# PARC: Cross-Lingual Retrieval Augmented Prompt

**Ercong Nie**[*] [1,2] **Sheng Liang**[*] [1,2] **Helmut Schmid**[1] **Hinrich Schütze**[1,2]

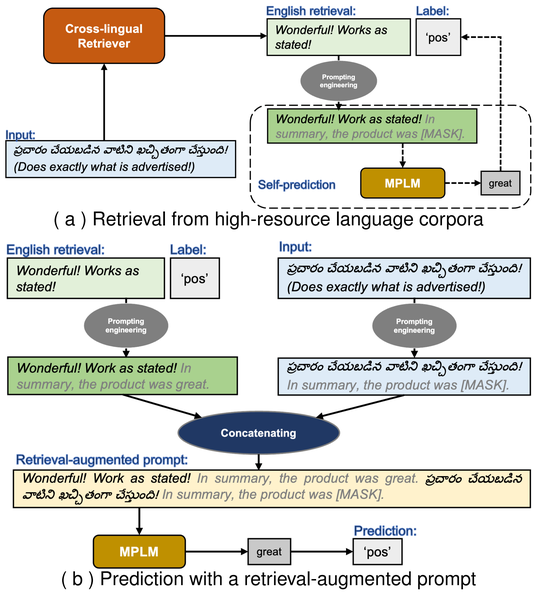
[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Germany
{nie, shengliang}@cis.lmu.de

## 1. Introduction

Multilingual pretrained language models (MPLMs), pretrained on multilingual corpora with >100 languages, exhibit strong multilinguality on downstream tasks. Low-resource languages (LRLs), for which little text data is available for pretraining monolingual pretrained language models (PLMs), benefit from MPLMs. However, LRLs are under-represented in the pretraining corpora, resulting in suboptimal performance for these languages. On the other hand, the scarcity of task-specific annotated data makes it difficult for LRLs to employ the pretraining-finetuning paradigm. To mitigate the high demand for labeled data, another line of research – prompt-based learning – emerges, focusing on exploiting large PLMs by reformulating the input. The prompt is designed to help PLMs "understand" the task better and "recall" what has been learned during pretraining. Prompt-based methods provide 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 is an effective method for LRLs lacking annotated data.

## 2. Motivation

Our work aims to improve the zero-shot transfer learning performance of LRLs on natural language understanding tasks by taking advantage of cross-lingual information retrieval and the multilinguality of MPLMs. Specifically, we retrieve semantically similar cross-lingual sentences from high-resource languages (HRLs) as prompts and use the cross-lingual retrieval information to benefit the LRLs from the multilinguality of MPLMs. To this end, we made the following contributions in this work:




( a ) Retrieval from high-resource language corpora

( b ) Prediction with a retrieval-augmented prompt

**Figure 1:** *The pipeline of our proposed PARC method*

1. We propose **P**rompts **A**ugmented by **R**etrieval **C**rosslingually (**PARC**), a pipeline for integrating retrieved cross-lingual information into the prompts for improved zero-shot learning (Figure 1).
2. We conduct experiments on three different multilingual classification tasks: binary sentiment analysis of product reviews, news topic classification, and natural language inference task. The results show 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.

[*] Equal Contribution.

## 3. PARC: Prompts Augmented by Retrieval Crosslingually

We enhance zero-shot learning for low-resource languages (LRLs) by cross-lingual retrieval from **labeled**/**unlabeled** HRLs. 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.

### 3.1 Retrieval from HRL corpora

The cross-lingual retriever takes an LRL input sample as query and retrieves 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).

### 3.2 Prediction with a retrieval-augmented prompt

The retrieved HRL sample, its label and the input sample are transformed into a prompt. For that, we need

(i) a *pattern* $P(.)$ which converts the input sentence into a cloze-style question with a mask token, e.g.:

$$P(X) = X \circ \text{“In summary, the product was [MASK].”}$$

where $X$ is the input sentence, $\circ$ refers to the string concatenation operator, and

(ii) a representative word for each possible class. e.g.:

$$\{\text{pos} \rightarrow \text{“great”}, \text{neg} \rightarrow \text{“terrible”}\}$$

(ii) a mapping (called verbalizer) from the class labels to representative HRL words. e.g.:

.

