







Neural Approaches to Multilingual Information Retrieval

Dawn Lawrie¹(✉) , Eugene Yang¹ , Douglas W. Oard^{1,2} , and James Mayfield¹ 

¹ HLTCOE, Johns Hopkins University, Baltimore, MD 21211, USA
{lawrie, eugene.yang, mayfield}@jhu.edu

² University of Maryland, College Park, MD 20742, USA
oard@umd.edu

Abstract. Providing access to information across languages has been a goal of Information Retrieval (IR) for decades. While progress has been made on Cross Language IR (CLIR) where queries are expressed in one language and documents in another, the multilingual (MLIR) task to create a single ranked list of documents across many languages is considerably more challenging. This paper investigates whether advances in neural document translation and pretrained multilingual neural language models enable improvements in the state of the art over earlier MLIR techniques. The results show that although combining neural document translation with neural ranking yields the best Mean Average Precision (MAP), 98% of that MAP score can be achieved with an 84% reduction in indexing time by using a pretrained XLM-R multilingual language model to index documents in their native language, and that 2% difference in effectiveness is not statistically significant. Key to achieving these results for MLIR is to fine-tune XLM-R using mixed-language batches from neural translations of MS MARCO passages.

Keywords: Multilingual ad-hoc retrieval · ColBERT-X · DPR-X · Multilingual training of MPLM

1 Introduction

With advances in neural models for machine translation (MT) and Information Retrieval (IR), it is time to revisit the problem of Multilingual IR (MLIR). Soon after Cross-Language IR (CLIR) was proposed as an information retrieval task, research began on MLIR [34]. MLIR seeks to produce a total ordering over retrieved documents, regardless of language, such that the most useful documents appear at the top of the ranking. Assuming a searcher can consume multilingual information (either directly or using MT), the search engine should be able to return useful information regardless of the language of the document.

Much prior work on MLIR has involved subsetting documents by language, performing CLIR on each document set, and merging the results [37]. The advent of neural machine translation and neural IR using Multilingual Pretrained Language Models (MPLMs) creates new opportunities for MLIR that we study here.

If MT were perfect, translating all documents into the query language and searching monolingually might suffice. Indeed, our experiments confirm that for the high-resource languages with which we have experimented (English, French, German, Italian, and

Spanish), using neural machine translation to convert each document into the query language is effective when used with neural ranking (in our experiments, ColBERT [26]) fine-tuned on MS MARCO [2]. However, using neural MT in that way incurs substantial indexing costs because a GPU is required first to translate the document and then again to encode it into dense vectors for neural IR. Alternatively, we can use translations of MS MARCO to fine-tune an MPLM; that approach is nearly as effective, not statistically different, and considerably faster at indexing time. Our use of MS MARCO makes English a natural choice as the query language, but our approach is extensible to any query language for which suitable fine-tuning data exists.

This paper makes the following contributions: (1) Using a collection containing five high-resource European languages, we show that neural MT with neural IR achieves higher MAP and Precision at 10 scores than any other known MLIR technique, but that reliance on neural MT greatly increases the time required to index a collection. (2) We show that extending the ColBERT-X [32] Translate-Train (TT) CLIR model to multiple languages achieves equivalent retrieval effectiveness with less than half the indexing time when used with mixed-language fine-tuning. (3) We show that some language bias in favor of query-language documents is present with all approaches, but that query-language bias is smaller with our Multilingual Translate-Train (MTT) implementation of ColBERT-X.

2 Background

We provide an overview of MLIR, followed by a brief review of traditional and neural IR. The term “multilingual” has been used in several ways in IR. Hull and Grefenstette [22], for example, note that it has been used to describe monolingual retrieval in multiple languages, as in Biloshmi et al. [5], and it has also been used to describe CLIR tasks that are run separately in several languages [7–9, 27, 31]. We adopt the Cross-Language Evaluation Forum (CLEF)’s meaning of MLIR: using a query to construct one ranked list in which each document is in one of several languages [36]. We note that this definition excludes mixed-language queries and mixed-language documents, which are yet other cases to which “multilingual” has been applied.

Five broad approaches to MLIR have been tried. Among the earliest, Rehder et al. [39] represented English, German and French documents in a learned trilingual embedding space, represented the query in the same embedding space, and then computed query-document similarity in the embedding space. The techniques and training data for creating multilingual embeddings were, however, too limited at the time to get good results from that technique. More recently, Sorg and Cimiano [44] garnered substantial attention by training embeddings on topically-related Wikipedia pages in English, German, French and Spanish. This paper extends this line of work.

A second approach by Nie and Jin [33] indexed terms from all documents in their original language then created queries containing translations of the query terms in all target languages. With many document languages, this can lead to long queries. A third approach is to translate indexed terms into the query language at indexing time; the original queries can then be used directly to find similar (translated) content [18, 29, 38]. We experiment with this approach as well. This approach is, however, only practical

when just a few query languages are to be supported. To address that limitation, the second and third approaches can be combined to create a fourth approach in which documents and query terms are each converted into one of a small number of indexing languages. This has been called a “pivot language” approach, because in the limit all documents and queries can be translated into a single language.

The fifth, and most widely studied, approach is to first use monolingual or bilingual retrieval to create a ranked list for each document language, and then to merge those ranked lists to construct a single result list [37,43,45]. While this approach is architecturally similar to collection sharding, a widely-used approach to address efficiency, differences in collection statistics result in incompatible scores that require normalization prior to late fusion. Unfortunately, normalizing scores for collections across languages has been shown to be challenging [37].

