Distillation for Multilingual Information Retrieval

Eugene Yang
eugene.yang@jhu.edu
HLTCOE, Johns Hopkins University
Baltimore, Maryland, USA

Dawn J. Lawrie
lawrie@jhu.edu
HLTCOE, Johns Hopkins University
Baltimore, Maryland, USA

James Mayfield
mayfield@jhu.edu
HLTCOE, Johns Hopkins University
Baltimore, Maryland, USA

ABSTRACT

Recent work in cross-language information retrieval (CLIR), where queries and documents are in two different languages, has demonstrated the benefit of training a cross-language neural dual-encoder model with translation and distillation using a framework called Translate-Distill. However, Translate-Distill only supports a single document collection language. Multilingual information retrieval (MLIR), which ranks a multilingual document collection based on a query in yet another language, is significantly harder to train than CLIR, since the model needs to order documents in different languages based only on relevant information no matter the language in which the information is expressed. In this work, we extend Translate-Distill and propose Multilingual Translate-Distill (MTD) for MLIR. We show that ColBERT-X models trained with MTD outperform their counterparts trained with Multilingual Translate-Train, which is the previous state-of-the-art training approach, by 5% to 25% in nDCG@20 and 15% to 45% in MAP. We also show that the model is robust to the way languages are mixed in the training batches. Our implementation is available on GitHub.

CCS CONCEPTS

• **Information systems → Language models**; **Multilingual and cross-lingual retrieval**; Retrieval effectiveness.

KEYWORDS

Dense retrieval, multilingual training, dual encoder architecture

ACM Reference Format:
Eugene Yang, Dawn J. Lawrie, and James Mayfield. 2024. Distillation for Multilingual Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington D.C., USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Multilingual Information Retrieval (MLIR) refers to search over a document collection written in several languages to produce a single ranked list [29, 42, 44, 49, 50]. The retrieval system must retrieve and rank documents based only on query relevance, independent of the document language. Such search problems are challenging in part because cross-language systems may not be able to exploit surface forms. While there is a range of multilingual search problems, here we use the term MLIR to refer to queries in one language searching a *multilingual* collection of *monolingual* documents. Our evaluation includes CLEF data [5] with English queries and French, German, Spanish, and English documents; CLEF data [2–5] with English queries and French, German, and Italian documents; and TREC NeuCLIR data [26, 27] with English queries and Chinese, Persian, and Russian documents.

Dual-encoder retrieval models such as ColBERT [23] that matches token embeddings, and DPR [22] that matches query and document embeddings, have shown good results in both monolingual [46] and cross-language [32, 37, 53, 55] retrieval. These approaches use pre-trained language models like multilingual BERT [10] and XLM-RoBERTa [6] as text encoders to place queries and documents into a joint semantic space; this allows embedding distances to be calculated across languages. Multilingual encoders are generally trained monolingually on multiple languages [7, 10], which leads to limited cross-language ability. Therefore, careful fine-tuning, such as Translate-Train [37], C3 Pretraining [54] and Native-Train [38], are essential to be able to match across languages [32, 48, 53].

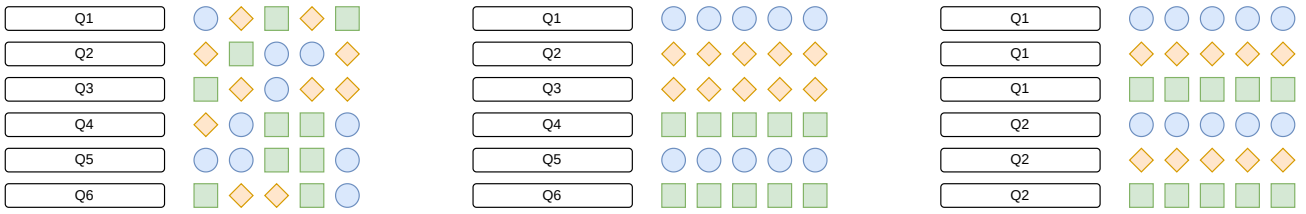
Generalizing from one to multiple document languages is not trivial. Prior work showed that training multilingual ColBERT-X using training data translated into all document languages, with Multilingual Translate-Train (MTT) [29], is more effective than BM25 search over documents translated into the query language. However, using the English ColBERT model to search translated documents is still more effective than using MTT, which still incurs a high translation cost on each search collection at indexing time compared to the amortizable cost on the translating the training corpus for MTT. In this work, we aim to close this gap by developing training that produces more effective models than its monolingual English counterparts.

