Unsupervised Multilingual Dense Retrieval via Generative Pseudo Labeling

Chao-Wei Huang[†‡] **Chen-An Li**[†*] **Tsu-Yuan Hsu**[†*] **Chen-Yu Hsu**[†] **Yun-Nung Chen**[†]
[†]National Taiwan University, Taipei, Taiwan
[‡]Taiwan AI Labs, Taipei, Taiwan
f07922069@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

Dense retrieval methods have demonstrated promising performance in multilingual information retrieval, where queries and documents can be in different languages. However, dense retrievers typically require a substantial amount of paired data, which poses even greater challenges in multilingual scenarios. This paper introduces **UMR**, an Unsupervised Multilingual dense Retriever trained without any paired data. Our approach leverages the sequence likelihood estimation capabilities of multilingual language models to acquire pseudo labels for training dense retrievers. We propose a two-stage framework which iteratively improves the performance of multilingual dense retrievers. Experimental results on two benchmark datasets show that UMR outperforms supervised baselines, showcasing the potential of training multilingual retrievers without paired data, thereby enhancing their practicality.[1]

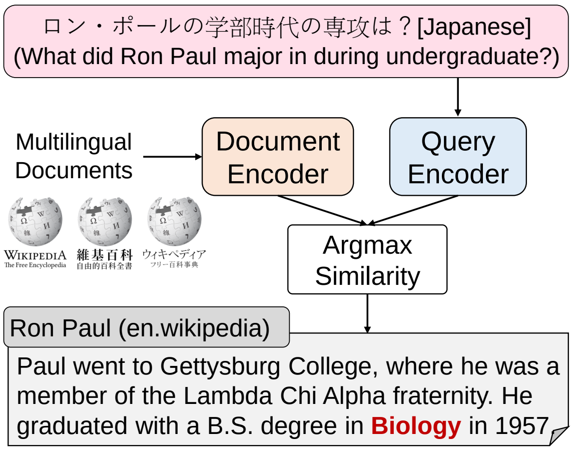

ロン・ポールの学部時代の専攻は？[Japanese]
(What did Ron Paul major in during undergraduate?)
Multilingual Documents
WIKIPEDIA The Free Encyclopedia
維基百科 自由的百科全書
ウィキペディア フリー百科事典
Document Encoder
Query Encoder
Argmax Similarity
Ron Paul (en.wikipedia)
Paul went to Gettysburg College, where he was a member of the Lambda Chi Alpha fraternity. He graduated with a B.S. degree in Biology in 1957

Figure 1: Illustration of the multilingual dense retrieval process. Given a query, the goal is to retrieve relevant documents in any language. Dense retrieval achieves this by encoding the query and documents into dense representations and performing vector similarity search.

1 Introduction

Multilingual information retrieval (mIR) has attracted significant research interest as it enables unified knowledge access across diverse languages. The task involves retrieving relevant documents from a multilingual collection given a query, which may be in a different language. Traditional sparse retrieval methods that rely on lexical matching often yield inferior performance due to the different scripts used (Asai et al., 2021b). On the other hand, dense retrieval methods have shown promising results in multilingual retrieval by capturing semantic relationships between queries and documents (Shen et al., 2022; Zhang et al., 2022; Ren et al., 2022; Sorokin et al., 2022). Figure 1 illustrates the process of multilingual dense retrieval.

Nevertheless, training dense retrievers requires a large amount of paired data, which is costly and time-consuming to collect. This challenge is particularly pronounced for low-resource languages where the availability of annotated data is limited. Consequently, there is a growing demand for more efficient techniques to build multilingual dense retrievers, such as leveraging unsupervised learning and transfer learning, to alleviate the data requirement.

The advance of large-scale language model pretraining (Devlin et al., 2019; Conneau et al., 2020) presents a compelling avenue to explore, namely leveraging the multilingual capabilities of pretrained multilingual language models. In this paper, we propose **UMR**, an unsupervised approach to multilingual dense retrieval that only relies on multilingual queries *without requiring any paired data*. Our method leverages the sequence likelihood estimation capabilities of multilingual language models to obtain pseudo labels by estimating the conditional probability of generating the query given the document. This allows training of multilingual

[*] Equal contribution
[1] All of our source code, data, and models are available: https://github.com/MiuLab/UMR

Findings of the Association for Computational Linguistics: EACL 2024, pages 736–746
March 17-22, 2024 ©2024 Association for Computational Linguistics

dense retrievers in a fully unsupervised manner.

