CL-ReKD: Cross-lingual Knowledge Distillation for Multilingual Retrieval Question Answering

Anonymous ACL submission

Abstract

Cross-Lingual Retrieval Question Answering (CL-ReQA) is concerned with retrieving answer documents or passages to a question written in a different language. A common approach to CL-ReQA is to create a multilingual sentence embedding space such that question-answer pairs across different languages are close to each other. In this paper, we propose a novel CL-ReQA method utilizing the concept of knowledge distillation and a new cross-lingual consistency training technique to create a multilingual embedding space for ReQA. To assess the effectiveness of our work, we conducted comprehensive experiments on CL-ReQA and a downstream task, machine reading QA. We compared our proposed method with the current state-of-the-art solutions across three public CL-ReQA corpora. Our method outperforms competitors in 19 out of 21 settings of CL-ReQA. When used with a downstream machine reading QA task, our method outperforms the best existing language-model-based method by 10% in F1 while being 10 times faster in sentence embedding computation.

1 Introduction

Cross-lingual question answering allows a question posed in one language to be answered using materials written in a different language. As exemplified in Figure 1, one may ask, "Who was the first king of Hongsawadee?" and have their answer retrieved from a collection of historical documents in Burmese or other languages. To support the given example application, we require a retrieval system that can handle documents and questions in multiple languages at the same time. That is, we want to map questions and answers from multiple languages into the same space for easy retrieval. This functionality is also known as *Cross-Lingual Retrieval Question Answering* (CL-ReQA).

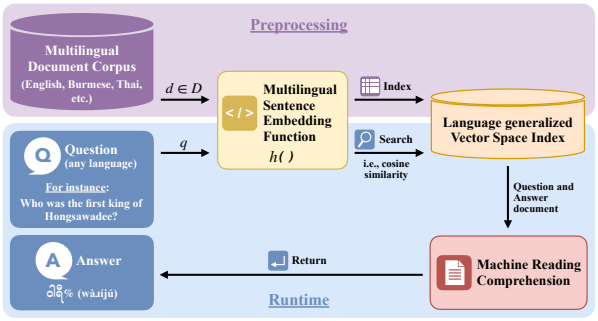

Figure 1: Overview of CL-ReQA. A user wishes to retrieve the answer to the question "Who was the first king of Hongsawadee?" from a collection of multilingual documents.

1.1 Existing Methods

One prominent approach to CL-ReQA is *multilingual sentence embedding*, i.e., creating an embedding space that can handle questions and answers from different languages. This approach can be further categorized into (i) *LM-Based:* finetuning a language model (LM), e.g., mBERT and XLM-R; (ii) *USE-Based:* finetuning the Universal Sentence Encoder (USE) for QA.

LM-based. Devlin et al. (2019) and Conneau et al. (2020) proposed a pretrained large-scale language model (LM) with multiple languages (100+ languages) called mBERT and XLM-R, respectively. Both solutions rely on finetuning the LM part to the target task. Reimers and Gurevych (2020) showed an accuracy improvement from 11.6% to 88.6% after finetuning with a bilingual text mining task. Finetuning LMs has been explored by many recent works, e.g., triplet loss with various supervised learning tasks (Reimers and Gurevych, 2019), knowledge distillation (Reimers and Gurevych, 2020), dense network QA encoder (Karpukhin et al., 2020), and providing initial word embeddings for the translation task (Feng et al., 2020). Nonetheless, finetuning these models requires a large number of training samples (more than 100,000 sentences in some cases (Reimers and Gurevych, 2020; Zhang et al., 2021; Wang et al.,

2021)) to give the best performance in multilingual settings. On the other hand, cross-lingual QA training corpora are usually smaller with only 1,000 to 1,500 questions per language. We need a method that can operate with a limited amount of data.

Multilingual Universal Sentence Encoding (mUSE). Based on the Universal Sentence Encoder (USE) architecture (Cer et al., 2018), Yang et al. (2020) proposed a training method utilizing a multilingual corpus with 16 different languages and multiple training objectives. They call their pretrained network multilingual USE or mUSE.

Experimental results from Trijakwanich et al. (2021) show that mUSE provides superior performance over the LM-based methods. However, this method performs poorly on languages outside the mUSE training corpus, i.e., *unsupported languages*. This limitation hinders the adoption of mUSE on limited-resource languages.

1.2 Our Work

Proposed Method. In this paper, our goal is to improve the robustness of multilingual sentence embedding that works with a wide range of languages, including those with a limited amount of training data. Leveraging the generalizability of knowledge distillation, we propose a *Cross-Lingual Retrieval Knowledge Distillation* (CL-ReKD) framework. Figure 1 illustrates how cross-lingual retrieval can be conducted through a multilingual embedding function $h()$. Given a question-document pair (q, d) in any language, $h(d)$ is closer to $h(q)$ than any other documents using any similarity measure, e.g., cosine similarity.

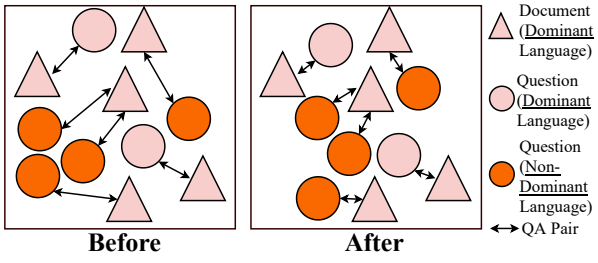

Figure 2: QA vector representations before and after performing the CL-ReKD framework. The main goal of our framework is to improve the consistencies between document-question pairs from different languages in the embedding space so that they can be correctly retrieved.

Learning Objective. As shown in Figure 2, the proposed CL-ReKD framework is designed to improve the embedding space by making cross-lingual question-answer pairs closer to each other. The crux of our proposed framework lies in the following two parts. First, we formulate a distillation process to create a language-generalized student. In particular, we leverage the fact that there is likely to be one language in a large multilingual corpus that dominates all others. We use that language to help improve the embedding quality of other languages. Second, we formulate a new loss function designed to improve the cross-lingual consistency between question-answer pairs in a multilingual environment. We aim to improve the consistencies between the teacher (dominant language) and student (other languages) for the following teacher-student output pairs: question-question, document-document, and document-question.

Experimental Studies. To determine the effectiveness of our approach, we compared the proposed methods with the current best practices (discussed in Section 1.1) on the CL-ReQA task across three datasets in 15 languages. Experimental results show that the CL-ReKD framework outperformed all competitive methods on languages supported by mUSE in all cases. The results on unsupported languages, i.e., languages outside of the mUSE training corpus, show that the CL-ReKD framework improved the performance of the mUSE encoder significantly $(p < 0.05)$ in all cases. Moreover, on a downstream task of machine reading QA (MR-QA), our method obtained better F1 and exact match scores than those of the best existing LM-based method in seven out of eight cases. Last but not least, our method is also 10 times faster than the state-of-the-art LM-based competitor in sentence embedding computational cost.

Summary of Contributions.

- We propose a new knowledge distillation method called *Cross-Lingual Retrieval Knowledge Distillation (CL-ReKD)* to transfer knowledge from the dominant language to non-dominant languages and build a language-generalized encoder.
- We design a new loss function to enforce cross-lingual consistency between dominant and non-dominant language vector representations.
- To assess the performance and efficiency of the models, we conducted an extensive set of experimental studies involving 2 tasks, 15 languages, and 8 competitors. Experimental results show the benefits of our proposed CL-ReKD framework. Moreover, we found that retrieving answers at the document level yields a significant improvement over the passage-level methods.

