Knowledge Enhanced Pre-training for Cross-lingual Dense Retrieval

Hang Zhang[1*], **Yeyun Gong**, **Dayiheng Liu**[1],
Shunyu Zhang[2], **Xingwei He**[3†], **Jiancheng Lv** [1†], **Jian Guo**[4†]

[1] College of Computer Science, Sichuan University
[1] Engineering Research Center of Machine Learning and IndustryIntelligence
[2] Beihang University [3] The University of Hong Kong [4] IDEA Research, China
hangzhang_scu@foxmail.com, hexingwei15@gmail.com

Abstract

In recent years, multilingual pre-trained language models (mPLMs) have achieved significant progress in cross-lingual dense retrieval. However, most mPLMs neglect the importance of knowledge. Knowledge always conveys similar semantic concepts in a language-agnostic manner, while query-passage pairs in cross-lingual retrieval also share common factual information. Motivated by this observation, we introduce KEPT, a novel mPLM that effectively leverages knowledge to learn language-agnostic semantic representations. To achieve this, we construct a multilingual knowledge base using hyperlinks and cross-language page alignment data annotated by Wiki. From this knowledge base, we mine intra- and cross-language pairs by extracting symmetrically linked segments and multilingual entity descriptions. Subsequently, we adopt contrastive learning with the mined pairs to pre-train KEPT. We evaluate KEPT on three widely-used benchmarks, considering both zero-shot cross-lingual transfer and supervised multilingual fine-tuning scenarios. Extensive experimental results demonstrate that KEPT achieves strong multilingual and cross-lingual retrieval performance with significant improvements over existing mPLMs.

Keywords: Cross-lingual Retrieval, Dense Retrieval, Pre-training

1. Introduction

Cross-lingual retrieval aims to search relevant documents given queries and large document collections in different languages (Peters et al., 2012). It is fundamental in various cross-language downstream tasks, such as open-domain question answering (Asai et al., 2021b; Liu et al., 2019), dialogue generation (Kim et al., 2021), fact-checking (Huang et al., 2022), etc. Meanwhile, it is essential in real-world search engines, for example, Google Search provides services across more than 100 languages. Nowadays, dense retrieval-based methods have shown strong performance and potential in cross-lingual retrieval (Zhang et al., 2022b; Asai et al., 2021b). These methods typically map queries and multilingual documents into a low-dimensional language-agnostic dense space and utilize the vector similarity between them to measure semantic relevance (Nair et al., 2022).

For obtaining a high-quality multilingual dense space, multilingual pre-trained language models (mPLMs) have been widely applied. General mPLMs, e.g., mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), usually adopt token-level tasks (multilingual or translation mask language modeling), which are not suitable for cross-lingual retrieval. To address this limitation, recent works propose several sentence-level pre-training tasks and achieve remarkable improvements, such as mContriever (Izacard et al., 2022) and CCP (Wu et al., 2022) leverage contrastive learning with positive pairs mined by randomly cropping, InfoXLM (Chi et al., 2021) and LaBSE (Feng et al., 2022) encourage the alignment of bilingual sentence pairs with large-scale private parallel data.

Despite the success, they neglect the importance of knowledge. On the one hand, knowledge conveys similar semantic concepts and meanings in a language-agnostic manner (Vulic and Moens, 2013), while queries and matched documents also share similar underlying semantics. This perspective highlights the potential of leveraging knowledge to construct high-quality simulated query-document pairs. On the other hand, Wikipedia [‡] and Wikidata [§] (Vrandečić and Krötzsch, 2014), the publicly large-scale corpora, contain abundant knowledge annotation within and across languages. This provides us with the opportunity to collect a vast amount of relevant text pairs through efficient mining. A good pre-training task should be relevant to the downstream task and cost-efficient to collect data (Chang et al., 2019). Considering these advantages, leveraging knowledge to construct text pairs for pre-training cross-lingual dense retrieval models holds great promise.

Based on the motivation above, we propose a **k**nowledge **e**nhanced **p**re-**t**rained (KEPT) model for cross-lingual dense retrieval. KEPT learns language-agnostic semantic representations

[*] Work is done during internship at IDEA Research
[†] Corresponding author

[‡] https://www.wikipedia.org/
[§] https://www.wikidata.org/

LREC-COLING 2024, pages 9810–9821
20-25 May, 2024.

through intra- and cross-language text pairs constructed based on knowledge. Specifically, we first build a multilingual knowledge base by utilizing hyperlink annotations from Wikipedia and cross-language page alignment annotations from Wikidata. In order to construct intra-language positive pairs, we extract symmetrically linked segments, which tend to contain similar facts. As for cross-language positive pairs, we extract descriptions of the same entity in different languages from Wikipedia pages, which generally provide an overall understanding of the corresponding entity. After obtaining intra- and cross-language pairs, we adopt contrastive learning as the pre-training objective to train KEPT. Compared to previous works that utilize cropping (Wu et al., 2022; Izacard et al., 2022) or expensive parallel corpus (Chi et al., 2021; Feng et al., 2022), our knowledge-based text pairs are closer to downstream tasks and more cost-effective.

To verify the effectiveness of KEPT, we conduct extensive experiments on three cross-lingual retrieval tasks, including Mr. TyDi (Zhang et al., 2021b), XOR Retrieve (Asai et al., 2021a), Mewsli-X (Ruder et al., 2021). Experimental results show that our approach brings a significant improvement over other advanced mPLMs, and achieves strong performance in cross-lingual dense retrieval. Our ablation studies and analysis demonstrate the effectiveness of proposed intra- and cross-language knowledge-based positive pairs.

2. Related Works

Multilingual Pre-training: mPLMs have become the fundamental model on various multilingual tasks for their superior performance (Liang et al., 2020; Hu et al., 2020). mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) pre-train models with multilingual masked language modelling (MMLM) task. Some works utilize parallel corpora to improve the cross-language capability of the model, such as XLM (Conneau and Lample, 2019), Unicoder (Huang et al., 2019), ERNIE-M (Ouyang et al., 2021).. Recently, integrating knowledge into multilingual pre-training has received increasing interest (Jiang et al., 2022; Liu et al., 2022; Ri et al., 2022; Xu et al., 2023). They focus on entity-knowledge-related tasks such as MLQA (Lewis et al., 2019) and NER (Liang et al., 2020).

