LEXA: LANGUAGE-AGNOSTIC CROSS-CONSISTENCY TRAINING FOR QUESTION ANSWERING TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-lingual information retrieval (CLIR) is a knowledge-intensive NLP task that requires a lot of domain-specific data in different languages. In previous works, authors were mostly using machine translation and iterative training for data mining. We considered the problem from another angle and present a novel cross-lingual pre-training and fine-tuning approach for CLIR tasks based on cross-lingual alignment. We present a new model LEXA-LM significantly improving cross-lingual knowledge transfer thus achieving new state-of-the-art in cross-lingual and monolingual question answering and cross-lingual sentence retrieval. Moreover, we show that our pre-training technique LEXA is a very powerful tool for a zero-shot scenario allowing to outperform some supervised methods.

1 INTRODUCTION

In the modern world, it is often when someone cannot find an information they search for in the same language they query. Cross-lingual information retrieval (CLIR) is an NLP task that requires a lot of domain-specific multilingual data. In previous works, authors were using machine translation and iterative training for data mining (Asai et al., 2021b; Sorokin et al., 2022; Bonifacio et al., 2021). We think this problem could be solved on a lower level, essentially using cross-lingual text vector representations. We present a novel cross-lingual pre-training and fine-tuning approach for CLIR tasks based on Wikipedia cross-lingual alignment. We call it LanguagE-agnostic cross-consistency trAining (LEXA). Our method significantly improves cross-lingual knowledge transferring, and allows a model LEXA-LM trained with it to outperform or achieve comparable quality to previous state-of-the-art approaches in cross-lingual and monolingual question answering, multilingual passage ranking, and cross-lingual sentence retrieval. Moreover, we show that our pre-training method is very powerful in the zero-shot scenario and can outperform some previous supervised methods. We define two entities:

Weak Alignment For each item in language L_1, the closest neighbor in language L_2 is the most semantically relevant item.

Strong Alignment Regardless of their language, all semantically relevant items are closer than all irrelevant items, for each item. Importantly, relevant items in different languages are closer than irrelevant items in the same language.

Our method LEXA allows a language model to improve its cross-lingual understanding ability and convert its existing weak alignment to the strong one. The sample for such conversion is shown on Fig. 1.

The contribution of our work is as follows: (i) we present a pre-training method LEXA improving cross-lingual alignment; (ii) we present results for a language model trained with LEXA for ranking on XOR-Retrieve, Mr. TyDi, and BUCC tasks; (iii) results LEXA-LM trained for answer generation in XOR-Full task; and also results in zero-shot MKQA question answering task.[1]

[1] We are going to release LEXA code after the review process is over.

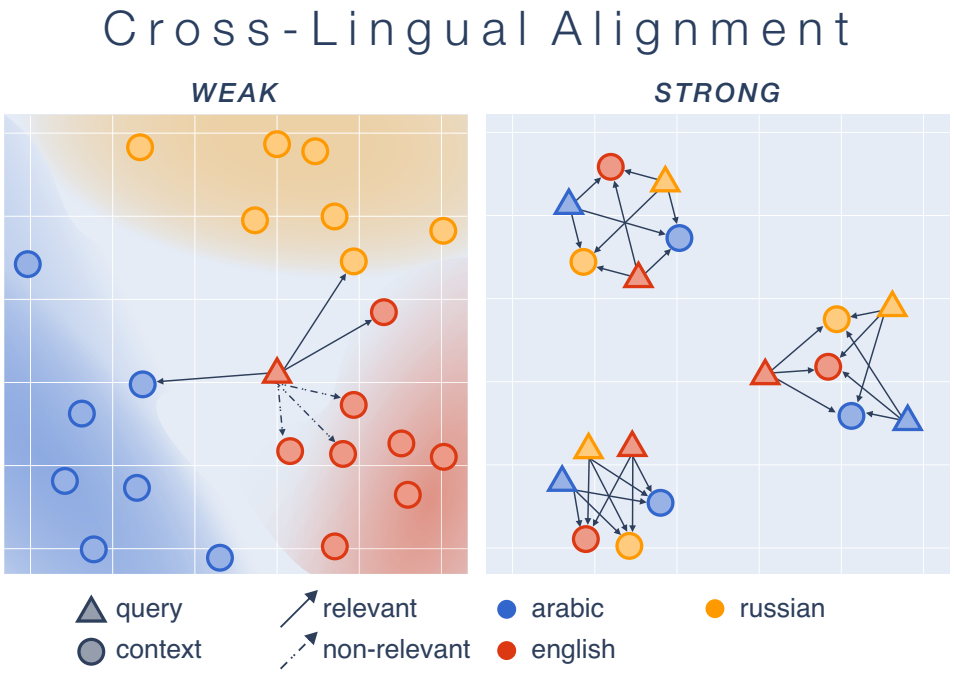


Figure 1: Here is a difference between strong and weak cross-lingual alignment. In a strongly aligned embedding space, the most semantically relevant pairs are always the closest, regardless of language. A weakly aligned multilingual embedding space just enables zero-shot transfer between languages, but incorrect answers in the same language are preferred over correct answers in a different language.