2 Background

2.1 Dominant Language

In a multilingual dataset, the distribution of languages tends to be imbalanced. As shown in Figure 3, the number of sentences in English is approximately 50% of all sentences in the corpus used to construct mUSE (Yang et al., 2020).

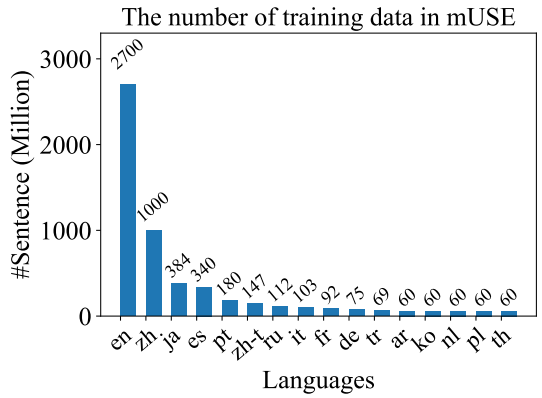

Figure 3: The distribution of QA training data used by mUSE (Yang et al., 2020)[1].

Due to the stated language imbalance, the model performance in languages with a large amount of data tend to be substantially better than that in other languages (Arivazhagan et al., 2019; Wang et al., 2020). This issue can be problematic when we want the model performance to be consistent across multiple languages.

For the case of mUSE, as shown in Figure 3, we can see that English is the *dominant language* in terms of training data available. Hence, the English-to-English retrieval performance tends to be better than all other language pairs. To verify this performance gap, we conducted a CL-ReQA experimental study using questions in non-English and answer documents in English; mUSE was used to encode the questions and documents. Experimental results show a significant performance improvement when the questions are translated into English instead of using the original non-English questions, i.e., translated questions from Russian to English improving the precision-at-1 from 43.3% to 52.8%. For the full results, see Appendix A.2.

2.2 Language Knowledge Transfer

There are many techniques to boost a model performance on low-resource languages using the structure obtained from rich-resource ones. Transfer and multitask learning have been popular paradigms for leveraging rich-resource languages. These methods usually rely on the shared-encoder strategy so that the language pattern learned in one language can be shared across all other languages using the same model (Lin et al., 2019; Nooralahzadeh et al., 2020; Zoph et al., 2016; Schwenk and Douze, 2017; Neubig and Hu, 2018; Yang et al., 2020; Feng et al., 2020). These classes of techniques are commonly known as *Language Knowledge Transfer*.

With a shared encoder, improvements on one language tend to benefit other languages as well. Let us consider a scenario where we have a large number of question-answer pairs in English and a significantly smaller number of pairs in other languages, e.g., Russian, French, and German. By letting other languages share the same encoder as we update the encoder weights while training with English data, we can also improve the general encoding performance of the model in other languages.

3 Proposed Method

In this section, we formulate our proposed methods by leveraging the two concepts discussed in the previous section, *dominant language* and *language knowledge transfer*. In particular, we perform knowledge distillation to transfer the knowledge from the dominant languages to other languages. Our proposed method consists of two stages: teacher model preparation and Cross-Lingual Retrieval Knowledge Distillation (CL-ReKD), which are described as follows.

3.1 Stage 1: Teacher Model Preparation

The purpose of this stage is to create a strong teacher for knowledge distillation in the next stage. For a base model, we use $\text{mUSE}_{\text{small}}$ for efficiency and performance reasons [2].

To create the teacher model, $\text{mUSE}_{\text{teacher}}$, we use Triplet loss $\mathcal{L}_{\text{tp}}$ (Equation 1). A training objective that maximizes the cosine similarity $\cos(\cdot)$ between anchor-positive pairs (a, p) and makes similarity between anchor-negative pairs (a, n) smaller than a given threshold α for all the training data M.

$$\mathcal{L}_{\text{tp}} = \sum_{i=0}^{|M|} [\max((1 - \cos(h(a_i), h(p_i))) - (1 - \cos(h(a_i), h(n_i))) + \alpha, 0)] \quad (1)$$

For detailed information about teacher model preparation such as the training strategy and comparison with other finetuning approaches, see Appendices A.6 and A.7.1.

[1] For brevity, we use the ISO-639 standard to refer to the languages used in this paper.

[2] See Sections 5.2 and 5.4 for further details.

3.2 Stage 2: CL-ReKD

We now describe our method to improve the general CL-ReQA performance using the concept of knowledge distillation. Knowledge distillation is a paradigm where a target model (a student) is trained to mimic the general behavior of a source model (a teacher). For example, one can construct a smaller model that behaves in a similar fashion as a larger one by minimizing the discrepancy between their outputs (Sanh et al., 2019; Jiao et al., 2020; Fang et al., 2021). Applying the same concept to our problem, we can set the knowledge distillation process to improve the embedding consistency between the dominant language and other languages. In particular, we setup the distillation environment as follows: (i) the teacher operates in the dominant language, i.e., English; (ii) the student operates in non-dominant languages; (iii) the student tries to mimic the embedding outputs of the teacher. In what follows, we describe the teacher and student models, inputs, and the loss function for the training process.

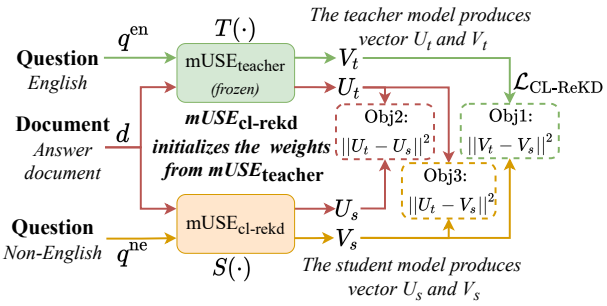

Figure 4: The training process of Cross-lingual Retrieval Knowledge Distillation (CL-ReKD) comprising (i) a teacher model, mUSE$_{teacher}$; (ii) a student model, mUSE$_{cl\text{-}rekd}$; (iii) three training objectives, *Obj 1-Obj 3*.

Teacher and Student Models. As illustrated in Figure 4, the *Cross-Lingual Retrieval Knowledge Distillation (CL-ReKD)* process consists of a teacher, student, and loss function. Initially, the student's parameters are initialized to the same values as those of the teacher trained in Stage 1. During the training process, the teacher's parameters are fixed; we only adjust the student's parameters according to the loss function.[3]

Inputs. Let us now consider input questions and answer documents of the training process. As illustrated in Figure 4, both teacher and student models accept the same document input d. However, there are two different versions for each question, *English* q^{en} and *non-English* q^{ne}. The English question q^{en} is a translation of the original one q^{ne}. This gives us a question pair (q^{ne}, q^{en}) for knowledge distillation between different languages. For simplicity, we use GNMT to translate q^{ne} into q^{en}. Note that if available, one may also use human-translated parallel questions.

The teacher model $T()$ accepts q^{en} as input, while the student model $S()$ accepts q^{ne} as input. In other words, q^{en} functions as the "reference" of the distillation process. According to our assessment (Appendix A.7.3), English provides the best performance and hence is chosen as the dominant language for the training process. Note that this finding also conforms with the data distribution shown in Figure 3.

Loss Function. The goal of our CL-ReKD loss function $\mathcal{L}_{\text{CL-ReKD}}$ is to let the student mimic the teacher's knowledge from the dominant language to the student's target language. As shown in Figure 5, our loss function $\mathcal{L}_{\text{CL-ReKD}}$ has three consistency objectives, namely, question-question, document-document, and document-question. We describe them as follows.