To evaluate the effectiveness of our approach, we conduct experiments on XOR-TyDi QA (Asai et al., 2021a), a widely used benchmark for multilingual information retrieval. Our results demonstrate that **UMR** outperforms or performs comparably to existing supervised baselines on both XOR-Retrieve and XOR-Full. Additionally, we conduct comprehensive ablation studies to analyze the impact of different components of our approach. Our approach shows great potential for being applied to a broad range of multilingual information retrieval tasks, where it can reduce the dependence on costly paired data.

Our contributions can be summarized in 3-fold:

- We propose **UMR**, the first unsupervised method for training multilingual dense retrievers without any paired data.
- Experimental results on two benchmark datasets show that our proposed method performs comparable to or even outperforms strong supervised baselines.
- The detailed analysis justifies the effectiveness of individual components in our **UMR**.

2 Related Work

Dense Retrieval Dense retrieval has garnered significant attention for its potential to enable retrieval in the semantic space. A prominent method in this area is the dense passage retriever (DPR) (Karpukhin et al., 2020), which comprises a query encoder and a passage encoder. Several studies have also explored efficient training approaches, such as RocketQA (Qu et al., 2021) and alternative architectures for dense retrieval, e.g., ColBERT (Khattab and Zaharia, 2020). A common technique for training performant dense retrievers is knowledge distillation from cross encoders. BERT-CAT (Hofstätter et al., 2020) proposed cross-architecture knowledge distillation to improve dense retrievers and rankers. Izacard and Grave distilled knowledge from the reader model to the retriever model, thus improving its performance on open-domain question answering. However, the majority of previous work has primarily focused on English retrieval, limiting its applicability to other languages.

Multilingual Dense Retrieval Multilingual information retrieval has been an active research area for several decades. Early work in this field primarily focused on cross-lingual information retrieval (CLIR), aiming to retrieve relevant documents in a different language from the query language (Nasharuddin and Abdullah, 2010). Traditional CLIR systems relied on aligning bilingual dictionaries or parallel corpora to translate queries or documents into a common language for retrieval. However, these systems often faced limitations in translation quality, vocabulary coverage, and handling domain-specific expressions (Ballesteros and Croft, 1996; Vulić and Moens, 2015; Sharma and Mittal, 2016).

In recent years, dense retrieval has emerged as a promising approach for multilingual information retrieval. Various studies have demonstrated the effectiveness of dense retrieval methods in cross-lingual and multilingual scenarios. Models such as XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019) have achieved remarkable performance on diverse natural language processing tasks, including similarity-based retrieval tasks. The success of these models has spurred researchers to explore their application in multilingual information retrieval (Jiang et al., 2020).

Supervised mIR Most existing multilingual retrieval models rely on supervised training, where paired data consisting of queries and corresponding relevant documents in different languages is required. These methods use popular datasets such as Mr. TyDi (Zhang et al., 2021) and XOR-TYDI QA (Asai et al., 2021a). DR.DECR proposes to leverage the knowledge of an English retriever to improve cross-lingual retrieval (Li et al., 2022). It uses paired data for machine translation to align multilingual representations. Quick proposes to leverage supervised question generation to improve cross-lingual dense retrieval (Ren et al., 2022). However, these methods still rely on question-document pairs and paired translation data. The requirement for paired training data can be a significant bottleneck for multilingual information retrieval, especially for low-resource languages, where it is challenging to obtain large amounts of data. In contrast, our method does not require any paired data or paired translation data, eliminating the requirement for annotation resources.

Unsupervised Dense Retrieval There have been recent efforts to develop unsupervised or weakly supervised approaches to dense retrieval. InPars (Bonifacio et al., 2022), Promptagator (Dai et al., 2022), and CONVERSER (Huang et al., 2023) all propose to generate synthetic queries with LLMs from few-shot examples, which achieved comparable performance to supervised methods in dense retrieval. However, synthetic query generation is less suitable for the multilingual setting as multilingual query generation remains a hard problem for multilingaul LLMs, which is demonstrated in our experiments. UPR and ART are the most closely related work to our work (Sachan et al., 2022a,b). UPR proposes to rerank passages with zero-shot question generation, which only requires a base LLM. ART proposes to train a retriever without paired data with unsupervised reranking by language models. Our method is similar to the framework proposed in ART, while we focus on multilingual scenarios where supervised data is even harder to collect.