Finally, one can simply show one ranked list per language to the user, as is done in the 2lingual search engine.¹ This approach does not scale well beyond a small number of languages, but it has the advantage of making it fairly clear to the searcher what the search engine has done.

Every MLIR ranked retrieval model must rank the indexed documents given a query. Traditional ranking methods such as computing inner products between the query and each indexed document containing a query term using sparse BM25 [40] term weights are fast, but neural IR methods yield better rankings [24,26,32] with more relevant documents earlier in the ranked list.

This paper focuses on tradeoffs between effectiveness and efficiency. Each technique described in this paper achieves ranking latency sufficient for interactive use (below 300 ms) on the collections that we experiment with, but the time required to index the documents varies. Indexing time consists of three components: text processing (e.g., casing and tokenization), machine translation, and representation (e.g., McCarley [30] and Magdy and Jones [29]). Of these, neural MT is the slowest, so IR methods that do not require neural MT at indexing time have a substantial indexing time advantage (e.g., Aljlal and Frieder [1]). Our principal MLIR result is that MPLMs can achieve MAP close to the best results while producing substantial savings in indexing time.

We achieve this by extending the ColBERT-X [32] CLIR model to perform MLIR. ColBERT-X combines three key ideas. First, drawing insight from BERT [15], it represents documents using contextual embeddings, which better represent meaning than simple term occurrence. Second, using both multilinguality and improved pretraining from either multilingual BERT [47] or XLM-R [11], ColBERT-X generates similar contextual embeddings for terms with similar meaning, regardless of language. Third, drawing its structure from ColBERT [26], ColBERT-X limits ranking latency by separating query and document transformer networks, allowing offline indexing. ColBERT scores documents by focusing query term attention on the most similar contextual embedding in each document. Our experiments confirm that this approach yields better MLIR MAP than does computation of inner products between classification tokens for the query and each document, an approach known as Dense Passage Retrieval (DPR) [24].

¹ <https://www.2lingual.com/>.

3 Fine-Tuning MPLMs for MLIR

Following Nair et al. [32] we consider two high-level approaches to fine-tuning for generalizing neural retrieval models to MLIR. Both approaches use existing MPLMs such as XLM-R [11] to encode queries and documents in multiple languages. We adapt the MPLM to MLIR via task-specific fine-tuning. These approaches are applicable to any retrieval model that is able to encode text using an MPLM.

Consider a set of queries in a source language \mathbf{L}_s and a set of documents in m target languages $\mathbf{L}_t = \cup_{i=1}^m \mathbf{L}_i$. We want to train a scoring function $\mathcal{M}_\Theta(q_{(s)}, d_{(t)}) \rightarrow \mathbb{R}$ for ranking documents with respect to a query. This paper denotes the language of an instance as a subscript $\bullet_{(l)}$.

3.1 English Training (ET)

Since MPLMs can encode text from many languages, we follow Nair et al. [32] and only fine-tune the model monolingually. When processing queries, we transfer the model to MLIR zero-shot. Specifically, consider a loss function \mathcal{L} (for example, cross-entropy),

$$\Theta = \arg \min_{\theta} \sum_{q,d} \mathcal{L}_\theta(q_{(s)}, d_{(s)}, r_{q,d})$$

where $q_{(s)}$ and $d_{(s)}$ are representations of the queries and documents and $r_{q,d}$ is the relevance judgment of document d on query q , both in language \mathbf{L}_s , encoded by an MPLM. We use English as our query language because that is the language of MS MARCO. We refer to this approach as ‘‘English Training’’ or ET. However, this approach could equally well use any language for which similar extensive training data is available.

Despite only exposing the model to text in \mathbf{L}_s during fine-tuning, the multilingual model can transfer its task model to other languages, as has been seen in prior CLIR work [32]. However, such zero-shot language transfer is suboptimal because of (1) the lack of alignment objectives between languages during pretraining [48]; and (2) differences in the representation of each language by the MPLM, which has been called *the curse of multilinguality* [11,46]. As we show in Sect. 6.1, such zero-shot transfer not only produces suboptimal retrieval effectiveness, it can also lead to language bias.

3.2 Multilingual Translate Training (MTT)

To mitigate those issues, we propose a Multilingual Translate-Train (MTT) approach that generalizes the CLIR Translate-Train (TT) approach to MLIR [32,42]. To expose target languages $\mathbf{L}_1 \dots \mathbf{L}_m$ to the model, we translate the monolingual training documents into each target language using MT. Specifically, the training objective can be expressed as

$$\Theta = \arg \min_{\theta} \sum_{q,d} \sum_{l=1}^m \mathcal{L}_\theta(q_{(s)}, d_{(l)}, r_{q,d})$$

This objective exposes the retrieval model to language pairs that it might see when processing queries, resulting in a more effective, better-balanced model. We experiment with two batching approaches. In Mixed-language (MTT-M), each batch contains

Table 1. Dataset statistics of CLEF 2001, 2002, and 2003. CLEF 2001 and 2002 share the document collection but have different queries. Numbers in parentheses are the number of topics in each query set. We report the number of documents judged relevant over all the topics in a particular year.