As knowledge distillation has shown success monolingually [12, 43, 46], we adapt this concept to train MLIR models. In Translate-Distill [53], a way to train CLIR ColBERT-X models, a teacher model scores monolingual training data using text in the language that produces its best results. Then when training the student ColBERT-X model, training data is translated into the languages that match the final CLIR task. Prior work [53] showed that the student model is on par with or more effective than a retrieve-and-rerank system that uses the same teacher model as a reranker. We propose Multilingual Translate-Distill (MTD), a multilingual generalization of Translate-Distill. Instead of training with a single document language, we translate training passages into all document languages. This opens a design space of how to mix languages in training batches.

This paper contributes the following: (1) an effective training approach for an MLIR dual-encoder that combines translation and distillation; (2) models trained with the proposed MTD are more

SIGIR '24, July 14–18, 2024, Washington D.C., USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) Mix Passages **(b) Mix Entries** **(c) Round Robin Entries**

Figure 1: Three language mixing strategies for Multilingual Translate-Distill. Each row indicates an entry with a query and a list of sampled passages in the training mini-batch. Circles, diamonds, and squares represent different document languages.

effective than the previously reported state-of-the-art MLIR model, which is ColBERT-X trained with MTT, and (3) a robustness analysis of mini-batch passage mixing strategies. Models and implementation are available on Huggingface Models[1] and GitHub[2].

2 BACKGROUND

An IR problem can be "multilingual" in several ways. For example, Hull and Grefenstette [18] described a multilingual IR problem of monolingual retrieval in multiple languages, as in Blloshmi et al. [1], or alternatively, multiple CLIR tasks in several languages [3–5, 28, 36]. We adopt the Cross-Language Evaluation Forum (CLEF)'s notion of MLIR: using a query to construct one ranked list across documents in several languages [41]. We acknowledge that this definition excludes mixed-language or code-switched queries and documents, other cases to which "multilingual" has been applied.

Prior to neural retrieval, MLIR systems generally relied on cross-language dictionaries or machine translation models [9, 24, 35]. Translating documents into the query language casts MLIR as monolingual in that language [14, 33, 44]. While translating queries into each document language is almost always computationally more economical than translating the documents, it casts the MLIR problem as multiple monolingual problems whose results must be merged to form the final MLIR ranked list [42, 49, 50]. Moreover, quality differences between translation models could bias results by systematically ranking documents in some languages higher [17, 29].

Recent work in representation learning for IR [12, 13, 45] and fast dense vector search algorithms [19, 21, 34] spawned a new class of models called dual-encoders. These models encode queries and documents simultaneously into one or more dense vectors representing tokens, spans, or entire sequences [22, 23, 30, 31]. While replacing the underlying language model with a multilingual one, such as multilingual BERT [10] and XLM-RoBERTa [7], produces systems that accept queries and documents in multiple languages, zero-shot transfer of a model trained only monolingually to a CLIR or MLIR problem is suboptimal; it leads to systems even less effective than BM25 over document translations [29, 37]. Therefore, designing an effective fine-tuning process for transforming multilingual language models into multilingual IR models is critical.

Various retrieval fine-tuning approaches have been explored, such as contrastive learning [22, 23, 46], hard-negative mining [12, 16], and knowledge distillation [12, 43, 46]. Knowledge distillation has demonstrated more effective results in both monolingual and cross-language IR [32, 53] than the others. The recently proposed Translate-Distill approach decoupled the input languages of the teacher and student models. This allowed large English rerankers to train ColBERT-X for CLIR, leading to state-of-the-art CLIR effectiveness measured on the NeuCLIR 22 benchmark [26]. Recent work by Huang et al. [17] proposes a language-aware decomposition for prompting (or augmenting) the document encoder. In this work, we explore the simple idea of relying on translations of MS MARCO and distilling the ranking knowledge from a large MonoT5 model with mT5XXL underneath [20, 40, 52].

