Negative Sampling Techniques for Dense Passage Retrieval in a Multilingual Setting

Thilina C. Rajapakse
University of Amsterdam
The Netherlands
t.c.r.rajapakse@uva.nl

Andrew Yates
University of Amsterdam
The Netherlands
a.c.yates@uva.nl

Maarten de Rijke
University of Amsterdam
The Netherlands
m.derijke@uva.nl

ABSTRACT

The bi-encoder transformer architecture has become popular in open-domain retrieval, surpassing traditional sparse retrieval methods. Using hard negatives during training can improve the effectiveness of dense retrievers, and various techniques have been proposed to generate these hard negatives. We investigate the effectiveness of multiple negative sampling methods based on lexical methods (BM25), clustering, and periodically updated dense indices. We examine techniques that were introduced for finding hard negatives in a monolingual setting and reproduce them in a multilingual setting. We discover a gap amongst these techniques that we fill by proposing a novel clustered training method. Specifically, we focus on monolingual retrieval using multilingual dense retrievers across a broad set of diverse languages. We find that negative sampling based on BM25 negatives is surprisingly effective in an in-distribution setting, but this finding does not generalize to out-of-distribution and zero-shot settings, where the newly proposed method achieves the best results. We conclude with recommendations on which negative sampling methods may be the most effective given different multilingual retrieval scenarios.

CCS CONCEPTS

• **Information systems** → *Retrieval models and ranking*; Document representation.

KEYWORDS

Dense retrieval, Generalizability

ACM Reference Format:
Thilina C. Rajapakse, Andrew Yates, and Maarten de Rijke. 2024. Negative Sampling Techniques for Dense Passage Retrieval in a Multilingual Setting. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3657854

1 INTRODUCTION

Dense retrieval architectures consisting of two transformer models (bi-encoders) have become the state-of-the-art architecture for passage retrieval [6, 8, 9, 25]. The dense passage retriever (DPR) model consists of two transformer models that encode the queries and passages separately. Bi-encoder architectures for dense retrieval are typically employed to pre-compute passage representations at indexing time, on top of which computationally more costly re-rankers such as the cross-encoder architecture [4] can be run.

Hard negative mining or negative sampling techniques have been used in prior work to improve the effectiveness of bi-encoder models [6, 23, 25]. However, recent work has demonstrated that reported results can vary significantly based on multiple factors that can be easily overlooked. For example, Lassance and Clinchant [10] show that some previous work uses the titles from the MS MARCO dataset leading to unfair comparisons with methods that do not use the titles. In light of this, we implement all the methods compared in this work from scratch, using the same libraries and library versions, and evaluate the methods using the same evaluation framework to provide a fair comparison. All code is publicly available on Github.

Effectiveness of negative sampling strategies on MS MARCO (English). The choice of negative sampling strategy has a significant effect on the effectiveness of the final retrieval model. Based on the work of Xiong et al. [25], Hofstätter et al. [6], and Wang and Zuccon [23], we find that clustering-based negative sampling methods and iterative negative sampling methods offer comparable performance while outperforming lexical negative sampling methods (details on the different methods can be found in Section 3.3). Our goal in this work is to extend the analysis of negative sampling methods to multilingual retrieval and determine which negative sampling strategy is best-suited for this understudied setting.

Monolingual retrieval beyond English. Information access in languages other than English is a topic with a long history in information retrieval, with resources, benchmarking activities and algorithm development going back decades; see, e.g., [15] for an early survey. In contrast, research on DPR has mainly been focused on English [8, 9, 25], even though some work has been done on monolingual DPR for other languages, such as Arabic, Japanese, and Russian [31]. These models have been trained on monolingual corpora and have achieved high performance on monolingual retrieval tasks.

Multilingual DPR for monolingual retrieval. Using a *multilingual* DPR model for *monolingual* purposes has some clear advantages. It allows for leveraging cross-lingual information transfer, which can improve performance on low-resource languages [31]. Furthermore, a multilingual model can perform zero-shot retrieval on languages for which it has not been explicitly trained. For example, Zhang et al. [31] show that a multilingual dense retrieval model can be used on a new language in a zero-shot manner with some success, enabling retrieval even in languages for which no retrieval training data is available. Other work [e.g., 1, 2, 11, 13, 19, 30]

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657854

supports this finding. Using a single *multilingual* model for *monolingual* retrieval for many different languages is more cost-effective and scalable than training a separate *monolingual* model for each language of interest. Thus, the *zero-shot* setting is of particular interest in this work. We explore the capabilities of *multilingual* bi-encoders in a *monolingual* setting. We train our models to perform retrieval for multiple languages (one model for many languages, i.e., *multilingual*), and we test them on *monolingual* datasets (queries and passages in the same language, i.e., *monolingual*).

Generalizability. We address the generalizability of dense retrieval models to new data and new languages. This is of particular interest because dense retrieval models are known to struggle with out-of-distribution data [20, 30], often falling behind traditional sparse methods when tested in zero-shot settings. Given this drawback, we consider the retrieval performance across three settings: (i) *in-distribution*, (ii) *out-of-distribution*, and (iii) *zero-shot*. The in-distribution setting gives us an idea of how well a model learns the distribution of data similar to the training data. Out-of-distribution testing demonstrates how we may expect a model to perform when exposed to new types of queries and passages. The zero-shot performance of the models is of particular interest as this represents the real-world use-case of using a multilingual retrieval model on a language that it is not trained on as no training data was available for that language.

Negative sampling. Negative sampling is the process of selecting negative examples (passages that are not relevant to a given query) for training a dense retrieval model. Negative examples are used to train the model to differentiate between relevant and non-relevant passages. Negative sampling that includes *hard* negatives (passages that are similar to the query but are not relevant) is crucial for the effectiveness of dense retrieval models [6, 25]. Given the importance of negative sampling, we study the effectiveness of negative sampling methods in a multilingual setting.

Simple methods to select negative examples for training dense retrieval models include random selection from the corpus (DPR_{base}) or from BM25's top-ranked documents (DPR_{BM}). However, these approaches do not ensure that the negative examples are hard negatives, which has motivated other work.

Hofstätter et al. [6] cluster queries and select queries from the same cluster for a given batch to increase the probability of in-batch negatives being hard negatives (TAS-Q). Similarly, passages can be clustered, and training batches can be built from the same cluster of passages (TAS-P). Xiong et al. [25] iteratively update a dense index of the full collection by periodically re-computing representations of all passages and select passages that are ranked highly (but not at the top) for each query as negative examples (ANCE).