- *Obj 1: Question-Question.* The first objective is to enforce the consistency between $S()$ and $T()$ when encoding the same question expressed in English q^{en} and non-English q^{ne}, respectively.
- *Obj 2: Document-Document.* While adjusting the student $S()$ for the first objective, we also want to keep its answer document encoding unchanged. Hence, we want to maintain the consistency between $T(d)$ and $S(d)$.
- *Obj 3: Document-Question.* To accommodate the lookup process, the embedding space should also keep question-answer pairs consistent with each other. As our third objective, we minimize the discrepancy between the student's question vector $S(q^{ne})$ and the teacher's document vector $T(d)$.

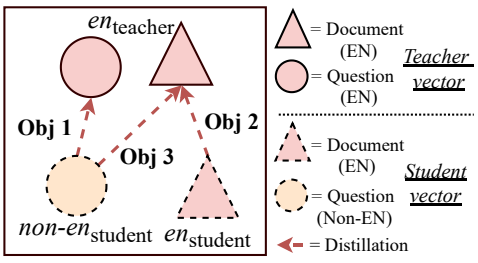

Figure 5: Illustration three objectives of Cross-Lingual Retrieval Knowledge Distillation (CL-ReKD) loss function.

[3]Note that the student model can be of any architecture and can be initialized using any method. In this work, we choose the self-model for simplicity. See Section 5.5 and Appendix A.5 for more information.

We formulate the loss function $\mathcal{L}_{\text{CL-ReKD}}$ as a linear combination of these three consistency objectives. Using the squared L2 norm as the discrepancy measure, we obtain the following loss function:

$$\mathcal{L}_{\text{CL-ReKD}} = \frac{\gamma}{|M|}\sum_{i=0}^{|M|}[\beta||T(q_i^{\text{en}}) - S(q_i^{\text{ne}})||^2 + \lambda||T(d_i) - S(d_i)||^2 + \omega||T(d_i) - S(q_i^{\text{ne}})||^2], \quad (2)$$

where M is the set of training samples used in a given batch, and β, λ, and ω are the weighting coefficients.

Discussion. As stated earlier, the goal of the loss function is to transfer the teacher's knowledge to the student operating in target languages. Since the performance of the teacher's dominant language is generalized, after the distillation, other distilled languages will have the same properties. The experimental results show that the student can better handle unsupported languages and improve the performance of supported languages than the teacher model. This improvement comes from the cross-lingual consistency objectives, Obj 1 and Obj 3, in the loss function, while Obj 2 maintains the monolingual consistency. Moreover, $\mathcal{L}_{\text{CL-ReKD}}$ does not require the teacher and student models to be of the same architecture; it can be applied to any pre-trained models. (For more information, see Appendix A.5)

4 Experimental Setup

4.1 Datasets

To evaluate the effectiveness of our method, we conduct our experiments on three well-known CL-ReQA corpora: XORQA, XQuAD, and MLQA. All experiments were done by XX→EN where XX is the question language (15 languages), and EN is answer passages or documents.

XORQA (Asai et al., 2021a) is a benchmark dataset for multilingual open-retrieval question answering. The dataset contains questions in a diverse set of seven non-English languages and answer documents in English. We use the *Gold Paragraph* part of the corpus, which contains 12,895 documents and 8,949 question-answer pairs. The authors, however, did not provide a test dataset. Thus, we divide the samples into train/dev/test (0.7/0.1/0.2).

XQuAD (Artetxe et al., 2020) is a dataset for evaluating cross-lingual question answering performance. XQuAD comprises 48 documents and 13,090 question-answer pairs obtained from the development set of SQuAD v1.1 (Rajpurkar et al., 2016) with 11 languages. Since XQuAD is too small for model training, we used it for testing only. Translated questions from SQuAD v1.1 (training set) were used instead for training the models (the same setup as XQuAD (Artetxe et al., 2020)).

MLQA (Lewis et al., 2020) is also a dataset for evaluating cross-lingual question answering performance. The dataset contains 15,806 documents and 33,706 question-answer pairs in seven different languages. However, the authors did not provide any training dataset. As a result, we combined the development and test datasets and divided them into train/dev/test (0.7/0.1/0.2).

4.2 Competitive Methods

We compare the performance of our method with two groups of competitive methods as follows:

LM-based. As discussed in Section 1, one approach to CL-ReQA is to use an embedding space based on some language model. In the experimental studies, we compare our methods to the following LM-based competitors.

- *XLM-R-nli-stsb*: A RoBERTa-based cross-lingual model trained using the NLI and STS benchmark datasets (Reimers and Gurevych, 2019).
- *mBERT-triplet*: A BERT-based multilingual model finetuned with a QA dataset using triplet loss (Reimers and Gurevych, 2019).
- *XLM-R←SBERT*: A XLM-RoBERTa model trained by distilling from the sentence BERT model (Reimers and Gurevych, 2020).
- *DPR*: A dense-network solution using the multilingual BERT model to provide the cross-lingual QA capability (Karpukhin et al., 2020).
- *CORA*: An adaptation of DPR on multilingual Wikipedia QA data (Asai et al., 2021b).
- *LaBSE*: A multilingual sentence encoder using mBERT to provide initial word embedding vectors (Feng et al., 2020).

We retrained mBERT-triplet, XLM-R←SBERT and DPR with the CL-ReQA training set following previous work (Reimers and Gurevych, 2019; Asai et al., 2021b; Zhang et al., 2021).

Multilingual Universal Sentence Encoding (mUSE). As an alternative to LM, we can also construct a QA embedding space from a well-known multilingual sentence encoder, mUSE. In particular, we consider the following mUSE variants.

- *mUSE$_{small}$*: The mUSE$_{\text{small}}$ encoder was based on Convolution Neural Network (Kim, 2014).

- *$mUSE_{large}$*: The $\text{mUSE}_{\text{large}}$ encoder was on the transformer architecture (Vaswani et al., 2017).

Note that although there exists a QA variant of mUSE, mUSE_{qa}, we found that this QA variant does not provide any performance improvement over $\text{mUSE}_{\text{large}}$. As a result, we omit mUSE_{qa} from our study.

Our proposed methods. As previously discussed, we construct our proposed methods based on $\text{mUSE}_{\text{small}}$. The first method, $\text{mUSE}_{\text{teacher}}$, is constructed from triplet loss where each triplet consists of a question, its corresponding answer document, and a non-answer document. The second method, $\text{mUSE}_{\text{cl-rekd}}$, is constructed from the process of Cross-Lingual Retrieval Knowledge Distillation.

4.3 Hyperparameter and Evaluation Settings

Hyperparameter. In these experiments, we use grid search on the following hyperparameters: learning rates, triplet loss margin (α), $\mathcal{L}_{\text{CL-ReKD}}$'s coefficients ($\gamma$,$\beta$,$\omega$), and the number of negative samples for triplet loss. The hyper-parameter configurations are given in Appendix A.9. For the Cross-Lingual Retrieval Knowledge Distillation settings, we use a batch size of 8 with a total number of 10 epochs. Since the student model receives the initial weights from the teacher, the CL-ReKD's loss value ranges between $[10^{-3}, 10^{-5}]$ and the loss value of Obj 3 is lower than those of other objectives. To prevent the CL-ReKD's loss value from being too small, we multiply the value by γ and set the coefficient of Obj 2, λ, to 1. In addition, we evaluate the precision score on the development set every 100 steps. If the precision score does not improve, the learning rate is halved.

Evaluation. We use precision at k where we set k to 1 (P@1) which is a common practice for the CL-ReQA task (Ahmad et al., 2019; Yang et al., 2020; Guo et al., 2021) and cross-lingual retrieval tasks (Reimers and Gurevych, 2020; Feng et al., 2020). We also provide precision at 5 and 10 results in Appendix A.4. Furthermore, we used McNemar's test as the significant statistical measurement ($p < 0.05$) for all experiments.