Multilingual Evidence for Fact Checking The power of generative models has made it easier for misleading information to spread, posing challenges in its detection (Shu et al., 2017; Wang, 2017). Previous fact-checking research has considered single-language evidence, often lacking sufficient cues for verification. Dementieva et al. (2023) proposed the use of multilingual evidence as features for fake news detection, resulting in improved performance. While our method does not specifically focus on fact checking, it can be applied to assist in finding multilingual evidence, thereby enhancing the verification process.

In this paper, we introduce an unsupervised multilingual dense retrieval approach that leverages the generative capabilities of multilingual language models to obtain pseudo labels for training the dense retriever. Our method eliminates the need for paired training data, making it particularly suitable for low-resource languages.

3 Our Method: UMR

The goal of multilingual information retrieval is to retrieve relevant documents, denoted as D^+, from a collection of multilingual documents $\mathcal{D} = d_1, \cdots, d_n$. We adopt a widely used dense retrieval architecture, DPR (Karpukhin et al., 2020), comprising a query encoder E_q and a document encoder E_d. The documents are pre-encoded using the document encoder and then indexed for efficient vector search. Given a query q, the relevance score of a query-document pair is computed as their vector similarity:

$$r(q, d_i) = E_q(q)^\top E_d(d_i)$$

This section introduces our proposed framework **UMR** for training unsupervised multilingual retrievers iteratively. The framework consists of two stages: 1) unsupervised multilingual reranking and 2) knowledge-distilled retriever training, as illustrated in Figure 2.

3.1 Unsupervised Multilingual Reranking

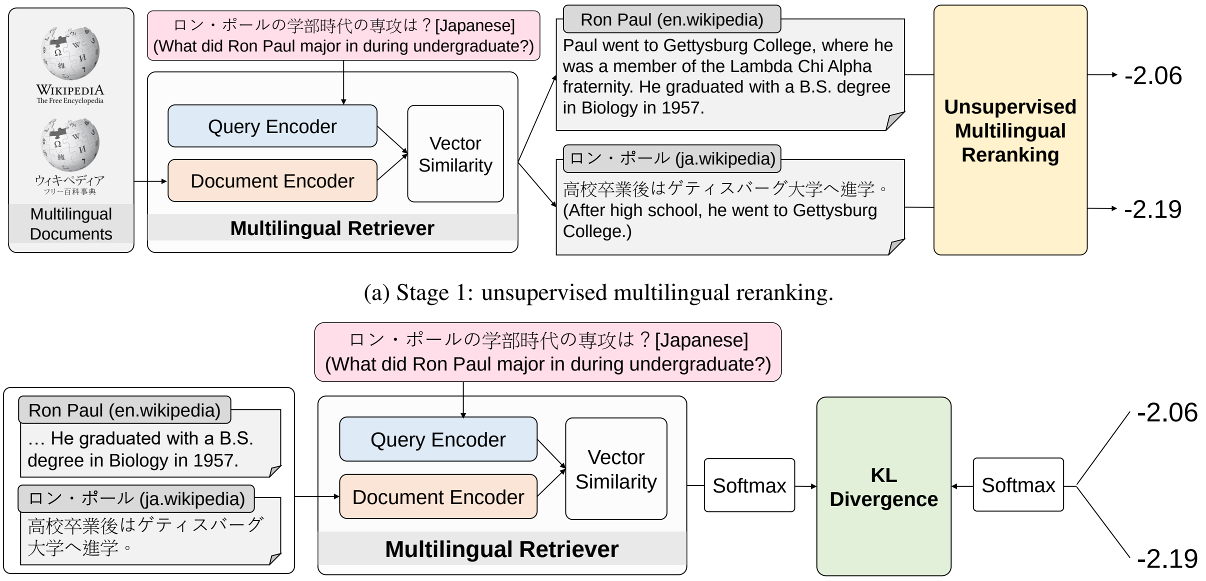
In the first stage, we leverage the generative capabilities of multilingual language models to rerank retrieved passages and obtain pseudo labels for training the dense retriever in an unsupervised manner. This stage is depicted in Figure 2a.

