Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models

Suraj Nair$^{1,2[0000-0003-2283-7672]}$, Eugene Yang$^{2[0000-0002-0051-1535]}$,
Dawn Lawrie$^{2[0000-0001-7347-7086]}$, Kevin Duh$^{2[0000-0001-8107-4383]}$,
Paul McNamee$^{2[0000-0002-0548-5751]}$, Kenton Murray$^{2[0000-0002-5628-1003]}$,
James Mayfield$^{2[0000-0003-3866-3013]}$, and Douglas W. Oard$^{1,2[0000-0002-1696-0407]}$

[1] University of Maryland, College Park MD 20742, USA {srnair,oard}@umd.edu
[2] HLTCOE. Johns Hopkins University, Baltimore MD 21211, USA
{eugene.yang,lawrie,mcnamee,kenton,mayfield}@jhu.edu,kevinduh@cs.jhu.edu

Abstract. The advent of transformer-based models such as BERT has led to the rise of neural ranking models. These models have improved the effectiveness of retrieval systems well beyond that of lexical term matching models such as BM25. While monolingual retrieval tasks have benefited from large-scale training collections such as MS MARCO and advances in neural architectures, cross-language retrieval tasks have fallen behind these advancements. This paper introduces ColBERT-X, a generalization of the ColBERT multi-representation dense retrieval model that uses the XLM-RoBERTa (XLM-R) encoder to support cross-language information retrieval (CLIR). ColBERT-X can be trained in two ways. In *zero-shot* training, the system is trained on the English MS MARCO collection, relying on the XLM-R encoder for cross-language mappings. In *translate-train*, the system is trained on the MS MARCO English queries coupled with machine translations of the associated MS MARCO passages. Results on ad hoc document ranking tasks in several languages demonstrate substantial and statistically significant improvements of these trained dense retrieval models over traditional lexical CLIR baselines.

Keywords: CLIR, ColBERT, ColBERT-X, Dense Retrieval

1 Introduction

BERT-style neural ranking models that use cross-attention between query and document terms [7, 16] define the state of the art for monolingual English retrieval. Such models are typically used as rerankers in a retrieve-and-rerank pipeline, due to the quadratic time and space complexity of self-attention in the transformer architecture [30]. Reranking using these models is effective but time-consuming, so the number of documents to be reranked must be tuned to balance the trade-off between effectiveness and efficiency. In contrast to the reranking approach, dense retrieval models encode query and document representations independently and match them with custom similarity functions (e.g., cosine similarity). Dense retrieval complements the lexical first phase retrieval by using an approximate nearest neighbor search over contextualized representations.

While the retrieve-and-rerank framework has been adapted and explored in cross-language information retrieval (CLIR) [37, 11, 38, 3, 36], most approaches translate queries into the language of the documents and perform monolingual retrieval [28, 29]. Dense retrieval models, on the other hand, remain under-explored in CLIR. In this work, we develop an effective dense retrieval model for CLIR.

Dense retrieval models can be broadly categorized into two variants: single-representation and multi-representation [15]. Single-representation models encode queries and documents separately to create a single aggregate representation. However, that can lead to loss of information. Multi-representation models use multiple representations of queries and documents to predict relevance. One such model is ColBERT [13], which computes a similarity between each query term representation and each document term representation. Yet ColBERT is exclusively monolingual. This paper presents *ColBERT-X*, a generalization of the ColBERT approach that supports CLIR. ColBERT-X uses a translate and train fine-tuning approach to exploit existing CLIR training resources.

This generalization poses two challenges: enabling the encoders to process multiple languages, and identifying appropriate resources with which to train the model. To address the former, we adapt XLM-R [5], a multilingual pretrained transformer language model, to initialize the dense retrieval model. For the latter challenge, we use translations of MS MARCO [2], a widely-used passage ranking collection for training monolingual neural retrieval models.

We evaluate ColBERT-X on ad hoc document ranking tasks using English queries to retrieve documents in other languages, exploring two ways to cross the language barrier. In the zero-shot setting, where we lack cross-language training resources, we train the model only on English MS MARCO. In the translate-train setting, the model is trained on machine-generated translations of MS MARCO passages paired with English queries. This paper additionally investigates the effect of machine translation on ColBERT-X retrieval results.

