Parameter-Efficient Neural Reranking for Cross-Lingual and Multilingual Retrieval

Robert Litschko[1] **Ivan Vulić**[2] **Goran Glavaš**[1,3]
[1] Data and Web Science Group, University of Mannheim, Germany
[2] Language Technology Lab, University of Cambridge, UK
[3] Center for AI and Data Science (CAIDAS), University of Würzburg, Germany

Abstract

State-of-the-art neural (re)rankers are notoriously data-hungry which – given the lack of large-scale training data in languages other than English – makes them rarely used in multilingual and cross-lingual retrieval settings. Current approaches therefore commonly transfer rankers trained on English data to other languages and cross-lingual setups by means of multilingual encoders: they fine-tune *all* parameters of pretrained massively multilingual Transformers (MMTs, e.g., multilingual BERT) on English relevance judgments, and then deploy them in the target language(s). In this work, we show that two *parameter-efficient* approaches to cross-lingual transfer, namely Sparse Fine-Tuning Masks (SFTMs) and Adapters, allow for a *more lightweight* and *more effective* zero-shot transfer to multilingual and cross-lingual retrieval tasks. We first train language adapters (or SFTMs) via Masked Language Modelling and then train retrieval (i.e., reranking) adapters (SFTMs) on top, while keeping all other parameters fixed. At inference, this modular design allows us to compose the ranker by applying the (re)ranking adapter (or SFTM) trained with source language data together with the language adapter (or SFTM) of a target language. We carry out a large scale evaluation on the CLEF-2003 and HC4 benchmarks and additionally, as another contribution, extend the former with queries in three new languages: Kyrgyz, Uyghur and Turkish. The proposed parameter-efficient methods outperform standard zero-shot transfer with full MMT fine-tuning, while being more modular and reducing training times. The gains are particularly pronounced for low-resource languages, where our approaches also substantially outperform the competitive machine translation-based rankers.

1 Introduction

In recent years, neural rankers (Nogueira et al., 2019b; MacAvaney et al., 2019; Khattab and Zaharia, 2020), trained on large-scale datasets (Bajaj et al., 2016; Dietz et al., 2017; Craswell et al., 2021), have substantially pushed the performance on various retrieval benchmarks. Since such models are generally too computationally involved (i.e., too slow) for ad-hoc retrieval on large document collections, they are commonly leveraged as rerankers, i.e., they rerank the output of some fast model (e.g., BM25) that produces the initial ranking. Large-scale datasets for training neural rerankers, however, exist only in English, which impedes their adoption in retrieval scenarios that involve other languages: (a) monolingual retrieval in other languages and (b) cross-lingual information retrieval (CLIR) in which, for a given query in one language, one needs to determine relevance of documents written in one or more other languages.

While CLIR is often instantiated in the form of standalone tasks (e.g., to allow users from different countries to search over the aggregated global collection of COVID-19 news and findings in their native language (Casacuberta et al., 2021)), it also supports a range of IR-backed NLP tasks such as cross-lingual question answering (Asai et al., 2021), entity linking (Liu et al., 2021a), and cross-lingual summarization (Zhu et al., 2019; Vitiugin and Castillo, 2022). A truly multilingual search engine requires reliable estimation of both monolingual (for a wide range of languages) as well as cross-lingual query-document relevance, which both crucially rely on the alignment of text representations across different languages (Nie, 2010). The lack of large-scale retrieval datasets in languages other than English means that monolingual reranking for those languages has to be achieved by means of cross-lingual transfer of a reranking model trained on English relevance judgments.

Pretrained massively multilingual Transformers (MMTs) like multilingual BERT (mBERT) (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) have been leveraged to this effect, but were shown to require substantial task-specific (i.e., ranking-

Proceedings of the 29th International Conference on Computational Linguistics, pages 1071–1082
October 12–17, 2022.

oriented) fine-tuning for reliable prediction of semantic similarity and relevance scores (Reimers and Gurevych, 2020; Litschko et al., 2021). MMTs offer zero-shot cross-lingual transfer of neural (re)ranking models out of the box – an MMT is fine-tuned on English relevance judgments and then employed in (monolingual or cross-lingual) retrieval tasks that involve other languages. Conceptually, via such transfer, no fine-tuning data (i.e., relevance judgments) is required for the target language(s).

This procedure, in principle, enables downstream zero-shot transfer to any language seen by the MMT in pretraining (e.g., for mBERT, 104 languages). However, in language understanding tasks (Hu et al., 2020), massive performance drops have been observed when transferring between distant languages, and especially in transfer to low-resource languages, underrepresented in MMT pretraining (Lauscher et al., 2020). Our results (§4) confirm these findings for ad-hoc IR. This is the consequence of the effect known as the *curse of multilinguality* (Conneau et al., 2020): sharing MMT parameters (i.e., its fixed parameter budget/capacity) across more and more languages makes text representations for individual languages less accurate. This effect is especially detrimental to low-resource languages, those least represented in multilingual pretraining corpora. What is more, large-scale full fine-tuning on the source language data (e.g., English) is likely to lead to catastrophic forgetting and interference effects (McCloskey and Cohen, 1989; Mirzadeh et al., 2020) that further bias the multilingual representation space towards the source language, at the expense of representation quality for low-resource languages. Besides the standard zero-shot cross-lingual transfer (MacAvaney et al., 2020; Huang et al., 2021a), other cross-lingual transfer approaches, commonly applied in other NLP tasks, such as training data translation (Shi et al., 2020), or leveraging external word-level alignments (Huang et al., 2021b), as well as distant supervision (Yu et al., 2021) have been explored as means to improve the cross-lingual transfer of neural rankers in IR. While translation-based approaches are competitive for high-resource languages, they may not be as effective for low-resource languages for which reliable MT models are missing; also, translation-based cross-lingual transfer has been shown to suffer from unwanted artifacts, such as "translationese" (Zhao et al., 2020; Vanmassenhove et al., 2021).