Cross-lingual Dense Retrieval: Recent years have witnessed the remarkable progress of cross-lingual retrieval (Tran et al., 2020). Dense retrieval-based methods have shown strong performance and potential on this task (Jiang et al., 2020; Artetxe et al., 2020). The studies of cross-lingual dense retrieval can be divided into two categories, (1) pre-training with large corpora; (2) more effective fine-tuning. In the first category, mContriever (Izacard et al., 2022) and CCP (Wu et al., 2022) pre-train the dual-encoder model with pseudo-query-document pair constructed by randomly cropping a document. MSM (Zhang et al., 2023) proposes a masked sentence prediction task, which enforces the model to produce an information-rich sentence representation. Some works focus on learning sentence embeddings with parallel data, such as InfoXLM (Chi et al., 2021) and LaBSE (Feng et al., 2022) collect large private bilingual sentence pairs and learn bilingual alignment; m-USE (Yang et al., 2020) trained models with translation pairs, QA pairs, and SNLI (Bowman et al., 2015) corpus. Some works (Sun and Duh, 2020; Yang et al., 2022) also explore the use of Wiki corpus to train cross-language retrievers, but our knowledge mining method is more effective. As for effective fine-tuning strategies, some studies improve cross-lingual retriever with distillation (Reimers and Gurevych, 2020; Li et al., 2022; Ren et al., 2023; Zhuang et al., 2023), iterative self-supervised training (Tran et al., 2020), and parallel semantic contrastive learning (Hu et al., 2022) etc.

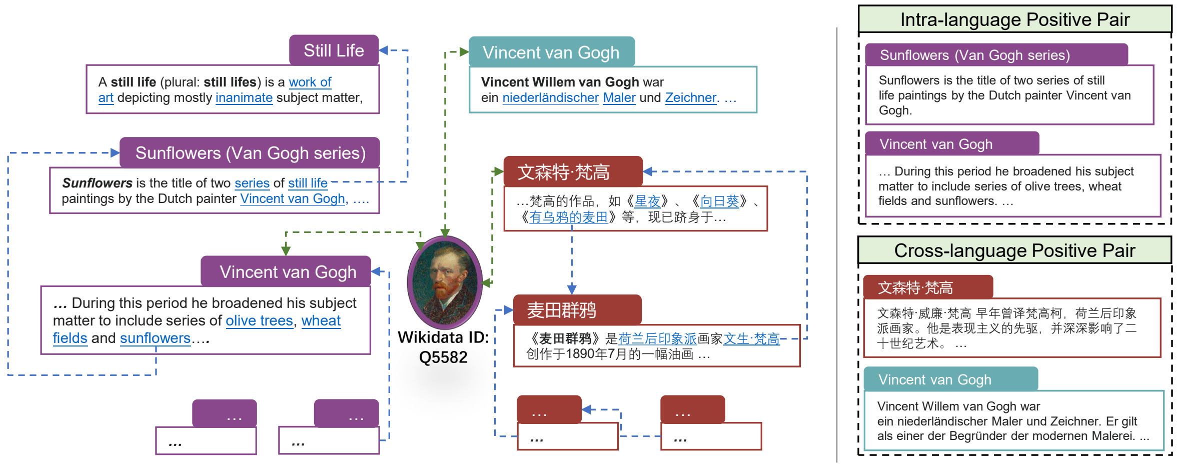
Monolingual Dense Retrieval Pre-training: Monolingual retrieval focuses on querying relevant documents in a single language (Karpukhin et al., 2020; Lin et al., 2022; Zhang et al., 2022a; He et al., 2022; Zhang et al., 2021a; Sun et al., 2023, 2022). Some studies explore the pre-training techniques tailored for dense retrieval. One line of work utilize contrastive learning with text pairs mined in different ways, such as ICT (Lee et al., 2019), WLP (Chang et al., 2019), HLP (Zhou et al., 2022), CROP (Izacard et al., 2022; Ma et al., 2022), etc. Another line enforces the model to produce an information-rich sentence representation via the autoencoding tasks, such as SEED-Encoder (Lu et al., 2021), Condenser (Gao and Callan, 2021), SimLM (Wang et al., 2022), CDMAE (Li et al., 2023), RetroMAE (Liu and Shao, 2022), etc.

3. Methodology

3.1. Task Definition and Architecture

Given a query q in any language l, cross-lingual dense retrieval aims to find the most relevant M documents $\{d_i^+\}_{i=1}^M$ from a large candidate corpus $\mathbb{C} = \{d_1, d_2, \ldots, d_N\}$ with N documents ($M \ll N$), which can be in any language, even if it differs from the query language l.

To achieve this goal, a typical cross-lingual dense retrieval model adopts a dual-encoder architecture. The query q and the document d are mapped into k-dimensional language-agnostic dense embedding space, respectively. Then the semantic relevance score $f_\theta(q, d)$ between q and d is measured by the similarity of their dense representations, which can


Still Life
A still life (plural: still lifes) is a work of art depicting mostly inanimate subject matter,
Vincent van Gogh
Vincent Willem van Gogh war ein niederländischer Maler und Zeichner. …
Sunflowers (Van Gogh series)
Sunflowers is the title of two series of still life paintings by the Dutch painter Vincent van Gogh, ….
文森特·梵高
…梵高的作品，如《星夜》、《向日葵》、《有乌鸦的麦田》等，现已跻身于…
Vincent van Gogh
… During this period he broadened his subject matter to include series of olive trees, wheat fields and sunflowers….
Wikidata ID: Q5582
麦田群鸦
《麦田群鸦》是荷兰后印象派画家文生·梵高创作于1890年7月的一幅油画 …
…
…
…
…
…
…
…
…
Intra-language Positive Pair
Sunflowers (Van Gogh series)
Sunflowers is the title of two series of still life paintings by the Dutch painter Vincent van Gogh.
Vincent van Gogh
… During this period he broadened his subject matter to include series of olive trees, wheat fields and sunflowers. …
Cross-language Positive Pair
文森特·梵高
文森特·威廉·梵高 早年曾译梵高柯，荷兰后印象派画家。他是表现主义的先驱，并深深影响了二十世纪艺术。 …
Vincent van Gogh
Vincent Willem van Gogh war ein niederländischer Maler und Zeichner. Er gilt als einer der Begründer der modernen Malerei. …