Our main contributions can be summarized as follows:

- We generalize ColBERT to support CLIR and develop a fine-tuning task that leverages translations of existing monolingual retrieval collections.
- We demonstrate significant effectiveness gains over query translation baselines on news in several languages, showing the ability of term-level Approximate Nearest Neighbor (ANN) search to overcome vocabulary mismatch.
- We analyze components of ColBERT-X and techniques to improve effectiveness, including effects of different machine translation models, alternative multilingual encoders, and relevance feedback.
- We release our code to train and evaluate ColBERT-X, and our new machine translations of MS MARCO into Chinese, Persian and Russian.[3]

[3] `https://github.com/hltcoe/ColBERT-X`

2 Related Work

In this section, we briefly review related work on neural retrieval and its extension to cross-lingual settings. For many years, sparse retrieval models such as BM25 [26] and Query Likelihood [24] were the dominant models for ad hoc retrieval tasks. Only in recent years, with the rise of BERT [7] and the availability of large scale retrieval collections such as MSMARCO [2] for training, have neural information retrieval (neural IR) models emerged as the state of the art.

Similar to sparse retrieval models, neural IR models take as input the query and documents, and produce a relevance score. For each query and document pair, matching mechanisms, such as DRMM [9], KNRM [6] or PACCR [10], construct the interaction matrix between the distributed term representations of the query and the documents, and aggregate them into a relevance score. Alternatively, the BERT passage pair classification model [7] considers the query and the document as the input pair, and uses the final classification score as the relevance score [35]. CEDR [18] incorporates contextualized embeddings such as ELMo [23] or BERT [7] into the matching, providing significant effectiveness improvements by taking advantage of contextualization. However, due to the high computational cost, these models are used to rerank top-ranked documents from a sparse retrieval system.

ColBERT [13] further improves efficiency by keeping separate the query-document interaction until the end of the neural architecture. This is called *late interaction*. As opposed to matching mechanisms that require both the query and the document to be present simultaneously, late interaction allows offline encoding of the documents into bags-of-vectors. Document representations are combined with query representations by an efficient *MaxSim* operator, which significantly reduces computation at inference time. This decoupling enables the documents to be encoded offline and indexed to support approximate nearest neighbor search. Further details are discussed in Section 3.

Cross-language transfer learning is important for CLIR. Due to the lack of training data for ad hoc neural retrieval models other than in English, prior work explored zero-shot model transfer to other languages, trained with only English retrieval examples [17, 28]. Model initialization with a multilingual language model such as mBERT [7] has been shown to be effective in zero-shot evaluations. However, this approach requires both queries and documents to be in the same language, resulting in evaluation based either on monolingual non-English retrieval [17], or on query translation into the target language [28].

With the availability of translations of the widely-used English ad hoc retrieval resource MS MARCO [4], translate-train (training the retrieval model on a translated collection) using large ad hoc retrieval collections becomes feasible. Prior work explored a dense retrieval approach to translate-train, showing effectiveness gains on monolingual non-English retrieval tasks [29]. However, this approach relied on a single-representation dense retrieval model with an mBERT encoder, combined with sparse retrieval methods such as BM25. Lacking an end-to-end CLIR dense retrieval model that does not require the help of a sparse retrieval system, we bridge the gap by generalizing ColBERT to support directly querying non-English documents with English queries.

3 ColBERT-X

ColBERT is a multi-stage dense retrieval model that uses monolingual BERT [7] to encode both query and document terms. It employs a late-interaction mechanism, MaxSim, that computes the similarity between the encoded query and document term representations. Computing MaxSim for every query and document term pair in the collection is not feasible, so ColBERT has two ways to reduce the number of required similarity comparisons: reranking or end-to-end retrieval. In reranking, a retrieval system such as BM25 generates an initial ranked list, which is then reranked using ColBERT's MaxSim operation. The disadvantage of such a cascaded pipeline is that the overall recall of the system is limited to the recall of the initial ranked list. In the context of CLIR systems, we face the additional complexity of crossing the language barrier that further affects recall. We thus restricted our work to end-to-end (E2E) retrieval.