Contributions. Even if one would have sufficient amounts of labelled data in target languages, training language- or language-pair specific neural rerankers for all languages and language pairs would be prohibitively computationally expensive and unsustainable (Strubell et al., 2019). In this work we additionally remedy for this by composing (re)rankers in a modular way that enables more sustainable cross-lingual transfer. Concretely, we introduce neural (re)ranking models for cross-lingual and multilingual document retrieval based on MMTs that enable much more parameter efficient fine-tuning and more effective cross-lingual transfer for relevance prediction. Our (re)rankers are based on two styles of modular components: **1)** *Adapters* (Rebuffi et al., 2017; Houlsby et al., 2019; Pfeiffer et al., 2020) and **2)** *Sparse Fine-Tuning Masks* (SFTMs) (Ansell et al., 2022). When integrated into the architecture of a pretrained MMT, both allow for (1) the pretrained multilingual knowledge to be fully preserved, alleviating the negative interference and forgetting effects, and (2) offer additional language-specific model capacity which is used to improve the MMTs' representations for target languages, thus remedying for the curse of multilinguality.

We provide an extensive evaluation of both approaches in (i) zero-shot transfer for monolingual retrieval and (ii) CLIR, on two established benchmarks (Braschler, 2003; Lawrie et al., 2022). As an additional contribution, we expand the CLEF dataset (Braschler, 2003) with three query languages from the Turkic family (Turkish, Kyrgyz, and Uyghur, the latter two being low-resource languages), typologically and etymologically distant from the Indo-European languages.[1] Our results show that our modular neural (re)rankers are not only faster to train, but also outperform standard zero-shot transfer based on full MMT fine-tuning, and especially so in retrieval tasks that involve linguistically distant and low-resource languages. Moreover, our adapter- and SFTM-based rerankers generally outperform a strong preranker that utilizes state-of-the-art machine translation.

2 Methodology

We first introduce the general multi-stage ranking (i.e., preranking-reranking) framework, com-

[1] In this manner, our work addresses the calls for more linguistic diversity in NLP and IR research (Bender, 2011; Joshi et al., 2020; Ponti et al., 2020; Ruder et al., 2021).

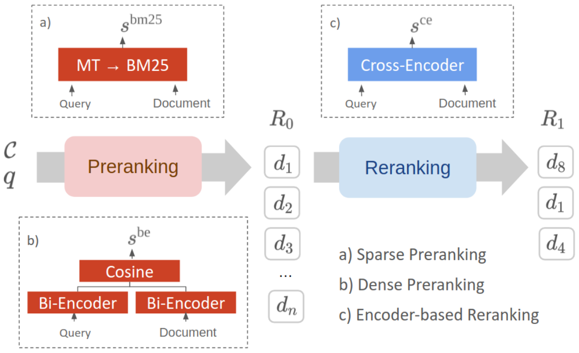

Figure 1: Overview of the multi-stage ranking approach to ad-hoc retrieval. **Stage 1 - Preranking:** We rank the document collection $\mathcal{C}$ by (a) running *sparse* BM25 retrieval on translated queries, or (b) according to the cosine similarity between *dense* query and document representations yielding an initial ranking R_0. **Stage 2 - Reranking:** We refine R_0 by reranking the top-k documents according to relevance scores predicted by a Cross-Encoder, yielding the refined ranking R_1.

monly used in information retrieval tasks, within which our work is embedded. We then introduce adapters and sparse fine-tuning masks (SFTMs), and present how to leverage them as crucial vehicles of the parameter-efficient cross-lingual transfer of the reranking component.

2.1 Multi-Stage Ranking

Pretrained Transformers like BERT (Devlin et al., 2019) are often used as *Cross-Encoder (CE)* scoring models: the Transformer encodes a query-document concatenation fed as input to the model, and the encoding is then fed to a dense layer that predicts the relevance score (MacAvaney et al., 2020; Jiang et al., 2020; Nogueira et al., 2019b). Computing scores for all query-documents pairs with Cross-Encoders is too slow for practical IR applications: they are thus primarily used as rerankers in a multi-stage ranking approach (MacAvaney et al., 2020; Geigle et al., 2021). In this work we adopt this paradigm for cross-lingual ad-hoc retrieval: Figure 1 illustrates its workflow.[2]

Preranking, based on a fast and efficient ranking method, is applied to every document from the document collection in order to provide a good initial ranking, targeting high recall. Let q^{l1} be a query in language l_1 and $\mathcal{C}^{l2} = \{d_i\}_{i=1}^{n}$ be a document collection containing n documents in language l_2. Associating and ranking documents w.r.t. relevance scores s_i we obtain an initial ranking

$$R_0 = [(d_1, s_1), (d_2, s_2) \dots (d_n, s_n)], \qquad (1)$$

where $s_1 > s_2 > \dots s_n$. We transfer our rerankers based on MMTs – and trained on English relevance judgments – to (i) CLIR tasks as well as to (ii) monolingual IR tasks in target languages. The latter task, termed MoIR, is effectively zero-shot cross-lingual transfer for monolingual retrieval. In MoIR, we opt for a lexical preranker and score documents with $s^{\text{bm25}} = \text{BM25}(q, d)$.[3] In CLIR we follow the widely used approach of machine translating the query (Bonifacio et al., 2021; Lawrie et al., 2022): this process effectively translates CLIR into a noisy variant of MoIR. In addition, we experiment with a representation-based approach based on pretrained multilingual *Bi-Encoders* (*BE*): here, we embed the query and documents independently, and then use the cosine similarity between their embeddings $s^{be} = cos(BE(q), BE(d))$. In the preranking stage, unlike later in reranking, we use the encoders merely as general-purpose text encoders, without any additional retrieval-specific training.