As part of our reproducibility work, we identify a gap left by these methods and consider a combination of these two approaches that combines clustered training with iterative updates produced using a subset of the collection, which we refer to as *iterative clustered training* (ICT). Unlike the work in [6], this method uses the representations from the model being trained to perform clustering instead of a separate teacher model. The passages are clustered at the start of every training epoch to ensure that the training objective remains challenging even as the model learns to differentiate between similar passages better (ICT-P). Similarly, this method can also be applied to query representations (ICT-Q). This method is complementary to existing methods and combines insights from the methods proposed by Hofstätter et al. [6] and Xiong et al. [25]. Sections 3.2 and 3.3 provide detailed descriptions of these negative sampling methods.

Main findings. We find that the use of negative sampling methods yields significant improvements in a multilingual retrieval setting, reproducing the lessons from prior work in English. The ICT methods perform the best overall, showing the best results in both out-of-distribution and zero-shot conditions, while achieving the second-highest scores under the in-distribution condition. ICT-P performs best out of the two ICT methods. DPR_{BM} shows the best results under the in-distribution conditions.

Furthermore, we see that TAS style clustering is less effective in a multilingual setting than the other methods. This contradicts the lessons learned from English only retrieval, where TAS is competitive with other negative sampling methods (such as ANCE). Thus, we find that ANCE generalizes better to our new multilingual setting than TAS.

Our results demonstrate that the clustered training method we propose leads to the best overall retrieval quality in a multilingual retrieval setting. Finally, we provide recommendations on which negative sampling method should be used in different scenarios.

2 TASK DEFINITION

Our task is to use *multilingual* (the same model used with multiple languages) DPR models to perform *monolingual* (queries and passages in the same language) dense retrieval. We study the effectiveness of existing negative sampling techniques as well as our proposed technique, clustered training, for this task. Extending beyond the findings of prior work on English language (monolingual) retrieval with English models, we explore *monolingual* retrieval with *multilingual* models and test whether these findings can be reproduced in this new setting. We investigate the effectiveness of each negative sampling technique under three conditions: (i) in-distribution, (ii) out-of-distribution, and (iii) zero-shot.

- The *in-distribution condition* uses test data from the same datasets used to train the models. The in-distribution test datasets consists of languages the models have been trained on for retrieval. As the training and test data were all gathered using the same methods at the same time, we consider this to be the in-distribution setting.
- The *out-of-distribution condition* uses test data from datasets that are different from the models' training datasets. The test sets employed under this condition solely consist of languages the models have been trained on for retrieval. As these test datasets were built using different methods, at different times, and by different contributors compared to the training data for the models, we call this the out-of-distribution setting. This setting is out-of-distribution with respect to the testing datasets.
- Similar to the out-of-distribution setting the *zero-shot testing condition* uses test datasets that were built using different methods, at different times, and by different contributors compared to the training data for the models. However, the test sets under this condition consist solely of languages the models have not been trained on for retrieval. This setting is out-of-distribution with

respect to both the test datasets *and* the languages being tested. Hence, this is our zero-shot test setting.

3 NEGATIVE SAMPLING FOR DENSE RETRIEVAL

We recall DPR and negative sampling techniques that have been considered for DPR. We discover a natural but "missing" approach for negative sampling, which we then describe in detail.

3.1 Dense passage retriever (DPR)

Our work uses the DPR model [8]; one of the first effective dense retrieval models. The DPR model consists of two BERT encoders, a passage encoder $E_p(\cdot)$ and a query encoder $E_q(\cdot)$, used to encode passages and queries separately. The passage encoder $E_p(\cdot)$ is used to encode all passages into d-dimensional vectors, and a dense retrieval index is built with FAISS [7] for all M passages [8].

During retrieval, the query encoder $E_q(\cdot)$ is used to encode a query to a d-dimensional vector and a desired number of passages are retrieved from the index where the passage vectors are most similar to the query vector. The similarity is simply defined as the dot product of two vectors [8]:

$$\text{sim}(q, p) = E_Q(q)^\intercal E_p(p). \tag{1}$$

The training goal is to learn encoders $E_p(\cdot)$ and $E_q(\cdot)$ such that the encoded representations for relevant queries and passage pairs have higher similarity relative to irrelevant query and passage pairs. Consider $D = \{\langle q_i, p_i^+, p_{i,1}^-, \ldots, p_{i,n}^-\rangle\}_{i=1}^m$, where D is a training batch consisting of m instances. Each such instance contains a question q_i and a relevant passage p_i^+, as well as n irrelevant passages $p_{i,j}^-$ [8].

In our work, we use the in-batch negative [8] strategy when training the models. Therefore, the irrelevant passages are the relevant passages for the other queries in the batch.

The loss function is optimized as the negative log likelihood of the relevant passage:

$$L(q_i, p_i^+, p_{i,1}^-, \ldots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_{(i,j)}^-)}}. \tag{2}$$

The original DPR model was initialized from BERT [4] (English). However, the models in this work are initialized with Multilingual BERT (mBERT), following [31], to facilitate multilingual retrieval.

3.2 Key characteristics of negative sampling methods

We observe three key dimensions of negative sampling techniques for dense retrieval:

Iterative or non-iterative: whether the negative samples are updated periodically during training;

Negative mining model: the model used to find the hard negatives; and

Hard negative source: what is the source of the hard negatives, i.e., whether the hard negatives are sampled from the corpus or from the training passages or queries (the size of the corpus is much bigger than the number of training queries/passages).

Table 1 summarizes the negative sampling methods we have listed so far and the design decisions made for the three dimensions listed above. The top part of the table shows the decisions of the existing negative sampling techniques. We observe that there is a gap in the existing methods: they do not consider the use of clustering self-generated (the model being trained) representations periodically to generate hard negatives. The bottom part of the table characterizes these gaps, which we describe in detail in Section 3.3.

Table 1: Overview of negative sampling methods used for dense retrieval DPR and their features. Top: previously published. Bottom: newly proposed.