Formally, given a query q in language L, we retrieve the top-k documents $d_1, \cdots, d_k$ from the multilingual document collection using a multilingual dense retriever, forming k query-document pairs. We then utilize a pre-trained autoregressive multilingual language model (mLM) for unsupervised multilingual reranking. For each query-document pair (q, d_i), the relevance score is reestimated as:

$$\hat{r}(q, d_i) = \frac{1}{|q|} \sum_{j=1}^{|q|} -\log p(q_j \mid d_i, q_{<j}, I),$$

where q_j denotes the j-th token of q, $|q|$ denotes the length of q, $q_{<j}$ denotes the first $(j-1)$ tokens of q, and I represents an instruction. Note that the language model does not actually perform generation, as we are only estimating the joint probability since the actual query q is given. Therefore, we can directly employ pre-trained mLMs, without requiring any instruction tuning. In our framework, we employ the prefix "*Based on the passage, please write a question in L*" for reranking.

This relevance score can be interpreted as the negative log-likelihood of the mLM generating the query q given the document d_i. Intuitively, the more relevant d_i is to q, the more likely the mLM will generate q. Thus, we leverage this property to rerank multilingual passages, even though the mLM is pre-trained without any ranking supervision. Notably, while this step does not require any paired data, we need a set of multilingual queries, which is comparatively easier to collect than query-document pairs.


WIKIPEDIA
The Free Encyclopedia
ウィキペディア
フリー百科事典
Multilingual Documents
ロン・ポールの学部時代の専攻は？[Japanese]
(What did Ron Paul major in during undergraduate?)
Query Encoder
Document Encoder
Vector Similarity
Multilingual Retriever
Ron Paul (en.wikipedia)
Paul went to Gettysburg College, where he was a member of the Lambda Chi Alpha fraternity. He graduated with a B.S. degree in Biology in 1957.
ロン・ポール (ja.wikipedia)
高校卒業後はゲティスバーグ大学へ進学。
(After high school, he went to Gettysburg College.)
Unsupervised Multilingual Reranking
-2.06
-2.19
ロン・ポールの学部時代の専攻は？[Japanese]
(What did Ron Paul major in during undergraduate?)
Ron Paul (en.wikipedia)
… He graduated with a B.S. degree in Biology in 1957.
ロン・ポール (ja.wikipedia)
高校卒業後はゲティスバーグ大学へ進学。
Query Encoder
Document Encoder
Vector Similarity
Multilingual Retriever
Softmax
KL Divergence
Softmax
-2.06
-2.19

(a) Stage 1: unsupervised multilingual reranking.

(b) Stage 2: knowledge-distilled retriever training.

Figure 2: Illustration of our proposed UMR, unsupervised multilingual dense retrieval.

3.2 Knowledge-Distilled Retriever Training

Previous work has demonstrated that distilling knowledge from a strong reranker can significantly enhance the performance of the retriever (Rosa et al., 2022; Li et al., 2022). In the second stage, we employ the mLM reranker from the first stage as the teacher model to improve the performance of the e performance of the dense retriever. We initialize the student model with the multilingual retriever used in the first stage and train it to mimic the outputs of the teacher model by minimizing the Kullback-Leibler (KL) divergence.

Specifically, the relevance of a document d_i to a query q predicted by the student model can be defined as:

$$P(d_i \mid q) = \frac{\exp(r(q, d_i))}{\sum_{d_j \in \mathcal{D}_\mathcal{B}} \exp(r(q, d_j))},$$

where $\mathcal{D}_\mathcal{B}$ denotes the documents in the current batch. Similarly, the relevance predicted by the teacher model can be defined as:

$$\hat{P}(d_i \mid q) = \frac{\exp(\hat{r}(q, d_i)/\tau)}{\sum_{d_j \in \mathcal{D}_\mathcal{B}} \exp(\hat{r}(q, d_j)/\tau)},$$

where τ is the temperature parameter for controlling the sharpness of the distribution. Finally, the loss function is the KL divergence between two distributions:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{q \in \mathcal{B}} \mathrm{KL}(\hat{P}(d \mid q) \| P(d \mid q)),$$

where $|\mathcal{B}|$ denotes the size of the batch. Note that we do not convert rankings into hard labels as done in previous work, where only the top-ranked passage is labeled as positive and the rest are treated as hard negatives. The prior approach disregards the fine-grained scores of the negatively labeled documents, potentially leading to suboptimal knowledge transfer. Instead, we use KL loss to enable the retriever to learn the predicted distribution of the reranker, which we observed improves retrieval performance.