2 DATASETS

XOR TyDi (Asai et al., 2021a) is a multilingual open-retrieval QA dataset that enables cross-lingual answer retrieval. The dataset is based on questions from TyDi QA (Clark et al., 2020) and consists of three new tasks that involve finding documents in different languages using multilingual (including English ones) resources. In contrast to the TyDi QA dataset XOR TyDi consists of questions in only 7 languages, which are still typologically diverse: Arabian, Bengali, Finnish, Japanese, Korean, Russian, and Telugu. In our work we evaluate the models on 2 relevant to our work tasks: *XOR-Retrieve* is a cross-lingual retrieval task where a question is written in a target language (e.g., Japanese), and a system is required to retrieve an English document that answers the question; *XOR-Full* is a cross-lingual retrieval task where a question is written in the target language (e.g., Korean), and a system is required to output a short answer in the target language (i.e. Korean in our example).

Mr. TyDi (Zhang et al., 2021) dataset for monolingual retrieval that consists of eleven topologically diverse languages, designed to evaluate ranking with learned dense representations. Mr. TyDi is also constructed from TyDi dataset (Clark et al., 2020). Authors annotated every question from 11 languages with snippets (100 tokens of text) from Wikipedia which contain the answer to the question. Thus, at a high level, Mr. TyDi can be viewed as an open-retrieval extension to TyDi QA.

MKQA dataset (Longpre et al., 2020) consist of 10 thousand question-answer pairs from Natural Questions dataset (Kwiatkowski et al., 2019) translated to 26 different languages ending with 260 thousand questions. In the original paper, the dataset was purposed for a multilingual extractive question answering task. However, in recent works (Asai et al., 2021b; Sorokin et al., 2022) it was adopted for zero-shot cross-lingual question answering. We are concentrating on the latter task.

BUCC The BUCC task (Zhang et al., 2017) consists of 95k to 460k sentences in each of 4 languages, namely, German, French, Russian, and Mandarin Chinese, with around 3% of such sentences being English-aligned. The task is to match the pairs of sentences being the translations of each other.

Natural Questions (NQ) (Kwiatkowski et al., 2019) dataset is designed for end-to-end question answering. The questions are mined from real Google search queries and the answers are spans in Wikipedia articles identified by annotators. We use this dataset in two ways. One way is for training and another one is for zero-shot evaluation.

Joshi et al. (2017) presented **Trivia QA**, a large-scale question-answering dataset that includes so-called evidence documents, allowing one to state a task of information retrieval. Trivia QA includes 95 thousand question-answer pairs authored by trivia enthusiasts and independently gathered evidence documents, six per question on average ending with 650 thousand total triples.

Probably-Asked Questions (**PAQ**) dataset was presented in (Lewis et al., 2021b). It is a large dataset of 65 million automatically generated question-answer pairs in the English language.

3 METHOD

In this section, we will describe our language-agnostic cross-consistency training (**LEXA**), which is potentially applicable to any language model. The goal of our pre-training is to robustly learn the embedding space of passages for cross-lingual information retrieval. To achieve this, we introduce a new contrastive learning objective: for a randomly masked document from Wikipedia, we train the CLS vector such that the CLS embedding will be closer to the document in another language but same topic or the same document but masked in a different way than random or hard negatives documents. More formally, for a given random topic from Wikipedia we have n relevant documents $D = [d_1, d_2, ..., d_n]$ all in different languages from $L = [l_1, l_2, ..., l_n]$. We uniformly choose document d_j from the set D. Finally, optimizing noise contrastive estimation loss with batch size m:

$$\mathcal{L} = \sum_{i=1}^{n} -\log \frac{e^{\mathrm{sim}(d_i, d_j)}}{\sum_{k=1}^{m} e^{\mathrm{sim}(d_j, d_k)} + e^{\mathrm{sim}(d_i, d_j)}}, \tag{1}$$

So in general, passages with the same topics should have similar representations and the ones with different topics should have different representations.

3.1 PARALLEL DATA MINING

For the LEXA method to work, there is a need for a specific cross-lingually aligned dataset. Parallel data is often used for machine translation tasks (Fan et al., 2020; Pan et al., 2021). However, for our approach translation is not required to be accurate, the rough alignment is enough. Thus, we use language links between Wikipedia pages on the same topics to obtain such alignment. We use only the first paragraphs of the pages, since on Wikipedia these are supposed to summarize information of the whole article. In the end, we mined more than 16 million paragraphs in 28 languages using MKQA and XOR QA datasets as seed ones.