Model	Source	Negative mining model	Hard negative source	Iterative updates
Base	[8]	N/A	N/A	N/A
BM25	[8]	BM25	Full corpus	No
TAS-Q	[6]	Teacher model	Training queries	No
TAS-P	[6]	Teacher model	Training passages	No
ANCE	[25]	Self	Full corpus	Yes
ICT-Q	This paper	Self	Training queries	Yes
ICT-P	This paper	Self	Training passages	Yes

3.3 Current negative sampling techniques

Random negatives (DPR_{base}). A dense retrieval model can easily be trained with random negatives in an inbatch-negative contrastive loss training scheme [8]. Here, a single training sample s consists of a query q, out of the full set of queries Q, and its relevant passage p_q. Then, a single training batch B of batch size b out of the full set of n training batches D is built as follows:

$$D = \sum_{i=1}^{n} B_i \tag{3}$$

$$B = \{(q, p_q) \mid q \in random(Q, b)\}. \tag{4}$$

Here, B_i is the i-th batch of the full set D and $random(Q, b)$ are b queries randomly sampled from Q without replacement. Then the dense retriever can be trained as described in Section 3.1.

BM25 negatives (DPR_{BM}). Negatives can be sampled from the corpus using the BM25 algorithm [16] by sampling passages from the top k retrieved passages. The BM25 algorithm is a bag-of-words retrieval function that ranks a set of documents (or passages) based on the query terms appearing in each document. We refer the reader to [8, Section 3.2] for a detailed description of using BM25 to sample negatives for training dense retrievers.

Topic aware sampling (TAS-Q and TAS-P). Topic aware sampling (TAS) [6] is a technique that aims to improve the effectiveness of dense retrievers by building training batches where all in-batch negatives are hard negatives for any given query. TAS achieves this by clustering queries once at the beginning of training and then sampling from those clusters to build training batches (TAS-Q). TAS style negative sampling requires a teacher model (any model trained to generate representations) to generate the initial representations used for clustering. For completeness, we also study the effect of sampling from passage clusters (TAS-P) in addition to query clusters. See [6, Section 3] for a detailed description of the TAS algorithm. Note that, the original TAS method also used knowledge distillation in a dual-teacher setup (see [6, Section 2.3]),

but we only consider the negative sampling strategy proposed in the same work.

ANCE. Approximate nearest neighbor negative contrastive estimation (ANCE) [25] is a technique that aims to improve the effectiveness of dense retrievers by using hard negatives during training. ANCE accomplishes this by periodically identifying false positive examples using the retrieval model currently being trained. As the hard negatives are periodically updated, we refer to this as an *iterative method*. The false positive examples are then used as hard negatives for the next training epoch. ANCE requires maintaining a continuously updated dense index, which requires significant compute resources. Further details on the ANCE algorithm can be found in [25, Section 4].

Iterative clustered training (ICT-Q and ICT-P). Considering the established need to ensure the presence of hard negatives [25] when training DPR models, we combine intuitions from ANCE and TAS to use clustering to place similar training samples in each training batch. However, unlike with TAS style negative sampling, text representations are generated by the model itself, thus eliminating the need for a teacher model. The representations used in clustering can be either passage or query representations. Similar to ANCE, we also iteratively update the representations used to perform clustering. But, clustered training methods are more efficient than ANCE since they only cluster the training queries or passages (unlike ANCE where a full index of the corpus is built to update the hard negatives). In a typical information retrieval setup, the number of training queries or passages is much smaller than the total number of documents in the corpus.

We provide a formal description of the clustered training method below. Note that we provide the method for clustering passages ICT-P, but the process for clustering queries remains the same with queries ICT-Q replacing passages in the method.

Before each training epoch, we group all training samples S into k clusters with k-means clustering based on the passage representations generated by the passage encoder $E_p(\cdot)$. The objective of the clustering is to minimize the following:

$$\arg_C \min \sum_{i=1}^{k} \sum_{p \in C_i} \|p - \mu_i\|^2. \tag{5}$$

Here, μ_i is the centroid of the cluster C_i and p is a passage representation generated by $E_p(\cdot)$. Now, the training samples S are grouped into k clusters C_i where $i \in \{1, \ldots, k\}$.

Next, we split each cluster C containing $|C|$ samples, where $|C| > b$ into sub-clusters c_j such that $|c_j| \leq b$. For a cluster C_i:

$$C_i = \left\{ c \in \{1, \ldots, j\} \mid j = \left\lceil \frac{|C|}{b} \right\rceil \right\}. \tag{6}$$

Then, $|c_j| \leq b$. Finally, we combine all sub-clusters containing less than b samples such that each combined cluster contains b or fewer samples until no further combinations are possible. The set of all sub-clusters of size b, all combined sub-clusters, and any sub-clusters that could not be combined becomes the set of training batches for a training epoch.

Then, the training dataset consisting of the set of training batches, built according to the above procedure, is used to train a dense retrieval model. The clustering representations are refreshed periodically during training. We refer to this method as *iterative clustered training (ICT)*, and it comes in two flavors: ICT-Q for clustered training on queries and ICT-P for clustered training on passages.

3.4 Summary of negative sampling techniques

We first looked at the in-batch negative sampling technique used in the original DPR paper [8]. Then, we summarized the BM25 negative sampling technique [8]. Next, we described the topic-aware sampling technique [6] and, finally, the ANCE technique [25]. We also introduced iterative clustered training, which combines ideas from [6] and [25] and fills a gap left by previously proposed methods. Our next step is to perform a systematic comparison of these negative sampling methods for dense retrieval under three conditions: *in-distribution*, *out-of-distribution*, and *zero-shot*.

4 EXPERIMENTAL DESIGN

We now describe the training and evaluation processes, the datasets used at query and passage level, and the models that we used in the systematic comparison of negative sampling methods for dense retrieval promised at the end of the previous section,

4.1 Process

We describe the process of training and evaluating the models below.

Training. The DPR models discussed in this work are trained in two steps. First, the model is pre-finetuned on English and then finetuned on the combined training sets of all available languages. We follow this procedure to train models using each of the negative sampling techniques discussed in Section 3.

Evaluation. Each model is evaluated in the three settings described in Section 2: in-distribution, out-of-distribution, and zero-shot.

The specific implementation details for each of these processes are discussed in the remainder of this section.

4.2 Datasets

Training datasets. All models evaluated in this work are trained on the same datasets. The MS MARCO (MAchine Reading COmprehension) dataset (English) [14] is used for pre-finetuning, followed by fine-tuning on the Mr. TyDi collection of datasets [30].

The Mr. TyDi [30] dataset is a multilingual retrieval benchmark based on the TyDi dataset [3]. Mr. TyDi contains data from eleven typologically diverse languages, some of which are written in Latin script, while the others are written in other scripts (with no two languages sharing the same non-Latin script) [30]. Table 2 shows the languages and the number of associated queries and passages for each language.