In the retriever training process, in-batch negative examples play a critical role in dense retrieval performance, enabling larger batch sizes while remaining efficient (Karpukhin et al., 2020). We incorporate this technique in our knowledge distillation process by considering documents from other queries in the same batch as in-batch negatives. The scores of the in-batch negatives are set to a very small number, effectively zeroing their probability after the softmax operation. Specifically, with a batch size of b and n documents per query, each query has n associated reranking scores and $n \times (b-1)$ negative documents.

3.3 Iterative Training

To prevent overfitting on the same top-k passages and optimize the retriever's performance, we introduce an iterative training approach. In each iteration, we use the trained retriever to build an index, retrieve the top-k documents, and perform unsupervised multilingual reranking. We then fine-tune the trained retriever using knowledge-distilled retriever training. The fine-tuned retriever becomes the retriever for the next iteration. This iterative training allows for refreshing the retrieval index in each iteration, avoiding training solely on the same documents. Notably, in the first iteration where no trained retriever is available, we employ the unsupervised pretrained multilingual retriever, mContriever (Izacard et al., 2021).

4 Experiments

Our proposed framework, **UMR**, can be applied to various multilingual information retrieval tasks, such as *cross-lingual passage retrieval* and *multilingual open-domain question-answering*. We evaluate our approach on XOR-TYDI QA (Asai et al., 2021a), a popular benchmark for multilingual information retrieval. We also conduct ablation studies to analyze the impact of different components of our approach.

4.1 Datasets

XOR-TYDI QA (Asai et al., 2021a) is a multilingual open QA dataset consisting of 7 typologically diverse languages, Arabic, Bengali, Finnish, Japanese, Korean, Russian, and Telugu. The questions are originally from TYDI QA (Clark et al., 2020) and posed by native speakers in a naturally information-seeking scenario. There are two subtasks in XOR-TYDI QA:

- **XOR-Retrieve** requires a system to retrieve English passages given a query in language L other than English. The evaluation metrics used are R@2kt and R@5kt, which measure the recall by computing the fraction of the questions for which the minimal answer is contained in the top $\{2000, 5000\}$ tokens retrieved.

- **XOR-Full** requires a system to retrieve either English documents or documents in the query language L in order to generate an answer in L. The answers are annotated by 1) extracting spans from Wikipedia in the same language as the question (in-language) or 2) translating English spans extracted from English Wikipedia to the target language (cross-lingual). The evaluation metrics used are F1, EM, and BLEU. Note that since **UMR** is only responsible for retrieving relevant documents, we use the reader model from CORA to generate an answer given the retrieved documents.

For the multilingual passage collection, we directly use the preprocessed passage collection released by CORA (Asai et al., 2021b), which consists of February 2019 Wikipedia dumps of 13 diverse languages from all XOR-TYDI QA languages. The collection has 44 million passages.

4.2 Baseline Systems

- **BM25** retrieves passages from the target language only. We use a BM25-based lexical retriever implemented in CORA (Asai et al., 2021b), which uses the implementation from Pyserini (Lin et al., 2021). The retrieved passages are fed to a multilingual QA model to extract final answers.

- **MT+DPR** first translates the question into English and retrieves English documents with DPR (Karpukhin et al., 2020), which is a monolingual retriever.

- **mGenQ** generates multilingual questions with `mT0`[2], a multilingual instruction-tuned language model. The generated questions are used to train a multilingual retriever. We generate the same amount of questions as the training set of XOR-Retrieve for each language.

- **mDPR**(Asai et al., 2021a) is a supervised multilingual retriever based on the popular DPR model. It is initialized from mBERT and trained on the training set of XOR-Retrieve and NaturalQuestions (Kwiatkowski et al., 2019).

- **CORA** (Asai et al., 2021b) consists of mDPR and mGEN, which follows the *retrieve-and-generate* recipe. The models are trained on the training set of XOR-Full with iterative data mining.

- **Sentri+mFiD** (Sorokin et al., 2022) is the state-of-the-art system of XOR-Full, which utilizes multilingual translations of the training set and self-training.

[2] TyDi QA is part of mT0's training data, which gives this baseline a slight advantage.

Model	R@2kt								R@5kt							
	Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg	Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg
Supervised																
mDPR	41.2	43.9	50.3	29.1	34.5	35.3	37.2	38.8	50.4	57.7	58.9	37.3	42.8	44.0	44.9	48.0
MT+DPR	48.3	54.4	56.7	41.8	39.4	39.6	18.7	42.7	52.5	63.2	65.9	52.1	46.5	47.3	22.7	50.0
Unupervised																
UMR	36.7	33.6	51.6	33.2	38.3	37.2	35.8	38.1	45.0	48.8	61.9	43.4	47.3	46.9	44.4	48.2

Table 1: Performance on XOR-Retrieve test set (%).