Reranking: This stage refines the initial ranking obtained via preranking. It relies on a *CE* model which captures fine-grained (but more costly to model and run) semantic interactions between queries and documents. The ranking is then:

$$R_1 = [(d_1, s_1^{ce}), (d_2, s_2^{ce}) \dots (d_k, s_k^{ce})] \qquad (2)$$

To this end, we rely on multilingual CEs to compute the binary relevance score s^{ce} on the concatenation of query and document pairs: $s^{ce} = CE(\texttt{[CLS]}\, q\, \texttt{[SEP]}\, d_i\, \texttt{[SEP]})$. We adopt a common practice (MacAvaney et al., 2019; Craswell et al., 2020; Naseri et al., 2021) of reranking the top $k = 100$ pre-ranked documents, yielding the final ranking R_1. Finally, it is also possible to ensemble the preranker's and reranker's ranked lists via simple rank averaging. In our experiments (4), we evaluate such preranking-reranking ensembles as well and show that such interpolations often bring additional performance gains.

[2] Alternative approaches that leverage pretrained encoders for IR include late interaction models (Khattab and Zaharia, 2020; Gao et al., 2021; Nair et al., 2022; Santhanam et al., 2022), embedding-based retrieval (Hofstätter et al., 2021; Litschko et al., 2021), and augmentation (Nogueira et al., 2019c,a).

[3] We used the `pyserini` implementation of BM25 (Lin et al., 2021) with the suggested (i.e., default) parameter configuration.

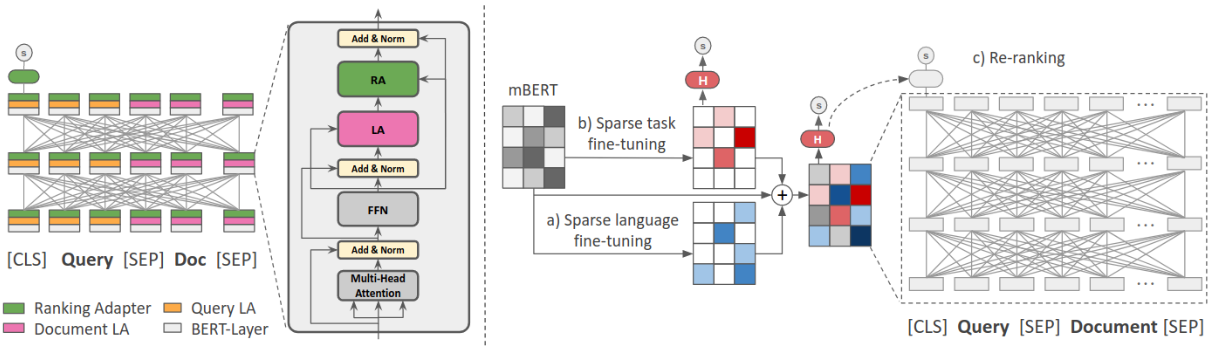

Figure 2: Overview of parameter-efficient transfer learning for neural (re)ranking. *Left*: A reranker is composed by stacking a pretrained target Language Adapter (LA) and a Ranking Adapter (RA; trained with source language data) on top of the original Transformer layers of an MMT (e.g., mBERT). *Right*: Sparse fine-tuning of a Ranking Mask (RM) and a Language Mask (LM) from mBERT parameters; rerankers are composed by adding the RM and LM values to the original mBERT parameters.

2.2 Parameter-Efficient Cross-Lingual Ranker Transfer

In this work, we propose a modular and parameter-efficient framework that allows faster training and more effective cross-lingual transfer of neural rerankers, that enhances both CLIR and MoIR. We first learn language-specific Adapters (LAs) or Sparse Fine-Tuning Masks (SFTMs) via Masked Language Modelling (MLM) on unannotated monolingual corpora of respective languages, while keeping the original MMT parameters intact. We then train Ranking Adapters (or Ranking SFTMs) using source-language data on top of the source-language LAs (language SFTMs), while keeping all other parameters frozen. At inference time, for a given IR (MoIR or CLIR) task, we compose our reranker by placing the Ranking Adapters (Ranking SFTMs) on top of the LAs (language SFTMs) of the query and/or document languages of that concrete retrieval task. The modular framework is illustrated in Figure 2.

Adapters. We train Ranking Adapters (RA) and Language Adapters (LA) based on the architecture of Pfeiffer et al. (2020). In the Transformer architecture, each layer l consists of a multi-head attention block (i.e., sub-layer) and a feed-forward network (FFN), both followed by a residual connection and layer normalization. We denote the residual connection (output of FFN) with $\mathbf{r}_l$ and the hidden state after the layer norm with $\mathbf{h}_l$.

$$\mathrm{LA}(\mathbf{h}_l, \mathbf{r}_l) = \mathbf{U}_l(\psi(\mathbf{D}_l(\mathbf{h}_l)) + \mathbf{r}_l \tag{3}$$

$$\mathrm{RA}(\mathbf{h}_l, \mathbf{r}_l) = \mathbf{U}_l(\psi(\mathbf{D}_l(\mathrm{LA}_l))) + \mathbf{r}_l \tag{4}$$

Adapters are parameterized by the down-projection matrix $\mathbf{D} \in \mathbb{R}^{h \times d}$ and the up-projection matrix $\mathbf{U} \in \mathbb{R}^{d \times h}$, where h and d denote the hidden size of the Transformer and the bottleneck dimension of the adapter, respectively. The ratio between h and d is also called the *reduction factor*, and corresponds to the level of parameter compression (i.e., how many times fewer parameters are updated if we train adapters instead of updating all Transformer parameters). The forward pass of a Language Adapter consists of a down-projection of h_l, a non-linear activation function $\psi(\cdot)$ and an up-projection. Ranking Adapters are stacked on top of LAs and process their output. Both adapters have residual connections to the output of the FFN.[4] We train LAs using the standard MLM objective (Devlin et al., 2019), whereas we train RAs together with the dense scoring layer by means of minimizing the standard binary cross-entropy loss.