5 Experimental Results

5.1 Passage- vs Document-bases on Machine Reading QA (MR-QA)

To determine the best answer retrieval unit for MR-QA, we compare two scenarios. (i) *Passage-based*: Retrieving answers as passages; (ii) *document-based*: Retrieving answers as documents. For conciseness, we chose DPR, which is the state-of-the-art LM-based competitor, for comparison. For testing, we chose XORQA, which is the newest MR-QA benchmark. For all test cases, we used the same machine reading comprehension model constructed from XLM-R. In particular, we finetuned XLM-R using the same training portion of XORQA described in Section 4.1.

Results. Table 1 displays the MR-QA scores as F1 and exact match (EM) and provides a comparison between the two MR-QA input options: passage-based and document-based. In all cases, the document-based option improves over the passage-based one. We can also see that our method significantly outperformed DPR for both input options on average.

Model	XORQA									
	RU		KO		JA		FI		AVG	
	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
Passage-based										
DPR	17.4	12.9	6.1	0.3	14.6	10.4	21.7	15.3	15.0	9.7
$\text{mUSE}_{\text{cl-rekd}}$	21.1	16.8	27.8	20.4	26.1	20.2	20.7	16.1	23.9	18.4
Document-based										
DPR	16.8	13.0	6.2	0.3	15.3	11.5	**22.5**	16.4	15.2	10.3
$\text{mUSE}_{\text{cl-rekd}}$	**24.3**	**18.8**	**28.5**	**21.7**	**26.3**	**20.6**	21.2	**16.7**	**25.1**	**19.5**

Table 1: F1 and EM scores on the cross-lingual machine reading QA.

Discussion. The document-based representation has advantages and drawbacks in comparison to the passage-based one. In particular, by grouping passages associated with the same document together, the retrieval unit becomes larger, making it harder to miss an answer. However, operating at the document level also means that we have to handle a larger input. That is, for the three corpora used in the experimental studies, the model has to handle 128 tokens on average when the input is a passage, while the input size can be up to 1,996 tokens when the input is a document. The results provide empirical evidence that the benefits outweigh the drawback. We believe that as machine reading models improve over time, longer input passages will provide even better results. For more information about the retriever's performance, see Appendix A.1.

5.2 CL-ReQA: Supported Languages

In this experiment, we report the effectiveness of our proposed methods on mUSE's supported languages where the answer retrieval unit is document-based. We evaluated our methods against the competitors discussed in Section 4.2.

Model	XORQA			XQuAD							MLQA			
	RU	KO	JA	AR	DE	ES	RU	TH	ZH	TR	AR	DE	ES	ZH
LM-based														
mBERT-triplet	41.0	19.0	46.4	23.5	53.4	52.1	44.1	6.7	36.6	43.3	10.1	28.9	32.8	14.7
XLM-R-nli-stsb	28.7	26.6	28.0	39.1	43.3	45.8	42.9	41.2	44.1	44.5	27.9	34.0	31.2	29.4
XLM-R←SBERT	21.5	19.8	20.7	39.5	39.9	41.6	41.6	40.8	40.3	43.3	24.8	33.4	30.8	34.7
DPR	33.8	2.0	26.9	38.7	51.3	58.0	52.1	10.9	29.2	41.2	35.5	56.6	59.0	55.0
CORA	18.9	11.5	10.4	21.0	39.9	36.1	34.5	4.6	20.6	24.8	18.2	31.2	35.6	19.4
LaBSE	29.8	26.7	33.2	41.2	43.7	47.1	42.4	13.0	44.5	40.8	33.8	35.4	38.4	40.3
Multilingual Universal Sentence Encoding (mUSE)														
$\text{mUSE}_{\text{small}}$	43.3	35.5	41.2	64.7	73.1	75.6	66.8	72.7	71.0	70.2	44.5	60.4	57.0	53.2
$\text{mUSE}_{\text{large}}$	52.1	41.1	47.7	57.1	65.1	68.5	59.7	63.4	62.2	61.8	35.6	42.3	39.5	31.8
Our proposed methods														
$\text{mUSE}_{\text{teacher}}$	54.2	44.5	47.7	68.5	79.8	82.4	72.3	75.2	82.3	72.7	49.1	**64.8**	62.8	57.1
$\text{mUSE}_{\text{cl-rekd}}$	**58.2**	**47.7**	**49.5**	**79.4**	**83.2**	**84.0**	**83.6**	**86.1**	**82.4**	**80.3**	**49.5**	**64.8**	**63.4**	**57.9**

Table 2: Precision at 1 (P@1) on the CL-ReQA task in *supported languages*

Results. As shown in Table 2, our proposed models $mUSE_{teacher}$ and $mUSE_{cl\text{-}rekd}$ provide significant improvements from the base model, $\text{mUSE}_{\text{small}}$. Moreover, our models also outperformed the largest pre-trained variant of mUSE, $\text{mUSE}_{\text{large}}$. All of our proposed models also significantly performed better than the LM-based competitors. The results also show that our consistency enhancement method, CL-ReKD, were effective in all cases except DE for the MLQA dataset.

Discussion. Experimental results verify that for languages supported by mUSE, our approach based on the language knowledge transfer concept (Section 2.2) can provide significant improvements over the teacher model. However, the improvements were less significant than ours when the language knowledge transfer concept is applied to an LM to create mBERT-triplet and XLM-R-nli-stsb from mBERT and XLM-R, respectively.

Notice that methods based on mBERT performed poorly in Thai (TH). We can also see that finetuning mBERT with triplet loss (mBERT-triplet) and multilingual dense retrieval (CORA) did not provide any improvements on Thai. This is because Thai was not included in the construction process of mBERT (uncased-version), and the amount of the training data is insufficient to improve the model.

5.3 CL-ReQA: Unsupported Languages

Let us consider how well our proposed models performed when used with languages not supported by the base model, $\text{mUSE}_{\text{small}}$, i.e., FI, RO, EL, HI, and VI. Similar to the study presented in the previous subsection, we used XORQA, XQuAD, and MLQA as our test corpora.

Results. Table 3 presents the P@1 scores of our methods and the competitors. We can see that the original mUSE models, $\text{mUSE}_{\text{small}}$ and $\text{mUSE}_{\text{large}}$, did not perform well in these languages. As expected, $\text{mUSE}_{\text{teacher}}$ had a tendency to provide some improvements over $\text{mUSE}_{\text{small}}$. This is because these languages were not included in the original training process, and the amount of data is insufficient to improve the performance of these languages. In contrast, we obtained significant improvements through the CL-ReKD methods, $\text{mUSE}_{\text{cl-rekd}}$. For five out of seven cases, $\text{mUSE}_{\text{cl-rekd}}$ were the best performer compared to other models. Two LM-based methods, XLM-R←SBERT and LaBSE, were the best performer in HI with the test corpora of XQuAD and MLQA, respectively.