3 MULTILINGUAL TRANSLATE-DISTILL

Our proposed Multilingual Transalte-Distill (MTD) training approach requires a monolingual training corpus consisting of queries and passages; no relevance labels are required.

3.1 Knowledge Distillation

To train a student dual-encoder model for MLIR, we first use two teacher models: a query-passage selector and a query-passage scorer. Following Yang et al. [53], the query-passage selector retrieves k passages for each query. This can be replaced by any hard-negative mining approach [16, 43] or by adapting publicly available mined passages.[3] The query-passage scorer then scores each query-passage pair with high accuracy. The scorer is essentially a reranker from which we would like to distill ranking knowledge implicit in an expensive model such as MonoT5 [40] that is generally too slow to apply by itself. The final product from the two teachers is a set of tuples, each containing a query, a passage, and the associated teacher score. We use these data to train the student dual-encoder model. Specifically, for each training mini-batch of size n, we select n training queries and sample m retrieved passage IDs. To teach the student model to rank documents across languages, we translate each passage into all of the target languages. When constructing the mini-batch, we determine the language for each passage ID, which we discuss in more detail in the next section. Finally, the loss function is the KL divergence between the teacher and student scores on the query and the translated passages.

[1] https://huggingface.co/collections/hltcoe/multilingual-translate-distill-66280df75c34dbbc1708a22f
[2] https://github.com/hltcoe/colbert-x

[3] For example, https://huggingface.co/datasets/sentence-transformers/msmarco-hard-negatives.

Table 1: Collection Statistics

	CLEF		NeuCLIR	
	Subset[17]	2003	2022	2023
Languages	de, fr, it	de, fr, es, en	zh, fa, ru	
# of Docs	0.24M	1.05M	10.04M	
# of Passages	1.90M	6.96M	58.88M	
# of Topics	113	60	41	65
Avg. Rel/Topic	40.73	102.42	125.46	67.77

3.2 Language Mixing Strategies

To train an effective ColBERT-X model for MLIR, each training batch must include documents in more than one language [29]. Training with MTD opens a design space for selecting languages for the mini-batch passages. We experiment with three mixing strategies (see Figure 1):

Mix Passages. In each training batch entry, all passages are randomly assigned to one of the document languages. In this case, each language is equally likely to be present during training. Each language also has an equal probability of being assigned to any passage in such a way that language representation is balanced, thus a language is just as likely to be assigned to a passage with a high score as a low score. This mixing method directly trains the student model to rank passages in different languages.

Mix Entries. Alternatively, we can assign the same randomly selected language to all passages associated with a query. This method ensures the translation quality does not become a possible feature that the student model could rely on if there is a language with which the machine translation model struggles. While not directly learning MLIR, this model jointly learns multiple CLIR tasks with distillation and eventually learns the MLIR task.

Round Robin Entries. To ensure the model equally learns the ranking problem for all languages, we experiment with training query repetition to present passages from all languages. In this case, the model learns the CLIR tasks using the same set of queries instead of a random subset when mixing entries. However, this reduces the number of queries per mini-batch given some fixed GPU memory size. Given this memory constraint, round robin may not be feasible if the number of document languages exceeds the number of entries the GPU can hold at once.

4 EXPERIMENTS

We evaluate our proposed model on four MLIR evaluation collections: a subset of CLEF00-03 curated by Huang et al. [17][4]; CLEF03 with German, French, Spanish, and English [5]; and NeuCLIR 2022 [26] and 2023 [27]. Collection statistics are summarized in Table 1. Queries are English titles concatenated with descriptions.

We use MS MARCO [39] to train the MLIR ColBERT-X models with MTD, for which we adopt the PLAID-X implementation released by Yang et al. [53].[5] We use the English ColBERTv2 model released by Santhanam et al. [46] that was also trained with knowledge distillation[6] and MonoT5 with mT5XXL released by Jeronymo et al. [20][7] as query-passage selector and scorer, respectively. Both the selector and the scorer received English MS MARCO queries and passages to generate training teacher scores.

To support MTD training, we translated the MS MARCO passages with Sockeye v2 [11, 15] into the document languages. Student ColBERT-X models are fine-tuned from the XLM-RoBERTa large models [7] using 8 NVidia V100 GPUs (32GB memory) for 200,000 gradient steps with a mini-batch size of 8 entries each associated with 6 passages on each GPU. We use AdamW optimizer with a 5×10^{-6} learning rate and half-precision floating points.