In CLIR setups, queries and documents are in different languages. It is thus, in principle, possible to stack the RA on top of (i) the query language adapter LA^{Q}, (ii) document language adapter LA^{D}, or by using (iii) *split adapters* LA^{S}: here, we encode query tokens up to the separator token (`[SEP]`) using the LA of the query language and the document tokens (after `[SEP]`) with the LA of language of the document collection (cf. Fig. 2).

Sparse Fine-Tuning Masks. Like adapters, SFTMs (Ansell et al., 2022) aim to decouple task knowledge from language knowledge, but instead of introducing additional parameters, the idea is

[4] To alleviate the mismatch between the multilingual vocabulary of the MMT and the target language vocabulary, Pfeiffer et al. (2020) also additionally place invertible adapters INV on top of the embedding layer along with their inverses INV^{-1} placed before the output layer. In our experiments we adopt this variant; for more details we refer the reader to the work of Pfeiffer et al. (2020).

to directly update only small subsets of MMT's original parameters. Sparse Fine-Tuning (SFT) consists of two phases. In *Phase 1* we fine-tune all mBERT's parameters $\theta^{(0)}$, resulting in updated parameter values $\theta^{(1)}$. We then select the top K parameters with the largest value change, i.e., those with the largest values $|\theta_i^{(0)} - \theta_i^{(1)}|$. We then construct a binary mask: the selected K parameters remain trainable, whereas all other parameters are frozen. In *Phase 2* all parameters are reset to $\theta^{(0)}$ and training restarts, but this time only the selected parameters of the mask are updated, yielding $\theta^{(2)}$. The final update (i.e., the SFTM) is then obtained as the difference vector $\mathrm{M} = \theta^{(2)} - \theta^{(0)}$. As is the case with Language Adapters, we obtain the Language Masks (LM) by means of (additional) MLM training on language-specific corpora; whereas the Ranking Mask (i.e., the mask for the ranking task, RM) is learned via binary cross-entropy objective on source-language (English) relevance judgments. At inference, the reranker is composed as $\theta^{(0)} + \mathrm{RM} + \mathrm{LM}$ (cf., Figure 2). In our CLIR settings (§3), we explore using (i) the query language mask (LM^{Q}), (ii) document language mask (LM^{D}) or (iii) the combination of both masks ($\mathrm{LM}^{\mathrm{B}} = \mathrm{LM}^{\mathrm{Q}} + \mathrm{LM}^{\mathrm{D}}$). Note that SFTMs represent a more computationally efficient solutions at inference time: unlike adapters, they do not extend (i.e., deepen) the Transformer architecture.

3 Experimental Setup

Adapter and SFTM Training. We train adapters following the recommendations from Pfeiffer et al. (2020). Unless noted otherwise, we train LAs with the reduction factor of 2 (i.e., $h/d = 2$) on Wikipedias of respective languages, for 250K steps with batch size 64 and learning rate of 1e-4. For RAs we experimented with the different reduction factors: 1, 2, 4, 8, 16, 32 (cf. §4). Following Ansell et al. (2022), for fair comparisons between adapters and SFTMs, we set the mask size K for SFTMs to the same number of parameters that adapters with a certain reduction factor have.[5]

Reranking Training. We train mBERT-based[6] rerankers on MS-MARCO (Craswell et al., 2021), with a linear warm-up over the first 5K updates, in batches of 32 instances with a maximum sequence length of 512, and using a learning rate of 2e-5. We evaluate the model on the validation data every 25K updates and choose the checkpoint with the best validation performance.

Evaluation Data. We evaluate the models on the standard CLEF-2003 benchmark (Braschler, 2003)[7] as well as on the recently introduced HC4 benchmark (Lawrie et al., 2022). With CLEF, we use monolingual test collections in EN, DE, IT, RU, and FI for MoIR, and experiment with the following cross-lingual directions: EN-{FI, DE, IT, RU}, DE-{FI, IT, RU}, FI-{IT, RU}. Each experimental run covers 60 queries, whereas the document collection sizes are as follows: RU – 17K, FI – 55K, IT – 158K, and DE – 295K.

We additionally evaluate the models in CLIR tasks with CLEF queries posed in lower-resource languages. To this end, (i) we leverage Swahili (SW) and Somali (SO) queries (Bonab et al., 2019), where the queries were obtained via manual translation of English queries; (ii) we create another set of translated CLEF queries in three languages: Turkish (TR), Kyrgyz (KG), and Uyghur (UG). The new set covers one high-resource and two low-resource languages and is intended to facilitate and diversify evaluation of CLIR with low-resource languages in future work. The queries were constructed via the standard post-editing procedure borrowed from other data collection tasks (Glavaš et al., 2020; Hung et al., 2022): we obtained initial query translations via Google Translate, which were then post-edited by native speakers.

HC4 comprises queries *and* document collections in three languages: Persian (FA), Russian (RU) and Chinese (ZH). Compared to CLEF, HC4 collections are considerably larger, spanning 646K, 486K and 4.72M documents per each respective language, associated with 50 test queries in each language. We use *title* and *description* fields as queries following Lawrie et al. (2022). HC4 is used in MoIR experiments.

Baseline Models. The primary baseline for our adapter- and SFTM-based transfer is the standard and well established method for zero-shot transfer of English-trained rerankers (MacAvaney et al., 2020), termed *MonoBERT*. This is the reranking Cross-Encoder where we allow for full-tuning of the underlying monolingual or multilingual BERT

[5] Leading to the number of trainable parameters (sparsity) of 14M (8.5%), 7.1M (4.2%), 3.6M (2.1%), 1.8M (1.1%), 894K (0.52%) and 452K (0.27%) respectively.

[6] Pretrained `bert-base-multilingual-uncased` weights from the HuggingFace Transformers library (Wolf et al., 2020) are used.