3.2 SELF-TRAINING

Previous works (Qu et al., 2021; Izacard & Grave, 2020) show the effectiveness of iterative hard negatives mining. In a work (Gao & Callan, 2021) the authors also use hard negatives in the training process, but by design it cannot be done during the pre-training, only during training. In our setup the hard negative usage for pre-training is possible. And not only possible, but it is an important feature of our architecture. The mined hard-negatives for self-training on pretraining stage make pre-training harder and closer to a downstream task that significantly improves final metrics. We separate our pre-training into two stages. Firstly, we train our model using only in-batch negative examples similar to (Karpukhin et al., 2020). Secondly, we use the model from the previous iteration for mining similar passages for training documents that are not linked with it.

We use the Memory Efficient Pre-training method described in (Gao & Callan, 2021) to make our pre-training more stable.

3.3 XPAQ

The authors of (Oğuz et al., 2021) show that pre-training on automatically-generated questions from the PAQ dataset can significantly improve the quality of information retrieval. As mentioned earlier, our pre-training procedure significantly improves knowledge transfer between languages, which allows us to effectively use more data from other, not target languages. However, PAQ data is generated using only the questions asked by English-speaking people and a lot of data is out-of-domain for other languages. To make the distribution of PAQ questions closer to one of the multilingual data we filtered them using trained LEXA-LM (E_{uns}); we are filtering basing on $[CLS]$ token vector similarity (sim) between questions from PAQ (q_{paq}) and multilingual data from XOR TyDi (q_x).

$$\mathrm{sim}(q_{paq}, q_x) = E_{uns}(q_{paq})^\intercal \cdot E_{uns}(q_x). \tag{2}$$

In this way, the dataset after filtration consists of the semantically closest questions in English to the multilingual questions. We filter from PAQ only a small fraction, which is intended to be about twice as big as our multi-lingual data. Thus we ended up with 400 thousand questions in XPAQ. We use XPAQ for additional pre-training and mark the models using it as + *XPAQ*.

3.4 Unsupervised and Zero-shot Retrieval Setups

Pre-trained multilingual alignment allows using our model effectively for unsupervised retrieval. All the tasks were evaluated in a similar to supervised setup way by either cosine distance or dot product on $[CLS]$ token embedding. However, for the question-answering task, we can adapt the PAQ dataset for a cross-lingual setup. A model trained with LEXA in an unsupervised way can be used for a cross-lingual similarity search that allows finding the nearest question from PAQ for every question in languages L. In zero-shot setup related to (Sorokin et al., 2022; Asai et al., 2021b), firstly, model fine-tuning on English datasets and testing on other languages.

4 Experiments

Model	R@2000 tokens								R@5000 tokens							
	Ar	Bn	Fi	Ja	Ko	Ru	Te	**Avg**	Ar	Bn	Fi	Ja	Ko	Ru	Te	**Avg**
Dev set																
LEXA-LM$_{large+XPAQ}$	**61.1**	**76.9**	**72.6**	**60.9**	**69.1**	**69.1**	**75.6**	**69.3**	**70.2**	**83.8**	**79.6**	**69.7**	**73.6**	**75.5**	**83.1**	**76.5**
LEXA-LM$_{base+XPAQ}$	55.6	66.1	69.4	55.6	65.9	61.1	72.6	63.8	62.7	72.6	75.1	66.3	72.6	70.8	81.9	71.7
DPR + BM25 + MT	43.4	53.9	55.1	40.2	50.5	30.8	20.2	42.0	52.4	62.8	61.8	48.1	58.6	37.8	32.4	50.6
CORA (Asai et al., 2021b)	32.0	42.8	39.5	24.9	33.3	31.2	30.7	33.5	42.7	52.0	49.0	32.8	43.5	39.2	41.6	43.0
Sentri$_{base}$ (Sorokin et al., 2022)	37.8	37.5	47.1	33.6	37.5	32.4	49.1	39.3	47.5	48.0	56.0	43.1	48.7	43.0	58.4	49.2
Sentri$_{large}$ (Sorokin et al., 2022)	47.6	48.1	53.1	46.6	49.6	44.3	67.9	51.0	56.8	62.2	65.5	53.2	55.5	52.3	80.3	60.8
DrDecr (Li et al., 2021)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	73.1
Zero-shot setup																
LEXA-LM$_{base+XPAQ}$	46.2	50.3	56.6	41.4	48.7	52.3	54.6	50.0	53.0	60.5	66.2	49.7	56.1	60.7	63.8	58.6
Unsupervised setup																
LEXA-LM$_{large}$	51.1	50.2	48.6	35.1	57.3	32.2	64.4	48.4	61.0	58.4	52.6	40.5	66.7	40.8	70.1	55.7
Test set																
LEXA-LM$_{large+XPAQ}$	**66.3**	**75.6**	**65.8**	**57.8**	**62.9**	**64.7**	**62.0**	**65.0**	**69.7**	**82.1**	**71.3**	**64.5**	**67.9**	**69.5**	**68.4**	**70.5**
LEXA-LM$_{base}$	59.1	67.2	62.5	53.4	56.4	56.0	54.8	58.5	65.0	73.6	68.3	61.4	61.7	61.8	60.1	64.6
DPR + BM25 + MT	48.3	54.4	56.7	48.1	39.4	39.4	18.7	42.7	52.5	63.2	65.9	52.1	46.5	47.3	22.7	50.0
GAAMA	-	-	-	-	-	-	-	52.8	-	-	-	-	-	-	-	59.9
Sentri$_{large}$ (Sorokin et al., 2022)	53.8	66.7	55.4	42.9	46.8	55.1	48.7	52.8	63.0	72.4	63.5	55.1	56.9	61.8	56.4	61,0
CCP	-	-	-	-	-	-	-	54.8	-	-	-	-	-	-	-	63.0
DrDecr (Li et al., 2021)	-	-	-	-	-	-	-	63.0	-	-	-	-	-	-	-	70.3