Model	Target Language F1							Macro Average		
	Ar	Bn	Fi	Ja	Ko	Ru	Te	F1	EM	BLEU
Supervised										
MT + DPR	7.6	5.9	16.2	9.0	5.3	5.5	0.8	7.2	3.3	6.3
CORA	59.8	40.4	42.2	44.5	27.1	45.9	44.7	43.5	33.5	31.1
Sentri + mFiD	-	-	-	-	-	-	-	46.2	39.0	33.7
Unsupervised										
BM25	31.1	21.9	21.4	12.4	12.1	17.7	–	–	–	–
UMR + CORA Reader	59.8	41.0	41.4	44.3	30.4	46.4	50.9	44.9	34.7	32.5

Table 2: Performance on XOR-Full test set (%).

4.3 Implementation Details

For the reranking stage, we retrieve top-100 documents with the trained retriever using a highly-efficient vector search engine, faiss (Douze et al., 2024). All top-100 documents are reranked by the language modeling-adapted variant of **mt5-xl**, which has 3 billion parameters (Xue et al., 2021). Note that it is neither fine-tuned on supervised data nor instruction-tuned.

For the knowledge distillation stage, we use **mContriever** as the initial retriever (Izacard et al., 2021). In order to reduce memory consumption, we employ the gradient cache technique (Gao et al., 2021). All experiments are conducted on 4xNVIDIA V100 GPUs. Detailed hyperparameters for training retrievers are shown in Appendix A. We run two iterations of iterative training.

4.4 Main Results

4.4.1 XOR-Retrieve

The experimental results on the test set of XOR-Retrieve are shown in Table 1. Compared to the supervised baseline mDPR, our proposed **UMR** achieves comparable or even slightly better performance (48.0% vs. 48.2%) despite not using any paired data. This demonstrates the effectiveness of utilizing mLM for generative pseudo labeling, providing supervision of similar quality compared to human annotation. The results for each language show that **UMR** underperforms mDPR significantly in Arabic (Ar) and Bengali (Bn) while achieving comparable or superior performance in other languages.

4.4.2 XOR-Full

The experimental results on the test set of XOR-Full are shown in Table 2. Our proposed **UMR** outperforms a strong supervised baseline CORA and only slightly underperforms the state-of-the-art system Sentri+mFiD. This result further demonstrates the effectiveness of our proposed method, which requires neither paired data nor query translations. The performance could be further improved by combining **UMR** with mFiD, which was shown to be very crucial to the state-of-the-art performance of Sentri (Sorokin et al., 2022). Results for each language show that **UMR** outperforms CORA significantly in Telugu while achieving similar performance in other languages.

5 Analysis and Discussion

In this section, we conduct analytical experiments on the dev set of XOR-Retrieve and XOR-Full since the test sets are not publicly available.

	R@2kt	R@5kt
mDPR	40.50	50.20
mGenQ	29.08	38.67
mContriever	25.50	35.06
+ rerank	34.24	41.88
UMR (iter=1)	41.23	51.50
UMR (iter=2)	41.68	51.94
+ rerank	42.34	52.36

Table 3: Performance of unsupervised multilingual reranking on XOR-Retrieve dev set (%). We conduct analyses on the dev set as the test set is not publicly available.

	R@2kt	R@5kt
UMR (iter=1)	41.23	51.50
- in-batch negative	39.56	49.41

Table 4: Performance on XOR-Retrieve dev set with or without using in-batch negatives (%).

Temperature	R@2kt	R@5kt
1	29.58	38.82
0.1	37.38	46.70
0.04	37.12	46.55
0.02	38.43	46.45

Table 5: Performance on XOR-Retrieve dev set when varying the value of temperature (%).

Batch size	R@2kt	R@5kt
4	36.45	46.02
8	38.94	49.38
16	40.07	50.30
32	40.41	50.48

Table 6: Performance on XOR-Retrieve dev set when varying the value of batch size (%).