Query set	English		German		Spanish		French		Italian		Total	
	# Rel.	# Docs	# Rel.	# Docs	# Rel.	# Docs	# Rel.	# Docs	# Rel.	# Docs	# Rel.	# Docs
2001 (50)	856	113,005	2,130	225,371	2,694	215,738	1,212	87,191	1,246	108,578	8,138	749,883
2002 (50)	821		1,938		2,854		1,383		1,072		8,068	
2003 (60)	1,006	169,477	1,825	294,809	2,367	454,045	946	129,806	–	–	6,144	1,048,137

documents in multiple languages, which encourages the model to learn similarity measures for all languages simultaneously.² With Single-language (MTT-S), each batch contains only documents in one language, helping the model to learn retrieval for one language pair at a time. We found that MTT-M yields better retrieval effectiveness; thus, we present MTT-M as our main result. Section 5.1 compares the two approaches. In Sect. 6.1, we also demonstrate that MTT-M reduces language bias in MLIR. Implementation details can be found in Appendix A

4 Experiments

One of the few test collections that currently supports MLIR evaluation with relevance judgments across multiple languages is from the Cross-Language Evaluation Forum (CLEF). Following Rahimi et al. [38] we use five document languages in the CLEF 2001–2002 collections [7, 8] and four languages in the CLEF 2003 collection [9]. Table 1 shows collection statistics. We report performance for both title and title+description queries, also following Rahimi et al. [38]. Because the number of query elements (subwords) is limited when encoding a query for dense retrieval, we remove *stop structure* to ensure that no query exceeds the length limit. Stop structure includes phrases such as “Find documents” and a limited stop-word list including “on,” “the,” and “and.”³

4.1 Neural Retrieval Models

We evaluate our proposed training approaches on two retrieval models – ColBERT-X [32] and DPR-X [48, 49], which are multilingual variants of ColBERT [26] and DPR [24]. Nair et al. [32] generalized the ColBERT [26] model to CLIR, calling it ColBERT-X, by modifying the vocabulary space and replacing the monolingual pretrained language model with the MPLM XLM-RoBERTa (XLM-R) Large (550M parameters) [11]. With proper training, ColBERT-X achieves state-of-the-art effectiveness in CLIR. In this study, we integrate our proposed fine-tuning approaches with the ColBERT-X XLM-R implementation, which is based on the ColBERTv1 code base.

² Batches include the same query paired with document passages translated into each language.

³ For a complete list: <https://github.com/hltcoe/ColBERT-X/blob/main/scripts/stopstructure.txt>.

We similarly adapted DPR [24, 48], a neural retrieval model that matches a single dense query vector to a single dense document vector. We name this model DPR-X. We use Tevatron [17], an open-source implementation of several neural end-to-end retrieval models in Python, for training, indexing, and retrieval.

For training data, we use MS MARCO-v1 [2], a commonly-used question-answering collection in English for fine-tuning neural retrieval models. For MTT, we use the publicly available mMARCO translations of MS MARCO [6], fine-tuning using the “small training triple” (query, positive and negative document) file released by mMARCO’s creators. We trained all retrieval models with four GPUs (NVIDIA DGX and v100 with 32 GB Memory) with a per-GPU batch size of 32 triples for 200,000 update steps. All models are trained with half-precision floating points and optimized by the AdamW optimizer with a learning rate of 5×10^{-6} .

During indexing, documents are separated into overlapping spans of 180 tokens with a stride of 90 [32]. We aggregate by MaxP [3, 13], which takes the maximum score among the passages in a document as the document score.

4.2 Evaluation

We report previously published results for the state-of-the-art MULM [38] system as a baseline for models that do not perform MT on the full collection. MULM is essentially an MLIR version of Probabilistic Structured Queries (PSQ) [14]. PSQ maps term frequency vectors from document to query language using a matrix of translation probabilities generated using statistical machine translation. For MLIR, a translation matrix is created for each query-document language pair. The query likelihood model is used to score documents. Three key decisions led to good performance: (1) estimating collection statistics based on translation probabilities; (2) estimating document length based on the translation and using that for smoothing; and (3) truncating the translation list at three. As another baseline, we use BM25 ($b = 0.4$, $k_1 = 0.9$) as implemented in PatapSCO [12] over neural machine translated documents (abbreviated ITD for Indexed Translated Documents). For BM25, English queries and documents are tokenized by spaCy [21] and stemmed by the NLTK [4] Porter stemmer (all supported by PatapSCO).

For approaches that require document translation, we use directional MT models built on a transformer architecture (6-layer encoder/decoder) using Sockeye 2 [16, 19]. Measured by BLEU [35], Sockeye 2 achieves state-of-the-art effectiveness in each translation direction. Optimizations cut decoding time in half compared to Sockeye 1 [20]. We chose Sockeye 2 for its good trade-off between efficiency and effectiveness.

To evaluate effectiveness on multiple languages in CLEF 2001–2002 and CLEF 2003, we combine the relevance judgments (qrels) for all languages for each query. In general, different languages have different numbers of relevant documents for each query. To evaluate models trained with English training data, we also translate the document sets into English with MT for indexing. Our main effectiveness measures are Mean Average Precision (MAP) and Precision at 10 (P@10). Both measures focus on the top of the rankings, and both were used by Rahimi et al. [38], facilitating comparison between the neural approaches presented herein and prior state-of-the-art results.

To evaluate language bias, we count the number of relevant documents for a query across all languages, and calculate recall at that level. To compute the measure for a

Table 2. Configurations of experiments identifying the pre-trained language model when applicable, the fine tuning data and process, the retrieval model, and the language of the indexed documents. Under Fine-Tuning Data, MS MARCO refers to English MS MARCOv1, while mMARCO includes the translations into the various languages as well as the original English MS MARCOv1. A model that lists *either* under its Indexing Language can index either machine translated document (translation) or native documents in their various languages.