Testing datasets. Three collections of datasets/benchmarks are used for testing the models in three conditions: *in-distribution*, *out-of-distribution*, and *zero-shot*. For the in-distribution setting, we use the test sets from the Mr. TyDi dataset, as all models were trained on the Mr. TyDi train sets.

Table 2: Mr. TyDi languages and the associated number of queries and passages.

Language	# Train queries	# Test queries	# Corpus size
Arabic	12,377	1,081	2,106,586
Bengali	1,713	111	304,059
English	3,547	744	32,907,100
Finnish	6,561	1,254	1,908,757
Indonesian	4,902	829	1,469,399
Japanese	3,697	720	7,000,027
Korean	1,295	421	1,496,126
Russian	5,366	995	9,597,504
Swahili	2,072	670	136,689
Telugu	3,880	646	548,224
Thai	3,319	1,190	568,855

Table 3: The datasets used at each stage of the study.

Stage	Condition	Dataset
Training	pFT (pre-finetuning)	MS MARCO
	FT (finetuning)	Mr. TyDi
Testing	In-distribution	Mr. TyDi
	Out-of-distribution	mMARCO (known languages)
	Zero-shot	mMARCO (unknown languages) BSARD (French) GerDaLIR (German) Multi-CPR E-com (Chinese) Multi-CPR video (Chinese)

The mMARCO [2] dataset consists of 13 different languages created using machine translation from the MS MARCO dataset. Four of these languages (Arabic, Indonesian, Japanese, Russian) are common to both mMARCO and Mr. TyDi. These four languages are used to evaluate the models in the out-of-distribution setting as they represent languages the models are trained on but created using different methods.

The remaining nine languages from mMARCO (Chinese, Dutch, French, German, Hindi, Italian, Portuguese, Spanish, Vietnamese) are not found in Mr. TyDi, and thus, the models have not been trained on these languages for retrieval. Since mMARCO consists of machine translated datasets, we include human annotated datasets in our test datasets for the languages where we were able to find a retrieval dataset. These datasets are Multi-CPR (E-commerce and entertainment), BSARD (Legal IR), and GerDaLIR (Legal IR) for Chinese, French, and German respectively. In addition to being unknown languages, these datasets are out-of-domain in terms of data distribution as the retrieved documents are from different domains. Therefore, these languages are used to evaluate the models in the zero-shot setting.

The datasets that are used in this work, and their purpose, are summarized in Table 3.

Analysis datasets. Finally, we also use two additional datasets for further analysis (see Section 6) beyond our main results. We use the unknown languages from MIRACL [32], an updated version of the Mr. TyDi dataset, to form an in-distribution, unknown language setting. Then, we use the nine smallest datasets (for faster evaluation) from BEIR [20], designed to evaluate out-of-domain performance of retrieval models, to form an out-of-domain, known language setting.

4.3 Models

We train DPR models with different negative sampling methods using the Simple Transformers[1] framework, which is based on Huggingface Transformers [24]. All models we train use mBERT[2] as the starting point, are then pre-finetuned on MS MARCO, and finetuned on the complete training set of Mr. TyDi. They consist of a DPR transformer bi-encoder, with distinct encoders for the queries and passages, initialized from mBERT (*bert-base-multilingual-cased*).

The models trained with TAS negative sampling require two teacher models to perform negative sampling. The first, used for the English pretraining step, is a publicly available DistilBERT[3] model. The second, used for multilingual finetuning, is a publicly available BERT model[4] trained on the Mr. TyDi training set.

To specify the DPR models that we train, we use the abbreviations and acronyms introduced for the corresponding negative sampling methods in Section 3.3 and 3.3, and summarized in Table 1:

- **DPR$_{\text{base}}$**: A DPR model trained without any negative sampling.
- **DPR$_{\text{BM}}$**: A DPR model trained with BM25[5] negatives.
- **TAS-Q**: A DPR model trained with TAS negative sampling on queries.
- **TAS-P**: A DPR model trained with TAS negative sampling on passages.
- **ANCE**: A DPR model trained with ANCE negative sampling.
- **ICT-Q**: A DPR model trained with ICT using training queries.
- **ICT-P**: A DPR model trained with ICT using training passages.

4.4 Implementation

Training pipeline. Using the Adam optimizer, each model is trained for 40 epochs with a learning rate of 1e-5 and a batch size of 16. Negative log likelihood loss is used as the loss function. This procedure is followed separately for both the pre-finetuning (pFT) and the finetuning (FT) steps. The model, initialized from mBERT, is pre-finetuned on the MS MARCO dataset for 40 epochs and is then finetuned for another 40 epochs on the combined training sets of Mr. TyDi following the setup in [31]. The representations are updated every 10 epochs for the iterative methods.

Evaluation and testing. Following [30, 31], we report the MRR and Recall@100 scores for each test dataset. We test the seven DPR models on two datasets, the Mr. TyDi benchmark and the mMARCO dataset, under three settings. We report results under the three conditions, in-distribution (Mr. TyDi test sets), out-of-distribution (mMARCO languages that are present in Mr. TyDi), and zero-shot (mMARCO languages that are not present in Mr. TyDi).

We consider observed differences to be statistically significant if $p < 0.05$ in a paired t-test. We write ** to indicate $p < 0.01$ and * to indicate $p < 0.05$. Statistical significance is computed between each dataset's highest and second-highest scores.

[1] https://github.com/ThilinaRajapakse/simpletransformers
[2] https://github.com/google-research/bert
[3] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b
[4] https://huggingface.co/castorini/mdpr-tied-pft-msmarco-ft-all
[5] https://github.com/castorini/pyserini

Table 4: Results on the Mr Tydi (in-distribution) datasets.