[7] http://catalog.elra.info/en-us/repository/browse/ELRA-E0008/

Model	TR-EN	TR-IT	TR-DE	TR-FI	TR-RU	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	**AVG**	**ENS**
DISTIL$_{\text{DmBERT}}$ (PR)	.183	.251	.190	.252	.260	.294	.290	.313	.247	.300	.267	.284	.221	.302	.261	-
MonoBERT	.235	.197	.208	.285	.217	.339	.315	.248	.295	.329	.270	.246	.197	.174	.254	.274
+RA +LA^{S}	.269	**.253**	**.252***	**.362**	.186	.363	.352	.197	.317*	.329	.300	.223	**.266**	.207	.277	.287
+RA +LA^{D}	.252	.234	.222	.267	**.267**	.366*	**.366***	**.248**	.314*	.350	.302	**.315**	.220	**.234**	**.283**	**.298**
+RA +LA^{Q}	**.270**	.243	.242	.293	.191	.370	.355	.189	.318	.325	.279	.223	.247	.182	.266	.285
+RM +LM^{B}	.229	.228	.197	.244*	.168	.299	.344	.181*	.303	.309	.302	.191*	.206	.108*	.236	.269
+RM +LM^{D}	.231	.226	.229	.317	.149*	**.394***	.359	.173*	.320*	.376	.304	.187	.239	.166*	.262	.279
+RM +LM^{Q}	.239	.252	.232	.316	.162*	.359	.349	.191	.310*	**.391**	**.323***	.195	.255*	.160	.267	.280

Table 1: CLIR results (Mean Average Precision, MAP) with DIST$_{\text{DmBERT}}$ as Stage 1 preranker. **Bold**: Best neural retrieval model for each language pair. *: significance tested against MonoBERT at $p \leq 0.05$, computed via paired two-tailed t-test. Ranking and Language Adapters have a reduction factor of 16 and 2 (see §2), respectively. Ranking and Language Masks both correspond to a reduction factor of 2 (see §3). We report average results (AVG), and also averaged ensemble (ENS) results where we combine ranking lists from Stage 1 and Stage 2 rankings; see §2.1. Superscripts over LAs and LMs denote query language (Q), document language (D), split adapters (S) for LAs, and '(B)oth masks' for LMs (see §2.2).

Model	TR-EN	TR-IT	TR-DE	TR-FI	TR-RU	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	**AVG**	**ENS**
NMT+BM25 (PR)	.392	.353	.308	.307	.227	.378	.446	.285	.355	.367	.385	.272	.364	.271	.336	-
MonoBERT	.415	.375	.339	.345	.307	.386	.411	.351	.371	.409	.380	.322	.367	.340	.366	.360
+RA +LA	**.448**	.408*	.353	.371	.327	.388	**.435**	**.367**	.385	.413	.405	.348	.381	**.365**	.385	**.374**
+RM +LM	.447	**.414**	**.356**	**.386**	**.336**	**.413**	.429	.345	**.390**	**.468**	**.407**	**.363**	**.395**	.364	**.394**	.371

Table 2: CLIR results (Mean Average Precision, MAP) with NMT+BM25 as Stage 1 preranker. For modular rerankers, we report the numbers with the best-performing configurations from CLEF experiments: $+$RA $+$LA^{D} and $+$RM $+$LM^{Q}; see also the caption of Table 1.

model on MS-MARCO. For CLIR experiments, we opt for DISTIL$_{\text{DmBERT}}$ as our Bi-Encoder preranker (PR), as it showed strong performance in our recent comparative empirical study (Litschko et al., 2021). In brief, DISTIL$_{\text{DmBERT}}$ is trained via knowledge distillation where sentence-similarity features are distilled from a monolingual English teacher, specialized for semantic encoding of sentences, into a multilingual student model; see (Reimers and Gurevych, 2020) for further details.

Finally, also for CLIR, we couple a state-of-the-art NMT system of Fan et al. (2020) (FAIR-MT), which we use to translate queries to the document collection language, with the BM25 ranker in the target language. For Kyrgyz and Uyghur, we use another NMT model, provided by the Turkic Interlingua (TIL) community[8] (Mirzakhalov et al., 2021), because we failed to obtain meaningful $\{$KG, UG$\}\rightarrow l_2$ translations with FAIR-MT.

4 Results and Discussion

Cross-Lingual Retrieval (CLIR). Tables 1 and 2 show the CLIR results, for fourteen language pairs from the augmented CLEF 2003 benchmark[9] using DISTIL$_{\text{DmBERT}}$ and NMT+BM25 as Stage 1 prerankers, respectively. With DISTIL$_{\text{DmBERT}}$ as the preranker (Table 1), Adapter- and SFTM-based rerankers consistently improve the initial preranking results, with gains of up to 2.7 MAP points, and EN-RU as the only exception. Importantly, compared to full fine-tuning (MonoBERT), our modular reranking variants bring gains between 1 and 4 MAP points on average, across all language pairs. Interestingly, the best adapter configuration (RA $+$LA^{D}), where at inference we stack the RA on top of the LA of the document collection language) outperforms the best SFTM-based reranker (RM $+$LM^{Q} and RM $+$LM^{D}) by 1.6 MAP points. Somewhat surprisingly, adapting only to the language of the document collection (LA^{D}; LM^{D}) yields better performance than adapting to both the query and collection language of the target task (LA^{S}; LM^{B}).

The language pairs in Tables 1 and 2 consist of high-resource languages for which large parallel corpora and, consequently, reliable NMT models exist. However, even when starting from a more

[8] https://turkic-interlingua.org

[9] We add TR-* pairs to the evaluation, enabled by our EN→TR translations of the queries.