Model	XORQA	XQuAD				MLQA	
	FI	RO	EL	HI	VI	HI	VI
LM-based							
mBERT-triplet	18.7	48.3	30.3	23.9	39.9	7.1	24.1
XLM-R-nli-stsb	30.7	45.4	44.5	39.1	40.3	31.4	28.6
XLM-R←SBERT	25.3	42.4	42.9	**40.3**	40.3	29.4	28.4
DPR	39.1	52.9	36.1	15.5	10.1	26.2	32.1
CORA	15.1	27.7	26.9	18.5	28.6	15.8	23.9
LaBSE	40.6	42.0	42.9	37.8	39.9	**31.8**	27.2
Multilingual Universal Sentence Encoding (mUSE)							
$\text{mUSE}_{\text{small}}$	18.3	41.6	10.9	4.2	25.6	2.0	25.4
$\text{mUSE}_{\text{large}}$	27.2	49.6	13.0	3.4	25.6	1.4	16.8
Our proposed methods							
$\text{mUSE}_{\text{teacher}}$	25.0	42.4	13.4	4.2	29.4	2.0	27.4
$\text{mUSE}_{\text{cl-rekd}}$	**48.2**	**76.9**	**64.3**	34.0	**72.3**	3.2	**44.2**

Table 3: Precision at 1 (P@1) on the CL-ReQA task in *unsupported languages*.

As an alternative to cross-lingual retrieval, one can convert the problem to a monolingual retrieval one using a machine translation (MT) model. We found that using an MT model (i.e., GNMT, MBART) with DPR following Asai et al. (2021a) helps improve the performance of LM-based models. However, the performance decreases in some languages, i.e., in Finish (XORQA), DPR's perfor-

mance is dropped from 39.1 to 32.0 and dropped to 36.4 when GNMT and MBart were applied to the DPR, respectively. For the full discussion and results see Appendix A.2.

Discussion. The performance gap between $\text{mUSE}_{\text{small}}$ and $\text{mUSE}_{\text{cl-rekd}}$ demonstrates the effectiveness of the proposed CL-ReKD framework. In particular, we can use CL-ReKD to generalize a base sentence embedding model to handle languages that were not originally included in the training process.

Regarding the CL-ReKD performance on Hindi (HI), one important observation is that Hindi is the only language in this study whose family is not represented in the original training data, which results in a lot of OOV tokens from unknown characters. We provide more explanation in Appendix A.8.

5.4 Run-time Efficiency on Query Encoding

Let us now consider the efficiency of the methods. Since this investigation focuses on the embedding methods, we consider only the sentence embedding computational time. We used a DGX-1 machine using one Intel Xeon E5-2698 and one NVIDIA Tesla V100 GPU to benchmark the models.

Results. As shown in Appendix A.3, the experimental result shows that $\text{mUSE}_{\text{small}}$ is the fastest. The result shows that $\text{mUSE}_{\text{small}}$ took only $\sim$7.9 ms on average to encode one query at a time. Since our method is based on the $\text{mUSE}_{\text{small}}$ architecture, we also obtain a similar run time. For the LM-based methods, we found that LaBSE is the quickest one. However, the method is still slower than $\text{mUSE}_{\text{teacher}}$ and $\text{mUSE}_{\text{cl-rekd}}$ by at least 80%. We can also find that MT-assisted, i.e., GNMT and MBart, are significantly slower than our methods. For the MT-assisted results, GNMT used 258.3 ms for one query while MBart used 9,132 ms. For the full results, see Appendix A.3.

Discussion. Since $\text{mUSE}_{\text{small}}$ is a much smaller model than BERT-base and XLM-R-base, it is advantageous to use our proposed methods when efficiency is a concern, e.g., edge deployment. While GNMT and MBart were effective in improving the performance of LM-based models (Appendix A.2), the additional machine translation cost renders the approach less desirable.

5.5 Ablation Studies on the Training Objective for Knowledge Distillation

This study compares our Cross-Lingual Retrieval Knowledge Distillation method with other knowledge distillation techniques using the same baseline model. To directly assess the effect of the distillation method, we use the original $\text{mUSE}_{\text{small}}$ instead of the $\text{mUSE}_{\text{teacher}}$ as the starting model in this study. We compare four training objectives:

- We apply the training objective following Reimers and Gurevych (2020)'s work denoted mUSE_{mse} where the training objective contains the first CL-ReKD's objective with an additional loss term that minimizes the difference between English and English embeddings from the previous iteration;
- As mentioned in Section 5.3 the CL-ReKD loss has three objectives. $\text{mUSE}^{\text{q}}_{\text{cl-rekd}}$ uses only the first objective;
- $\text{mUSE}^{\text{qd}}_{\text{cl-rekd}}$ uses the first and the second objective; and
- Lastly, $\text{mUSE}_{\text{cl-rekd}}$ uses the full version of the CL-ReKD loss function.

Results. The experiment results are given in Table 4. As expected, our training objective outperformed competitive training objectives. The performance of $\text{mUSE}_{\text{cl-rekd}}$ outperformed other training objectives from six out of seven cases. Especially Reimers and Gurevych (2020)'s training objective, our method outperformed with significant results on six out of seven cases except for HI→EN in MLQA.

Model	XORQA	XQuAD				MLQA	
	FI	RO	EL	HI	VI	HI	VI
$\text{mUSE}_{\text{small}}$	18.3	41.6	10.9	4.2	25.6	2.0	25.4
mUSE_{mse}	31.0	65.5	17.6	8.0	35.3	1.4	26.0
$\text{mUSE}^{\text{q}}_{\text{cl-rekd}}$	33.7	67.2	16.8	5.5	34.5	2.2	37.0
$\text{mUSE}^{\text{qd}}_{\text{cl-rekd}}$	38.5	62.6	18.9	8.4	34.5	**3.7**	38.0
$\text{mUSE}_{\text{cl-rekd}}$	**39.4**	**73.9**	**58.0**	**27.0**	**71.4**	**3.7**	**38.7**

Table 4: Comparison of different knowledge distillation training objectives. Precision at 1 (P@1) on the CL-ReQA task in unsupported languages.

6 Conclusion

In this paper, we propose a novel Cross-Lingual Retrieval Knowledge Distillation framework for CL-ReQA. Our framework is designed to improve the general performance and enable the baseline model to handle unsupported languages by exploiting the concepts of *Dominant Language* and *Language Knowledge Transfer*. Our method outperformed competitive methods in all cases of supported languages and five out of seven cases of unsupported languages. Furthermore, we demonstrated that grouping passages associated with the same document together could benefit machine reading QA.

References

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. ReQA: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*.

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

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *NAACL-HLT*.

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

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations*.

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*.

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)*.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. 2021. SEED: self-supervised distillation for visual representation. In *9th International Conference on Learning Representations, ICLR 2021*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. Multireqa: A cross-domain evaluation for retrieval question answering models. *Proceedings of the Second Workshop on Domain Adaptation for NLP*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

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

Mahmut Kaya and Hasan Sakir Bilge. 2019. Deep metric learning: A survey. *Symmetry*, 11(9):1066.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Association for Computational Linguistics*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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. *Trans. Assoc. Comput. Linguistics*.

Zhuoyuan Mao, Prakhar Gupta, Chenhui Chu, Martin Jaggi, and Sadao Kurohashi. 2021. Lightweight cross-lingual sentence representation learning. In *ACL*. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

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 (EMNLP)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*.

Nattapol Trijakwanich, Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, and Sarana Nutanong. 2021. Robust fragment-based framework for cross-lingual sentence retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Liang Wang, Wei Zhao, and Jingming Liu. 2021. Aligning cross-lingual sentence representations with dual momentum contrast. In *EMNLP*. Association for Computational Linguistics.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021. Bootstrapped unsupervised sentence representation learning. In *ACL*. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

A Appendix

A.1 CL-ReQA: Passage- vs Document-bases

This study reports the effect of the passage-based and document-based input on retrieval performance. The experiment was conducted in supported and unsupported languages of three datasets: XORQA, XQuAD, and MLQA.

Results. As shown in Table 5, the document-based input substantially outperform the passage-based one in all languages and all datasets. The performance of $\text{mUSE}_{\text{cl-rekd}}$ on passage-based for supported and unsupported languages is lower than document-based significant on every unsupported language. These results conform with those of the downstream task presented in Table 1.