In the first stage of the E2E setting, a candidate set of documents is generated by ANN search using every query term. Specifically, the k nearest document tokens are retrieved from the ANN index for every query term representation. These tokens are mapped to document IDs, and the union of these IDs creates the final set of candidate documents. In the next stage, these documents are reranked using the late-interaction "MaxSim" operation. For every query term, MaxSim finds the closest document token using the dot product of the encoded query and document term representation. The final score of the document is the summation of individual query term contributions, as shown in Equation 1. η denotes the monolingual BERT encoder.

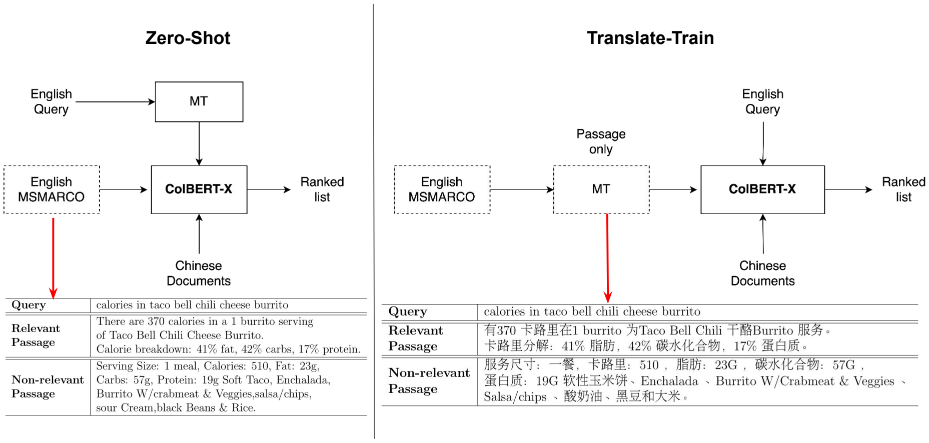
$$s_{q,d} = \sum_{i=1}^{|q|} \max_{j=1..|d|} \eta(q_i) \ . \ \eta(d_j)^T \tag{1}$$

To generalize ColBERT to CLIR, we replaced monolingual BERT with XLM-R. We call the resulting model ColBERT-X. Initializing the encoder to a multilingual model allows retrieval in any language supported by the embeddings. However, these models must be trained before they can be used for CLIR.

3.1 CLIR Training Strategies

ColBERT was trained using pairwise cross-entropy loss on MS MARCO [2] triples, which consist of an English query, a relevant English passage, and a non-relevant English passage. To train ColBERT-X for CLIR, we explored two strategies from the cross-language transfer learning literature:

1. Zero-Shot: This is a common technique in which a multilingual model (e.g., mBERT or XLM-R) is trained in a high-resource language (usually English)



Query	calories in taco bell chili cheese burrito
Relevant Passage	There are 370 calories in a 1 burrito serving of Taco Bell Chili Cheese Burrito. Calorie breakdown: 41% fat, 42% carbs, 17% protein.
Non-relevant Passage	Serving Size: 1 meal, Calories: 510, Fat: 23g, Carbs: 57g, Protein: 19g Soft Taco, Enchalada, Burrito W/crabmeat & Veggies,salsa/chips, sour Cream,black Beans & Rice.

Query	calories in taco bell chili cheese burrito
Relevant Passage	有370 卡路里在1 burrito 为Taco Bell Chili 干酪Burrito 服务。 卡路里分解：41% 脂肪，42% 碳水化合物，17% 蛋白质。
Non-relevant Passage	服务尺寸：一餐，卡路里：510 ，脂肪：23G ，碳水化合物：57G ，蛋白质：19G 软性玉米饼、Enchalada 、Burrito W/Crabmeat & Veggies 、Salsa/chips 、酸奶油、黑豆和大米。

Fig. 1: Two ColBERT-X Transfer Learning Pipelines: Zero-Shot (left) and Translate-Train (right). Dashed boxes denote components used during the training step. In zero-shot, ColBERT-X trained on English MS MARCO is applied on the machine translated queries. With translate-train, the training set consists of translated passages to enable ColBERT-X to cross the language barrier.

and then applied to the document language. In this paper, we first train a ColBERT-X model initialized with an XLM-R encoder on English MS MARCO passage ranking triples. At query time, we use machine translation (MT) to translate the English query to the document language, and use the trained ColBERT-X model to perform retrieval in the document language using Equation 2. $\hat{q}$ is the translated query. Multilingual language models have demonstrated good cross-language generalization in many other natural language processing tasks; we hypothesized it would also work well for CLIR.