	CLEF 2003				HC4				
Model	SW–EN	SO–EN	KG–EN	UG–EN	EN–FA	EN–ZH	EN–RU	**AVG**	**ENS**
NMT+BM25 (PR)	.325	.157	.228	.091	.183	.113	.186	.183	-
MonoBERT	.362	.158	.255	.157	.246	.172	.218	.224	.216
$+\text{RA} + \text{LA}^{\text{D}}$	**.407**	**.166**	.305	.155	.259	.189	.234	.245	**.228**
$+\text{RM} + \text{LM}^{\text{D}}$	.389	.161	**.311**	**.165**	**.267**	**.196**	**.241**	**.247**	.225

Table 3: CLIR results on extended CLEF pairs with low-resource query languages (Swahili, Somali, Kyrgyz, and Uyghur) and three language pairs from the HC4 benchmark.

competitive MT-based preranker (NMT+BM25; Table 2), our modular cross-lingual transfer of the reranker yields performance gains. In fact, with this stronger preranker, the gains from modular reranking are even more pronounced: +5/+6 MAP points for Adapters and SFTMs, respectively, compared to preranker and +2/+3 MAP points, respectively, compared to MonoBERT. This could explain why interpolating between the preranking and reranking (ENS, last column) yields further gains with $\text{DISTIL}_{\text{DmBERT}}$ as the preranker (Table 1), but not when we prerank with NMT+BM25 (Table 2).

Table 3 shows CLIR results for (a) language pairs from extended CLEF with queries written in low-resource languages – Swahili and Somali queries created by Bonab et al. (2019), as well as Kyrgyz and Uyghur queries that we created; and (b) three cross-lingual pairs of arguably distant languages (EN-{Farsi, Chinese, Russian}) from the HC4 benchmark. The gains that our SFTM- and Adapter-based modular rerankers bring for language pairs involving low-resource languages, over the MT-based preranker and the full fine-tuning (MonoBERT), are generally more substantial than those for high-resource language pairs: e.g., +8 and +4 MAP points w.r.t. NMT+BM25 and MonoBERT, respectively for SW-EN, and +8 and +5 points for KG-EN. The gains are similarly prominent for more distant language pairs from the HC4 dataset (+8 MAP points over the NMT+BM25 preranker for EN-FA and EN-ZH). With such prominent gains of the modular reranking over the preranker, it is no surprise that averaging the preranking and reranking document ranks (ENS) reduces the performance of the reranker. We believe that these results in particular emphasize the effectiveness of modular cross-lingual transfer that allows to increase the capacity of MMTs for individual languages, by means of LMs or LAs. The representations of low-resource languages, for which MMTs have seen little data in pretraining, particularly suffer from the curse of multilinguality (Conneau et al., 2020; Lauscher et al., 2020) – this is why particularly prominent gains are achieved for those languages when we increase the MMTs capacity for their representation via LMs/LAs.

Cross-Lingual Transfer for MoIR. Table 4 displays the results of monolingual retrieval with our best-performing modular rerankers for EN (as the source language) and four target languages (DE, IT, FI, RU).[10] Unlike the fully fine-tuned reranker (MonoBERT), our modular Adapter- and SFTM-based rerankers improve the initial rankings produced by BM25. These results strengthen the finding that our modular rerankers are not just more parameter-efficient (i.e., faster to train), but also lead to better cross-lingual transfer due to decoupling of language- and ranking-specific knowledge. In MoIR tasks the SFTM-based transfer outperforms its Adapter-based counterpart, same as in the case of CLIR with NMT+BM25 preranking (Table 1). Also as in the case of the latter CLIR results (Tables 1 and 3), interpolating between preranking and reranking results does not bring any gains.

It is worth noting that all MoIR scores are substantially higher than CLIR results from Tables 1 and 2. This is expected and reflects the fact that matching representations within a language – where models can still rely on exact lexical matches between queries and documents – is easier than aligning text representations across languages.

Effectiveness vs Efficiency. Adapters increase query latency because they deepen the Transformer. Rücklé et al. (2021) show that one can drop adapters from lower layers with small-to-negligible effect on performance. Table 5 shows the results of a similar analysis, where we drop the adapters from the first N layers at inference. Dropping adapters from only the first two layers (row 1-2) only slightly decreases the MoIR performance whereas it even slightly increases the

[10] Note that in MoIR, the actual reranking is always monolingual (albeit in the target language). Both queries and documents are thus encoded with the same target language LA/LM.

	CLEF 2003					HC4				
Model	EN	FI	DE	IT	RU	FA	ZH	RU	**AVG**	**ENS**
BM25 (PR)	.480	.505	.434	.494	.361	.279	.196	.228	.372	-
MonoBERT	.464	.528	.444	.463	.363	.356	.283	.245	.398	.402
+RA + LA	.512	.537	.457	.495	**.389**	.372	.284	.261	.413	.410
+RM + LM	**.515**	**.564**	**.459**	**.502**	.379	**.398**	**.307**	**.264**	**.423**	**.417**

Table 4: Results of zero-shot cross-lingual transfer for monolingual retrieval (MoIR) on CLEF 2003 and HC4 datasets. Results with reduction factors of 16 and 2 for Adapters and SFTMs, respectively.

Layer	CLIR	MoIR	AVG	Latency	Δ Speed-Up	Δ MAP
None	.282	.418	.331	34.6 ms	-	-
1-2	.295	.412	.337	33.7 ms	+2.6%	+.006
1-4	.269	.395	.314	32.8 ms	+5.0%	−.017
1-6	.229	.375	.281	31.9 ms	+7.7%	−.050
1-8	.134	.284	.187	31.0 ms	+10.4%	−.143
1-10	.086	.210	.130	30.0 ms	+12.9%	−.200
1-12	.086	.208	.129	29.5 ms	+14.2%	−.201

Table 5: Trade-off between efficiency and effectiveness when dropping adapters in +RA + LA^{D}. Average over all CLIR/MoIR setups and all reduction factors.