Table 1: Performance on XOR-Retrieve task. The best result is given in **bold**.

In this section, we will describe details about parallel data pre-training and pipeline for fine-tuning for cross-lingual open domain question answering and cross-lingual sentence retrieval tasks.

4.1 Pre-training

Generally, we initialize our model with pre-trained XLM-RoBERTa (Conneau et al., 2020), $_{base}$ and $_{large}$ variants, more experiments with model initialization you can find in section

Model	Target Language, F1							Macro Average		
	Ar	**Bn**	**Fi**	**Ja**	**Ko**	**Ru**	**Te**	**F1**	**EM**	**BLEU**
	Dev Set									
DPR + BM25 + MT	9.2	15.8	14.4	4.8	7.9	5.2	0.5	8.3	4.6	7.5
CORA (Asai et al., 2021b)	42.9	26.9	41.4	36.8	30.4	33.8	30.9	34.7	25.8	23.3
Sentri (Sorokin et al., 2022)	52.5	31.2	45.5	44.9	43.1	41.2	30.7	41.3	34.9	30.7
LEXA-LM$_{base+XPAQ}$	51.3	45.8	48.5	43.8	48.2	43.3	35.7	45.2	37.5	33.9
LEXA-LM$_{large+XPAQ}$	**53.4**	**50.2**	**49.3**	**44.7**	**49.5**	**49.3**	**38.9**	**47.8**	**38.7**	**35.5**

Table 2: End-to-end performance on XOR-Full task. Here Sentri and LEXA-LM uses MFiD (Sorokin et al., 2022) as reader.

sec:Ablation$_{S}tudy.WecallthesemodelsLEXA-LM_{base}$ and $_{large}$ respectively. We use an AdamW optimizer with a learning rate 1e-4, weight decay of 0.01, and linear learning rate decay. We also use gradient cashing for pre-training, similar to (Gao & Callan, 2021). The accumulated batch size was equal to 500. We train our models on 4 NVIDIA Tesla V100 GPUs.

4.2 Cross-lingual Open Domain Question Answering

To evaluate the effectiveness of our LEXA pre-training approach in the Cross-lingual ODQA task, we use XOR TyDi dataset. We evaluate the models trained with our method in supervised, unsupervised, and zero-shot scenarios. For the training, we use a method similar to the one described in (Sorokin et al., 2022). The system consists of question encoder $E_q(\cdot)$ and passage encoder $E_p(\cdot)$. We replace question and context token vectors ($< q >$ and $< ctx >$ respectively) with pre-trained semantically rich $[CLS]$ vectors from LEXA-LM without supervised training. For fine-tuning we use the XOR-Retrieve train set, Natural Questions, and Trivia QA datasets. For XOR-Retrieve there is used Recall metric in the form of searching the answer in the first n tokens, not the documents themselves.

	Recall@100											
	Ar	**Bn**	**En**	**Fi**	**Id**	**Ja**	**Ko**	**Ru**	**Sw**	**Te**	**Th**	**Avg**
	Zero-shot											
mDPR (Zhang et al., 2021)	62.0	67.1	47.5	37.5	46.6	53.5	49.0	49.8	26.4	35.2	45.5	47.3
LEXA-LM$_{base+XPAQ}$	**89.4**	**95.0**	**81.2**	**83.6**	**90.6**	**78.1**	**76.0**	**80.2**	**74.4**	**93.0**	**92.8**	**84.9**
	Supervised											
Hybrid (Zhang et al., 2021)	86.3	93.7	69.6	78.8	88.7	77.8	70.6	76.0	78.6	82.7	87.5	80.9
LEXA-LM$_{base+XPAQ}$	**91.1**	**95.4**	**81.5**	**88.3**	**90.8**	**82.6**	**82.3**	**84.5**	**75.2**	**93.8**	**92.1**	**87.0**

Table 3: Performance on Mr. TyDi (Zhang et al., 2021) test set. The best result is given in **bold**.

The results of the evaluation are presented in Tab. 1. LEXA-LM$_{large}$ model fine-tuned with the XPAQ dataset has outperformed all the other approaches on development and test sets. On the development set, it shows 4% improvement, while on the test set it is only 0.2%. In our opinion that could be explained by the existing domain shift between development and test sets. Interestingly, LEXA-LM$_{base}$ outperforms Sentri$_{large}$ model, which is also XLM-RoBERTa based.