(a) Multilingual Knowledge Base

(b) Knowledge-based Text Pairs

Figure 1: Illustration of multilingual knowledge base and knowledge-based positive pairs. Different colored boxes represent Wikipedia pages in various languages, i.e., red for Chinese, purple for English, and purple for German. The blue text represents anchor text. The blue arrow connects the anchor text to the hyperlinked page, and the green arrow connects the page to the corresponding entity.

be formulated as follows:

$$f_\theta(q,d) = sim\left(E(q;\theta), E(d;\theta)\right), \qquad (1)$$

where $E(\cdot;\theta)$ denotes the encoder module parameterized with θ, and sim is the similarity function, e.g., euclidean distance, cosine distance. As queries and documents can encompass multiple languages, it is essential for encoders to possess robust cross-lingual comprehension capabilities. One widely employed strategy for achieving this is the utilization of mPLMs.

In practice, we utilize a shared transformer network for both queries and documents. The embedding $E(q;\theta)$ (resp. $E(d;\theta)$) for a query (resp. document) is obtained by averaging the hidden representations of the last layer. The similarity function sim adopted in KEPT is the inner product.

3.2. Multilingual Knowledge Construction

In our approach, we utilize Wikipedia and Wikidata (Vrandečić and Krötzsch, 2014) as the primary source of knowledge. To demonstrate the structure of intra- and cross-language knowledge, we show a simple knowledge graph in Figure 1(a).

To obtain intra-language knowledge, we utilize hyperlinks within Wikipedia pages. Given a Wikipedia page $p_i = [s_1, s_2, ..., s_K]$ composed of K segments s, we denote the title of p_i as $T(p_i)$ which serves as the unique identification of the page. Wikipedia provides annotations of hyperlinks in the form of $(a, h(a))$, where a is an anchor text and $h(a)$ is the hyperlinked Wikipedia page. As demonstrated in Figure 1(a), the segment in Wikipedia page *"Sunflowers (Van Gogh series)"* contains various anchors such as *"still life"* and *"Vincent van Gogh"*. Taking the anchor *"still life"* as an example, we link it to its corresponding Wikipedia page *"Still Life"*. Through these hyperlink annotations, we can effectively leverage the cross-page connections to acquire valuable intra-language knowledge.

For cross-language knowledge, we utilize Wikidata to align Wikipedia pages written in different languages but corresponding to the same entity. Wikidata comprises over 100 million entities and provides annotations linked to Wikipedia pages in the form of (e, l, p), where e represents the entity identifier, l denotes the language type, and p denotes the Wikipedia page of entity e in language l. As demonstrated in Figure 1(a), For the entity *Q5582*, its corresponding Wikipedia page in the Chinese language is "文森特·梵高" and its corresponding page in the English language is *"Vincent van Gogh"*. Through these annotations, we can link Wikipedia pages across languages.

3.3. Building Positive Pairs with Knowledge

It is acknowledged that a good pre-training task should be relevant to the downstream task. We observe that in cross-lingual dense retrieval tasks, positive pairs (q, d^+) usually exhibit similar semantics. Drawing from our constructed multilingual knowledge base, we further discover that text pairs with link annotations inherently contain comparable semantic concepts, including both intra-language page hyperlinks and cross-language entity links. Inspired by these observations, we leverage the knowledge base to mine text pairs that describe similar factual information.

For **intra-language positive pairs**, we find that symmetrically linked segments tend to contain similar facts. Specifically, given two segments s_i and s_j from Wikipedia pages $T(s_i)$ and $T(s_j)$, respectively, the set of Wikipedia pages they linked to are denoted as $H(s_i)$ and $H(s_j)$, respectively. We define s_i and s_j to be symmetrically linked if they satisfy the following condition:

$$T(s_i) \in H(s_j) \wedge T(s_j) \in H(s_i). \quad (2)$$

As demonstrated in Figure 1(b), the segment *"Sunflowers is . . . Vincent van Gogh"* in the Wikipedia page *"Sunflowers (Van Gogh series)"* and the segment *"During this . . . sunflowers"* in the Wikipedia page *"Vincent van Gogh"* are symmetrically linked. Both sentences contain information corresponding to the same fact of *"Vincent van Gogh drew Sunflowers"*. We mine these symmetrically linked pairs as pseudo-positive instances to pre-train KEPT. Considering that queries and documents are typical of different lengths, we limit one segment to the sentence level while the other to the passage level during constructing these pairs.

For **cross-language positive pairs**, we leverage Wikipedia pages of the same entity in different languages. Specifically, we annotate (s_i, s_j) as a cross-language positive pair if segments s_i and s_j satisfy the following condition:

$$E(s_i) = E(s_j) \wedge L(s_i) \neq L(s_j), \quad (3)$$

where $E(s_i)$ denotes the entity identifier corresponding to segment s_i, and $L(s_i)$ denotes the language of segment s_i. In practice, we select the top three passages of the respective Wikipedia pages as the segments s_i and s_j. The top passages typically provide a comprehensive overview of the entity, encompassing important information and characteristics that are often similar across different languages. As shown in Figure 1(b), both segments within our constructed English-Chinese text pair contain the comparative summaries of the entity "Q5582".

In addition to the intra- and cross-language pairs based on knowledge, we also employ the widely-used *CROP* strategy (Wu et al., 2022; Izacard et al., 2022). This strategy constructs positive pairs by sampling two independent spans from a document chunk. It encourages the model to learn patterns of lexical co-occurrence between matched pairs.

3.4. Pre-training Objective

After obtaining pseudo-positive pairs, we train KEPT with contrastive loss to learn language-agnostic dense representations of queries and documents. Formally, given a paired query-document training sample (q, d^+), the contrastive loss is defined as:

$$L(q, d^+) = -log \frac{exp(f_\theta(q, d^+)/\tau)}{\sum\limits_{d' \in \mathbb{D}} exp(f_\theta(q, d')/\tau)}, \quad (4)$$

where τ is the temperature, and following Izacard et al. (2022), we set $\tau = 0.05$ during pre-training. The set $\mathbb{D}$ consists of the paired document d^+ and in-batch negative documents [¶].