	MRR@100								Recall@100							
Dataset	ANCE	BM25	ICT-P	ICT-Q	DPR_{BM}	DPR_{base}	TAS-P	TAS-Q	ANCE	BM25	ICT-P	ICT-Q	DPR_{BM}	DPR_{base}	TAS-P	TAS-Q
ar	*0.524*	0.247	0.457	0.417	**0.586****	0.305	0.422	0.413	0.868	0.636	*0.884*	0.882	**0.907****	0.856	0.865	0.859
bn	0.446	0.333	0.492	0.469	**0.563**	0.398	0.454	*0.494*	0.847	0.730	**0.919**	**0.919**	0.901	0.892	0.901	*0.910*
fi	0.419	0.161	*0.431*	0.396	**0.471****	0.259	0.373	0.357	0.828	0.507	0.856	*0.866*	**0.881**	0.823	0.836	0.854
id	*0.475*	0.288	0.427	0.413	**0.502***	0.321	0.402	0.382	0.870	0.742	0.877	*0.878*	**0.903****	0.864	0.876	0.876
ja	0.333	0.173	*0.362*	0.319	**0.430****	0.215	0.314	0.284	0.794	0.624	0.835	*0.844*	**0.864***	0.808	0.814	0.815
ko	*0.354*	0.196	0.343	0.323	**0.399****	0.239	0.316	0.306	0.753	0.360	0.753	*0.760*	**0.805****	0.720	0.732	0.724
ru	*0.410*	0.209	0.350	0.326	**0.436***	0.241	0.317	0.294	0.821	0.462	0.843	*0.859*	**0.871**	0.813	0.826	0.838
sw	0.397	0.363	*0.530*	0.492	**0.530**	0.386	0.507	0.438	0.785	0.743	0.863	**0.881**	*0.870*	0.852	0.854	0.867
te	0.677	0.186	*0.703*	0.579	**0.774****	0.469	0.594	0.461	0.921	0.426	**0.967**	0.964	*0.966*	0.947	0.966	0.966
th	0.416	0.161	*0.423*	0.384	**0.489****	0.288	0.392	0.353	0.807	0.489	0.887	**0.905**	0.842	0.863	0.873	*0.893*

Table 5: Results on the MMARCO OOD datasets.

	MRR@100								Recall@100							
Dataset	ANCE	BM25	ICT-P	ICT-Q	DPR_{BM}	DPR_{base}	TAS-P	TAS-Q	ANCE	BM25	ICT-P	ICT-Q	DPR_{BM}	DPR_{base}	TAS-P	TAS-Q
ar	0.092	0.106	**0.138****	*0.125*	0.105	0.103	0.124	0.116	0.355	0.375	**0.461****	*0.441*	0.370	0.409	0.428	0.433
id	0.114	**0.154****	*0.140*	0.124	0.118	0.099	0.117	0.115	0.425	**0.541****	*0.509*	0.487	0.439	0.435	0.474	0.470
ja	0.128	0.136	**0.167****	*0.157*	0.139	0.129	0.153	0.147	0.462	0.469	**0.545**	*0.540*	0.474	0.506	0.522	0.524
ru	0.134	0.102	**0.157****	*0.143*	0.139	0.124	0.137	0.136	0.480	0.354	**0.541****	*0.531*	0.482	0.498	0.503	0.509

5 RESULTS

5.1 In-distribution results (Known language)

Table 4 shows the MRR@100 and Recall@100 scores obtained by each model on the Mr. TyDi test sets (the in-distribution setting).

We find that DPR_{BM} outperforms all other methods across all languages (statistically significant for all but two languages) in the in-distribution setting. This indicates that simple BM25 negatives are surprisingly effective when training multilingual dense retrievers. While Hofstätter et al. [6], who introduced TAS-style negative sampling, demonstrated impressive retrieval effectiveness, our results indicate that most of the improvements possibly came from the other techniques used in [6] (e.g., knowledge distillation).

ICT-P, obtains the second-best performance with the passage clustering approach outperforming query clustering across the board. We see the same pattern with TAS clustering where TAS-P outperforms TAS-Q. We believe the better performance of clustering passages instead of queries is likely due to passages being longer and containing more information, leading to better clusters and harder negatives.

ANCE performance is close to ICT-P performance, with ICT-P obtaining higher MRR@100 on six languages while ANCE obtains higher MRR@100 on four languages. In terms of Recall@100, ICT-P gets higher scores than ANCE on all languages except Korean (tie).

DPR_{base} with random in-batch negatives performs the worst out of all methods, confirming that effective negative sampling methods are essential to train good dense retrievers in a multilingual setting.

Based on these results, we recommend using negative sampling based on BM25 hard negatives when training a multilingual dense retrieval model if the model is primarily tasked with retrieval in an in-distribution setting. We further analyze the effectiveness of DPR_{BM} in the in-distribution setting in Section 6.1.

5.2 Out-of-distribution results (Known language)

Table 5 shows the MRR@100 and Recall@100 scores for each model on the out-of-distribution language datasets from mMARCO. In this setting, the two variants of iterative clustered training, ICT-P and ICT-Q, outperform all other negative sampling methods. Similar to the in-distribution setting, we again see that passage-based clustering yields better results than query-based clustering, with the ICT-P model obtaining the highest scores of the negative sampling methods on all four languages (statistically significant).

We compare ICT-P and DPR_{BM} on BEIR datasets (English language) to confirm our findings in the out-of-distribution setting free from machine translation artifacts in Section 6.2.

These results indicate that the clustered training methods (both ICT-P and ICT-Q) provide superior out-of-distribution results compared to the other negative sampling methods.

5.3 Zero-shot results (Unknown language)

Next, Table 6 shows the MRR@100 and Recal@100 scores obtained by each model on the zero-shot languages. This is the setting that we are most interested in as it represents the real-world scenario of using a multilingual dense retrieval model for monolingual retrieval in a language that it has not been trained on for retrieval.

Similar to the out-of-distribution setting, the ICT methods outperform all other methods on all zero-shot languages (statistically significant). Again, we see that clustering passages yields better results compared to clustering queries for both ICT and TAS-style clustering.

We also report results on four additional retrieval datasets in zero-shot languages (French, German, and Chinese) to confirm that the results on the mMARCO datasets are not due to machine translation artifacts. In addition to these datasets being zero-shot

Table 6: Results on the MMARCO zero-shot datasets.

	MRR@100								Recall@100							
Dataset	ANCE	BM25	ICT-P	ICT-Q	DPR_{BM}	DPR_{base}	TAS-P	TAS-Q	ANCE	BM25	ICT-P	ICT-Q	DPR_{BM}	DPR_{base}	TAS-P	TAS-Q
zh	0.136	0.119	**0.169****	*0.164*	0.145	0.135	0.154	0.154	0.499	0.451	*0.576*	**0.578**	0.515	0.529	0.543	0.547
nl	*0.155*	0.142	**0.172****	0.154	0.150	0.128	0.148	0.142	0.518	0.488	**0.582****	*0.567*	0.513	0.517	0.534	0.535
fr	0.159	0.149	**0.186****	*0.167*	0.157	0.139	0.159	0.154	0.548	0.519	**0.611**	*0.608*	0.551	0.560	0.567	0.573
de	0.156	0.135	**0.175****	*0.158*	0.158	0.137	0.156	0.150	0.517	0.464	**0.575****	*0.561*	0.519	0.530	0.534	0.538
hi	0.087	*0.134*	**0.141**	0.130	0.107	0.111	0.131	0.128	0.324	*0.470*	**0.470**	0.462	0.387	0.435	0.453	0.449
it	0.154	0.145	**0.179****	*0.166*	0.154	0.137	0.155	0.151	0.543	0.499	**0.604****	*0.593*	0.541	0.546	0.561	0.560
pt	0.156	0.158	**0.185****	*0.168*	0.152	0.140	0.160	0.160	0.537	0.544	**0.604****	*0.593*	0.534	0.542	0.563	0.568
es	0.170	0.159	**0.196****	*0.177*	0.164	0.147	0.168	0.166	0.566	0.551	**0.635****	*0.624*	0.558	0.578	0.590	0.594
vi	0.118	*0.140*	**0.141**	0.126	0.121	0.106	0.120	0.118	0.423	**0.508**	*0.498*	0.481	0.439	0.444	0.461	0.459