$$s_{\hat{q},d} = \sum_{i=1}^{|\hat{q}|} \max_{j=1..|d|} \eta(\hat{q}_i) * \eta(d_j) \tag{2}$$

2. Translate-Train: In this setting, an existing high-resource language (e.g., English) collection is translated to the document language. As in zero-shot training, we choose training triples from the MS MARCO passage ranking collection and use a trained MT model to translate them. Since our focus here is using English queries to retrieve content in non-English languages, we pair the original English queries with machine translations of relevant and non-relevant MS MARCO passages to form new triples.[4] We then train

[4] If we had wanted to experiment with using non-English queries to find English content, we could have instead translated only the MS MARCO queries.

Table 1: Test collection statistics for the CLEF and HC4 newswire collections.

Collection	HC4 Chinese	HC4 Persian	CLEF French	CLEF German	CLEF Italian	CLEF Russian	CLEF Spanish
#documents	646K	486K	129k	294k	157k	16k	454k
#passages	3.6M	3.1M	0.7M	1.6M	0.8M	0.1M	2.7M
#queries	50	50	200	200	200	62	160

ColBERT-X on these newly constructed triples in the same manner as ColBERT.

Figure 1 shows these two pipelines. The key difference is that in the zero-shot setting we have a single ColBERT-X model for a given query language (in this case English) that is used for retrieval in multiple document languages. In the translate-train setting, we train a ColBERT-X model for each query-document language pair. We might also combine translations in multiple languages to train a single multilingual ColBERT-X model, but we leave that for future work.

3.2 Retrieval

While we train ColBERT-X on passages, our goal is to rank documents. We split large documents into overlapping passages of fixed length with a stride. During indexing, we use the trained ColBERT-X model to generate term representations from these passages. These representations are stored in a FAISS-based ANN index [12] and are saved to disk for subsequent MaxSim computation. At query time, we use the trained ColBERT-X model to generate a ranked list of passages for each query using the approaches discussed in the section above and then use a document's maximum passage score as its document score.

4 Experiments

Collection Statistics. Table 1 provides details for the test collections used in our experiments. We worked with several languages from the 2000 to 2003 Cross-Language Evaluation Forum (CLEF) evaluations [22], using ad hoc news collections for French, German, Italian, Russian and Spanish. We also conducted experiments using the new CLIR Common Crawl Collection (HC4) [14], where the documents are newswire articles from Common Crawl in Chinese or Persian. Throughout, English queries are used to search a collection in a non-English language. We experiment with title and description queries. The MS MARCO [2] passage ranking dataset, which we use for training ColBERT-X, consists of roughly 39M training triples, spanning over 500k queries and 8.8M passages.

ColBERT-X Training and Retrieval. Our two ColBERT-X model strategies, zero-shot (ZS) and translate-train (TT), are trained using mostly the same hyperparameters used to train the original ColBERT model.[5] We replaced the BERT encoder with the XLM-RoBERTa (large) encoder provided by the HuggingFace transformers [33] library (but see Section 5.2 for mBERT results). To generate passages from documents, we use a passage length of 180 tokens with a stride of 90 tokens. We index these passages using the trained ColBERT-X model in the same way as the original ColBERT model in the E2E setting.[6]

Machine Translation. For CLEF languages, we use MS MARCO passage translations[7] from Bonifacio *et al.* [4], and the same MT model to translate queries. For the HC4 languages, we use directional MT models built on top of a transformer base architecture (6-layer encoder/decoder) using Sockeye [8]. To produce translations of MS MARCO, the original passages were split using *ersatz* [32], and sentence-level translation was performed using the trained MT model.

Baselines. We compare these strategies with several baselines:

- Human Translation: Monolingual retrieval using Anserini BM25 [34] with the document-language queries provided in the test collection.
- Query Translation: BM25 retrieval using translated queries produced by a specific MT model and original documents in the target language.[8]
- Reranking: We rerank query translation baseline results using the public mT5 reranker[9] trained on translated MS MARCO in 8 languages [4].