4. Experiment

4.1. Evaluation Benchmarks

To verify the effectiveness of KEPT, we conduct experiments on three commonly-used cross-lingual retrieval benchmarks, including Mr. TyDi (Zhang et al., 2021b), XOR Retrieve (Asai et al., 2021a), Mewsli-X (Ruder et al., 2021). The detailed statistics of these datasets are shown in Table 1.

Mr. TyDi is a multilingual benchmark that comprises 11 diverse languages for mono-lingual retrieval, aiming to find relevant passages in the language corresponding to the given question. Consistent with previous works (Zhang et al., 2021b, 2022b), we employ MRR@100 and Recall@100 as the evaluation metrics.

XOR-Retrieve is a sub-task of cross-lingual open-domain QA benchmark XOR-QA (Asai et al., 2021a). Unlike Mr. TyDi, XOR-Retrieve specifically targets English passages as retrieval candidates, regardless of the question's language. As suggested by the original paper, we take R@2kt and R@5kt (kilo-tokens) as the metrics, where R@nt represents the proportion of top n retrieved tokens that contain the answers.

Mewsli-X consists of 15K queries in 11 languages. Given a query, it requires the model to retrieve the target passage from a candidate pool across 50 languages. This multilingual retrieval task emphasizes the model's capability to handle diverse languages. Following Ruder et al. (2021), we report mean average precision at 20 (mAP@20).

4.2. Evaluation Settings

Zero-shot Cross-lingual Transfer: In this setting, the model is fine-tuned only on the English training corpus and evaluated on the test set in

[¶] To ensure a fair comparison with baselines (Zhang et al., 2023; Izacard et al., 2022), we employ commonly-used In-batch negative sampling (Karpukhin et al., 2020) and refrain from using more effective techniques such as "Hard Negative Mining" (Xiong et al., 2020). These alternative negative sampling strategies are orthogonal to our KEPT.

	Query-passage Data Pairs				Details of Passage Pool		
	Train	Dev	Test	*#Lang of q*	*#Passage*	Type	*# Lang*
Mr. TyDi	48.7k	12.3k	8.6k	11	58M	Mono-lingual	11
XOR	15.2k	2.1k	-	7	18M	English-only	1
Mewsli-X	167.7k (*EN*)	14.1k	9.6k	11	1M	Language-agnostic	50

Table 1: Statistics of datasets. "*#Lang of q*" means the number of languages in questions during evaluation, "#Passage" means the number of passages, and "#Lang" means the number of languages in the Passage Pool. Note that there is no multilingual training data available in Mewsli-X, all training pairs are in English.

multiple languages. It is widely used in cross-lingual scenarios since most tasks only have labelled data in English and lack training data in low-resource languages. For Mr. TyDi, we follow the original paper (Zhang et al., 2021b), adopting Nature Questions (NQ) (Kwiatkowski et al., 2019) as the fine-tuning corpus. For XOR-Retrieve, we fine-tune all models on the NQ dataset following Asai et al. (2021a). For Mewsli-X, we adopt the setting in Ruder et al. (2021), fine-tuning models on training set consisting of English-only pairs.

Supervised Multilingual Fine-tuning: In this setting, the model is fine-tuned with multilingual training data. For Mr. TyDi and XOR-Retrieve, we further report the performance of models fine-tuned on the combined training data of all languages. Following previous works (Izacard et al., 2022; Asai et al., 2021a; Zhang et al., 2023), the models are firstly pre-fine-tuned with English-only training data and then fine-tuned on the multilingual train set. For Mewsli-X, since there is no multilingual training data, we skip this setting.

4.3. Implement Details

Pre-training: Considering the expensive cost of pre-training from scratch, we initialize KEPT with the pre-trained mContriever checkpoint (Izacard et al., 2022). To construct the multilingual knowledge base and mine knowledge-related text pairs, we collect the Wikipedia dump from August 1, 2022, and the Wikidata dump from September 26, 2022. For the *CROP* objective, we adopt the same text pair construction procedure as conducted by Izacard et al. (2022). KEPT is trained using a learning rate of 2e-5 and an Adam optimizer with a linear warm-up. The batch size of each task and temperature is set to 512 and 0.05, respectively. To ensure a fair comparison with Izacard et al. (2022), we limit the max length of input text to 128, which is consistent with mContriever. We perform pre-training on 8 NVIDIA Tesla A100 GPUs with 40GB memory for 300k steps. With automatic mixed precision, the process takes about 2 days.

Fine-tuning: For a fair comparison, we mainly follow the setting in previous works (Izacard et al., 2022; Zhang et al., 2023). For Mr. TyDi and XOR-Retrieve, When training on NQ, the batch size, learning rate, temperature, and max epoch number is set as 128, 2e-5, 1, and 40, respectively. When further fine-tuning with XOR-Retrieve's multilingual data, the learning rate is set as 1e-5 and others keep unchanged. Following Izacard et al. (2022), we set the temperature as 0.05, and perform hard negative mining when fine-tuning with MARCO and continue-fine-tuning with Mr. TyDi. For Mewsli-X, we follow Ruder et al. (2021), setting the learning rate, batch size, and epoch number as 2e-5,64 and 2, respectively. All fine-tuning experiments are implemented with the HuggingFace Transformers library (Wolf et al., 2019) on 8 NVIDIA Tesla V100 GPUs with 32 GB memory.

4.4. Experimental Results

In diverse cross-lingual retrieval settings, we compare KEPT with state-of-the-art methods: **BM25** (Robertson et al., 2009) is a traditional sparse retriever based on the exact term matching; **mBERT** (Devlin et al., 2019) and **XLM-R** (Conneau et al., 2020) are pre-trained with token-level multilingual MLM task; **InfoXLM** (Chi et al., 2021) adopts the momentum contrast and translation language modelling for pre-training with large-scale private parallel data pairs. **mContriever** (Izacard et al., 2022) and **CCP** (Wu et al., 2022) leverage MoCo (He et al., 2020) algorithm to pre-train dual-encoder where positive pairs are randomly cropping from a document. **LaBSE** (Feng et al., 2022) adopts a larger vocabulary and a translation ranking loss to increase model's cross-lingual transfer ability. **MSM** (Zhang et al., 2023) utilizes a masked sentence prediction task to pre-train a cross-lingual retriever.