For XOR-Full task there are several measures used, these are per-token F1 measure comparing the answer given and the ground truth; exact match of these two, and BLEU metric (Papineni et al., 2002). In XOR-Full task, there are retrieval and reader parts of the task. Retrieval part is solved by trained LEXA-LM model used in XOR-Retrieve task. In the reader part, in which the answer is generated from retrieved documents, we used MFiD model following (Sorokin et al., 2022; Izacard & Grave, 2020). The results of evaluation for the named models are presented in Tab. 2. LEXA-LM variants showed state-of-the-art results in this task also. LEXA-LM$_{large}$ fine-tuned on XPAQ and using MFiD has improved the previous best result by almost 5%.

4.3 MONO-LINGUAL OPEN DOMAIN QUESTION ANSWERING

In this section, we evaluate our method in mono-lingual setup. We chose Mr. TyDi (Zhang et al., 2021) dataset for mono-lingual retrieval that consists of eleven topologically diverse languages, designed to evaluate ranking with learned dense representations. The task is formulated as follows: for given a question in language L, need to retrieve a ranked list of passages from C_L, the Wikipedia collection in the same language. In previous work authors used mDPR, multilingual version of DPR model, and BM25 baselines (Zhang et al., 2021). Mr. TyDi uses classic Recall metric, searching for the ground truth document in top-100 retrieved ones.

We evaluate LEXA-LM, fine-tuned similarly to the XOR-Retrieve task, in a zero-shot setup like the mDPR model. The results are presented in Tab. 3. As one can see, our model significantly outperforms the hybrid approach in supervised setup, being better on average and in all particular languages.

4.4 ZERO-SHOT CROSS-LINGUAL TRANSFER

To test the transfer ability of the models trained with LEXA across the languages we evaluated our models with different types of fine-tuning: without any task-specific fine-tuning, with monolingual fine-tuning across languages from different language families, with cross-lingual fine-tuning. Here we discuss results achieved on different tasks. Firstly, XOR-Retrieve ones, presented in Tab. 1. As one can see, our method outperformed several strong supervised baselines. We also see that fine-tuning with XPAQ added about 3% for the final quality. Secondly, LEXA-LM has significantly outperformed the mDPR model in zero-shot setup, as shown in Tab. 3. More on that, LEXA-LM in an unsupervised setup shows only 2% drop in quality. Interestingly, in zero-shot setup, performance on Thai 0.7% better than in supervised one. We leave the investigation of this peculiar fact for future research

And last but not least, we report zero-shot evaluation results for the MKQA dataset. This setup has been presented in the previous works (Asai et al., 2021b; Sorokin et al., 2022). For the evaluation, Recall in tokens is used here. We report three models trained with LEXA here. All of them are LEXA-LM$_{base}$, but pre-trained on different datasets. *XPAQ* has been pre-trained on XPAQ dataset; *(En)* has been trained on English data from XOR TyDi; while *(multi)* has been pre-trained on XOR TyDi data in languages not included in MKQA.

	F1 score				
Model	**avg**	**De**	**Fr**	**Ru**	**Zh**
LEXA$_{large}$ unsupervised	**83.5**	81.2	**84.5**	82.9	**85.5**
(Tien & Steinert-Threlkeld, 2021)	82.4	**91.4**	75.6	**86.1**	76.5
LEXA$_{base}$ unsupervised	71.3	69.3	72.1	70.7	72.9
XLM-R L16-boe	68.7	75.4	65.0	75.6	59.0
(Artetxe et al., 2020)	75.8	80.1	78.8	77.2	67.0
(Keung et al., 2020)	69.5	74.9	73.0	69.9	60.1
(Kiros, 2020)	51.7	59.0	59.5	47.1	41.1
XLM-R B6-boe	23.5	18.4	18.8	27.0	30.0

Table 4: F1 scores on the BUCC mining task.

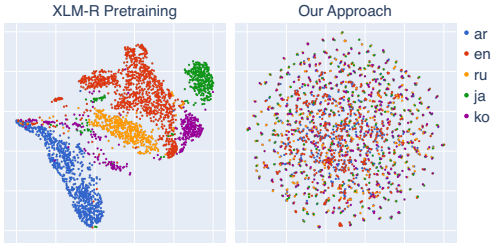
As it can be seen in Tab. 5, the pre-training using parallel data creates cross-lingual alignment that significantly improves zero-shot transfer. Even (En) variant has outperformed the same-sized previous model, while (multi) variant outperformed the larger one.