Table 7: Results on the other zero-shot datasets.

	MRR@100								Recall@100							
Dataset	ANCE	BM25	ICT-P	ICT-Q	DPR_{BM}	DPR_{base}	TAS-P	TAS-Q	ANCE	BM25	ICT-P	ICT-Q	DPR_{BM}	DPR_{base}	TAS-P	TAS-Q
BSARD	0.150	**0.225***	0.161	*0.168*	0.146	0.161	0.147	0.148	0.310	**0.466**	0.368	0.398	0.352	*0.430*	0.383	0.348
GerDaLIR	0.120	**0.199****	*0.163*	0.151	0.104	0.148	0.158	0.144	0.349	**0.650****	*0.422*	0.401	0.319	0.401	0.409	0.387
Multi-CPR Ecom	0.118	**0.293****	0.191	0.192	0.118	0.190	0.188	*0.203*	0.399	**0.711****	0.530	0.549	0.409	0.550	0.542	*0.552*
Multi-CPR Video	0.112	**0.230**	0.203	*0.210*	0.124	0.188	0.199	0.204	0.449	**0.735****	0.634	0.648	0.469	0.601	0.650	*0.666*

languages, they are out-of-domain datasets as described in Section 4.2. Table 7 shows that the ICT methods outperform the other negative sampling methods on these human annotated datasets. However, we see that the baseline BM25 model outperforms the dense retrieval method on these datasets. This agrees with findings from Thakur et al. [20] that the lexical-based BM25 method can be superior to dense retrievers in an out-of-domain setting.

The results from the zero-shot language tests shows that the iterative ICT methods (ICT-P and ICT-Q) show superior domain adaptability as well as adaptability to new languages compared to the other negative sampling methods.

5.4 Summary of results

Finally, we summarize the findings from this section.

- DPR_{BM} demonstrates impressive results on in-distribution test sets, outperforming all other methods.
- ICT outperforms the other negative sampling methods in out-of-distribution and zero-shot settings.
- TAS-style clustering using an external model underperforms other negative sampling techniques in a multilingual setting.

6 ANALYSIS

We look at two variants of the three main settings from Section 5 and consider the two best-performing negative sampling methods, namely, ICT-P and DPR_{BM}.

6.1 In-distribution data, unknown language

We introduce a variant of the in-distribution setting to further analyze the effectiveness of DPR_{BM} under in-distribution testing conditions. We compare the performance of the DPR_{BM} and ICT-P models (best and second-best results in the in-distribution setting, respectively) on the MIRACL languages that do not appear in Mr. TyDi. In this setting, we test on data that is *in-distribution* in terms of data collection, annotation, and sources, but *zero-shot* in terms of the language.

Table 8: Results on the MIRACL zero-shot languages.

	MRR@100		Recall@100	
Dataset	ICT-P	DPR_{BM}	ICT-P	DPR_{BM}
German	0.435	**0.458**	0.767	**0.772**
Spanish	0.512	**0.572****	**0.712**	0.700
Persian	0.434	**0.461**	**0.805****	0.764
French	0.389	**0.459****	0.782	**0.798**
Hindi	0.412	**0.451***	**0.732**	0.727
Yoruba	0.540	**0.577**	**0.866**	0.861
Chinese	**0.517**	0.511	**0.854****	0.815

In Table 8 we see that DPR_{BM} outperforms ICT-P on in-distribution data in terms of ranking metrics even when the language is new to the model. Interestingly, ICT-P gets better recall than DPR_{BM}, unlike what we saw in Section 5.1. This suggests that the better generalizability of the ICT-P model helps it adapt to the newer languages. However, DPR_{BM} still has better overall performance in the in-distribution setting which strengthens our recommendation to use DPR_{BM} for in-distribution multilingual retrieval scenarios.

6.2 Out-of-domain data, known language

Now, we compare the performance of ICT-P and DPR_{BM} on BEIR which presents a setting where the data is *out-of-domain* (*out-of-distribution* and new domains) but in a known language (English).

In Table 9, we see that the ICT-P model outperforms the DPR_{BM} model in the *out-of-domain*, known language setting. This serves to confirm our recommendation to use ICT-P models in a multilingual retrieval setting where generalizability to new data distributions or domains is needed. We also report nDCG@10 for the BEIR results as this is the official metric used in [20].

Table 9: Results on the BEIR datasets.

	nDCG@10		MRR@100		Recall@100	
Dataset	ICT-P	DPR_{BM}	ICT-P	DPR_{BM}	ICT-P	DPR_{BM}
ArguAna	**0.304****	0.235	**0.214****	0.165	**0.891****	0.852
CQA Dup Stack	**0.207****	0.147	**0.219****	0.154	**0.406****	0.320
DBPedia	0.230	**0.238**	0.510	**0.538**	0.324	**0.332**
FiQa	**0.205****	0.181	**0.265***	0.239	**0.475****	0.432
NFCorpus	**0.214****	0.192	**0.395**	0.386	**0.202****	0.182
Quora	**0.770****	0.264	**0.760****	0.256	**0.967****	0.613
SciDocs	**0.084***	0.077	**0.175***	0.156	0.203	**0.204**
SciFact	**0.420**	0.396	**0.399**	0.370	0.753	**0.775**
TREC-COVID	**0.464****	0.355	**0.696**	0.583	**0.057**	0.051

7 RELATED WORK

Dense retrieval. Traditionally, passage retrieval has been performed using sparse retrieval methods such as BM25 [22]. Karpukhin et al. [8] show that transformer-based [21] dual-encoder models can surpass traditional sparse methods by using the ability of transformer models to represent semantic meaning, unlike classic keyword-based methods. However, later work [17, 18, 20, 28] has shown that dense retrieval models, specifically dual-encoder models, struggle to generalize to out-of-distribution data.