Documents are split into 180 token passages with a stride of 90 before indexing. The number of resulting passages is reported in Table 1. We index the collection with PLAID-X using one residual bit. At search time, PLAID-X retrieves passages, and document scores are aggregated using MaxP [8]. For each query, we return the top 1000 documents for evaluation.

To demonstrate MTD effectiveness, we report baseline ColBERT models that are trained differently: English ColBERT [46], ColBERT-X with Multilingual Translate-Train (MTT) [29], and ColBERT-X with English Distillation (ED). Since English ColBERT does not accept text in other languages, we index the collection with documents machine-translated into English (marked "DT" in Table 2). ColBERT-X models trained with MTT use the training triples released by MS MARCO with hyperparameters similar to the MTD ones except for the number of queries per batch per GPU is increased to 32. Finally, the English Distillation models are only exposed to English queries and passages during fine-tuning instead of the translated text. It performs a zero-shot language transfer at indexing and search time.

We also compare our models to the recently published KD-SPD [17], which is a language-aware MLIR model that encodes the entire text sequence as a single vector. To provide a broader context, we report sparse retrieval baselines PSQ-HMM [9, 51] and BM25 with translated documents, which are two strong MLIR baselines reported in NeuCLIR 2023 [27].

We report nDCG@20, MAP, and Recall at 1000 for the CLEF03 and NeuCLIR collections. To enable comparison to Huang et al. [17], we report nDCG@10, MAP@100, and Recall@100 on the CLEF00-03 subset. To test statistical superiority between two systems, we use a one-sided paired t-test with 95% confidence on the per-topic metric values. When testing for statistical "equivalence" where the null hypothesis is that the effectiveness of the two systems differ, we use a paired Two One-sided T-Tests (TOST) [25, 47] with a threshold of 0.05 and 95% confidence.

5 RESULTS

Table 2 summarizes our experiments. ColBERT-X models trained with MTD are more effective than those with MTT across all four evaluation collections, demonstrating a 5% (CLEF03 0.643 to 0.675 with mix passages) to 26% (NeuCLIR22 0.375 to 0.474 with round robin entries) improvement in nDCG@20 and 15% (CLEF03 0.451 to 0.520 with mix passages) to 47% (NeuCLIR22 0.236 to 0.347 with mix entries) in MAP. MTD-trained ColBERT-X models over documents

[4]The collection is reconstructed by using the author-provided document IDs, which excludes a large portion of unjudged documents. Documents added in subsequent years are also excluded. Thus some judged relevant documents are also excluded.
[5]https://github.com/hltcoe/ColBERT-X

[6]https://huggingface.co/colbert-ir/colbertv2.0
[7]https://huggingface.co/unicamp-dl/mt5-13b-mmarco-100k

Table 2: MLIR system effectiveness. Numbers in superscripts indicate the system of the row is statistically better than the systems in the superscript with 95% confidence by conducting a one-sided paired t-test. Numbers in subscripts indicate the system of the row is statistically *identical* within 0.05 in value to the systems in the subscripts with 95% confidence by conducting paired TOSTs. Bonferroni corrections are applied to both sets of statistical tests.