Name	Language model	Fine-tuning data	Fine-tuning process	Retrieval model	Indexing language
MULM	–	–	–	PSQ	Native
BM25-ITD	–	–	–	BM25	Translation
ColBERT-X(ET)	XLM-R	MS MARCO	ET	ColBERT-X	Either
ColBERT-X(MTT-M)	XLM-R	mMARCO	MTT-M	ColBERT-X	Either
ColBERT-X(MTT-S)	XLM-R	mMARCO	MTT-S	ColBERT-X	Either
DPR-X(ET)	XLM-R	MS MARCO	ET	DPR-X	Either
DPR-X(MTT-M)	XLM-R	mMARCO	MTT-M	DPR-X	Either
ColBERT(ET)	BERT	MS MARCO	ET	ColBERT	Translation

specific language, we keep this level constant, but ignore all documents in other languages (both in the MLIR results and in the relevance judgments). We call the mean of this measure over all queries $Recall@MLIR-Relevant$. When computing the mean, we omit from the calculation cases in which no relevant documents in that language are known (recall is undefined in such cases). This measure lies between 0 and 1, and values across that full range are achievable. We use the open source `ir-measures` [28]⁴ package to compute all effectiveness measures.

5 Results

We experiment with the Multilingual Translation Training (MTT) using two retrieval models and compare them to two strong baseline retrieval models: BM25-ITD indexing translated documents and MULM indexing native documents; these represent the state of the art on our test collections. Since per-query results for MULM have not been published we perform significance tests only between our systems and the BM25+ITD baseline (the stronger of the two baselines). Table 2 summarizes the experiments that facilitate this analysis. We first compare the effectiveness of our two batching strategies for MTT before examining their effectiveness relative to the baselines. Finally, we consider the trade-off between effectiveness and indexing time.

5.1 Multilingual Batching for Fine-Tuning

We compare two alternatives for fine-tuning the MTT condition and summarize the results with title+description queries in Table 3. In all cases, mixed-language batches (MTT-M) produce more effective retrieval models than single-language (MTT-S). This is likely because, in MLIR, the model must rank documents from different languages together instead of transferring trained models to other languages. The outcome might be different if our goal were to perform CLIR over monolingual document collections.

⁴ <https://ir-measur.es/>.

Table 3. ColBERT-X MTT for Multiple or Single language training batches, indexing documents in their native language using title+description queries. † indicates significant improvement over MTT-S by paired *t*-test with 3-test Bonferroni correction ($p < 0.05$).

	MAP			P@10		
	2001	2002	2003	2001	2002	2003
MTT-M	0.462†	0.462†	0.461†	0.704	0.752	0.653
MTT-S	0.422	0.405	0.433	0.696	0.702	0.649

5.2 Effectiveness Relative to Baselines

Our main effectiveness results are shown in Table 4. For ColBERT-X and DPR-X, MTT-M consistently improves effectiveness when retrieving documents in their native language (i.e., *without document MT*) compared to English Training (ET). Such improvements are seen in all three query sets, and for both Title (T) and Title+Description (T+D) queries. Differences are larger for MAP than P@10, indicating that MTT-M affects more than just the top ranks.

ColBERT-X MTT-M numerically outperforms MULM for both query types and over all collections in MAP and nearly all collections in P@10. With longer, more fluent title+description queries, ColBERT-X MTT-M gives a larger improvement over MULM in both MAP and P@10, indicating that XLM-R favors queries with more context. Since DPR-X is less effective [48], MTT-M only brings its performance up to par with MULM.

With modern MT models, we can improve MLIR effectiveness. A common, yet strong, baseline of using BM25 to search over translated documents yields substantial improvement over MULM in both MAP and P@10 with both query types. We argue that BM25+ITD is a proper baseline to which future MLIR experiments should be compared.

We can also reduce neural IR to the monolingual case, training our retrieval model with English training and searching documents represented by English machine translations. For both ColBERT and DPR, an English-trained model (ET) indexing translated documents often yields better effectiveness than MTT-M indexing translated documents (ITD). Furthermore, an English trained model indexing translated documents yields better effectiveness than MTT-M indexing documents in their native language; however, these differences are only statistically significant for CLEF 2002 on Title queries using a paired *t*-test with 3-test Bonferroni correction ($p < 0.05$). We observe similar results with ColBERT using the BERT-Large pretrained LM trained under the same conditions except for using a learning rate of 3×10^{-6} (the value suggested by the authors). Compare Table 5 to ColBERT-X with English training, presented in Table 4.

5.3 Preprocessing and Indexing Time

Applying machine translation to entire document collections is expensive. Table 6 summarizes the cost for preprocessing and indexing the collection in GPU-hours for ColBERT-X and BM25. We omit consideration of query latency here since all of our

Table 4. MAP and P@10 on CLEF Title and Title+Description queries. Bold are best among a year; italics are best in a row (*i.e.*, with and without neural machine translation), † indicates significant difference from BM25+ITD by paired *t*-test with 16-test Bonferroni correction ($p < 0.05$).