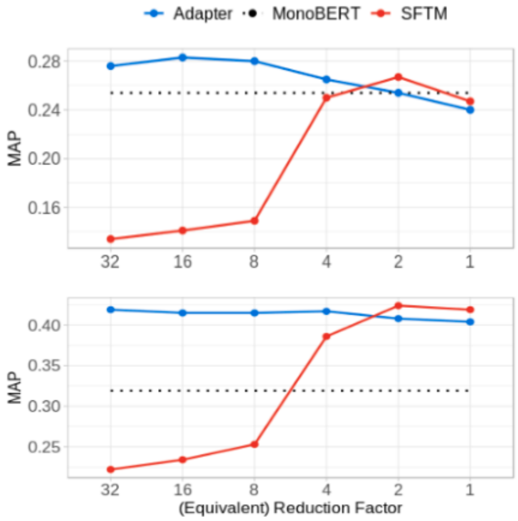

Figure 3: Retrieval performance at different parameter reduction factors; average MAP performance for CLIR (top) and MoIR (bottom).

CLIR results. Dropping adapters from more layers, however, substantially reduces the retrieval performance: e.g., removing adapters from the first 10 layers reduces CLIR performance by almost 20 MAP points, while reducing the query latency by only 13%. While Adapters and SFTMs yield comparable performance in our experiments, these observations favor SFTMs: for the *same query latency*,[11] SFTMs will yield better performance.

Parameter Efficiency. We also investigate the relation between various levels of parameter efficiency and retrieval performance. Figure 3 shows the performance of our modular rerankers for different parameter reduction factors. SFTMs exhibit stronger performance with smaller reduction factors (2 and 4), i.e., when we update a larger percentage of mBERT's original parameters. SFTMs shift the pretrained values of mBERT's parameters: this constrains the range of values that individual parameters can take, requiring the modification of the larger number of parameters for injecting complex language- and ranking-specific knowledge. In contrast, Adapters show better performance with higher reduction factor (8, 16, 32), i.e., when we add a relatively smaller number of Adapter parameters. This could be the consequence of the "unconstrained" initialization of the new Adapter parameters, which allows the complementary language- and ranking-specific knowledge to be compressed into a smaller number of parameters. Comparing those effects between CLIR and MoIR we observe the same trends. However, MAP gains compared to MonoBERT are larger in MoIR than in CLIR. This seems intuitive as ranking adapters (masks) are able to adapt for exact matches.

Impact of NMT on CLIR. In the cross-lingual setup the quality of retrieved documents crucially depends on the quality of query translations when NMT is used. In Table 6 we show original English queries together with their respective translations from Swahili and Somali. As expected, translations from Swahili are generally of higher quality compared to Somali, which explains the big performance gap reported in Table 3. In the best case the translation is semantically very close to the original query (cf., SW→EN; QID:172), or it contains only slight lexical (*flooding* vs. *floods*) and semantic variations, e.g., near-synonyms (*Holland* vs. *Netherlands*). In other cases, error prop-

[11] The query latency of an SFTM-based reranker is the same as that of MonoBERT as SFTMs do not increase the number of layers (nor parameters within layers) of the MMT.

QID	English Query *(original)*	NMT: Swahili → English	NMT: Somali → English
151	Wonders of **Ancient World** Look for **information** on the **existence** and/or the discovery of remains of the seven wonders of the **ancient world**.	Search for **information** about the **existence** and/or development of the **seventh** universe of the **ancient world**.	Thus, therefore, it is necessary to bear in mind that the truth is the truth, and that the truth is the truth, and that the truth is the truth.
172	**1995** Athletics **World Records** What **new world records** were achieved during the **1995** athletic world championships in **Gothenburg**?	What **new world records** were recorded at the **1995 World** Horses in **Gothenburg**?	The **1995 World** Trade Organization (WTO) announced that a **new** international trade agreement has led to a global trade agreement in **Gothenburg**.
187	**Nuclear** Transport in **Germany** Find reports on the protests against the transportation of **radioactive** waste with **Castor containers** in **Germany**.	**Nuclear** Delivery in **Germany** A report on the anti-trafficking of **radioactive** pollutants and **Castor containers** in **Germany**.	The Nugleerka department of Jarmalka Hel has been prepared for the development of the Nugleerka department of **Castor** district in Jarmalka.
200	Flooding in Holland and **Germany** Find statistics on flood disasters in Holland and **Germany** in **1995**.	The floods in the Netherlands and **Germany** have recorded the floods in the Netherlands and **Germany** in **1995**.	The Netherlands Federation and the United Nations have agreed with the Netherlands Federation and the Netherlands Federation in **1995**.

Table 6: Comparison between original CLEF queries and translations from Swahili and Somali to English. Tokens that occur both in the original query and translations are highlighted in **bold** (ignoring case, excluding stopwords).

agation from NMT impacts CLIR performance to different extents. Those include, e.g., missing keywords (*statistics*; QID:200), topic shifts (*sports* vs. *business*; SO→EN, QID:172) or queries consisting of unrelated text and repetitions (i.e., 'hallucinations'; SO→EN, QID:151, QID:200). Especially repetitions and hallucinations[12] are known unwanted artifacts in NMT (Fu et al., 2021; Raunak et al., 2021) and can cause retrieval models to emphasize unrelated keywords by inflating their term frequency.[13] Lastly, in cases where source words are copied instead of translated, e.g., *Nugleerka (Nuclear)* or *Jarmalka (Germany)* in QID:187, neural retrieval models need to rely on imperfect internal alignment of word translations (Cao et al., 2019).

5 Related Work

Next to Adapters and SFTMs there exist other parameter efficient transfer (PET) methods. For example, BitFit trains only bias vectors (Ben Zaken et al., 2022), LoRa trains low-rank decompositions of weight matrices in dense layers (Hu et al., 2022) and methods that learn continuous prompts (Liu et al., 2021b; Lester et al., 2021; Li and Liang, 2021, *inter alia*). In the context of retrieval for English, concurrent work focuses on the learning-efficiency (Ma et al., 2022) and out-of-domain generalization (Tam et al., 2022) of PET methods, whereas we investigate PET both on task- and language-level adaption for CLIR.