5.1 Unsupervised Multilingual Reranking

We conduct an analysis to validate the effectiveness of the unsupervised multilingual reranking stage. As shown in Table 3, reranking improves the unsupervised retriever mContriever significantly, improving the result from 25.50 to 34.24 in terms of R@2kt. This demonstrates that our unsupervised multilingual reranking is effective in reranking the results of the first-stage retriever. We also observe that the performance of **UMR** converges after two iterations. This could be explained by the result of reranking **UMR** (iter=2), where reranking only achieves a slight improvement. Given this result, we believe that the performance of **UMR** is bounded by the reranker. Future work could explore using more powerful or instruction-tuned LLM and developing superior reranking methods.

5.2 Question Generation

Previous work has shown that training a multilingual question generator for generating multilingual questions can improve the performance of multilingual retrieval (Ren et al., 2022). We aim to examine whether this method is feasible in an unsupervised scenario. We perform multilingual question generation via prompting an instruction-tuned multilingual LLM, *mT0* (Muennighoff et al., 2022). With randomly sampled passages, we generate the same amount of questions as the training set of XOR-Retrieve for each language. These question-passage pairs are then used to train a multilingual retriever, mGenQ, using the same hyperparameters as mDPR. The performance of mGenQ is reported in Table 3. mGenQ underperforms mDPR and **UMR** significantly, demonstrating the difficulty of applying question generation to a multilingual scenario where there is no training data. We manually examine the generated questions and find that roughly half of the questions are either nonsensical or not in the desired language. Future work could explore effective methods for unsupervised or few-shot multilingual question generation.

5.3 In-batch Negative

We conduct an ablation study to validate the effectiveness of the in-batch negative examples. The results are shown in Table 4. Removing in-batch negatives results in a slight degradation in performance, which is less pronounced compared to supervised dense retrieval methods. This could be explained by the fact that we include multiple documents per question with fine-grained scores for training, which already includes distinguishing between relevant documents and hard negatives.

5.4 Effect of Hyperparameters

Dense retrievers are known to be sensitive to hyperparameters, e.g., batch size. In this analysis, we examine how different hyperparameters affect the performance of **UMR**.

	English Answers Only			Target Language Answers			All		
	Top-1	Top-5	Top-20	Top-1	Top-5	Top-20	Top-1	Top-5	Top-20
Supervised									
CORA	10.8	26.9	41.8	37.0	55.0	64.9	27.1	45.7	58.1
Unsupervised									
mContriever	3.2	7.7	13.3	18.9	40.1	56.4	14.5	31.2	45.4
mContriever+rerank	4.4	9.4	15.1	29.1	50.1	61.5	20.5	37.5	49.1
UMR (iter=1)	5.2	10.8	18.1	27.7	48.6	64.6	20.2	37.6	52.1
UMR (iter=2)	4.7	11.4	17.9	26.2	49.2	64.6	19.1	38.5	52.1

Table 7: Retrieval performance on XOR-Full dev set (%).

5.4.1 Batch Size

Training dense retrievers requires a larger batch size. The results of varying batch sizes are shown in Table 6. When the batch size is under 16, we observe significant degradation in performance. Hence, in our experiments, we set the batch size to 16. Note that in our training framework, each question is associated with multiple documents. Therefore, with a batch size of 16 and 16 documents per question, each question is paired with 256 documents in a batch.

5.4.2 Temperature

The results of varying temperature values are shown in Table 5. We observe that **UMR** is highly sensitive to the value of temperature. When the temperature is set to 1, the performance is degraded significantly from 38.43% to 29.58% in terms of R@2kt. We hypothesize that the range of the negative log-likelihood of the reranker is the root cause of this phenomenon since higher temperature results in a more flat distribution, making it harder for the retriever to learn meaningful knowledge.

5.5 Retrieval Performance on XOR-Full

In order to evaluate the multilingual retrieval performance where the language of the relevant documents is not known apriori, we examine the retrieval performance on XOR-Full. Since there is no official evaluation of the retrieval performance, we take the answers from the dev set, where some of the questions have English answers. We split the questions into two categories: 1) questions with annotated English answers and 2) questions with only answers in the target language. We evaluate the retrieval performance by checking whether any of the answers are present in the top-k retrieved documents. The results are shown in Table 7.

We observe that despite outperforming CORA in downstream question-answering performance, **UMR** underperforms CORA significantly in terms of retrieval performance. This underperformance is especially pronounced in Top-1 recall, which aligns with the observation from ART (Sachan et al., 2022b). We hypothesize that while unsupervised reranking via estimating conditional probability can provide good supervision, it cannot distinguish the most relevant documents very well. We also note that since the reader model takes top-20 passages to generate the answer, Top-20 recall should be a better indicator for the downstream QA performance. This could explain why **UMR** achieves better QA performance while performing slightly worse in retrieval performance. In addition, this evaluation only considers the surface form of the answers, which might fail to capture the difference in surface forms.