The prompt pattern $P(.)$ transforms the retrieved HRL sample into the cross-lingual context.

$$C_k^i = P(X_k^{R_i}, v(y_k^{R_i})) \tag{1}$$

where $C_k^i$ is the cross-lingual context for the input $X_i^L$ with the $k$-th most similar HRL sample $X_k^{R_i}$ and $y_k^{R_i}$ is the label for $X_k^{R_i}$.
Next, the cross-lingual retrieval-augmented prompt $I_i$ is created by concatenating the cross-lingual context and the cloze-style question.

$$I_i = C_k^i \circ P(X_i^L) \tag{2}$$

The augmented prompt $I_i$ is fed to the MPLM $M$. $M$ performs masked token prediction and returns a probability distribution $p = M(I_i)$ over all HRL words. We predict the class $\hat{y}$ whose verbalizer $v(\hat{y})$ received the highest probability from model $M$:

$$\hat{y} = \arg\max_{y \in Y} p(v(y)) \tag{3}$$

## 4. Experimental Results and Analysis

### 4.1 Main results

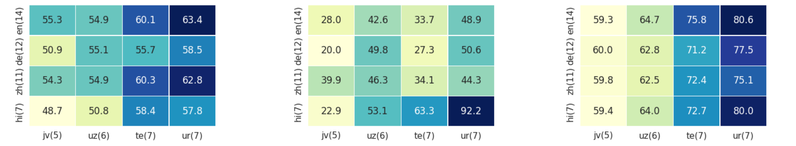
| | Amazon | AGNews | XNLI | Avg. |
|---|---|---|---|---|
| MAJ | 50.0 | 25.0 | 33.3 | 36.1 |
| Random | 48.2 | 25.6 | 32.4 | 35.4 |
| Direct | 53.8 | 36.3 | 33.1 | 41.1 |
| Finetune | 68.6 | 57.9 | 34.5 | 53.7 |
| PARC-unlabeled | 58.4 | 46.7 | 33.5 | 46.2 |
| PARC-labeled | **68.9** | **67.6** | **35.8**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

### 4.2 Effect of languages

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

**Table 2:** *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.*




( a ) Performance (Unlabeled) ( b ) Language Similarity ( c ) Performance (Labeled)

**Figure 2:** *Visualization of the correlation between zero-shot performance and language similarity, pretraining data size of source and target language.*

### 4.3 Effect of $k$

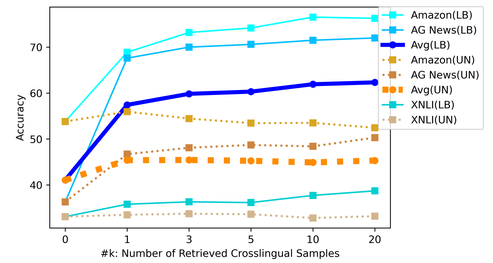



**Figure 3:** *Accuracy on three tasks with different $k$ in the labeled (LB) and unlabeled (UN) setup.*

### 4.4 Generalization to other retrievers and MPLMs

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

**Table 3:** *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.*

### 4.5 Robustness on unseen languages

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

**Table 4:** *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.*

## 5. Conclusion

In this work, we propose PARC, a pipeline for zero-shot transfer learning of LRLs. PARC gets and LRL input sentence, retrieves semantically similar sentences from HRL corpora, and uses them to augment prompts for improved zero-shot transfer learning.
Through extensive experiments on three different language understanding tasks, we show that

1. PARC performs better than the baseline without cross-lingual context in both unlabeled and labeled settings. PARC outperforms the finetuning baseline in the labeled setting.
2. LRL pretraining data size and language similarity both correlate positively with the transfer performance.
3. Increasing the number of retrieved prompts improves performance at first, but deteriorates it after a certain point.
4. PARC also performs well with other cross-lingual retrievers or MPLMs, and takes effect on unseen languages as well.