Model	XORQA			XQuAD			MLQA		
	RU	JA	FI	DE	ES	VI	DE	ES	VI
Passage-base									
$\text{mUSE}_{\text{cl-rekd}}$	56.4	45.9	45.6	51.7	51.3	23.5	61.9	62.4	32.3
Document-base									
$\text{mUSE}_{\text{cl-rekd}}$	**58.2**	**49.5**	**48.2**	**83.2**	**84.0**	**72.3**	**64.8**	**62.8**	**44.2**

Table 5: Comparison of different retrieval inputs such as passage- and document-bases on the CL-ReQA task in *supported* and *unsupported languages*.

Discussion. This experiment shows that the efficiency of document-based is more robust than passage-based significant. Moreover, when we applied both inputs to a downstream task, MR-QA, the experiment results showed that changing the input from passage-based to document-based is more robust.

A.2 CL-ReQA: Monolingual vs Cross-lingual Retrievals

In this study, we compare the CL-ReQA approach against the MT-assisted monolingual retrieval one. For CL-ReQA methods, we chose DPR and $\text{mUSE}_{\text{cl-rekd}}$. For the MT-assisted methods, we used two translators, GNMT and MBart (Liu et al., 2020), to translate all questions into English which is the documents' language. These two translators were then applied to assist DPR in the same manner as the XORQA investigation (Asai et al., 2021a).

Results. Table 6 shows that the two translators provide substantial improvements to DPR and mUSE-based. The GNMT-assisted tended to perform better than the MBart-assisted in almost all cases. As state in Section 3.2, our knowledge distillation comprise of GNMT in the training data process. Thus, the performance of our method with/without MT-assisted is similar. Our method was the best performer in six out of nine cases. Moreover, in seven out of nine cases, the performance of our method is decreased when MBart is applied. Since the performance of MBart is lower than GNMT, the performance of $\text{mUSE}_{\text{cl-rekd}}$+MBart is dropped significantly.

Model	XORQA			XQuAD			MLQA		
	RU	JA	FI	DE	ES	VI	DE	ES	VI
Cross-lingual retriever (LM-based)									
DPR	33.8	26.9	39.1	51.3	58.0	10.1	56.6	59.0	32.1
LM-based + MT-assisted									
DPR+GNMT	45.0	36.3	32.0	58.8	61.8	56.3	61.3	59.4	56.0
DPR+MBart	39.8	26.2	36.4	59.2	61.5	52.5	58.8	57.4	51.9
Multilingual Universal Sentence Encoding (mUSE)									
$\text{mUSE}_{\text{small}}$	43.3	41.2	18.3	73.1	66.8	25.6	60.4	57.0	25.4
mUSE-based + MT-assisted									
$\text{mUSE}_{\text{small}}$+GNMT	52.8	45.0	43.0	81.1	82.1	42.5	60.4	62.0	60.3
$\text{mUSE}_{\text{small}}$+MBart	51.6	37.6	42.3	79.1	78.6	68.0	59.4	58.8	53.6
Our proposed method									
$\text{mUSE}_{\text{cl-rekd}}$	58.2	**49.5**	48.2	**83.2**	**84.0**	**72.3**	**64.8**	**62.8**	44.2
Our proposed method + MT-assisted									
$\text{mUSE}_{\text{cl-rekd}}$+GNMT	**64.8**	46.2	**51.7**	82.1	83.6	62.2	64.5	**62.8**	**61.8**
$\text{mUSE}_{\text{cl-rekd}}$+MBart	48.2	37.7	38.2	**83.2**	83.5	34.0	63.0	62.4	58.5

Table 6: Precision at 1 (P@1) on the CL-ReQA task in *supported* and *unsupported languages*.

Discussion. Machine translators can provide a quick solution to improve the performance of cross-lingual retrieval. In this way, the problem of cross-lingual retrieval is converted into a monolingual one. This approach can be useful when the language pair has a reliable translator, but there is insufficient QA data to create a cross-lingual retrieval model. However, we consider the multilingual sentence embedding approach to be superior to the MT-assisted one due to the following reasons: (i) the computational cost benefits of skipping the MT process; (ii) the MT-assisted approach can be negatively affected by a poor MT performance; (iii) the reliance of MT models can be a limitation for some language pairs. See Appendix A.10 for a further analysis.

A.3 The Average Computational Sentence Encoding Time

This subsection shows the average running of time of all models in this paper. We used one Intel Xeon E5-2698 and one NVIDIA Tesla V100 GPU to evaluate the models.

Table 7 shows that the running time of mUSE-based models is efficient than every LM-based model. We can also see that as a two-step approach, the MT-assisted solutions incur substantially longer running times than other methods since the transformer requires more computation and resources,

which can be lead to memory limitation (Mao et al., 2021).

Model	XORQA	XQuAD	MLQA
MT-assisted			
DPR+GNMT	258.3±37.9	339.3±54.2	395.0±58.1
DPR+MBart	9,132±8,111	6,527±5,838	5,382±274
LM-based			
mBERT-triplet	20.2±1.0	30.6±1.5	30.2±1.7
XLM-R-nli-stsb	20.5±1.1	22.1±2.7	23.4±4.6
XLM-R←SBERT	22.2±3.7	31.6±1.9	31.3±2.4
DPR	58.3±28.5	110.3±47.6	103.0±52.5
CORA	197.7±9.9	369.7±16.4	274.7±145.3
LaBSE	15.2±2.6	14.3±2.2	14.6±2.3
Multilingual Universal Sentence Encoding (mUSE)			
$\text{mUSE}_{\text{small}}$	**8.4±1.0**	**7.9±1.1**	**7.6±1.2**
$\text{mUSE}_{\text{large}}$	23.8±2.5	30.3±5.0	27.0±4.7
Our proposed methods			
$\text{mUSE}_{\text{teacher}}$	8.6±1.0	8.3±1.1	8.1±1.0
$\text{mUSE}_{\text{cl-rekd}}$	8.5±1.5	8.3±1.5	8.5±1.4

Table 7: The average computational sentence encoding time and standard division in *ms*.

A.4 CL-ReQA: Precision at 5,10

In this experiment, we study the effectiveness of our method on mUSE's supported languages. We report the precision score at 5 and 10 on the XORQA dataset.

Results. As shown in Table 8, our proposed model, $\text{mUSE}_{\text{cl-rekd}}$, on P@5 and P@10 have the same narrative as P@1 (Table 2). That is, the cross-lingual retrieval knowledge distillation model improved the performance from the teacher model, $\text{mUSE}_{\text{teacher}}$, and it outperformed every LM-based model.

Model	XORQA					
	P@5			P@10		
	RU	KO	JA	RU	KO	JA
LM-based						
mBERT-triplet	59.3	37.6	63.3	68.2	46.6	70.4
XLM-R-nli-stsb	48.7	48.1	46.6	54.7	55.5	55.7
XLM←SBERT	43.8	46.2	43.0	54.7	55.7	53.4
DPR	61.6	4.0	56.7	74.2	15.0	67.9
CORA	48.1	34.0	31.1	62.2	48.4	41.7
LaBSE	50.1	45.9	51.8	59.6	57.0	59.8
mUSE-based						
$\text{mUSE}_{\text{small}}$	69.6	64.1	59.6	77.7	71.6	69.2
$\text{mUSE}_{\text{large}}$	75.1	68.1	71.0	82.2	74.3	78.5
Our proposed methods						
$\text{mUSE}_{\text{teacher}}$	78.2	69.6	69.2	82.8	78.1	78.5
$\text{mUSE}_{\text{cl-rekd}}$	**79.7**	**70.7**	**72.0**	**85.4**	**78.3**	**78.8**

Table 8: Comparison of different Precision at k (P@k) where k values are equal to 5 and 10 on the CL-ReQA task in *supported languages* on XORQA.