4.5 CROSS-LINGUAL SENTENCE RETRIEVAL

Another task, which we used for the evaluation of our LEXA method is BUCC, cross-lingual sentence retrieval. Following (Hu et al., 2020; Tien & Steinert-Threlkeld, 2022) our model is evaluated on the training split of the BUCC corpora and the threshold of the similarity score cutting off translations from non-translations is optimized for each language pair. Similarity scores are calculated based on dot products, also in pre-training. The BUCC task is interpreted as binary classification, thus

	Recall@2000 tokens										
	Da	**De**	**Es**	**Fr**	**He**	**Hu**	**It**	**Km**	**Ms**	**Nl**	**No**
CORA (Asai et al., 2021b)	44.5	44.6	45.3	44.8	27.3	39.1	44.2	22.2	44.3	47.3	48.3
BM25 + MT*	44.1	43.3	44.9	42.5	36.9	39.3	40.1	31.3	42.5	46.5	43.3
LEXA-LM$_{base}$ + PAQ	47.3	46.7	47.1	44.0	40.0	42.6	43.6	35.3	45.4	41.8	47.3
Sentri$_{base}$	52.5	50.5	51.9	51.4	40.5	44.9	49.3	38.1	51.0	52.9	52.2
LEXA-LM$_{base}$ (En)	56.6	55.3	54.4	55.6	38.7	44.1	51.1	33.1	54.5	56.1	55.9
Sentri$_{large}$	57.6	56.5	55.9	55.1	47.9	51.8	54.3	43.9	56.0	56.3	56.5
LEXA-LM$_{base}$ (multi)	**59.9**	**58.7**	**58.9**	**58.7**	**50.2**	**52.6**	**57.0**	**47.3**	**59.5**	**59.8**	**59.0**
	Pl	**Pt**	**Sv**	**Th**	**Tr**	**Vi**	**Zh-cn**	**Zh-hk**	**Zh-tw**	**Average**	
CORA (Asai et al., 2021b)	44.8	40.8	43.6	45.0	34.8	33.9	33.5	41.5	41.0	41.1	
BM25 + MT*	46.5	45.7	49.7	46.5	42.5	43.5	37.5	37.5	36.1	42.0	
LEXA-LM$_{base}$ + PAQ	48.7	47.2	53.1	44.2	43.7	45.3	39.1	40.6	38.8	44.1	
Sentri$_{base}$	50.1	51.2	52.7	48.5	47.3	49.7	39.9	44.5	44.2	48.2	
LEXA-LM$_{base}$ (En)	52.6	53.7	57.0	52.6	49.0	51.1	34.8	47.1	45.1	49.9	
Sentri$_{large}$	55.8	54.8	56.9	55.3	53.0	54.4	**50.2**	50.7	49.4	53.3	
LEXA-LM$_{base}$ (multi)	**57.9**	**58.7**	**60.2**	**57.7**	**55.6**	**57.9**	48.4	**53.1**	**51.5**	**56.1**	

Table 5: Zero-shot cross-lingual retrieval results on MKQA dataset.



(a) Projected embeddings of Wikipedia abstracts.

(b) The fraction of times where a passage in the same language is closest to the chosen passage.

Figure 2: Analysis of same-language bias in the cross-lingual embedding space.

classic F1 is used. Tab. 4 shows results for the previously presented approaches and the models used as initialization for LEXA-LM (RoBERTa variants). LEXA-LM large shows the best average performance, but also in comparison to the previous state-of-the-art model (Tien & Steinert-Threlkeld, 2022) the performance of our model is more stable across the languages.

5 Analysis

Fig. 1 shows the abstract representation of the idea that semantically aligned embedding space is language-agnostic. On Fig. 2a there are the actual embeddings for samples in five languages before and after LEXA training. As one can see, LEXA representations of passages are not aligned by the same language. Another view on this is presented in Fig. 2b. It investigates a same-language bias of representation space i.e. a percentage of the passage embeddings, which has an embedding of a passage in the same language as the closest neighbor. Thus we can conclude that LEXA training significantly improves the semantic closeness for different language passages. The samples for these tasks are taken from mined Wikipedia abstracts.

We also considered the following research question: *"Does the language-agnostically trained model forgets how to differentiate languages?"* Trying to understand this, we perform the linear probing on the language identification task, i.e. we take the representations of passages at layer k and evaluate how well a linear classifier (logistic regression) can be fitted to identify languages. We measure the accuracy and per-sample entropy of the fitted classifier for further analysis.

Probing shows that representations after the middle layers of LEXA (start with 2nd to 10th) languages can be differentiated more easily than by the same layers of XLM-R. While this discovery is surprising, output representations of the last layer do not follow the same pattern. LEXA representations contain less language-specific information than XLM-Roberta at this point. On Fig. 3b the per-language last layer probing is shown.

Overall, we think that LEXA-LM utilizes stronger language differentiation in the earlier layers to make the output of the whole model more language-agnostic. For these probings, we have used samples from OPUS-100 dataset (Zhang et al., 2020).