	CLEF00-03 Subset [17]			CLEF 2003			NeuCLIR 2022 MLIR			NeuCLIR 2023 MLIR		
Measure	nDCG	MAP	Recall	nDCG	MAP	Recall	nDCG	MAP	Recall	nDCG	MAP	Recall
Rank Cutoff	10	100	100	20	1000	1000	20	1000	1000	20	1000	1000
Baselines												
(0) KD-SPD[17]	0.416	0.220	0.469	–	–	–	–	–	–	–	–	–
(1) PSQ-HMM	0.529^{0}	0.339^{0}	0.617^{0}	0.445	0.282	0.711	0.315	0.193	0.594	0.289	0.225	0.693
(2) DT » BM25	0.568^{0}	0.388^{01}	0.662^{01}	0.636^{1}	0.453^{1}	0.857^{1}	0.338	0.215	0.633	0.316	0.275	0.756
(3) DT » ColBERT	0.609^{01}_{34}	0.422^{01}_{34}	0.700^{01}_{34}	0.669^{1}	0.497^{1}	0.889^{14}	0.403^{1}	0.285^{12}	0.708^{124}	0.361^{1}	0.298^{1}_{34}	0.786^{1}
(4) ColBERT-X MTT	0.613^{01}_{34}	0.411^{01}_{34}	0.687^{01}_{34}	0.643^{1}	0.451^{1}	0.827^{1}	0.375	0.236	0.612	0.330	0.281^{1}_{34}	0.760
(5) ColBERT-X ED	0.638^{012}_{58}	0.457^{01234}_{5678}	0.732^{01234}_{5678}	$\mathbf{0.699}^{14}_{58}$	0.530^{124}_{5678}	0.920^{124}_{578}	0.393	0.263	0.687^{14}	0.357^{1}	0.317^{1}	0.827^{124}
ColBERT-X MTD with Different Mixing Strategies												
(6) Mix Passages	0.666^{01234}_{678}	0.471^{01234}_{5678}	0.747^{01234}_{5678}	0.675^{1}	0.520^{14}_{567}	0.901^{14}_{67}	0.444^{12}	0.340^{1245}_{678}	0.762^{1245}_{678}	$\mathbf{0.404}^{1245}_{678}$	0.367^{12345}_{678}	0.868^{12345}_{678}
(7) Mix Entries	$\mathbf{0.674}^{012345}_{678}$	0.469^{01234}_{5678}	0.745^{01234}_{5678}	0.686^{1}	0.522^{14}_{567}	0.911^{124}_{5678}	0.461^{1245}	$\mathbf{0.347}^{12345}_{678}$	$\mathbf{0.768}^{12345}_{678}$	0.397^{124}_{678}	$\mathbf{0.372}^{12345}_{678}$	$\mathbf{0.877}^{12345}_{678}$
(8) Round Robin Entries	0.656^{01234}_{5678}	$\mathbf{0.476}^{01234}_{5678}$	$\mathbf{0.751}^{012345}_{5678}$	$\mathbf{0.699}^{12}_{58}$	$\mathbf{0.535}^{1234}_{58}$	$\mathbf{0.922}^{1234}_{578}$	$\mathbf{0.474}^{12345}$	0.341^{12345}_{678}	0.761^{1245}_{678}	0.388^{124}_{678}	0.347^{12345}_{678}	0.856^{1234}_{678}

Table 3: nDCG@20 on training with more languages

		Training Languages		
	Evaluation Collection	CLEF03	NeuCLIR	Both
Mix Passages	CLEF 2003	0.675	0.688	0.694
	NeuCLIR 2022 MLIR	0.437	0.444	0.431
	NeuCLIR 2023 MLIR	0.377	0.404	0.406
Mix Entries	CLEF 2003	0.686	0.679	0.680
	NeuCLIR 2022 MLIR	0.424	0.461	0.445
	NeuCLIR 2023 MLIR	0.359	0.397	0.379

in their native form are significantly more effective than translating all documents into English and searching with English ColBERT.

Since the languages in the two CLEF collections are closer to English than those in NeuCLIR, the ColBERT-X model trained with English texts (Row 5) still provides reasonable effectiveness using (partial) zero-shot language transfer during inference. MTD yields identical effectiveness to ED based on the TOST equivalence test in the two CLEF collections by measuring MAP (Table 2). In contrast, NeuCLIR languages do not benefit from this phenomenon. Instead, training directly with text in document languages enhances both the general language modeling and retrieval ability of the student models. In NeuCLIR 2022 and 2023, student ColBERT-X models trained with MTD (Rows 6 to 8) are 9% (NeuCLIR23 0.317 to 0.347 with round robin entries) to 32% (NeuCLIR22 0.263 to 0.347 with mix entries) more effective than ED (Row 5) by measuring MAP.

5.1 Ablation on Language Mixing Strategies

Since the TOST equivalence tests show that the three mixing strategies demonstrate statistically similar MAP and Recall for all collections except for a few cases in CLEF 2003 (CLEF 2003 may be an outlier because it has English documents, a known source of bias in MLIR [29]). We conclude that MTD is robust to how languages are mixed during training as long as multiple languages are present in each training mini-batch [29]. Such robustness provides operational flexibility to practitioners creating MLIR models. Since passage translation might not be available for all languages, mixing passages allows selecting passages only from a subset of languages. Mixing entries also allows training entries to be filtered for specific languages if relevance is known to drop after translation.