Query set	ITD	MAP						P@10					
		MULM	BM25	ColBERT-X		DPR-X		MULM	BM25	ColBERT-X		DPR-X	
				MTT-M	ET	MTT-M	ET			MTT-M	ET	MTT-M	ET
Title queries													
2001	✓	–	0.398	0.377	0.391	0.338	0.344	–	<i>0.648</i>	0.612	0.596	0.548	0.584
	✗	0.349	–	<i>0.360</i>	0.322	0.327	0.298†	0.650	–	0.600	0.588	0.592	0.570
2002	✓	–	0.337	0.367	0.389	0.287	0.304	–	0.618	0.606	0.670	0.530	0.596
	✗	0.276	–	<i>0.352</i> †	0.333	0.282	0.277	0.592	–	<i>0.614</i> †	<i>0.622</i>	0.544	0.556
2003	✓	–	0.349	0.337	0.349	0.276†	0.266†	–	0.595	0.542	0.573	0.517	0.497†
	✗	0.305	–	<i>0.332</i> †	0.290	0.273†	0.247†	0.497	–	<i>0.546</i>	<i>0.541</i>	0.527	0.492†
All	✓	–	0.361	0.359	0.375	0.299†	0.302†	–	0.619	0.583	0.611	0.531†	0.554†
	✗	0.310	–	<i>0.347</i>	0.314†	0.293†	0.273†	0.575	–	<i>0.584</i> †	0.581	0.553†	0.536†
Title + Description queries													
2001	✓	–	0.436	0.472	0.477	0.365	0.356	–	0.704	0.718	0.754	0.658	0.650
	✗	0.387	–	<i>0.462</i>	0.405	0.358	0.324†	0.700	–	<i>0.704</i> †	<i>0.744</i>	0.658	0.644
2002	✓	–	0.398	0.470†	0.480 †	0.347	0.332	–	0.696	0.774	0.770	0.664	0.620
	✗	0.347	–	<i>0.462</i>	0.410	0.335	0.310	0.666	–	<i>0.752</i>	0.720	0.672	0.640
2003	✓	–	0.394	0.419	0.410	0.343	0.328†	–	0.615	0.646	0.661	0.620	0.600
	✗	0.376	–	<i>0.409</i>	0.358	0.338	0.302†	0.563	–	<i>0.653</i>	0.637	0.622	0.575
All	✓	–	0.408	0.451†	0.453 †	0.351†	0.338†	–	0.669	0.709	0.725 †	0.646	0.622
	✗	0.368	–	<i>0.442</i>	0.390	0.343†	0.312†	0.643	–	<i>0.700</i> †	0.697	0.639	0.617

Table 5. Monolingual ColBERT model using BERT-Large trained with ET and evaluated with translated documents.

Queries	MAP				P@10			
	2001	2002	2003	All	2001	2002	2003	All
T	0.397	0.367	0.362	0.375	0.592	0.646	0.583	0.606
T+D	0.439	0.413	0.420	0.424	0.736	0.714	0.673	0.706

systems are sufficiently fast at query time for interactive use on collections of this size. We refer the interested reader to Santhanam et al. [41].

This table reveals that differences in total indexing time between searching native and translated documents range from four to 6.5 times depending on collection size and model.⁵ Despite that searching translated documents with monolingual retrieval models is more effective, the computational cost of MT at indexing time is significantly higher; one might choose not to bear this cost in exchange for the small and not statistically significant numerical gain in measured effectiveness over searching documents in their native language with MTT-M fine-tuning for title+description queries.

⁵ Although Marian [23] is faster than Sockeye 2, benchmark results from Sockeye 1 [20] and Sockeye 2 [19] confirm that Sockeye 2 is within a factor of 2 to 3 of Marian’s speed, leaving our conclusions unchanged.

Table 6. ColBERT-X GPU hours for translating and indexing. BM25 does not use GPU.

Model	ITD	CLEF2001-2002			CLEF2003		
		Translation	Index	Total	Translation	Index	Total
BM25	✓	55.0	–	55.0	68.6	–	68.6
ET	✓	55.0	9.3	64.3	68.6	12.3	80.9
	✗	–	9.9	9.9	–	12.4	12.4
MTT-S	✓	55.0	16.9	71.9	68.6	19.0	87.6
	✗	–	16.7	16.7	–	21.9	21.9
MTT-M	✓	55.0	17.3	72.3	68.6	20.1	88.7
	✗	–	15.1	15.1	–	19.3	19.3

We also see this trade-off on a per-document basis. Figure 1 shows that ColBERT-X with English training searching translated documents (*ColBERT-X(ET)+ITD*) achieves the best effectiveness with both title (0.375 MAP) and title+description (0.453 MAP) queries. However, it has a high preprocessing cost of 0.32 s per document, whereas ColBERT-X trained with MTT-M searching documents in their native languages (*ColBERT-X(MTT-M)*) requires under 0.05 s per document. This is an 84% reduction in preprocessing cost at an apparent (but not statistically significant) cost of only 2% in MAP with title+description queries.

6 Analysis

This section investigates our experimental results by breaking down the collection in two ways – by document language, and by topic.

6.1 Language Bias

Since MPLMs are known to exhibit language biases [10, 25], we investigate how retrieval models fine-tuned with our training schemes inherit or alleviate these biases. In MLIR

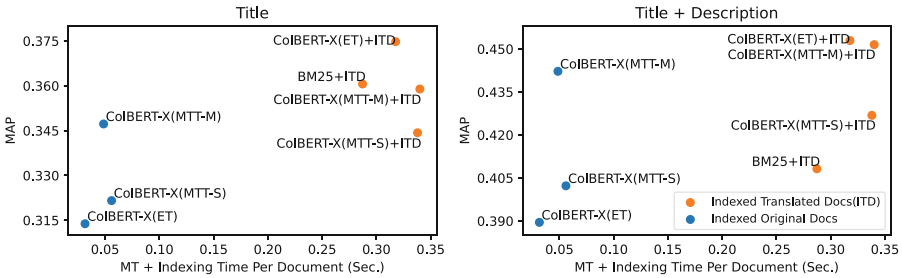


Fig. 1. Effectiveness (MAP) vs. efficiency (per-document GPU indexing time in seconds) trade-off on CLEF 2001–2003. MAP scores (y-axis) for Title and Title+Description queries are disjoint ranges. The upper left is the optimal part of the chart.

we consider a model biased if it ranks a language’s documents systematically higher or lower than those of another language. While MLIR is not a new task, we are not aware of prior work that has examined language bias. Therefore we introduce two approaches to studying this phenomenon. The first approach examines rates of relevant documents. Since relevant documents are unevenly distributed across languages (e.g., Spanish has more than three times as many known relevant documents as English among the CLEF 2001 topics, averaging 54 vs. 17 relevant documents per topic, respectively), meaningful comparisons require us to focus on rates rather than on counts. In this analysis, we focus on Recall@MLIR-Relevant (see Sect. 4.2), illustrating our analysis using the 100 title+description queries in CLEF 2001–2002 topics to characterize the coverage of relevant documents in each language (results on CLEF 2003 topics are similar).

