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

Robert Litschko¹  Ivan Vulić²  Goran Glavaš¹,³
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-
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). 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-

1In 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).
we adopt this paradigm for cross-lingual ad-hoc retrieval. **Stage 1 - Preranking:** We rank the document collection $C$ 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 $R_1$.

Commonly 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 $q_1$ be a query in language $l_1$ and $O^{l_2} = \{d_i\}_{i=1}^n$ 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

$$R_0 = [(d_1, s_1), (d_2, s_2), \ldots (d_n, s_n)], \tag{1}$$

where $s_1 > s_2 > \ldots > s_n$. 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 $s^{BM25} = BM25(q, d)$.\(^3\) 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:

$$R_1 = [(d_1, s^{CE}_1), (d_2, s^{CE}_2), \ldots (d_k, s^{CE}_k)] \tag{2}$$

To this end, we rely on multilingual CEs to compute the binary relevance score $s^{CE}$ on the concatenation of query and document pairs: $s^{CE} = CE([CLS]q[SEP]d_1[SEP])$. 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.
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 $r_l$ and the hidden state after the layer norm with $h_l$.

$\text{LA}(h_l, r_l) = U_l(\psi(D_l(h_l))) + r_l$ \hfill (3)

$\text{RA}(h_l, r_l) = U_l(\psi(D_l(\text{LA}_l))) + r_l$ \hfill (4)

Adapters are parameterized by the down-projection matrix $D \in \mathbb{R}^{h \times d}$ and the up-projection matrix $U \in \mathbb{R}^{d \times h}$, 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.\footnote{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 $\text{INV}$ on top of the embedding layer along with their inverses $\text{INV}^{-1}$ 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).}

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
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 $M = \theta^{(2)} - \theta^{(0)}$.

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 $\theta^{(0)} + RM + LM$ (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 (LM$^{B} = LM^{Q} + LM^{D}$). 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 482K (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 on MS-MARCO. For CLIR experiments, we opt for DISTIL-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-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 (Mirzakhalov et al., 2021), because we failed to obtain meaningful {KG, UG} → $l_2$ 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 using DISTIL-DmBERT and NMT+BM25 as Stage 1 rerankers, respectively. With DISTIL-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. Some-
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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT+BM25 (PR)</td>
<td>.325</td>
<td>.157</td>
<td>.228</td>
<td>.091</td>
<td>.183</td>
<td>.113</td>
<td>.186</td>
<td>.183</td>
<td></td>
</tr>
<tr>
<td>MonoBERT</td>
<td>.362</td>
<td>.158</td>
<td>.235</td>
<td>.157</td>
<td>.246</td>
<td>.172</td>
<td>.218</td>
<td>.224</td>
<td>.216</td>
</tr>
<tr>
<td>+RA + LA A</td>
<td>.407</td>
<td>.166</td>
<td>.305</td>
<td>.155</td>
<td>.259</td>
<td>.189</td>
<td>.234</td>
<td>.245</td>
<td>.228</td>
</tr>
</tbody>
</table>

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 (ENS, last column) yields further gains with DISTIL-DmBERT as the preranker (Table 1), but not when we prerank with NMT+BM25 (Table 2).

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). 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...
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.

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

Table 5: Trade-off between efficiency and effectiveness when dropping adapters in +RA + LA\textsuperscript{11}. 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,\textsuperscript{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-

\textsuperscript{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.
Table 6: Comparison between original CLEF queries and translations from Swahili and Somali to English. Tokens on which results improve from 0.157 to 0.280 MAP.

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

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).

Agitation 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 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. 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 pre-rankers 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](https://github.com/rlitschk/ModularCLIR).

Acknowledgments

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