5.1 Ablation Study

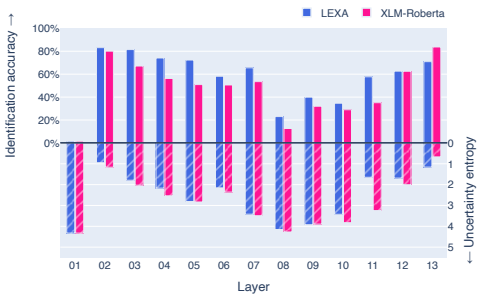
Model variation	Recall@2000 tokens		Recall@5000 tokens	
	MKQA	**XOR**	**MKQA**	**XOR**
XLM-R Base	48.2	39.3	55.1	49.2
XLM-R Base + LEXA	56.1	55.8	63.2	63.6
XLM-R Base + LEXA + XPAQ	57.2	58.8	65.3	66.1

Table 6: Ablation experiments on MKQA and XOR development sets.

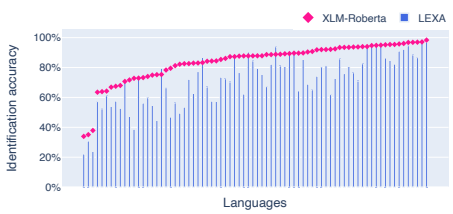
In this section, we will discuss the effectiveness of the proposed approach. Tab. 6 displays the result of retrieval on cross-lingual open domain question answering in two tasks, namely, MKQA and XOR TyDi. We compare here the XLM-RoBERTa base model without pre-training, the one with pretraining on parallel data (LEXA), and with additional fine-tuning on filtered PAQ data (XPAQ). As can be seen, LEXA pre-training improves results by more than 16% on XOR TyDi dataset and 8% in zero-shot setup on the MKQA dataset. XPAQ also adds up to 3% of quality in XOR TyDi and up to % on MKQA.

6 Related Work

Datasets The cross-lingual question answering datasets were scarce before recent years. Fortunately, these years left us with several publicly available datasets. Lewis et al. (2020) introduced MLQA dataset. It consists of parallel QA pairs in several languages. Liu et al. (2019) have presented an XQA dataset, with training set in English and validation and test sets in the other languages. Asai et al. (2021a) have presented a novel approach to cross-lingual QA introducing XOR TyDi dataset. The idea of this approach is to include in the dataset only the questions which have no answer in the target language, thus elaborating the usage of other language sources. Cross-lingual Question Answering Dataset (XQuAD) benchmark presented in Artetxe et al. (2020). It consists of a subset of 240 paragraphs and 1190 question-answer pairs from SQuAD v1.1 (Rajpurkar et al., 2016) together



(a) Per-layer linear probing results.



(b) Last layer language identification accuracy in details.

Figure 3: Linear probing in language identification task on OPUS-100.

with their translations into ten languages. Longpre et al. (2020) presented Multilingual Knowledge Questions and Answers (MKQA), an open-domain question answering evaluation set that contains 10 thousand question-answer pairs for 26 languages, as well as suggested Multilingual BERT, XLM, and XLM-RoBERTa baselines on it, in zero-shot and translation settings.

Systems Recent research was focused on creating non-English question answering datasets and applying cross-lingual transfer learning techniques, from English to other languages. Until recently, the availability of appropriate train and test datasets has been a key factor in the development of the field: however, in recent years, many works have focused on the collection of loosely aligned data obtained through automatic translation or by parsing similar multilingual sources. Lee & Lee (2019) have shown transfer learning applicability for cross-lingual QA with training on English data and evaluation on Chinese data. Artetxe et al. (2020) studied cross-lingual transferability of monolingual representations of a transformer-based masked language model. M'hamdi et al. (2021) examined a cross-lingual optimization-based meta-learning approach (meta-training from the source language to the target language(s) + meta-adaptation on the same target language(s) for more language-specific adaptation), to learn to adapt to new languages for question answering. (Gao & Callan, 2021) proposed unsupervised pre-training for dense passage retrieval, although the authors concentrated on retrieval itself, ignoring cross-lingual nature of the data.

In most previous approaches the authors use extractive models to generate the actual answer. This could be explained by the mental inertia from SQuAD-like datasets. By SQuAD-like we mean a dataset where labelled data includes an explicitly stated question, a passage, containing an answer, and a span markup for the answer. Such markup was presented for the question answering task called SQuAD in (Rajpurkar et al., 2016). Recently several works on cross-lingual generation of answers from raw texts has been presented. Kumar et al. (2019); Chi et al. (2019) studied cross-lingual question generation. Riabi et al. (2020) also suggested a method to produce synthetic questions in a cross-lingual way, using Multilingual MiniLM. Shakeri et al. (2020) proposed a method to generate multilingual question and answer pairs by a generative model (namely, a fine-tuned multilingual T5 model), it is based on automatically translated samples from English to the target domain. Generative question answering was mostly considered in previous work for long answers datasets. However, FiD model (Izacard & Grave, 2021) archives competitive results on SQuAD-like datasets, where an answer is supposed to be short text span. For open domain question answering, one of the first approaches named RAG used generative models was presented in (Lewis et al., 2021a). A key idea of this RAG model is to process several (top k) passages from the retriever in the encoder simultaneously. The produced dense representations of the passages are used in the decoder for the answer generation, this process is called fusion. Processing the passages independently in the encoder allows a model to scale to many contexts, as it only runs self-attention over one context at a time.