Mr. TyDi: The type of retrieval corpus in Mr. TyDi is monolingual. Specifically, given a query from language L, models aim to retrieve relevant passages from a candidate pool in language L. It mainly evaluates the model's intra-language semantic matching and multilingual compatibility. For the zero-shot cross-lingual transfer setting, we show the results after fine-tuning with NQ dataset in Table 2. We can find that: (1) despite having the fewest parameters, KEPT consistently outperforms all baselines, including those utilizing expensive parallel corpora or larger models; (2) KEPT im-

Methods	#Params	Metrics	AR	BN	EN	FI	ID	JA	KO	RU	SW	TE	TH	AVG
BM25 (Zhang et al., 2021b)	*	MRR @100	36.7	41.3	15.1	28.8	38.2	21.7	28.1	32.9	39.6	42.4	41.7	33.3
		Recall@100	80.0	87.4	55.1	72.5	84.6	65.6	79.7	66.0	76.4	81.3	85.3	74.3
mBERT (Devlin et al., 2019)	178M	MRR @100	28.7	33.4	28.1	24.0	32.0	24.3	21.8	29.0	18.7	14.5	19.3	24.9
		Recall@100	68.3	79.3	76.1	65.0	74.1	65.3	57.9	69.2	51.7	46.1	54.0	64.3
XLM-R$_{base}$ (Conneau et al., 2020)	279M	MRR @100	30.8	30.2	27.1	23.5	33.5	23.5	26.6	25.7	24.0	26.6	36.3	28.0
		Recall@ 100	72.3	78.4	74.3	67.0	79.8	66.5	64.7	65.0	57.2	72.7	84.6	71.1
XLM-R$_{large}$ (Conneau et al., 2020)	560M	MRR @100	36.5	37.4	27.5	31.8	39.5	29.9	30.4	30.6	27.4	34.6	40.1	33.3
		Recall@100	81.3	84.2	77.6	78.2	88.6	78.5	72.7	77.4	63.3	87.5	88.2	79.8
InfoXLM$_{base}$ (Chi et al., 2021)	279M	MRR@100	31.5	31.2	27.7	22.4	31.3	27.1	27.2	28.3	30.5	46.3	37.3	31.0
		Recall@100	74.8	81.5	77.2	62.6	78.4	72.8	66.2	70.3	64.0	84.5	85.1	74.3
InfoXLM$_{large}$ (Chi et al., 2021)	560M	MRR @100	37.2	50.4	31.4	30.9	37.6	27.1	30.9	32.5	39.4	46.5	37.4	36.5
		Recall@100	76.2	91.0	78.3	76.0	85.2	66.9	64.4	74.4	75.0	88.9	83.4	78.2
mContriever (Izacard et al., 2022)	178M	MRR@100	44.9	50.8	28.2	32.7	41.1	34.2	35.0	31.2	42.4	28.3	49.4	38.0
		Recall@100	87.2	92.8	78.6	85.6	89.4	83.2	78.5	79.5	89.3	84.2	91.5	85.4
CCP (Wu et al., 2022)	560M	MRR @100	42.6	45.7	35.9	37.2	46.2	37.7	34.6	36.0	39.2	47.0	48.9	41.0
		Recall@100	82.0	88.3	80.1	78.7	87.5	80.0	73.2	77.2	75.1	88.8	88.9	81.8
LaBSE (Feng et al., 2022)	471M	MRR@100	37.2	50.4	31.4	30.9	37.6	27.1	30.9	32.5	39.4	46.5	37.4	36.5
		Recall@100	76.2	91.0	78.3	76.0	85.2	66.9	64.4	74.4	75.0	88.9	83.4	78.2
MSM$_{base}$ (Zhang et al., 2023)	279M	MRR @100	37.9	39.4	29.9	27.7	38.3	28.0	26.8	28.5	32.1	43.1	42.0	34.0
		Recall@100	77.3	82.9	73.6	70.5	83.5	67.9	61.8	68.6	69.9	83.5	84.9	74.9
KEPT	178M	MRR @100	50.7	53.7	30.7	35.7	48.0	37.1	39.2	35.2	42.5	58.9	48.1	**43.6**
		Recall@100	90.3	95.5	78.6	85.5	91.2	84.8	80.6	80.7	85.8	95.5	92.8	**87.4**

Table 2: Zero-shot cross-lingual transfer results on Mr. TyDi after fine-tuning with NQ dataset. #Params means the number of model parameters. The results of mContriever, InfoXLM$_{base}$ are from our implementation by fine-tuning the released checkpoints since they were not reported previously. Other results are from published papers (Zhang et al., 2023; Wu et al., 2022; Hu et al., 2022). We keep the same fine-tuning pipeline and hyper-parameters as Zhang et al. (2023) to ensure a fair comparison.

Methods	#Params	Metrics	AR	BN	EN	FI	ID	JA	KO	RU	SW	TE	TH	AVG
BM25 (Zhang et al., 2021b)	*	MRR @100	36.7	41.3	15.1	28.8	38.2	21.7	28.1	32.9	39.6	42.4	41.7	33.3
		Recall@100	80.0	87.4	55.1	72.5	84.6	65.6	79.7	66.0	76.4	81.3	85.3	74.3
mBERT (Devlin et al., 2019)	178M	MRR @100	66.8	63.3	50.8	55.0	55.7	47.8	43.3	47.3	60.5	85.8	57.7	57.6
		Recall@100	90.3	95.0	89.0	85.8	89.3	81.6	80.2	85.6	86.7	96.3	87.2	87.9
XLM-R$_{base}$ (Conneau et al., 2020)	279M	MRR @100	66.4	66.0	48.5	53.7	58.3	45.4	44.3	47.4	61.4	86.2	65.7	58.5
		Recall@ 100	89.4	93.2	84.3	86.6	90.0	82.3	79.3	82.8	86.4	97.4	93.8	87.8
InfoXLM$_{base}$ (Chi et al., 2021)	279M	MRR@100	67.6	68.3	50.8	53.7	57.7	48.9	48.4	50.0	63.0	87.1	66.5	60.2
		Recall@100	91.1	92.8	87.8	88.0	89.6	84.1	79.0	85.4	89.4	97.1	94.3	89.0
mContriever (Izacard et al., 2022)	178M	MRR@100	72.4	67.2	56.6	60.2	63.0	54.9	55.3	59.7	70.7	90.3	67.3	65.2
		Recall@100	94.0	98.6	92.2	92.7	94.5	88.8	88.9	92.4	93.7	98.9	95.2	93.6
MSM$_{base}$ (Zhang et al., 2023)	279M	MRR @100	67.7	69.9	49.6	55.1	61.2	48.2	49.4	47.5	63.9	85.2	67.7	60.5
		Recall@100	90.5	95.9	86.5	88.0	90.1	82.0	81.6	82.5	88.9	97.6	95.0	89.0
KEPT	178M	MRR@100	72.0	67.9	56.8	61.4	64.1	55.1	58.4	59.7	69.4	88.7	67.7	**65.6**
		Recall@100	95.0	95.5	93.5	94.3	96.1	90.3	90.9	93.9	95.4	99.3	97.0	**94.6**