Improved training regimes to boost generalizability. Prior work has proposed a range of data generation, data augmentation, and data selection techniques for improving the effectiveness of dense retrieval models. Since the release of the BEIR benchmark [20], researchers have begun to specifically consider whether these techniques improve out-of-distribution generalization as well as in-domain effectiveness. Negative sampling techniques [6, 8, 25] represent one such approach to improve the generalizability of dense retrievers. We focus on their effectiveness in boosting the generalizability of dense retrievers in a multilingual setting.

Negative sampling for dense retrieval. Early dense retrieval models like DPR [8] select their negative training examples from a combination of false positives identified by BM25 and from other queries in the same training batch ("in-batch negatives"). ANCE [25] demonstrates the importance of selecting hard negative training examples, which it accomplishes by periodically identifying false positive examples using the retrieval model currently being trained. Although the original work focuses on in-distribution performance and does not explore out-of-distribution or zero-shot retrieval, later work [20] shows that out-of-distribution results also improve. ANCE requires maintaining a continuously updated dense index, which requires significant compute resources, though these requirements can be reduced by freezing the document encoder for part of training [27]. Alternatively, computational requirements can be reduced by caching negative examples rather than periodically recomputing the entire index [12, 26].

Rather than using a dense retrieval model to mine hard negative examples, TAS-B [6] creates difficult training batches by clustering queries once at the start of training and then samples from those clusters to build training batches. TAS-B combines this with knowledge distillation to improve the retrieval performance of dense retrievers. TAS-B uses a separately trained BERT model to generate the representations for the clustering. Therefore, the effectiveness of TAS-B is dependent on the availability of a teacher model, which can be a restrictive constraint in a multilingual setting.

A systematic comparison of negative sampling methods for dense retrieval under different generalizability conditions (in-distribution, out-of-distribution, zero-shot) is missing. This is the gap that we fill. We consider a rich multilingual setting that allows us to formulate all three conditions, and we discover a gap in the choices available for negative sampling for dense retrieval so far.

Multilingual retrieval. To understand the generalizability of dense retrievers we consider a multilingual setting, that naturally allows us to consider challenging in-distribution, out-of-distribution, and zero-shot settings. The literature on information retrieval in multiple languages is rich. Cross-lingual retrieval (queries in one language and passages in another), in particular, has been the focus of many publications, and we refer the reader to [5, 29] for recent surveys on this area. However, our work focuses on multilingual retrieval, where both queries and passages are in the same language (monolingual), but the models used support monolingual retrieval in many languages. However, cross-lingual and multilingual retrieval both benefit from cross-lingual transfer capabilities, particularly of large language models. The zero-shot knowledge transfer ability of large language models has previously been studied in [13, 19].

Zhang et al. [31] offer a comprehensive guide on training multilingual dense retrievers based on the Mr. TyDi benchmark and focuses on *monolingual retrieval* with *multilingual retrievers*. We also focus on the same task, however, we pay careful attention to the generalizability of the multilingual dense retrievers, both for out-of-distribution data and for new languages. We analyze the effectiveness of different negative sampling methods under the in-distribution, out-of-distribution, and zero-shot conditions in a multilingual setting, and investigate whether existing findings from English language research generalizes to these new conditions.

8 DISCUSSION

We discuss the strengths and weaknesses of the negative sampling techniques we have investigated and our central reproducibility question, viz. how existing findings from English language models and datasets generalize to the setting of monolingual retrieval with multilingual models. We also discuss the implications for the use of multilingual dense retrieval models in practice.

Generalizability of English language findings to the multilingual setting. Broadly speaking, we found that existing findings on English language retrieval (the importance of the presence of hard negatives) generalize to the multilingual domain. Good hard negative sampling methods yields significant improvements in multilingual retrieval quality. However, one of the most effective negative sampling methods, TAS, is less effective in the multilingual setting. We believe that this is due to the comparative lack of effective teacher models that can be employed to generate the query or passage representations for clustering. Therefore, the iterative negative mining methods (ICT-P, ICT-Q, and ANCE) demonstrate superior retrieval quality over the non-iterative methods as they do not require external models to perform negative sampling.

DPR_{base} (Random negatives) requires no additional data, models, or hardware resources to be used. It also requires the least training time and is the simplest to implement, but it is also the least effective

of the methods we have investigated. We recommend using DPR_{base} only when no additional data, models, or hardware resources are available and training time efficiency is the primary concern.

DPR_{BM} (Non-iterative BM25 negatives) requires an external BM25 model to find negatives. However, BM25 is a sparse retrieval method and is generally much faster and cheaper than dense retrieval methods. Therefore, DPR_{BM} can be used with little additional effort in most cases. DPR_{BM} does not require any additional hardware resources to use.

Using BM25 negatives is surprisingly effective as long as there is little distributional shift between the training and test data, demonstrating superior retrieval quality in the in-distribution setting compared to the other methods. Based on these factors, we recommend using DPR_{BM} when the model is used mostly for in-domain retrieval. This contradicts in-distribution English language findings where negative sampling methods like ANCE and TAS were developed in order to improve over BM25 negatives.

ICT-Q and ICT-P (Iterative, clustering-based negatives) ICT-P demonstrated superior performance in all three settings between the two ICT methods. While ICT-Q is marginally faster in the clustering phases during training due to queries usually being shorter than passages, we do not believe this makes a practical difference. Therefore, we recommend using ICT-P over ICT-Q.

Overall, ICT-P obtained the best results in two out of three settings (out-of-distribution and zero-shot). Based on this, we recommend using ICT-P as the negative sampling method in most multilingual retrieval scenarios except for the special case detailed above (the model is intended for use in an in-distribution setting).

TAS-Q and TAS-P (Non-iterative, clustering-based negatives) While both TAS-Q and TAS-P outperform DPR_{base} across all three settings, they perform similar or inferior to the other negative sampling methods. In addition to this, these two negative sampling methods require an external model to perform the clustering in order to sample similar queries/passages for training batches.

