PARC: 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), Germany
* Equal contribution

September 1, 2023

Outline

Introduction

Background:

- **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.

But...

1. Pretraining corpora of MPLMs are **imbalanced distributed** in languages. → LRLs are **under-represented**.
2. LRLs **lack annotated data** for finetuning. → LRLs are difficult to employ pretraining-finetuning paradigm.

Outline

Motivation

Our work aims to

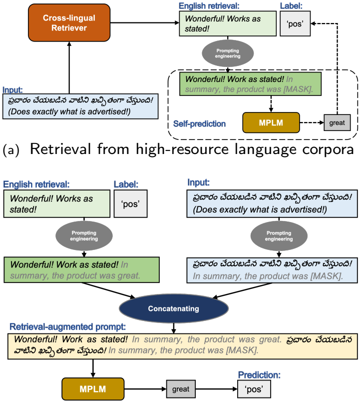
- improve the **zero-shot transfer performances** of **LRLs** on natural language understanding tasks
- leverage the **cross-lingual retrieval** and the **multilinguality** of MPLMs.

Specifically, we

- first retrieve **semantically similar** cross-lingual sentences from high-resource languages (**HRLs**)
- then use the cross-lingual retrieval information to benefit the LRLs from the **multilinguality** of MPLMs

To this end, we propose the **PARC** pipeline, **P**rompts **A**ugmented by **R**etrieval **C**rosslingually. It consists of two steps:

1. **Cross-lingual retrieval from HRL corpora**
   - 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).
2. **Prediction with a retrieval-augmented prompt**



(a) Retrieval from high-resource language corpora

(b) Prediction with a retrieval-augmented prompt

Figure: The pipeline of our proposed PARC method

PARC Pipeline

1. **Cross-lingual retrieval from HRL corpora**
2. **Prediction with a retrieval-augmented prompt**
   - The retrieved HRL sample together with its label and the input sample are reformulated as prompts. For that, we need a **pattern** $P(.)$ to convert the input sentence into a cloze-style question with a mask token, e.g.: $P(X) = X \circ$ "`In summary, the product was [MASK].`", and a **verbalizer** $v(.)$ to map each possible class onto a word, e.g.: $\{\texttt{pos} \rightarrow \text{"great"}, \texttt{neg} \rightarrow \text{"terrible"}\}$.
   - In this way, retrieved HRL sample is reformulated by the prompt pattern $P(.)$ as the cross-lingual context C_k^i:

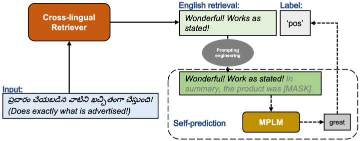
$$C_k^i = P(X_k^{R_i}, v(y_k^{R_i}))$$

   - Next, the cross-lingual retrieval-augmented prompt is created by the concatenation operator as the final input I_i.

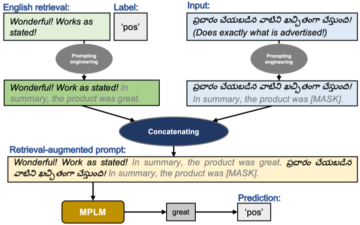
$$I_i = C_k^i \circ P(X_i^L)$$

   - At last, the prompted input augmented by cross-lingual retrieval I_i is taken by the MPLM M for prediction. 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:

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



(a) Retrieval from high-resource language corpora



(b) Prediction with a retrieval-augmented prompt

Figure: The pipeline of our proposed PARC method

Outline

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

	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: 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.

- PARC performs better than the direct baseline in both unlabeled and labeled settings.
- PARC in labeled setting outperforms the finetuning baseline.

Effect of Languages

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: 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.

- Pretraining data size of LRL and language similarity positively correlate to the transfer performance.

Effect of *k*

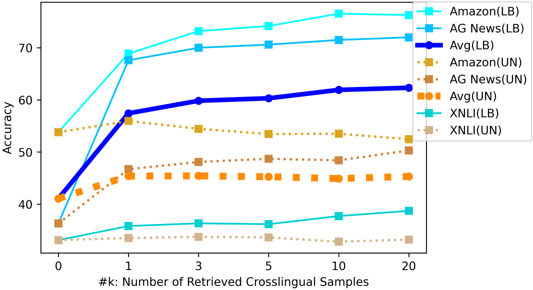


Figure: Accuracy on three tasks with different *k* in the labeled (LB) and unlabeled (UN) setup.

- Increasing the number of retrieved prompts improves performance at first, but deteriorates it after a certain point.

		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: 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.

- PARC shows strong generalization ability to different cross-lingual retrievers and MPLMs.

Robustness on unseen languages

		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: 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.

- PARC shows strong robustness to unseen languages.

Outline

Conclusions

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.
2. We conduct experiments on **three** different multilingual classification tasks: **binary sentiment analysis** of product reviews, news **topic classification**, and **natural language inference** task.
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.

Thanks for your attention!