Table 3: Supervised multilingual fine-tuning results on Mr. TyDi. #Params means the number of model parameters. Compared models' results are from previous papers (Izacard et al., 2022; Zhang et al., 2023).

proves over mContriever by **5.6** absolute points on the MRR@100 and **2.0** absolute points on the Recall@100, which demonstrates the effectiveness of our proposed knowledge-enhanced pretraining; (3) KEPT exhibits more improvements for most low-resource languages compared to English. For example, KEPT improves more than 5 points on MRR@100 on AR, ID, TE compared to mContriever. This clearly indicates KEPT's superior cross-lingual transfer ability, particularly in low-resource language settings. Table 3 shows the results under the supervised multilingual fine-tuning setting, We can find that the performance of all models can be further improved with the multilingual labeled data. Also, similarly to the results under the zero-shot setting, KEPT consistently outperforms all baselines. It further demonstrates the superior effectiveness of KEPT.

XOR-Retrieve: In the XOR-Retrieve task, the retrieval corpus is just English. It mainly evaluates the transfer ability of the model from English to other languages. We report the results under the zero-shot cross-lingual transfer setting in Table 4. One can be observed that KEPT achieves huge improvement over all baselines, e.g., compared to strong baseline mContriever, KEPT improves **9.6%** R@2k and **10.2%** R@5k. In Table 5, we can find that all models improve obviously with the multilingual training data, and KEPT achieves the best results among them. KEPT brings more improvement un-

Method	AR	BN	FI	JA	KO	RU	TE	AVG
				R@2k				
mBERT	31.1	26.6	38.5	32.4	38.6	24.9	29.1	31.6
XLM-R$_{base}$	39.9	25.7	41.1	27.8	31.9	22.4	25.2	30.6
InfoXLM$_{base}$	45.8	30.6	39.8	32.8	35.4	25.7	36.6	35.2
MSM$_{base}$	47.9	32.6	44.9	24.1	35.8	25.3	34.0	34.9
mContriever	50.8	28.3	50.0	31.1	38.6	28.3	37.9	37.9
KEPT	59.7	45.4	56.4	36.9	44.9	39.7	49.8	**47.5**
				R@5k				
mBERT	44.1	36.2	48.1	41.5	48.1	38.4	39.5	42.3
XLM-R$_{base}$	49.6	34.9	50.0	34.9	41.8	30.4	34.0	39.3
InfoXLM$_{base}$	55.0	42.4	46.2	42.7	47.7	33.8	47.2	45.0
MSM$_{base}$	58.4	41.8	51.9	36.9	44.6	37.6	42.1	44.7
mContriever	60.9	40.5	57.3	43.6	47.0	38.0	49.2	48.1
KEPT	71.8	59.5	63.4	46.5	55.4	51.1	60.5	**58.3**

Table 4: Zero-shot cross-lingual transfer results on XOR-Retrieve. All compared methods are base-sized unsupervised pre-trained models.

Method	AR	BN	FI	JA	KO	RU	TE	AVG
				R@2k				
mBERT	51.7	52.7	52.2	41.9	51.9	47.5	40.5	48.3
XLM-R$_{base}$	59.2	48.7	46.8	36.1	46.0	35.4	43.4	45.1
InfoXLM$_{base}$	58.4	52.3	51.9	39.8	51.9	41.4	48.9	49.2
MSM	61.8	55.3	51.3	36.9	50.5	41.4	43.0	48.6
mContriever	61.3	52.0	56.1	39.4	52.3	43.9	47.9	50.4
KEPT	64.7	60.5	59.2	47.3	54.7	50.6	52.8	**55.7**
				R@5k				
mBERT	58.4	61.5	58.6	50.6	63.2	54.9	51.8	57.0
XLM-R$_{base}$	67.2	58.6	53.5	45.6	56.1	45.1	51.5	53.9
InfoXLM$_{base}$	68.5	64.5	59.6	53.5	61.4	51.5	56.6	59.4
MSM	68.9	63.8	59.9	48.1	56.1	52.7	51.5	57.3
mContriever	72.7	66.4	62.7	50.6	61.1	53.2	57.0	60.5
KEPT	73.5	69.4	65.6	54.8	62.8	59.5	61.8	**63.9**

Table 5: Supervised multilingual fine-tuning results on XOR-Retrieve. All compared methods are base-sized unsupervised pre-trained models.

der the zero-shot cross-lingual transfer setting. For low-resource languages, there is usually a lack of available training data. In this case, KEPT can play a greater role in its strong cross-lingual transfer ability.

Mewsli-X: In the Mewsli-X task, the candidate corpus consists of passages in fifty languages. It evaluates the semantic understanding ability of the retriever over multi-languages, which is more challenging. Table 6 shows the results under the zero-shot cross-lingual transfer setting on Mewsli-X. The strong baselines mContriever and MSM obtain very limited gains compared to token-level mPLMs mBERT and XLM-R, showing the difficulty of this task. However, KEPT consistently improves by 5.7 absolute points compared to the second-best model, further demonstrating its superior capability. Furthermore, we observe a positive correlation between the number of cross-lingual knowledge pairs and the improvement in the target language compared to mContriever. languages such as AR, DE, ES, and PL, which have more cross-lingual knowledge pairs, show greater improvements compared to languages like JA and TR, which have fewer such pairs. It is interesting to note that some languages (e.g., FA, TA) that aren't covered by KEPT pre-training also see improvements with KEPT. That further indicates our knowledge-enhanced pre-training can encourage the model to learn universal representations.