Discussion. Experimental results verify that precision at other k's values does not change any conclusion from our work. The results show that our model outperformed other models in precision at 1, 5, and 10 settings.

A.5 CL-ReKD on Other Architectures

This study demonstrates our cross-lingual retrieval knowledge distillation method on LM-based models. For diversity, we select BERT-based (mBERT-triplet) and RoBERTa-based (XLM-R-nli-stsb) because these models trained only on the dominant language, English, same as $\text{mUSE}_{\text{teacher}}$.

Results. As shown in Table 9, we applied the CL-ReKD framework on other architectures, i.e., BERT and RoBERTa, on supported languages of XORQA. The experimental results show that CL-ReKD on LM-based models significantly improves XLM-R-nli-stsb. Furthermore, when we use mBERT-triplet as the teacher instead of XLM-R-nli-stsb. The result shows a small improvement over the student instead of using XLM-R as the teacher model.

Model		XORQA		
		RU	KO	JA
LM-based				
mBERT-triplet (1)		41.0	19.0	46.4
XLM-R-nli-stsb (2)		28.7	26.6	28.0
LM-based + CL-ReKD				
T=(1)	$\text{XLM-R-nli-stsb}_{\text{cl-rekd}}$	29.0	27.1	28.3
T=(2)	$\text{XLM-R-nli-stsb}_{\text{cl-rekd}}$	33.2	30.8	31.9

Table 9: CL-ReKD on different architecture between student and teacher models. Where T is the teacher model.

Discussion. Our CL-ReKD can be applied to any pre-trained model not limited to only in mUSE's architecture. However, to give the best results, let the student initialize the weight from the teacher. In addition, the results from both models are not over mUSE's performance.

A.6 Stage 1: Teacher Model Preparation

The purpose of this stage is to create a strong teacher for knowledge distillation in the second stage (Section 3.2). For a base model, we $\text{mUSE}_{\text{small}}$ for efficiency and performance reasons. Since the $\text{mUSE}_{\text{small}}$ encoder is shared across all languages, we can finetune the base model using questions and answers from one language and obtain performance improvements on other languages as well (as discussed in Section 2.2). In theory, we

can choose any supported languages to perform this stage. According to our assessment, English provides the best performance and hence is chosen as the language for the finetuning questions and answers (See Appendix A.7.2 and A.7.3 for further details).

Note that the datasets used in this work are all cross-lingual with answer documents in English and questions in non-English. Consequently, we need to translate all questions into English for the finetuning process. For simplicity, Google NMT (GNMT) was used to perform this task.

The process of transfer learning $\text{mUSE}_{\text{small}}$ consists of two main components: encoder and triplet loss.

- Encoder $h(\cdot)$. An encoder model is a function that maps a question q and a passage P into the 2vector representations $h(q)$ and $h(P)$, respectively.
- Triplet loss $\mathcal{L}_{\text{tp}}$ (Equation 1). A training objective that maximizes the cosine similarity $\cos(\cdot)$ between anchor-positive pairs (a, p) and makes similarity between anchor-negative pairs (a, n) smaller than a given threshold α for all the training data M.

$$\mathcal{L}_{\text{tp}} = \sum_{i=0}^{|M|}[\max((1-\cos(h(a_i), h(p_i)))- (1-\cos(h(a_i), h(n_i))) + \alpha, 0)] \quad (3)$$

Let us now consider the training sample mining process. At the initial step, we need to mine triplets (a: anchor, p: positive, n: negative). While the anchors a can be randomly sampled from the questions, we need the CL-ReQA model to choose positives p and negatives n. For negative sample categorization, we consider two options. First, we can directly use the original $\text{mUSE}_{\text{small}}$ model to categorize the negative samples according to the current embedding space (online fashion) (Kaya and Bilge, 2019). Second, we can apply the method proposed by Karpukhin et al. (2020), which utilizes BM25 (Trotman et al., 2014) to produce textual similarity scores. From the ablation study given in Appendix A.7.1, the results show that the first three epochs use the initial strategy for triplet mining (Kaya and Bilge, 2019) before proceeding to online mining (Kaya and Bilge, 2019) for five epochs.

What has the teacher learned? As mentioned in Section 2.1, the dominant language lifts the performance of multilingual representations. Since we finetune the teacher model with the dominant language question-document pairs, it allows us to handle supported languages more reliably. However, it relies on the fact that the base model's encoder has learned some structure of the target languages. As a result, we need a different method to improve the general performance of the model, which we described in Section 3.2.

A.7 Ablation Studies

This study presents the effect of each design decision in the triplet loss and cross-lingual retrieval knowledge distillation proposed. Here, we investigate the following components: (i) training strategies; (ii) training data settings; (iii) reference languages for cross-lingual retrieval knowledge distillation (CL-ReKD); and (iv) distance functions for CL-ReKD. In each investigation, we use the best setting from the previous steps. All experimental results were obtained from XX→EN retrieval on the XORQA test set across four languages, where XX can be one of these languages Russian (RU), Korean (KO), Japanese (JA), and Finnish (FI).

A.7.1 Training strategies

As shown in Table 10, we compare training strategies with/without each of the following components: initialization, online updates, and other deep metric learning techniques. As expected, the result shows that online negative sampling and initializing with BM25 helps improve the performance of triplet loss. Furthermore, we also study the effect of replacing triplet loss with contrastive learning as presented in the current state-of-the-art work (Karpukhin et al., 2020). We found that contrastive loss consistently provides a performance improvement over the original $\text{mUSE}_{\text{small}}$ model but still lags behind the triplet loss.

A.7.2 Training data settings

In terms of training data for the teacher training, there are two decisions we need to consider: (i) the answer representation unit: whether to use one passage or one document as the retrieval unit in the training process; (ii) the question language: whether to use the original questions (in multiple languages) or translate them all to English. In the case of English, all English questions were translated using GNMT. As shown in Table 10, the *passage* and *English* combination provides the best performance.

A.7.3 Teacher's language for CL-ReKD

In this study, we explore the choice of language to function as the reference (teacher) in the CL-ReKD

process. Intuitively, we want a language that is well-represented in the training corpora when constructing the original model. Consequently, we compare English, Spanish, and German. As expected, English, which is the dominant language in the training corpora of mUSE, provides the best performance. These results also conform with the discussion on the dominant language provided in Section 2.1.

A.7.4 Distance functions for CL-ReKD

The distance function is critical to the CL-ReKD performance. While there exists many distance functions we can apply to the distillation process, we consider two of the most widely used ones, cosine and squared L2. As we can see, squared L2 provides the best performance.

A.7.5 Discussion

From the results, we conclude that the default settings of our proposed methods are (i) triplet loss as the training objective (ii) English question + passages as the answer representation unit (iii) English as the teacher language, and (iv) square L2 distance as the distance function for CL-ReKD.