Evaluation. We evaluate ranking using Mean Average Precision (MAP). Differences in means are tested for significance using a paired t-test ($p<0.05$) with Holm-Bonferroni multiple test correction.

Results. Table 2 compares the effectiveness of our models to the baselines. Our main finding is that both ColBERT-X variants perform better than BM25 query translation in general. ColBERT-X trained using English MS MARCO alone performs better than query translation and fine-tuning ColBERT-X on translated MS MARCO data helps improve effectiveness further. These gains are statistically significant in both HC4 collections, and for many CLEF collections.

We also compare the ColBERT-X variants to the multilingual T5 reranker that reranks the query translation baseline output. In each collection, ColBERT-X performs consistently and significantly better than the reranker. This is particularly interesting in CLEF collections since both the mT5 reranker and ColBERT-X (TT) were trained on the same MS MARCO translations. However, the

[5] We increase our batch size from 32 to 128.

[6] `https://github.com/stanford-futuredata/ColBERT#indexing`

[7] `https://github.com/unicamp-dl/mMARCO`

[8] To compare the retrieval models fairly, we use the same MT model to translate the queries as the one used to translate the MS MARCO passages.

[9] `https://huggingface.co/unicamp-dl/mt5-base-multi-msmarco`

Table 2: Effectiveness results (mean average precision) for CLIR HC4 and CLEF collections using title queries. Statistically significant improvements over the query translation and reranking baselines are marked with * and † respectively. Bold indicates best MAP among the query translation and reranking methods.

Collection(→)	HC4	HC4	CLEF	CLEF	CLEF	CLEF	CLEF
Model(↓)	Chinese	Persian	French	German	Italian	Russian	Spanish
human translation							
BM25	0.301	0.276	0.403	0.304	0.350	0.452	0.452
ColBERT-X (ZS)	0.510	0.343	0.401	0.360	0.328	0.479	0.418
query translation							
BM25	0.237	0.211	0.387	0.263	0.275	0.377	0.405
reranker							
BM25+mT5-multi	0.312	-	0.333	0.297	0.279	0.303	0.370
our methods							
ColBERT-X (ZS)	**0.450***†	0.297*	0.382†	0.328*†	0.272	**0.418**†	0.379
ColBERT-X (TT)	0.408*†	**0.310***	**0.422**†	**0.397***†	**0.339***†	0.410†	**0.415**†

reranker was trained on a combined dataset in 8 languages, which might point to the curse of multilinguality [5].

When we compare the two variants of ColBERT-X, we observe that on average translate-train often does better than zero-shot, but these differences are only significant in CLEF collections except Russian and not in HC4 collections. The difference is likely a result of using different MT models in CLEF and HC4 collections, so we conduct further analysis in the next section.

5 Detailed Analysis

This section considers several aspects of ColBERT-X. First, different machine translation models are compared using both MT and CLIR measures. Second, effects of different multilingual encoders are explored. Third, the impact of pseudo-relevance feedback is examined. Then the influence of query length on performance is considered. Finally, ColBERT-X costs in terms of index size are noted.

5.1 Effect of Machine Translation

ColBERT-X utilizes machine translation in two different ways depending on whether it is trained using the zero-shot strategy or the translate-train strategy. In the zero-shot strategy, the queries are translated to the document language at query time, while the translate-train strategy requires an MT system to translate the monolingual training corpus (in this case, the MS MARCO passages) to the document language. The MT systems used to produce translations include:

Table 3: BLEU scores for translation systems using WMT'19 newstest for Chinese and Russian, and TICO-19 (from OPUS[11]) for Persian. These are computed on test sets distinct from the CLIR collections, so absolute BLEU score is not an exact reflection of quality of translations in CLIR experiments. Nevertheless, relative comparison of BLEU scores among MT systems is meaningful.

Language	Russian	Chinese	Persian
Benchmark	*newstest'19*	*newstest'19*	*tico-19*
OpusMT	26.3	14.6	-
SockeyeMT1	32.1	25.8	4.4
SockeyeMT2	**35.9**	**38.6**	**20.2**

- OpusMT – bidirectional MT model(s) with MarianNMT as the base architecture,[10] released by the Helsinki NLP group from Bonifacio *et al.* [4].
- SockeyeMT1 – MT model built on top of a transformer base architecture (6-layer encoder/decoder) trained on bitext. Depending on language, these include publicly available bitext such as OpenSubtitles, UN Corpus, Europarl, and WMT. The model is trained using AWS Sockeye v2 [8].
- SockeyeMT2 – identical model to SockeyeMT1 but trained with 2x–3x more bitext. The number of training sentence pairs for MT1 vs MT2 are 51M vs 120M for Russian, 36M vs 85M for Chinese, and 6M vs 11M for Persian.