4.5. Ablation Study

We conduct the ablation study on the Mr. TyDi and XOR-Retrieve datasets, under the zero-shot cross-lingual transfer setting. All models are fine-tuned three times on NQ using different random seeds, and the final scores are averaged. To analyze the impact of a longer pre-training step, we first continue training mContriever with the same step as KEPT. Then, we remove the pre-training tasks used in KEPT one by one to analyze their contribution.

Based on the results in Table 7, we have the following findings. (1) Continuing pre-training brings a slight gain, which is much smaller compared to KEPT. It indicates that the primary improvement in KEPT is not from longer training steps. (2) Removing any pre-training task leads to a performance decrease, suggesting that all pre-training tasks are beneficial for improving cross-lingual retrieval capabilities. (3) ILK has a more significant effect on Mr. TyDi which emphasizes the model's abilities in intra-language semantic matching and multilingual compatibility. On the other hand, CLK has a greater impact on XOR-Retrieval where the model's cross-lingual semantic alignment ability is crucial due to the queries and matched documents being in different languages. This observation aligns with the characteristics of the respective pre-training tasks. It suggests both knowledge-based tasks contribute to the model's multi-lingual retrieval ability but with different emphases. (4) Removing the *CROP* task shows the slightest performance decrease, indicating that *CROP* is not sufficient for cross-lingual retrieval tasks. Our proposed knowledge-based tasks ILK and CLK are the key to promoting the model's cross-lingual retrieval ability.

4.6. Uniformity and Alignment

To study the effect of KEPT on the language-agnostic representation space, we compute alignment and uniformity metrics (Wang and Isola, 2020) using the evaluation set of XOR-Retrieve. As shown in Table 8, the alignment loss and the uniformity loss of KEPT descend significantly which indicates KEPT can help align semantic representations of similar texts across languages and alleviate the embedding space anisotropy. It also demonstrates that our knowledge-based positive pairs can reduce the gap between pre-training and fine-tuning, compared with previous mPLMs.

Method	#Params	AR	DE	EN	ES	FA	JA	PL	RO	TA	TR	UK	AVG
mBERT (Devlin et al., 2019)	178M	14.2	65.5	57.4	55.7	11.3	44.8	57.2	38.2	5.8	42.7	36.2	39.0
XLM-R$_{base}$ (Conneau et al., 2020)	279M	15.5	62.7	57.2	53.5	12.5	46.2	59.3	34.1	7.5	51.8	36.0	39.7
InfoXLM$_{base}$ (Chi et al., 2021)	279M	21.3	65.6	58.9	55.2	14.9	45.5	61.1	35.7	8.4	52.8	41.4	41.9
mContriever (Izacard et al., 2022)	178M	20.9	65.2	57.3	49.4	11.6	41.7	59.0	32.6	12.1	46.3	34.8	39.2
MSM (Zhang et al., 2023)	279M	18.6	68.0	58.7	57.3	13.7	46.3	60.2	36.3	7.6	52.8	37.3	41.5
KEPT	178M	36.4	70.4	64.6	63.2	21.1	52.1	64.7	36.2	14.6	59.6	40.5	**47.6**

Table 6: Zero-shot cross-lingual transfer results on Mewsli-X. #Params means the number of model parameters.

Method	Mr. TyDi		XOR	
	MRR@100	Recall@100	R@2K	R@5K
mContriever	38.0	85.4	37.9	48.1
+ continue pre-train	38.3	85.5	38.1	48.5
KEPT	43.6	87.4	47.5	58.3
w/o *CROP*	42.9(-0.7)	87.2(-0.2)	46.9(-0.6)	57.6(-0.7)
w/o ILK	39.2(-4.4)	85.8(-1.6)	45.8(-1.7)	56.1(-2.2)
w/o CLK	41.8(-1.8)	86.5(-0.9)	41.1(-6.4)	51.3(-7.0)

Table 7: Ablation study, "ILK" means positive pairs built based on in-language knowledge, "CLK" means positive pairs built based on cross-language knowledge.

ID	Strategy	Mr. TyDi		XOR	
		MRR@100	Recall@100	R@2K	R@5K
	mContriever	38.0	85.4	37.9	48.1
1	KEPT(Defaulted)	43.6	87.4	47.5	58.3
2	*UHS-ILK*	39.5	85.9	43.5	54.8
3	*DS-CLK*	40.6	86.1	41.6	51.8
4	*CLHED*	42.5	87.1	47.0	58.2

Table 9: Various strategies to construct pre-training pairs based on the multilingual knowledge base.

Model	L_{align}	$L_{uniform}$
mBert (Devlin et al., 2019)	0.999	-1.668
mContriever (Izacard et al., 2022)	0.991	-1.788
KEPT	0.668	-1.869

Table 8: Alignment and uniformity analysis.

4.7. Variants of Data Construction

Besides the defaulted way of constructing knowledge-based pairs, as described in Section 3.3, we also explore alternative strategies, including:

(1) Unidirectionally Hyperlinked Segments for Intra-Language Knowledge (UHS-ILK): In our default setting, the pairs for intra-language knowledge are consisting of symmetrically hyperlinked segments. In this setting, we remove the constraint of symmetry. We treat the sentence containing the anchor text as a query and select the top passages from the hyperlinked Wikipedia page as the positive document. We replace the original intra-language knowledge positive pairs with this new data while keeping the other settings unchanged.

(2) Diverse Sampling for Cross-Language Knowledge (DS-CLK): In our default setting, the segments in cross-language positive pairs are selected from the top of Wikipedia pages. In this setting, we sample these segments from the entire Wikipedia pages, thereby introducing greater diversity to the cross-lingual knowledge pairs. This setting is consistent with C3 (Yang et al., 2022) so that we can make a fair comparison. We replace the original cross-language knowledge positive pairs with this new data while keeping the other settings unchanged.

(3) Cross-lingual Hyperlinked Entity Description (CLHED): In this setting, we consider the sentence containing the anchor text as a query and leverage the corresponding entity description from a different language Wikipedia page as the positive document. The entity description is selected from the top passages in the respective Wikipedia pages. We incorporate this task into the existing setting while preserving the other settings unchanged.