Figure 2 shows distributional statistics of Recall@MLIR-Relevant over topics by language and condition that have at least one known relevant document in that language (96 for German, 97 for Spanish, 94 for Italian, 90 for French, 73 for English). When transferring a ColBERT-X model fine-tuned zero-shot with English training (i.e., ColBERT-X(ET)) to other languages, the model favors English documents due to the fine-tuning condition. This results in a strong language bias in the retrieval results. Such biases can be ameliorated by fine-tuning with MTT. MTT-M appears to have more consistent behavior across languages compared to MTT-S, although the small apparent difference is not statistically significant. When indexing translated documents, Recall@MLIR-Relevant tends to be lower for English compared to other languages (though also not significantly). Since documents were translated sentence-by-sentence, we hypothesize that indexing translated documents provides more synonym variety when decoding similar terms, resulting in document expansion; this hypothesis requires more investigation, which we leave for future work.

An alternative approach to investigating language bias is to assume that in a bias-free approach to MLIR, the scores for relevant documents would be drawn from the same underlying distribution. Using the 2-sample Kolmogorov-Smirnov test, the null hypothesis is that the two samples are drawn from the same distribution. For this analysis, we chose English as a reference and tested each topic with at least three relevant

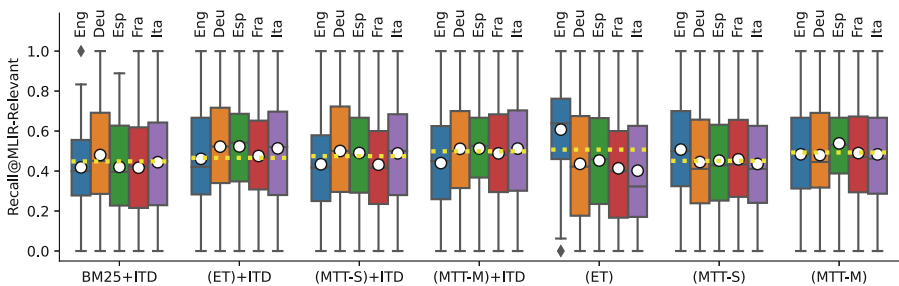


Fig. 2. R@MLIR-Relevant of BM25 and ColBERT-X variants for each language in CLEF2001-2002 with title+description queries. The yellow dashed line is the average over all languages, i.e., the R-Precision in MLIR. Outliers are defined as values beyond $1.5 \times$ interquartile range. Horizontal black bars indicate the median and white circles indicate the mean. (Color figure online)

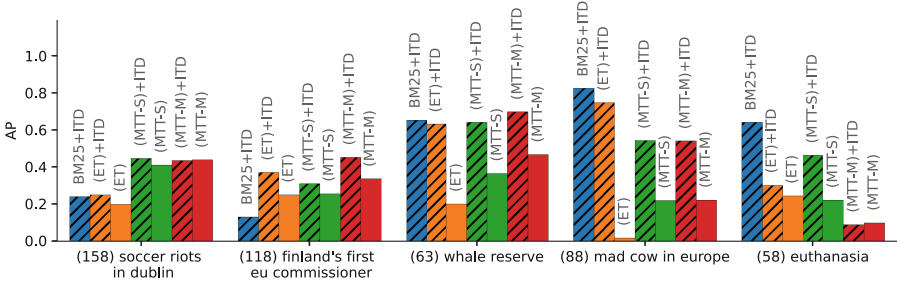


Fig. 3. Average Precision (AP) of BM25 and ColBERT-X on selected topics using title queries.

documents in each language. We then adjusted the p-values to account for multiple comparisons. We found that we could reject the null hypothesis for all languages and all configurations, indicating the document scores are not drawn from the same distribution based on language. Although some of this difference could result from differences in collection statistics (i.e., with some languages better supporting the queries than others based on the numbers of relevant documents), the differences we observe across retrieval models indicate that there are retrieval model effects as well. Notably, ColBERT-X(ET) retrieving documents in the native language has the largest percentage of topics with bias (from 15% to 30% depending on language pair), while all other configurations have no more than 12% of topics exhibiting biased scores. This confirms the qualitative analysis above, which revealed that ColBERT-X(ET) over the documents in their native language had the most skewed rates of relevant documents. Future research will need to address language bias in document scores.

6.2 Example Queries

For more insight into differences among the algorithms, we show effectiveness on individual queries in Fig. 3. Our query selection here is not meant to be representative, but rather illustrative of phenomena that we see. For two topics on which ColBERT-X outperformed BM25 (topics 158 and 118), the queries include terms that likely benefit from ColBERT-X soft term-matching – “soccer” and “commissioner” respectively. This term expansion effect has also been observed in monolingual retrieval with ColBERT.

MT is particularly helpful for topics 63 and 88, likely due to the quality of the translation for documents on these topics. Especially for topic 88, English monolingual retrieval produces strong results. Such behaviors indicate that the multilingual term matching in ColBERT-X is still not as effective on less common concepts like “mad cow” as is machine translation.