When evaluating with nDCG@20, the differences are larger but less consistent. For both CLEF collections and NeuCLIR 2022, topics are developed for a single language before obtaining relevance judgments across all languages. These topics may not be well-attested in all document languages, resulting in some CLIR topics with few relevant documents. For these three evaluation collections, models trained with mixed CLIR tasks (mix and round-robin entries) are more effective at the top of the ranking. High variation among the topics leads to inconclusive statistical significance results, suggesting opportunities for result fusion between these strategies. In NeuCLIR 2023 topics were developed bilingually, meaning that topics are not socially or culturally designed for a single language, leading to statistically equivalent nDCG@20 results.

5.2 Training Language Ablation

Finally, we explore training with languages beyond the ones in the document collection. Table 3 shows MTD-trained models for CLEF 2003, NeuCLIR, and both on each collection. Due to GPU memory constraints, we exclude the round-robin strategy from this ablation.

We observe that models trained with the mix passages strategy are more robust than the mix-entries variants when training on CLEF and evaluating on NeuCLIR and vice versa. This shows smaller degradation when facing language mismatch between training and inference. Surprisingly, training on NeuCLIR languages with the mix passage strategy yields numerically higher nDCG@20 than training on CLEF (0.675 to 0.688).

When training both CLEF and NeuCLIR languages, effectiveness is generally worse than only training on the evaluation languages. This trend suggests the models might be facing capability limits in the neural model, or picking up artifacts from the quality differences

in the translation. This observation demands more experimentation on MLIR dual-encoder models, which we leave for future work.

6 CONCLUSION

We propose Multilingual Translate-Distill (MTD) for training MLIR dual-encoder models. We demonstrated that ColBERT-X models trained with the proposed MTD are more effective than using previously proposed MLIR training techniques on four MLIR collections. By conducting statistical equivalence tests, we showed that MTD is robust to the mixing strategies of the languages in the training mini-batch.

REFERENCES

[1] Rexhina Blloshmi, Tommaso Pasini, Niccolò Campolungo, Somnath Banerjee, Roberto Navigli, and Gabriella Pasi. 2021. IR like a SIR: sense-enhanced information retrieval for multiple languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1030–1041.

[2] Martin Braschler. 2001. CLEF 2000 — Overview of Results. In *Cross-Language Information Retrieval and Evaluation*, Carol Peters (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 89–101.

[3] Martin Braschler. 2001. CLEF 2001—Overview of Results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 9–26.

[4] Martin Braschler. 2002. CLEF 2002—Overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 9–27.

[5] Martin Braschler. 2003. CLEF 2003–Overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 44–63.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://aclanthology.org/2020.acl-main.747

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451.

[8] Zhuyun Dai and Jamie Callan. 2019. 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*. 985–988.

[9] Kareem Darwish and Douglas W Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 338–344.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[11] Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. 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)*. Association for Machine Translation in the Americas, Virtual, 110–115.

[12] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2288–2292.

[13] Luyu Gao and Jamie Callan. 2022. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2843–2853.

[14] Ximo Granell. 2014. *Multilingual information management: Information, technology and translators*. Chandos Publishing.

[15] Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. SOCKEYE 2: A Toolkit for Neural Machine Translation. In *EAMT 2020*. https://www.amazon.science/publications/sockeye-2-a-toolkit-for-neural-machine-translation

[16] 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.

[17] Zhiqi Huang, Hansi Zeng, Hamed Zamani, and James Allan. 2023. Soft Prompt Decoding for Multilingual Dense Retrieval. *arXiv preprint arXiv:2305.09025* (2023).

[18] David A Hull and Gregory Grefenstette. 1996. 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*. 49–57.

[19] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2010), 117–128.

[20] Vitor Jeronymo, Roberto Lotufo, and Rodrigo Nogueira. 2023. NeuralMind-UNICAMP at 2022 TREC NeuCLIR: Large Boring Rerankers for Cross-lingual Retrieval. *arXiv preprint arXiv:2303.16145* (2023).

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

[22] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[23] 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.

[24] Wessel Kraaij, Jian-Yun Nie, and Michel Simard. 2003. Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics* 29, 3 (2003), 381–419.