6 Conclusion

In this paper, we propose **UMR**, the first unsupervised method for training multilingual dense retrievers without any paired data, which leverages the sequence likelihood estimation capability of pretrained multilingual language models. The proposed framework consists of two stages with iterative training. Experimental results on XOR-Retrieve and XOR-Full show that our proposed method performs comparable to or even outperforms strong supervised baselines. Finally, detailed analyses justify the effectiveness of individual components in our proposed **UMR**. We also identify that the performance of **UMR** might be bounded by the reranking performance of mLM. Hence, future work could explore better unsupervised reranking methods with large language models.

Limitations

While this paper demonstrates the promising performance of our fully unsupervised method for multilingual retrieval, it is important to acknowledge its limitations.

First, our approach assumes that the employed multilingual pre-trained language model already understands the languages present in our evaluated datasets. Consequently, the model's ability to estimate relevance for reranking in the first stage (unsupervised multilingual reranking) relies on this assumption. However, for low-resource languages that are not adequately covered by the language model, our proposed approach may struggle to achieve satisfactory performance due to inaccurate estimations. To address this limitation, we plan to conduct experiments on unseen languages in future work and explore alternative approaches, such as language adaptation techniques, to enhance the generalizability across diverse and even previously unseen languages.

It is crucial to address these limitations to ensure the applicability and effectiveness of our method across a wide range of languages, especially those with limited resources.

Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002- 013-MY3, 111-2628-E-002-016, and 112-2223-E002-012-MY5 and Google.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34:7547–7560.

Lisa Ballesteros and Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *DEXA*, pages 791–801. Citeseer.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2387–2392.

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

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

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

Daryna Dementieva, Mikhail Kuimov, and Alexander Panchenko. 2023. Multiverse: Multilingual evidence for fake news detection. *Journal of Imaging*, 9(4):77.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321, Online. Association for Computational Linguistics.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation.

Chao-Wei Huang, Chen-Yu Hsu, Tsu-Yuan Hsu, Chen-An Li, and Yun-Nung Chen. 2023. CONVERSER: Few-shot conversational dense retrieval with synthetic data generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 381–387.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

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

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with bert. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 26–31.

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)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2022. Learning cross-lingual IR from an English retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4428–4436, Seattle, United States. Association for Computational Linguistics.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Nurul Amelina Nasharuddin and Muhamad Taufik Abdullah. 2010. Cross-lingual information retrieval. *Electronic Journal of Computer Science and Information Technology*, 2(1).

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Houxing Ren, Linjun Shou, Ning Wu, Ming Gong, and Daxin Jiang. 2022. Empowering dual-encoder with query generator for cross-lingual dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3107–3121, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guilherme Moraes Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. No parameter left behind: How distillation and model size affect zero-shot retrieval. *arXiv preprint arXiv:2206.02873*.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022a. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. 2022b. Questions are all you need to train a dense passage retriever. *arXiv preprint arXiv:2206.10658*.

Vijay Kumar Sharma and Namita Mittal. 2016. Cross lingual information retrieval (clir): Review of tools, challenges and translation approaches. In *Information Systems Design and Intelligent Applications: Proceedings of Third International Conference INDIA 2016, Volume 1*, pages 699–708. Springer.

Tianhao Shen, Mingtong Liu, Ming Zhou, and Deyi Xiong. 2022. Recovering gold from black sand: Multilingual dense passage retrieval with hard and false negative samples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10659–10670, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Nikita Sorokin, Dmitry Abulkhanov, Irina Piontkovskaya, and Valentin Malykh. 2022. Ask me anything in your native language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 395–406, Seattle, United States. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

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

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022. Towards best practices for training multilingual dense retrieval models. *arXiv preprint arXiv:2204.02363*.

A Hyperparameters

The hyperparameters used for knowledge-distilled retriever training are listed in Table 8

hyperparameters	
max sequence length	256
batch size	16
gradient accumulation steps	1
# docs per question	16
train epochs	10
learning rate	2e-5
optimizer	AdamW
temperature τ	0.1

Table 8: Hyperparameters used in the knowledge distillation stage.