Topic 58 is an outlier. The term “euthanasia” is tokenized as a single token for BM25 but separated into `_eu`, `thana`, and `sia` by the XLM-R tokenizer; combined with the minimal context provided by a query, this prevents ColBERT-X from matching properly across languages. Such diverse behaviors suggest room for further MLIR improvements using system combination.

7 Conclusion and Future Work

This paper proposes the MTT training approach to MLIR that uses translated MS MARCO. When searching non-English documents, fine-tuning with MTT using mixed-language batches (MTT-M) enables neural models such as ColBERT and DPR to be more effective than if fine-tuned on English MS MARCO. ColBERT-X with MTT-M is not statistically different from monolingual English models applied to neural indexing-time translation of the collection into English, yet it achieves substantially better indexing time efficiency. These results may not hold for more diverse sets of languages or when MT is less effective; future work will examine the multilingual topics from the TREC 2022 the NeuCLIR track,⁶ which judges the relevance of documents written in Chinese, Persian, and Russian. Our observation that the retrieval method that yields the best retrieval effectiveness is query-dependent suggests future work on system combination, but our focus on efficiency and on language bias also calls attention to issues beyond retrieval effectiveness that will merit consideration in such a study.

A MTT Implementation Details

As described in Sect. 3.2, MTT-M consists of examples with different languages in the training batches. We implement it by mixing the translated MS-MARCO triples round-robin. Specifically, each triple consists of an English query and positive and negative passages translated into the target languages. We constructed such triples using the translated documents provided by mMARCO [6]. Each language results in a triple file of the same structure as `triples.train.small.tar.gz`.⁷ The following Bash command creates a combined triple file that mixes all languages:

```
paste -d '\n' <(cat ./original_msmarco/triples.train.small.tsv) \
          <(cat ./mmarco/french/triples.train.small.tsv) \
          <(cat ./mmarco/german/triples.train.small.tsv) \
          <(cat ./mmarco/italian/triples.train.small.tsv) \
          <(cat ./mmarco/spanish/triples.train.small.tsv) \
| cat > combined.tsv
```

Training with four GPUs and a per-GPU batch size of 32 triples guarantees that each batch consists of examples in different languages based on ColBERT-X's⁸ batching scheme.

For MTT-S, we modified the ColBERT-X batching mechanism to load multiple triple files and supply a batch of examples from only one source file whenever the training process requests one. After each request, we switch the source triple file to ensure all languages are presented equally to the model during training.

⁶ <https://neuclir.github.io/>.

⁷ <https://msmarco.blob.core.windows.net/msmarcoranking/triples.train.small.tar.gz>.

⁸ https://github.com/hltcoe/ColBERT-X/blob/main/xlmr_colbert/training/lazy_batcher.py.

References

1. Aljlal, M., Frieder, O.: Effective Arabic-English cross-language information retrieval via machine-readable dictionaries and machine translation. In: Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 295–302 (2001)
2. Bajaj, P., et al.: MS MARCO: a human generated machine reading comprehension dataset. arXiv preprint [arXiv:1611.09268](https://arxiv.org/abs/1611.09268) (2016)
3. Bendersky, M., Kurland, O.: Utilizing passage-based language models for document retrieval. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 162–174. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_17
4. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc., Sebastopol (2009)
5. Blloshmi, R., Pasini, T., Campolungo, N., Banerjee, S., Navigli, R., Pasi, G.: IR like a SIR: sense-enhanced information retrieval for multiple languages. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 1030–1041, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, November 2021. <https://doi.org/10.18653/v1/2021.emnlp-main.79>, <https://aclanthology.org/2021.emnlp-main.79>
6. Bonifacio, L.H., et al.: mMARCO: a multilingual version of MS MARCO passage ranking dataset. arXiv preprint [arXiv:2108.13897](https://arxiv.org/abs/2108.13897) (2021)
7. Braschler, M.: CLEF 2001 — overview of results. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 9–26. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45691-0_2
8. Braschler, M.: CLEF 2002 — overview of results. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 9–27. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45237-9_2
9. Braschler, M.: CLEF 2003 – overview of results. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 44–63. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30222-3_5
10. Choudhury, M., Deshpande, A.: How linguistically fair are multilingual pre-trained language models? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12710–12718 (2021)
11. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451. Association for Computational Linguistics, Online, July 2020. <https://aclanthology.org/2020.acl-main.747>
12. Costello, C., Yang, E., Lawrie, D., Mayfield, J.: Patapasco: a Python framework for cross-language information retrieval experiments. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 276–280. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_33
13. Dai, Z., Callan, J.: Deeper text understanding for IR with contextual neural language modeling. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 985–988 (2019)
14. Darwish, K., Oard, D.W.: Probabilistic structured query methods. In: Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 338–344 (2003)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186, Association for Computational Linguistics, Minneapolis, June 2019. <https://aclanthology.org/N19-1423>
16. Domhan, T., Denkowski, M., Vilar, D., Niu, X., Hieber, F., Heafield, K.: The Sockeye 2 neural machine translation toolkit at AMTA 2020. In: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pp. 110–115, Association for Machine Translation in the Americas, Virtual, October 2020
 17. Gao, L., Ma, X., Lin, J.J., Callan, J.: Tevatron: an efficient and flexible toolkit for dense retrieval. arXiv preprint [arXiv:2203.05765](https://arxiv.org/abs/2203.05765) (2022)
 18. Granell, X.: Multilingual Information Management: Information, Technology and Translators. Chandos Publishing, Cambridge (2014)
 19. Hieber, F., Domhan, T., Denkowski, M., Vilar, D.: Sockeye 2: a toolkit for neural machine translation. In: EAMT 2020 (2020). <https://www.amazon.science/publications/sockeye-2-a-toolkit-for-neural-machine-translation>
 20. Hieber, F., et al.: Sockeye: a toolkit for neural machine translation. arXiv preprint [arXiv:1712.05690](https://arxiv.org/abs/1712.05690) (2017)
 21. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: industrial-strength natural language processing in Python. Technical report, Explosion (2020)
 22. Hull, D.A., Grefenstette, G.: Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 49–57 (1996)
 23. Junczys-Dowmunt, M., Heafield, K., Hoang, H., Grundkiewicz, R., Aue, A.: Marian: cost-effective high-quality neural machine translation in C++. arXiv preprint [arXiv:1805.12096](https://arxiv.org/abs/1805.12096) (2018)
 24. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781. Association for Computational Linguistics, Online, November 2020. <https://aclanthology.org/2020.emnlp-main.550>
 25. Kassner, N., Dufter, P., Schütze, H.: Multilingual lama: investigating knowledge in multilingual pretrained language models. arXiv preprint [arXiv:2102.00894](https://arxiv.org/abs/2102.00894) (2021)
 26. Khattab, O., Zaharia, M.: 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, pp. 39–48 (2020)
 27. Lawrie, D., Mayfield, J., Oard, D.W., Yang, E.: HC4: a new suite of test collections for ad hoc CLIR. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13185, pp. 351–366. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99736-6_24
 28. MacAvaney, S., Macdonald, C., Ounis, I.: Streamlining evaluation with *ir-measures*. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 305–310. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_38
 29. Magdy, W., Jones, G.J.F.: Should MT systems be used as black boxes in CLIR? In: Clough, P., et al. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 683–686. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-5_70
 30. McCarley, J.S.: Should we translate the documents or the queries in cross-language information retrieval? In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 208–214 (1999)
 31. Mitamura, T., et al.: Overview of the NTCIR-7 ACLIA tasks: advanced cross-lingual information access. In: NTCIR (2008)
 32. Nair, S., et al.: Transfer learning approaches for building cross-language dense retrieval models. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13185, pp. 382–396. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99736-6_26

