	Languages
Sentiment Analysis	Amazon Reviews	1000	2	af, ur, jv,
Topic Categorization	AG News	2000	4	ta, mn, uz,
Sentence Pair Classification	XNLI	1500	3	tl, te, mn, sw

Table 18: Overview of the test sets for the three tasks. Size refers to the number of samples for each LRL.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The section after the conclustion and before the references.

☑ A2. Did you discuss any potential risks of your work?
In the limitation section.

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

☒ A4. Have you used AI writing assistants when working on this paper?
No AI writing assistants were used.

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?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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.?
Appendix

☑ 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.
Appendix

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Models used are introduced in Section 4.

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?
Experimental setup is discussed in Section 4. Hyperparameter search is not applicable.

☑ 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 5 Results

☐ 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.)?
Not applicable. Left blank.

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.