Component		XORQA			
		RU	KO	JA	FI
$\text{mUSE}_{\text{small}}$		43.3	35.5	41.2	18.3
Training strategies					
Triplet loss	+ online	45.0	35.7	42.5	23.4
	+ BM25	53.6	38.6	42.5	20.8
	+ BM25, + online	**54.2**	**44.5**	**47.7**	**25.0**
Contrastive loss	+ BM25, + online	52.1	39.9	42.5	21.8
Training data settings					
English questions + documents		53.0	40.7	44.3	21.8
English questions + passages		**54.2**	**44.5**	**47.7**	**25.0**
multilingual questions + documents		53.1	44.3	46.9	22.6
multilingual questions + passages		50.7	40.9	42.7	20.5
Teacher's language for CL-ReKD					
$\text{mUSE}_{\text{cl-rekd}}$	English as teacher	**58.2**	**47.7**	**49.5**	**48.2**
	Spanish as teacher	56.4	41.4	47.7	45.8
	German as teacher	56.4	41.3	48.7	43.1
Distance functions for CL-ReKD					
$\text{mUSE}_{\text{cl-rekd}}$	Squared L2 distance	**58.2**	**47.7**	**49.5**	**48.2**
	Cosine distance	57.3	44.7	49.0	47.8

Table 10: Comparison between training strategies, training data settings, teacher's language for CL-ReKD, and distance functions for CL-ReKD measured with P@1 score on XX→EN, XORQA test set.

A.8 Language Analysis

In this section, we discuss Hindi, which our method performs worse than other competitive methods, and examines other languages that we have not shown in the tables, such as Telugu and Bengali.

As shown in Table 3, the $\text{mUSE}_{\text{cl-rekd}}$'s performance is the best in every language except Hindi. This is due to the mUSE encoder's tokenizer (sentencepiece), which cannot handle Hindi well compared to other unsupported languages (i.e., FI, RO, EL, VI). For instance, we measured the tokenizer's out-of-vocabulary (OOV) rate of the mUSE's tokenizer on Hindi (XQuAD) and found that the OOV rate of Hindi is ∼14.5%. On the other hand, the OOV rate on Greek on the same dataset is only 2.2% which is ∼12.3% lower than Hindi. Since the language families of the languages used in mUSE are not Indo-Aryan (Hindi and Bengali) nor Dravidian (Telugu) [4], the mUSE's sentencepiece cannot handle Indo-Aryan and Dravidian language families well. To make the mUSE encoder handle these language families better, we might need to retrain the mUSE sentencepiece tokenizer by using other tokenizers (i.e., sentencepiece in mBERT). Since mBERT's tokenizer is trained on more than 100 languages, mBERT and XLM-R perform better than the $\text{mUSE}_{\text{cl-rekd}}$ on Indo-Aryan languages. Another solution would be to use the universal tokenizer (Gillick et al., 2016) which represents the input as bytes instead of characters.

A.9 Hyperparameter Configurations

Hyperparameters	Values for grid search
Learning Rates	1e-3,5e-4, 1e-4, 1e-5, 1e-6
α	0.1-1 (0.01/steps)
γ	[100, 1000, 10000]
β	1,1e-1, 1e-2, 1e-3, 1e-4
ω	1,1e-1, 1e-2, 1e-3, 1e-4
#negative samples	[1, 2, 3, 5, 10]

Table 11: Hyperparameter configurations.

A.10 Error Analysis: MT-assisted vs Ours

We provide some sample questions that our model and the MT-assisted model (DPR+GNMT) answer differently in Table 12. We notice that our method does better than MT-assisted when the questions are very specific, such as the question relating to Brothers Grimm. In general, we find that our model generally performs better when the question contains names. This is because machine translation can sometimes fail to translate names properly.

On the other hand, when the questions are not

[4] according to Ethnologue

Lan	Question	Predict context	Correct?
DE-EN (1)	Welchem Märchen der Gebrüder Grimm entspricht die Geschichte Diebstahl der Butter des Partners von Aarne Thompson? (Which fairy tale by the Brothers Grimm corresponds to the story of the theft of butter from Aarne Thompson's partner?)	$mUSE_{cl\text{-}rekd}$: Cat and Mouse in Partnership" (German: Katze und Maus in Gessellschaft) is a Brothers Grimm fairy tale. It is Aarne-Thompson type 15, Stealing the Partner's Butter.	✓
		DPR+GMT: Tubman and her brothers, Ben and Henry, escaped from slavery on September 17, 1849. Tubman had been hired out to Dr. Anthony Thompson, who owned a large plantation in an area called Poplar Neck in neighboring Caroline County [.....]	✗
VI-EN (1)	Edward ghét loại nhạc nào? (Edward hated any kind of music?)	$mUSE_{cl\text{-}rekd}$: [.....] Edward is musical, able to play the piano like a virtuoso. He enjoys a wide range of music, including classical, jazz, progressive metal, alternative rock, and punk rock, but dislikes country. [.....]	✓
		DPR+GMT: [.....] Born in Woodside, Dudley, Edwards signed for Manchester United as a teenager and went on to become the youngest player to play in the Football League First Division [.....]	✗
DE-EN (2)	Was war das Durchschnittseinkommen pro Person in der Stadt? (What was the median income per person in the city?)	$mUSE_{cl\text{-}rekd}$: The median income for a household in the city was $33,295, and the median income for a family was $39,250. Males had a median income of $31,875 versus $18,594 for females. The per capita income for the city was $14,606	✗
		DPR+GNMT: The median income for a household in the city was $46,795, and the median income for a family was $60,424. Males had a median income of $41,192 versus $29,454 for females. The per capita income for the city was $23,562.	✓
VI-EN (2)	Sự kiện nào diễn ra từ những năm 1793 đến 1802? (What events took place between 1793 and 1802?)	$mUSE_{cl\text{-}rekd}$: [.....] Tabinshwehti's brother-in-law, Bayinnaung, succeeded to the throne in 1550 and reigned 30 years, launching a campaign of conquest invading several states, including Manipur (1560) and Ayutthaya (1564). [.....]	✗
		DPR+GNMT: [.....] These wars were the War of the Austrian Succession (1740–1748), the Seven Years' War (1756–1763), the American Revolution (1765–1783), the French Revolutionary Wars (1793–1802) and the Napoleonic Wars (1803–1815). [.....]	✓
VI-EN (3)	Iron Man được phát hành vào năm nào? (Iron Man was released in what year?)	$mUSE_{cl\text{-}rekd}$: [.....] Created by Stan Lee, Larry Lieber and Jack Kirby, Ant-Man's first appearance was in Tales to Astonish #35 (September 1962). [.....]	✗
		DPR+GMT: [.....] After the successful release of Iron Man (2008) in May, the company set a July 2011 release date for The Avengers. [.....]	✓

Table 12: Examples from from $mUSE_{cl\text{-}rekd}$ and DPR+GNMT with the highest question-context similarity

specific, it is a toss-up whether the prediction is correct for both models. Rows 2 and 3 show examples of such vague questions. Lastly, we found that our model performed particularly worse than MT-assisted on contents related to numbers, as shown in the last two examples. Embedding numerical information is generally hard when the data is scarce. The model in our method has to map questions from multiple languages and numbers close together. This makes learning numerical concepts such as in example number four challenging. We believe this is a good avenue for further research.

A.11 Responsible NLP Research Checklist

Did you discuss the limitations of your work? The limitation of our work is out-of-domain problems. We strongly advise against using our model with out-of-domain data.

Did you discuss any potential risks of your work? There is a risk of retrieving incorrect documents causing the machine reading comprehension part to produce wrong answers.

Did you discuss the license or terms for use and/or distribution of any artifacts? The XORQA dataset is under the MIT License, while XQuAD and MLQA are under CC-BY-SA 4.0.

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)? XORQA, XQuAD, and MLQA were created for retrieval QA assessments; we use them for this exact purpose.

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? From our careful inspection, there is no offensive content or sensitive data included in the three datasets: XORQA, XQuAD, and MLQA.

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? We reported the mean average score for all experiments, which is a common practice for reporting the performance.

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.)? All experiments were done by Tensorflow because mUSE is only available on Tensorflow.