We believe that a possible reason for TAS negative sampling to underperform in a multilingual setting is that it relies heavily on the external model used for clustering to find good hard negatives. The external models available in the multilingual retrieval setting tend to be less effective and reliable compared to the models available for English such as ColBERT [9]. Furthermore, while the approach in [6] performs well in an English language setting, it uses other techniques, such as knowledge distillation, in addition to negative sampling explored in this work. Wang and Zuccon [23] also found that TAS-style negative sampling, without knowledge distillation, can underperform random negatives in an English language setting indicating that the success of [6] could largely have been influenced by the effectiveness of knowledge distillation. Due to these limitations, we do not recommend using TAS-Q or TAS-P for negative sampling in a multilingual retrieval setting.

ANCE (Iterative, full-corpus mined negatives) ANCE is fairly effective in multilingual retrieval in all three settings that we considered. However, it has the highest hardware resource requirements (for training) of all the methods considered in this work. Periodically building a dense index of the full document collection increases training time significantly compared to the other methods. Therefore, we recommend using ICT-P instead which builds on similar ideas as ANCE, but is more efficient to train, and also outperformed ANCE in all three settings.

9 CONCLUSION

We studied the generalizability of earlier insights into the effectiveness of negative sampling methods for multilingual retrieval under in-distribution, out-of-distribution, and zero-shot conditions. We identified a gap in the literature, and by combining earlier insights, introduced an iterative, clustering-based method, to fill this gap.

Our experiments confirmed the choice of the negative sampling method used to train dense retrievers has a significant impact on their multilingual retrieval effectiveness. This is in agreement with existing findings from English language research overall. However, TAS, a highly effective clustering-based negative sampling method in English, underperformed other negative sampling methods in a multilingual setting contradicting existing findings. On the other hand, iterative negative sampling methods performed well in the multilingual setting, maintaining their effectiveness from prior English language work. The comparative lack of effective teacher models in a multilingual setting poses a barrier for methods that rely on external representations, such as TAS. Iterative negative sampling methods do not require external representations; instead, they use the representations from the model being trained to find hard negatives and, therefore, succeed in finding good hard negatives even as the model learns even in a multilingual setting.

Interestingly, the best negative sampling method depends on whether the model is tested on in-distribution or out-of-distribution data. For the in-distribution setting, simple BM25 negatives (DPR_{BM}) obtained the best performance and, therefore, we recommend using DPR_{BM} for in-distribution multilingual retrieval tasks. For out-of-distribution and zero-shot settings, we found that ICT-P has the best performance. Based on this, we recommend using ICT-P for out-of-distribution or zero-shot scenarios. Unless the situation clearly calls for in-distribution performance only, our overall recommendation is also to use ICT-P due to its better generalizability.

As to limitations of our work, we have constrained ourselves to the DPR architecture and have not explored the benefits of clustered training for other architectures. We only considered a contrastive learning setup and did not experiment with other training methods such as knowledge distillation (consistently good teacher models are rarer in the multilingual retrieval setting than in English only retrieval). In future work we intend to generalize our findings to different architectures and training setups, and explore interactions between negative sampling methods and other training setups.

Acknowledgments. This research was supported by the Dreams Lab, a collaboration between Huawei Finland, the University of Amsterdam, and the Vrije Universiteit Amsterdam, by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.-1389.20.183, KICH3.LTP.20.006, and VI.Vidi.223.166, and by the European Union's Horizon Europe program under grant agreement No 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. *Advances in Neural Information Processing Systems* 34 (2021), 7547–7560.

[2] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mMarco: A Multilingual Version of the MS MARCO Passage Ranking Dataset. *arXiv preprint arXiv:2108.13897* (2021).

[3] 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 (2020), 454–470.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Petra Galuščáková, Douglas W Oard, and Suraj Nair. 2021. Cross-language Information Retrieval. *arXiv preprint arXiv:2111.05988* (2021).

[6] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 113–122.

[7] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale Similarity Search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[8] 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)*. Association for Computational Linguistics, 6769–6781.

[9] 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*. 39–48.

[10] Carlos Lassance and Stéphane Clinchant. 2023. The Tale of Two MS MARCO – And their Unfair Comparisons. *arXiv preprint arXiv:2304.12904* (2023).

[11] Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2021. Learning Cross-Lingual IR from an English Retriever. *arXiv preprint arXiv:2112.08185* (2021).

[12] Erik Lindgren, Sashank Reddi, Ruiqi Guo, and Sanjiv Kumar. 2021. Efficient Training of Retrieval Models Using Negative Cache. *Advances in Neural Information Processing Systems* 34 (2021), 4134–4146.

[13] Sean MacAvaney, Luca Soldaini, and Nazli Goharian. 2020. Teaching a New Dog Old Tricks: Resurrecting Multilingual Retrieval Using Zero-shot Learning. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer, 246–254.

[14] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. In *CoCo@ NIPs*.

[15] Carol Peters and Páraic Sheridan. 2001. *Multilingual Information Access*. Springer Berlin Heidelberg, Berlin, Heidelberg, 51–80. https://doi.org/10.1007/3-540-45368-7_3

[16] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (apr 2009), 333–389. https://doi.org/10.1561/1500000019

[17] Guilherme Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. In Defense of Cross-Encoders for Zero-Shot Retrieval. *arXiv preprint arXiv:2212.06121* (2022).

[18] 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* (2022).

[19] Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual Training of Dense Retrievers for Document Retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. 251–253.

[20] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv preprint arXiv:2104.08663* (2021).

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

[22] Ellen M. Voorhees. 1999. The TREC-8 Question Answering Track Report. In *Proceedings of TREC-8*. 77–82.

[23] Shuai Wang and Guido Zuccon. 2023. Balanced Topic Aware Sampling for Effective Dense Retriever: A Reproducibility Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2542–2551.

[24] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 38–45.

[25] 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* (2020).

[26] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2022. LaPraDoR: Unsupervised Pretrained Dense Retriever for Zero-Shot Text Retrieval. In *Findings of the Association for Computational Linguistics: ACL 2022*. 3557–3569.

[27] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1503–1512.

[28] Jingtao Zhan, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. Evaluating Extrapolation Performance of Dense Retrieval. *arXiv preprint arXiv:2204.11447* (2022).

[29] Liang Zhang and Xiaobing Zhao. 2020. An overview of cross-language information retrieval. In *Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, July 17–20, 2020, Proceedings, Part I 6*. Springer, 26–37.

[30] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A Multilingual Benchmark for Dense Retrieval. *arXiv preprint arXiv:2108.08787* (2021).

[31] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022. Towards Best Practices for Training Multilingual Dense Retrieval Models. *arXiv preprint arXiv:2204.02363* (2022).

[32] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages. *arXiv preprint arXiv:2210.09984* (2022).