The results of these strategies are reported in Table 9. We observe that these strategies outperform mContriever, demonstrating incorporating knowledge-related pre-training tasks can benefit the cross-lingual retrieval task. However, these strategies perform worse than our default strategy. Among them, Strategy 2, despite mining a greater number of data pairs, experiences a performance decline. We analyze it could be attributed to the lower semantic relevance of the text pairs. It indicates that symmetric hyperlinked segments have more semantic similarity and are more effective to cross-lingual retrieval tasks. The cross-lingual data pairs constructed by Strategy 3 show better diversity but poorer alignment quality with entities, resulting in a significant performance decrease in XOR-Retrieval. Strategy 4, which combines the additional task with our default tasks, shows a slight performance decrease. We speculate that the text pairs mined in this strategy are not factually consistent, which is quite different from cross-language retrieval tasks. As a result, a slight decrement in performance is observed. Overall, we have explored multiple approaches to utilize Wiki corpus. Our proposed knowledge-based pre-training tasks perform best and significantly improve the model's cross-lingual retrieval ability.

5. Conclusion

In this paper, we introduce KEPT, a mPLM that leverages intra- and cross-language knowledge to enhance cross-lingual dense retrieval. We first construct a multilingual knowledge base using hyperlinks and cross-language page alignment data annotated by Wiki, and then build intra- and cross-language pairs by extracting symmetrically linked segments and multilingual entity descriptions from the knowledge base. Finally, we adopt contrastive learning with mined pairs to pre-train KEPT. Extensive experiments show that KEPT achieves strong performance with significant improvement over existing mPLMs.

Acknowledgement

This work is supported by the Fundamental Research Funds for the Central Universities under Grant 1082204112364.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388.

Akari Asai, Jungo Kasai, Jonathan H 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.

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.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Wei-Cheng Chang, X Yu Felix, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2019. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

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*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

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)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Xingwei He, Yeyun Gong, A Jin, Hang Zhang, Anlei Dong, Jian Jiao, Siu Ming Yiu, Nan Duan, et al. 2022. Curriculum sampling for dense retrieval with document expansion. *arXiv preprint arXiv:2212.09114*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson.

2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*.

Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang, and Zhiheng Huang. 2022. Language agnostic multilingual information retrieval with contrastive learning. *CoRR*, abs/2210.06633.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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)*, pages 2485–2494.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035. International Committee on Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with bert. *arXiv preprint arXiv:2004.13005*.

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.

San Kim, Jin Yea Jang, Minyoung Jung, and Saim Shin. 2021. A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 352–365.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

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*.

Zehan Li, Yanzhao Zhang, Dingkun Long, and Pengjun Xie. 2023. Challenging decoder helps in masked auto-encoder pre-training for dense passage retrieval. *arXiv preprint arXiv:2305.13197*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Zhenghao Lin, Yeyun Gong, Xiao Liu, Hang Zhang, Chen Lin, Anlei Dong, Jian Jiao, Jingwen Lu, Daxin Jiang, Rangan Majumder, et al. 2022. Prod: Progressive distillation for dense retrieval. *arXiv preprint arXiv:2209.13335*.

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

Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq R. Joty, and Luo Si. 2022. Enhancing multilingual language model with massive multilingual knowledge triples. In *EMNLP*.

Zheng Liu and Yingxia Shao. 2022. Retromae: Pre-training retrieval-oriented transformers via masked auto-encoder. *arXiv preprint arXiv:2205.12035*.

Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is more: Pretrain a strong siamese encoder for dense text retrieval using a weak decoder. In *Empirical Methods in Natural Language Processing*.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*.

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38.

Carol Peters, Martin Braschler, and Paul Clough. 2012. *Multilingual information retrieval: From research to practice*. Springer.

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)*, pages 4512–4525.

Houxing Ren, Linjun Shou, Ning Wu, Ming Gong, and Daxin Jiang. 2023. Empowering dual-encoder with query generator for cross-lingual dense retrieval. *arXiv preprint arXiv:2303.14991*.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mluke: The power of entity representations in multilingual pretrained language models. In *ACL*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. *arXiv preprint arXiv:2307.07697*.

Jiashuo Sun, Hang Zhang, Chen Lin, Yeyun Gong, Jian Guo, and Nan Duan. 2022. Apollo: An optimized training approach for long-form numerical reasoning. *arXiv preprint arXiv:2212.07249*.

Shuo Sun and Kevin Duh. 2020. Clirmatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. *Advances in Neural Information Processing Systems*, 33:2207–2219.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Ivan Vulic and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 106–116. ACL; East Stroudsburg, PA.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Simlm: Pre-training with representation bottleneck for dense passage retrieval. *arXiv preprint arXiv:2207.02578*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Ning Wu, Yaobo Liang, Houxing Ren, Linjun Shou, Nan Duan, Ming Gong, and Daxin Jiang. 2022.

Unsupervised context aware sentence representation pretraining for multi-lingual dense retrieval. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4411–4417.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Weiwen Xu, Xin Li, Wai Lam, and Lidong Bing. 2023. mpmr: A multilingual pre-trained machine reader at scale. *arXiv preprint arXiv:2305.13645*.

Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W Oard. 2022. C3: Continued pretraining with contrastive weak supervision for cross language ad-hoc retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2507–2512.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.

Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021a. Poolingformer: Long document modeling with pooling attention. In *International Conference on Machine Learning*, pages 12437–12446. PMLR.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2022a. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*.

Shunyu Zhang, Yaobo Liang, MING GONG, Daxin Jiang, and Nan Duan. 2023. Modeling sequential sentence relation to improve cross-lingual dense retrieval. In *The Eleventh International Conference on Learning Representations*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021b. Mr. TyDi: A multi-lingual benchmark for dense retrieval. *arXiv:2108.08787*.

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

Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Lan Luo, Ke Zhan, Enrui Hu, Xinyu Zhang, Hao Jiang, Zhao Cao, Fan Yu, et al. 2022. Hyperlink-induced pre-training for passage retrieval in open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7135–7146.

Shengyao Zhuang, Linjun Shou, and Guido Zuccon. 2023. Augmenting passage representations with query generation for enhanced cross-lingual dense retrieval. *arXiv preprint arXiv:2305.03950*.