For question answering over knowledge graph, (Zhou et al., 2021) studied unsupervised bilingual lexicon induction for zero-shot cross-lingual transfer for multilingual question answering, in order to map training questions in the source language into those in the target language as augmented training data, which is important for zero-resource languages.

7 Conclusion

The multi-lingual understanding ability for the existing language models is now widely known. But the previous approaches for multi-lingual pre-training were not concentrated on the *cross*-linguality. That is an embedding for a passage on some topic should be closer to an embedding of another passage on the same topic, disregarding a language it has been written in. We proposed a novel pre-training technique LEXA, which allowed models trained with it to show state of the art results in several tasks, including retrieval (XOR-Retrieve, BUCC) and question answering (XOR-Full, MKQA). Our method is working as pre-training for supervised models (XOR, BUCC), and in zero-shot (MKQA). We also analyse the embeddings produces by the models and find out that they are more cross-lingual in the described sense, although there is a room for further research in this direction. We leave unanswered some questions on other models behaviour during the training, the quality and quantity of required for the training data, and so on. We hope that these and other questions will be answered in our (and not only our) future research.

REFERENCES

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. Xor qa: Cross-lingual open-retrieval question answering. *Proceedings of NAACL-HLT'2021*, 2021a.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. One question answering model for many languages with cross-lingual dense passage retrieval. *Proceedings of NeurIPS 2021*, 2021b.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897, 2021. URL https://arxiv.org/abs/2108.13897.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. Cross-lingual natural language generation via pre-training, 2019.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8: 454–470, 2020.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. 2020. doi: 10.48550/ARXIV.2010.11125. URL https://arxiv.org/abs/2010.11125.

Luyu Gao and Jamie Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. 2021. URL https://arxiv.org/abs/2108.05540.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. arXiv, 2020. doi: 10.48550/ARXIV.2003.11080. URL https://arxiv.org/abs/2003.11080.

Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. 2020.

Gautier Izacard and Édouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. Unsupervised bitext mining and translation via self-trained contextual embeddings, 2020. URL https://arxiv.org/abs/2010.07761.

Jamie Kiros. Contextual lensing of universal sentence representations, 2020. URL https://arxiv.org/abs/2002.08866.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4863–4872. Association for Computational Linguistics, July 2019.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Chia-Hsuan Lee and Hung-Yi Lee. Cross-lingual transfer learning for question answering. *arXiv preprint arXiv:1907.06042*, 2019.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7315–7330, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. 2021a.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. Paq: 65 million probably-asked questions and what you can do with them. 2021b.

Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. Learning cross-lingual ir from an english retriever. 2021. doi: 10.48550/ARXIV.2112.08185. URL https://arxiv.org/abs/2112.08185.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Xqa: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2358–2368, 2019.

Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *TACL*, 2020.

Meryem M'hamdi, Doo Soon Kim, Franck Dernoncourt, Trung Bui, Xiang Ren, and Jonathan May. X-METRA-ADA: Cross-lingual meta-transfer learning adaptation to natural language understanding and question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3617–3632, Online, 2021. Association for Computational Linguistics.

Barlas Oğuz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen-tau Yih, Sonal Gupta, and Yashar Mehdad. Domain-matched pre-training tasks for dense retrieval. 2021. doi: 10.48550/ARXIV.2107.13602. URL https://arxiv.org/abs/2107.13602.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. Contrastive learning for many-to-many multilingual neural machine translation. *ArXiv*, abs/2105.09501, 2021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *ArXiv*, abs/2010.08191, 2021.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. Synthetic data augmentation for zero-shot cross-lingual question answering. *arXiv preprint arXiv:2010.12643*, 2020.

Siamak Shakeri, Noah Constant, Mihir Sanjay Kale, and Linting Xue. Multilingual synthetic question and answer generation for cross-lingual reading comprehension. *arXiv preprint arXiv:2010.12008v1*, 2020.

Nikita Sorokin, Dmitry Abulkhanov, Irina Piontkovskaya, and Valentin Malykh. Ask me anything in your native language. *Proceedings of NAACL 2022*, 2022.

Chih-chan Tien and Shane Steinert-Threlkeld. Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining. 2021. doi: 10.48550/ARXIV.2104.07642. URL https://arxiv.org/abs/2104.07642.

Chih-chan Tien and Shane Steinert-Threlkeld. Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8696–8706, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.595. URL https://aclanthology.org/2022.acl-long.595.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, 2020.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1959–1970, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179. URL https://aclanthology.org/P17-1179.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. tydi: A multi-lingual benchmark for dense retrieval. 2021. doi: 10.48550/ARXIV.2108.08787. URL https://arxiv.org/abs/2108.08787.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5822–5834, Online, 2021. Association for Computational Linguistics.