Table 3 provides an intrinsic comparison of the systems translating from English on a translation task using BLEU scores [21]. For Russian and Chinese we evaluated using a recent WMT shared task (newstest'19); for Persian we evaluated with a collection of around 3000 sentences about COVID-19, as no WMT test is available. Scores were calculated with *sacrebleu* [25] using the `-lc` setting. The table reveals that SockeyeMT outputperforms OpusMT and that exposing SockeyeMT to more training data improves the BLEU score.

Table 4 shows that improving BLEU scores likely leads to improvements in CLIR for both training strategies. Table 4a shows the results of translating queries in the zero-shot strategy. While BLEU improvements tend to be realized downstream, this is not seen for HC4 Chinese where OpusMT has better MAP than SockeyeMT1. Note that asking MT to translate keyword queries may not align well with how the systems were trained with complete sentences.

Table 4b shows results for using different translation models on MS MARCO triples, and the effect this has on ColBERT-X retrieval as measured using MAP. Again, we see that the MAP scores tend to improve with improved BLEU; however, in this case the improvement in Russian BLEU from Table 3 between SockeyeMT1 and SockeyeMT2 does not carry over to ColBERT-X, where the performance is essentially the same. Generally, one can expect that improving MT quality will lead to improved effectiveness of ColBERT-X.

[10] `https://huggingface.co/Helsinki-NLP`

[11] `https://opus.nlpl.eu/`

Table 4: MAP using different MT models for ColBERT-X.

MT model	CLEF Russian	HC4 Chinese	HC4 Persian
OpusMT	0.418	0.411	-
SockeyeMT1	0.442	0.391	0.230
SockeyeMT2	**0.461**	**0.450**	**0.297**

(a) ColBERT-X zero-shot

MT model	CLEF Russian	HC4 Chinese	HC4 Persian
OpusMT	0.410	0.365	-
SockeyeMT1	**0.459**	0.389	0.287
SockeyeMT2	0.456	**0.408**	**0.310**

(b) ColBERT-X translate-train

Table 5: MAP scores for ColBERT-X initialized with the mBERT and XLM-R encoders, and trained on SockeyeMT1 MS MARCO translations.

Multilingual Model	CLEF Russian	HC4 Chinese	HC4 Persian
mBERT	0.341	0.284	0.173
XLM-R	**0.459***	**0.389***	**0.287***

5.2 Effect of Multilingual Language Models

Comparing different multilingual encoders to initialize ColBERT-X, we observe that XLM-R performs significantly better than mBERT, as shown in Table 5. While this might be unsurprising given that the XLM-R model is twice as large and was pretrained on more data than mBERT, tokenization differs across the languages. Considering the case of Chinese, mBERT tokenization produces character-level tokens, whereas the XLM-R tokenizer generates subwords (sentencepieces). This also implies that mBERT indexes are larger than XLM-R indexes, resulting from the term-level storage requirements of ColBERT-X model.

5.3 Pseudo-Relevance Feedback

Pseudo-relevance feedback (PRF) is a form of query expansion that adds discriminative terms extracted from retrieved documents. While PRF has been explored for pre and post translation query expansion [19], here we choose cross-language expansion terms using the ColBERT-X term representation, as suggested by Wang et al [31]. First, feedback documents (fb-docs) are selected from the top of a ColBERT E2E ranked list. Next, embeddings of terms from the feedback documents are clustered into k clusters. The top ranked centroids of these k clusters[12] by token IDF are used as feedback embeddings (fb-embs). These fb-embs are added to the original query and ColBERT E2E is run again to produce the final ranked list. We extend this approach to the ColBERT-X CLIR setting.

[12] Each centroid is mapped to the nearest document token using the ANN index.