6 Conclusion

In this work, we introduced modular and parameter-efficient neural rerankers for effective cross-lingual retrieval transfer. Our models, based on Adapters and Sparse Fine-Tuning Masks, allow for decoupling of language-specific and task-specific (i.e., ranking) knowledge. We demonstrate that this leads to more effective transfer to cross-lingual IR setups as well as to better cross-lingual transfer for monolingual retrieval in target languages with no relevance judgment improving over strong prerankers based on state-of-the-art NMT. Encouragingly, we observe particularly pronounced gains for low-resource languages included in our evaluation. We hope that our results will encourage a broader investigation of parameter-efficient neural retrieval in monolingual and cross-lingual setups. We make our code and resources available at: https://github.com/rlitschk/ModularCLIR.

Acknowledgments

RL and GG are supported by the EUINACTION grant from NORFACE Governance (462-19-010, GL950/2-1). IV work is supported by a Huawei research donation to the University of Cambridge.

[12] This phenomenon has been reported to occur in low-resource and out-of-domain settings (Müller et al., 2020). We confirm this finding as we find hallucinations appearing more often in EN→SO than in EN→SW query translations.

[13] Further investigation of NMT+BM25 on SO→EN reveals that manually filtering out queries containing more than two repetitions/hallucinations leaves us with 22 remaining queries on which results improve from 0.157 to 0.280 MAP.

References

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of ACL 2022*, pages 1778–1796.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of NAACL*, pages 547–564.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of ACL*, pages 1–9.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Hamed Bonab, James Allan, and Ramesh Sitaraman. 2019. Simulating clir translation resource scarcity using high-resource languages. In *Proceedings of ICTIR*, page 129–136.

Luiz Henrique Bonifacio, Israel Campiotti, Vitor Jeronymo, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

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

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual alignment of contextual word representations. In *Proceedings of ICLR*.

Francisco Casacuberta, Alexandru Ceausu, Khalid Choukri, Miltos Deligiannis, Miguel Domingo, Mercedes Garcıa-Martınez, Manuel Herranz, Vassilis Papavassiliou, Stelios Piperidis, Prokopis Prokopidis, et al. 2021. The covid-19 mlia@eval initiative: Overview of the machine translation task.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. Ms marco: Benchmarking ranking models in the large-data regime. In *Proceedings of SIGIR*, pages 1566–1576.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. In *TREC*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint*.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. *Proceedings of AAAI*, 35(14):12848–12856.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of NAACL*, pages 3030–3042.

Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. 2021. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *CoRR*, abs/2103.11920.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. Xhate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365.

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 SIGIR*, page 113–122.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of ICML*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

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

2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceeding of ICML*, pages 4411–4421. PMLR.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021a. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of EMNLP*, pages 1684–1697.

Zhiqi Huang, Hamed Bonab, Sheikh Muhammad Sarwar, Razieh Rahimi, and James Allan. 2021b. Mixed attention transformer for leveraging word-level knowledge to neural cross-lingual information retrieval. In *Proceedings of CIKM*, pages 760–770.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Ponzetto, and Goran Glavaš. 2022. Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703, Seattle, United States. Association for Computational Linguistics.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. In *Proceedings of LREC*, page 26.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of ACL*, pages 6282–6293.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of SIGIR*, pages 39—48.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of EMNLP*, pages 4483–4499.

Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022. Hc4: A new suite of test collections for ad hoc clir. In *Proceedigs of ECIR*, page 351–366. Springer-Verlag.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*, pages 3045–3059.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL*, pages 4582–4597.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of SIGIR*, pages 2356–2362.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Proceedings of ECIR*, pages 342–358.

Qian Liu, Xiubo Geng, Jie Lu, and Daxin Jiang. 2021a. Pivot-based candidate retrieval for cross-lingual entity linking. In *Proceedings of WWW*, page 1076–1085.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Scattered or connected? an optimized parameter-efficient tuning approach for information retrieval. *arXiv preprint arXiv:2208.09847*.

Sean MacAvaney, Luca Soldaini, and Nazli Goharian. 2020. Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. In *Proceedings of ECIR*, pages 246–254.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. Cedr: Contextualized embeddings for document ranking. In *Proceedings of SIGIR*, page 1101–1104.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. 2020. Understanding the role of training regimes in continual learning. In *Proceedings of NeurIPS*, volume 33, pages 7308–7320.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr, et al. 2021. A large-scale study of machine translation in turkic languages. In *Proceedings of EMNLP*, pages 5876–5890.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164.

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 *Proceedings of ECIR*, pages 382–396.

Shahrzad Naseri, Jeff Dalton, Andrew Yates, and James Allan. 2021. CEQE: contextualized embeddings for query expansion. In *Proceedings of ECIR*, pages 467–482.

Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to docttttquery. *Online preprint*, 6.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019b. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019c. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings EMNLP*, pages 7654–7673.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP*, pages 2362–2376.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of NAACL*, pages 1172–1183.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Proceedings of NeurIPS*, volume 30.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of EMNLP*, pages 4512–4525.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of EMNLP*, pages 7930–7946.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of EMNLP*.

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 NAACL*, pages 3715–3734.

Peng Shi, He Bai, and Jimmy Lin. 2020. Cross-lingual training of neural models for document ranking. In *Proceedings of EMNLP (Findings)*, pages 2768–2773.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Weng Lam Tam, Xiao Liu, Kaixuan Ji, Lilong Xue, Xingjian Zhang, Yuxiao Dong, Jiahua Liu, Maodi Hu, and Jie Tang. 2022. Parameter-efficient prompt tuning makes generalized and calibrated neural text retrievers. *CoRR*, abs/2207.07087.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Fedor Vitiugin and Carlos Castillo. 2022. Cross-lingual query-based summarization of crisis-related social media: An abstractive approach using transformers. *arXiv preprint arXiv:2204.10230*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45.

Puxuan Yu, Hongliang Fei, and Ping Li. 2021. Cross-lingual language model pretraining for retrieval. In *Proceedings of WWW*, page 1029–1039.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of ACL*, pages 1656–1671.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. In *Proceedings of EMNLP-IJCNLP*, pages 3054–3064.