33. Nie, J.-Y., Jin, F.: A multilingual approach to multilingual information retrieval. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 101–110. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45237-9_8
34. Oard, D.W., Dorr, B.J.: A survey of multilingual text retrieval. Technical report, UMIACS-TR-96019 CS-TR-3615, UMIACS (1996)
35. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, July 2002. <https://doi.org/10.3115/1073083.1073135>, <https://aclanthology.org/P02-1040>
36. Peters, C., Braschler, M.: The importance of evaluation for cross-language system development: the CLEF experience. In: LREC (2002)
37. Peters, C., Braschler, M., Clough, P.: Multilingual Information Retrieval: From Research to Practice. Springer, Heidelberg (2012). <https://doi.org/10.1007/978-3-642-23008-0>
38. Rahimi, R., Shakery, A., King, I.: Multilingual information retrieval in the language modeling framework. *Inf. Retrieval J.* **18**(3), 246–281 (2015). <https://doi.org/10.1007/s10791-015-9255-1>
39. Rehder, B., Littman, M.L., Dumais, S.T., Landauer, T.K.: Automatic 3-language cross-language information retrieval with latent semantic indexing. In: TREC, pp. 233–239. Cite-seer (1997)
40. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: BM25 and beyond. *Found. Trends® Inf. Retrieval* **3**(4), 333–389 (2009)
41. Santhanam, K., Khattab, O., Potts, C., Zaharia, M.: PLAID: an efficient engine for late interaction retrieval. arXiv preprint [arXiv:2205.09707](https://arxiv.org/abs/2205.09707) (2022)
42. Shi, P., Lin, J.: Cross-lingual relevance transfer for document retrieval. arXiv preprint [arXiv:1911.02989](https://arxiv.org/abs/1911.02989) (2019)
43. Si, L., Callan, J., Cetintas, S., Yuan, H.: An effective and efficient results merging strategy for multilingual information retrieval in federated search environments. *Inf. Retrieval* **11**(1), 1–24 (2008)
44. Sorg, P., Cimiano, P.: Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data Knowl. Eng.* **74**, 26–45 (2012). ISSN 0169-023X, <https://www.sciencedirect.com/science/article/pii/S0169023X12000213>, *Appl. Nat. Lang. Inf. Syst*
45. Tsai, M.F., Wang, Y.T., Chen, H.H.: A study of learning a merge model for multilingual information retrieval. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 195–202 (2008)
46. Xu, H., Van Durme, B., Murray, K.: BERT, mBERT, or BiBERT? A study on contextualized embeddings for neural machine translation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6663–6675. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, November 2021. <https://aclanthology.org/2021.emnlp-main.534>
47. Xu, Y.: Global divergence and local convergence of utterance semantic representations in dialogue. In: Proceedings of the Society for Computation in Linguistics 2021, pp. 116–124. Association for Computational Linguistics, Online, February 2021. <https://aclanthology.org/2021.scil-1.11>
48. Yang, E., Nair, S., Chandradevan, R., Iglesias-Flores, R., Oard, D.W.: 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 (SIGIR) (2022). <https://arxiv.org/abs/2204.11989>
49. Zhang, X., Ma, X., Shi, P., Lin, J.: Mr. TyDi: a multi-lingual benchmark for dense retrieval. In: Proceedings of the 1st Workshop on Multilingual Representation Learning, pp. 127–137. Association for Computational Linguistics, Punta Cana, Dominican Republic, November 2021. <https://aclanthology.org/2021.mrl-1.12>