Table 6: MAP for query translation BM25 and ColBERT-X translate-train. * or † denote significant improvement over BM25+PRF or ColBERT-X respectively

Retrieval Model	CLEF French	CLEF German	CLEF Italian	CLEF Spanish
baseline				
BM25	0.387	0.263	0.275	0.405
ColBERT-X	0.422	0.397	0.339	0.415
with PRF				
BM25	0.410	0.321	0.320	**0.438**
ColBERT-X	**0.459***†	**0.406***†	**0.371***†	0.436†

To better understand the effect of PRF, we compare ColBERT-X translate-train and query translation BM25, with and without PRF. For BM25, we use Anserini's RM3 to perform PRF, with default hyperparameters. For ColBERT-X PRF, we extend Terrier's [20] implementation[13] with default hyperparameters. Table 6 shows the effect of PRF on ColBERT-X translate-train MAP. Except in Spanish, applying PRF to ColBERT-X significantly improves effectiveness compared to ColBERT-X without PRF or compared to BM25 with PRF.

5.4 Effect of Longer Queries

Table 7 analyzes the effect of query type on ColBERT-X translate-train. We compare three representations: *title* (t), which is a short Web-like query; *description* (d), a well-formed sentence describing the information need, and *title+description* (td), the concatenation of the two. Longer queries pose a problem for ColBERT-X, however, since the model only supports queries up to 32 tokens long. To mitigate this problem, we use a list of "stop structures"[1] consisting of phrases (e.g., find documents on, reports of, etc.), which have been shown to work in the past, removing them from the td queries. We observe that td with stop structures removed leads to significant improvements over t or d alone.

5.5 Indexing Footprint

In addition to the FAISS-based ANN index, ColBERT-X requires access to the representation of each term to compute MaxSim. With each term embedded as a 128-dimensional vector and each dimension using 16-bits, that's 256 bytes per term. These are onerous requirements, with index size increasing with collection size. Table 8 provides statistics on storage requirements. ColBERTv2 [27] addresses this issue by clustering token embeddings. That approach could be extended to ColBERT-X for CLIR; we leave it for future work. An artifact of our

[13] `https://github.com/terrierteam/pyterrier_colbert`

Table 7: MAP results for ColBERT-X (TT) model using different query representations. * and † denote significant improvements over t and d queries respectively.

Query Representation	CLEF French	CLEF German	CLEF Italian	CLEF Spanish
title	0.422	0.397	0.339	0.415
description	0.434	0.410	0.380	0.456
title+description	**0.507***†	**0.466***†	**0.424***†	**0.500***†

design that affects index size is how passages are generated. We use a sliding window of document tokens, so most tokens have two representations. In the future, we will explore the effects of alternative document segmentation approaches.

6 Conclusion

We have developed ColBERT-X, a cross-language generalization of ColBERT that uses a multilingual query and document encoder to improve CLIR beyond what traditional systems such as BM25 can achieve. Using MT systems to translate MS MARCO, we create CLIR collections for training ColBERT-X. We additionally analyze the effect of MT on the CLIR task. In the future, we would like to create a single multilingual model that is trained on the data from many languages, and compare that with a separate models for each language. For pseudo-relevance feedback, it is important to understand which type of queries benefit from it; a per-query comparison could shed some light on that question.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9117. Views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, expressed or implied, of ODNI, IARPA, or the U.S. Government (USG). The USG is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Table 8: Collection-specific memory footprint.

Collection	HC4 Chinese	HC4 Persian	CLEF French	CLEF German	CLEF Italian	CLEF Russian	CLEF Spanish
#passages	3.6M	3.1M	0.7M	1.6M	0.8M	0.1M	2.7M
Disk Space	154GB	134GB	33GB	70GB	36GB	4.7GB	117GB