[25] Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science* 8, 4 (2017), 355–362.

[26] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldanini, and Eugene Yang. 2022. Overview of the TREC 2022 NeuCLIR Track. In *The Thirty-first Text REtrieval Conference (TREC 2022) Proceedings*.

[27] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldanini, and Eugene Yang. 2023. Overview of the TREC 2023 NeuCLIR Track. In *The Thirty-second Text REtrieval Conference (TREC 2023) Proceedings*.

[28] Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022. HC4: A New Suite of Test Collections for Ad Hoc CLIR. In *Proceedings of the 44th European Conference on Information Retrieval*.

[29] Dawn Lawrie, Eugene Yang, Douglas W Oard, and James Mayfield. 2023. Neural Approaches to Multilingual Information Retrieval. In *European Conference on Information Retrieval*. Springer, 521–536.

[30] Minghan Li, Sheng-Chieh Lin, Xueguang Ma, and Jimmy Lin. 2023. SLIM: Sparsified Late Interaction for Multi-Vector Retrieval with Inverted Indexes. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Taipei, Taiwan,) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1954–1959. https://doi.org/10.1145/3539618.3591977

[31] Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. CITADEL: Conditional Token Interaction via Dynamic Lexical Routing for Efficient and Effective Multi-Vector Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11891–11907. https://doi.org/10.18653/v1/2023.acl-long.663

[32] Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2022. Learning Cross-Lingual IR from an English Retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4428–4436.

[33] Walid Magdy and Gareth J.F. Jones. 2011. Should MT systems be used as black boxes in CLIR?. In *European Conference on Information Retrieval*. Springer, 683–686.

[34] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.

[35] Paul McNamee and James Mayfield. 2002. 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*. 159–166.

[36] Teruko Mitamura, Eric Nyberg, Hideki Shima, Tsuneaki Kato, Tatsunori Mori, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, Tetsuya Sakai, Donghong Ji, et al. 2008. Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access.. In *NTCIR*.

[37] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. 2022. Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway,*

April 10–14, 2022, Proceedings, Part I (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 382–396.

[38] Suraj Nair, Eugene Yang, Dawn Lawrie, James Mayfield, and Douglas W. Oard. 2023. BLADE: Combining Vocabulary Pruning and Intermediate Pretraining for Scaleable Neural CLIR. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1219–1229.

[39] 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. *arXiv preprint arXiv:1611.09268* (2016). arXiv:1611.09268 http://arxiv.org/abs/1611.09268

[40] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[41] Carol Peters and Martin Braschler. 2002. The Importance of Evaluation for Cross-Language System Development: the CLEF Experience.. In *LREC*.

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

[43] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5835–5847.

[44] Razieh Rahimi, Azadeh Shakery, and Irwin King. 2015. Multilingual information retrieval in the language modeling framework. *Information Retrieval Journal* 18, 3 (2015), 246–281.

[45] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

[46] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3715–3734.

[47] Donald J Schuirmann. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics* 15 (1987), 657–680.

[48] P Shi and J Lin. 2019. Cross-lingual relevance transfer for document retrieval. *arXiv preprint arXiv:1911.02989* (2019).

[49] Luo Si, Jamie Callan, Suleyman Cetintas, and Hao Yuan. 2008. An effective and efficient results merging strategy for multilingual information retrieval in federated search environments. *Information Retrieval* 11, 1 (2008), 1–24.

[50] Ming-Feng Tsai, Yu-Ting Wang, and Hsin-Hsi Chen. 2008. 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*. 195–202.

[51] Jinxi Xu and Ralph Weischedel. 2000. Cross-lingual information retrieval using hidden Markov models. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. 95–103.

[52] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

[53] Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W Oard, and Scott Miller. 2024. Translate-Distill: Learning Cross-Language Dense Retrieval by Translation and Distillation. In *Advances in Information Retrieval: 46th European Conference on IR Research, ECIR 2024*.

[54] Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W. Oard. 2022. C3: Continued Pretraining with Contrastive Weak Supervision for Cross Language Ad-Hoc Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2507–2512. https://doi.org/10.1145/3477495.3531886

[55] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A Multilingual Benchmark for Dense Retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 127–137. https://aclanthology.org/2021.mrl-1.12