References

1. Allan, J., Callan, J., Croft, W.B., Ballesteros, L., Byrd, D., Swan, R., Xu, J.: INQUERY does battle with TREC-6. NIST Special Publication 500-240 pp. 169–206 (1998)
2. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al.: MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv preprint arXiv:1611.09268v3 (2018)
3. Bonab, H., Sarwar, S.M., Allan, J.: Training effective neural CLIR by bridging the translation gap. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 9–18. Association for Computing Machinery, New York, NY, USA (Jul 2020)
4. Bonifacio, L.H., Campiotti, I., Lotufo, R., Nogueira, R.: mMARCO: A multilingual version of MS MARCO passage ranking dataset. arXiv preprint arXiv:2108.13897 (2021)
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: 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 (Jul 2020)
6. Dai, Z., Xiong, C., Callan, J., Liu, Z.: Convolutional neural networks for soft-matching n-grams in ad-hoc search. In: Proceedings of the 11th ACM International Conference on Web Search and Data Mining. pp. 126–134 (2018)
7. 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, Minnesota (Jun 2019)
8. 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 (Oct 2020)
9. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 55–64 (2016)
10. Hui, K., Yates, A., Berberich, K., de Melo, G.: PACRR: A position-aware neural IR model for relevance matching. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1049–1058. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
11. Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., Zhao, L.: Cross-lingual information retrieval with BERT. arXiv preprint arXiv:2004.13005 (Apr 2020)
12. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. arXiv preprint arXiv:1702.08734 (2017)
13. 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. Association for Computing Machinery, New York, NY, USA (Jul 2020)

14. Lawrie, D., Mayfield, J., Oard, D.W., Yang, E.: HC4: A new suite of test collections for ad hoc CLIR. In: Proceedings of the 44th European Conference on Information Retrieval (2021)
15. Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: Bert and beyond. Synthesis Lectures on Human Language Technologies **14**(4), 1–325 (2021)
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
17. MacAvaney, S., Soldaini, L., Goharian, N.: Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. Advances in Information Retrieval p. 246 (2020)
18. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: Cedr: Contextualized embeddings for document ranking. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1101–1104 (2019)
19. McNamee, P., Mayfield, J.: Comparing cross-language query expansion techniques by degrading translation resources. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 159–166 (2002)
20. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: European Conference on Information Retrieval. pp. 517–519. Springer (2005)
21. 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, Pennsylvania, USA (Jul 2002)
22. Peters, C., Braschler, M.: European research letter: Cross-language system evaluation: The CLEF campaigns. Journal of the American Society for Information Science and Technology **52**(12), 1067–1072 (2001)
23. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of NAACL-HLT. pp. 2227–2237 (2018)
24. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 275–281 (1998)
25. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Brussels, Belgium (Oct 2018)
26. Robertson, S.E., Walker, S., Jones, S., et al.: Okapi at TREC-3. In: Overview of the Third Text REtrieval Conference (TREC-3) (1995)
27. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: Colbertv2: Effective and efficient retrieval via lightweight late interaction. arXiv preprint arXiv:2112.01488 (2021)
28. Shi, P., Lin, J.: Cross-lingual relevance transfer for document retrieval. arXiv preprint arXiv:1911.02989 (2019)
29. Shi, P., Zhang, R., Bai, H., Lin, J.: Cross-lingual training with dense retrieval for document retrieval. arXiv preprint arXiv:2109.01628 (2021)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
31. Wang, X., Macdonald, C., Tonellotto, N., Ounis, I.: Pseudo-relevance feedback for multiple representation dense retrieval. arXiv preprint arXiv:2106.11251 (2021)

32. Wicks, R., Post, M.: A unified approach to sentence segmentation of punctuated text in many languages. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3995–4007. Association for Computational Linguistics, Online (Aug 2021)
33. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020)
34. Yang, P., Fang, H., Lin, J.: Anserini: Enabling the use of Lucene for information retrieval research. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1253–1256. SIGIR '17, Association for Computing Machinery, New York, NY, USA (Aug 2017)
35. Yang, W., Zhang, H., Lin, J.: Simple applications of BERT for ad hoc document retrieval. arXiv preprint arXiv:1903.10972 (2019)
36. Yu, P., Allan, J.: A study of neural matching models for cross-lingual ir. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1637–1640. Association for Computing Machinery, New York, NY, USA (2020)
37. Zhang, R., Westerfield, C., Shim, S., Bingham, G., Fabbri, A., Verma, N., Hu, W., Radev, D.: Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. arXiv preprint arXiv:1906.03492 (2019)
38. Zhao, L., Zbib, R., Jiang, Z., Karakos, D., Huang, Z.: Weakly supervised attentional model for low resource ad-hoc cross-lingual information retrieval. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). pp. 259–264. Association for Computational Linguistics, Hong Kong, China (Nov 2019)