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

COLING 2022

Outline

1. Introduction
2. Parameter-Efficient Reranking
3. Experimental Setup
4. Results
5. Conclusion

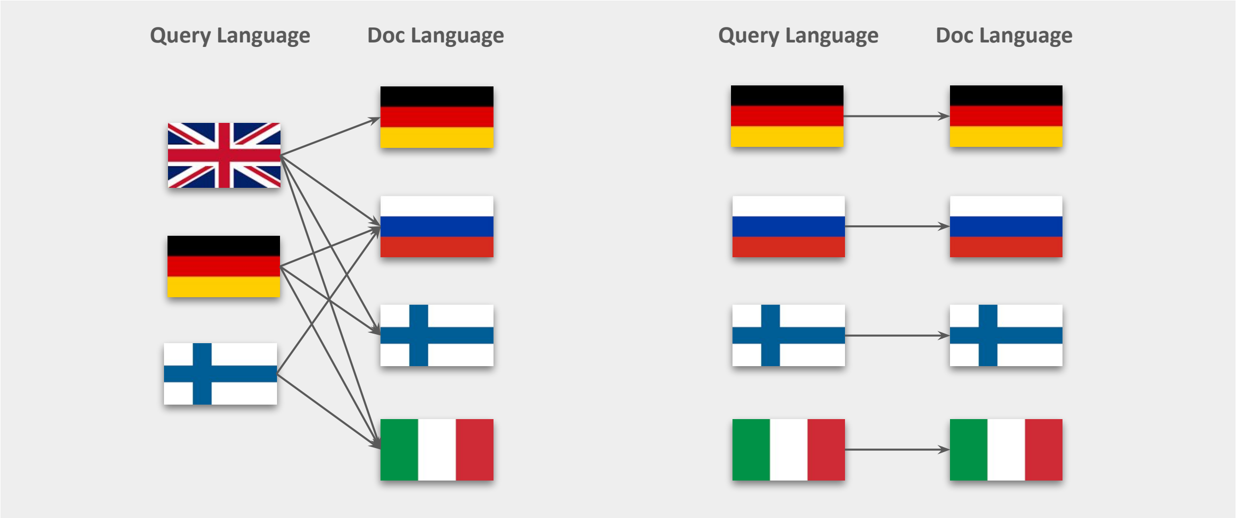
Outline

1. **Introduction**
    a. Cross-Lingual IR vs. Cross-Lingual Transfer
    b. Multi-Stage Ranking
    c. Problem Statement
2. Parameter-Efficient Reranking
3. Experimental Setup
4. Results
5. Conclusion

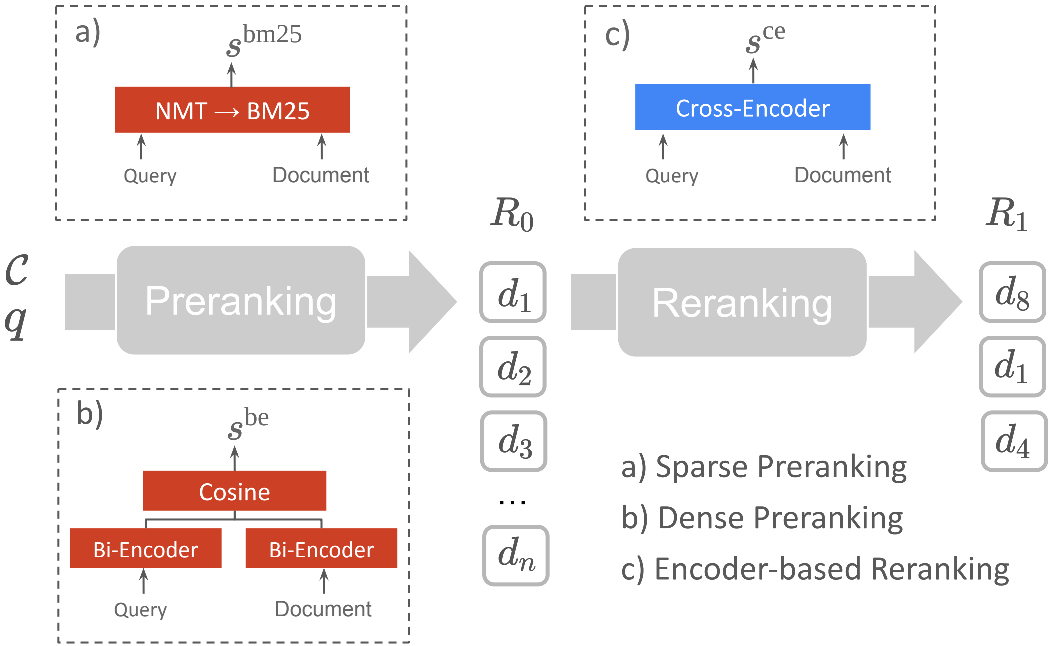
Cross-Lingual IR vs. Cross-Lingual Transfer

Cross-Lingual Information Retrieval (CLIR)

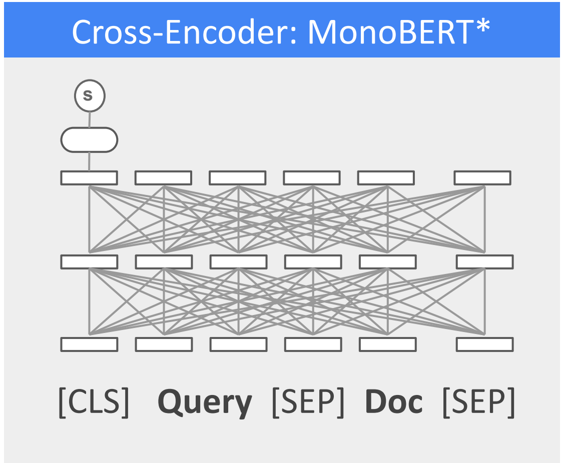
Cross-Lingual Transfer for Monolingual IR (MoIR)



Cross-Lingual Multi-Stage Ranking



Problem Statement



Transferring MonoBERT to new languages

- Maintain one model each language pair.
  - Space inefficient
  - Impossible without large scale training data.
- Alternative: Multilingual encoder.
- **Curse of Multilinguality** (Conneau et al., ACL'20). Model capacity restricts the number of languages that can practically be encoded with a multilingual LM.
- **Catastrophic Forgetting** (Mirzadeh et al., NeurIPS'20). Training MonoBERT on EN-EN might overwrite features important for other languages.

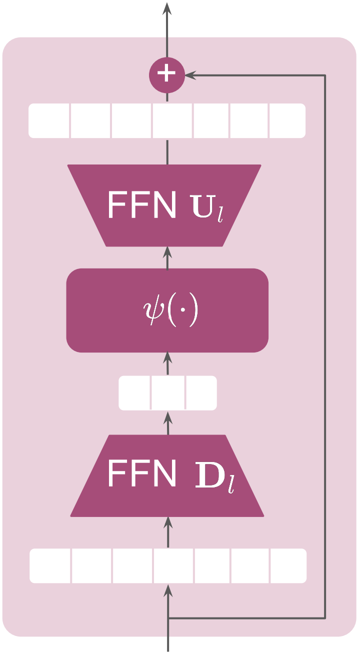
*Nogueira et al. (2019)

Outline

1. Introduction
2. **Parameter-Efficient Reranking**
    a. Adapters
    b. Sparse Fine-Tuning Masks
3. Experimental Setup
4. Results
5. Conclusion

Adapters

Adapters



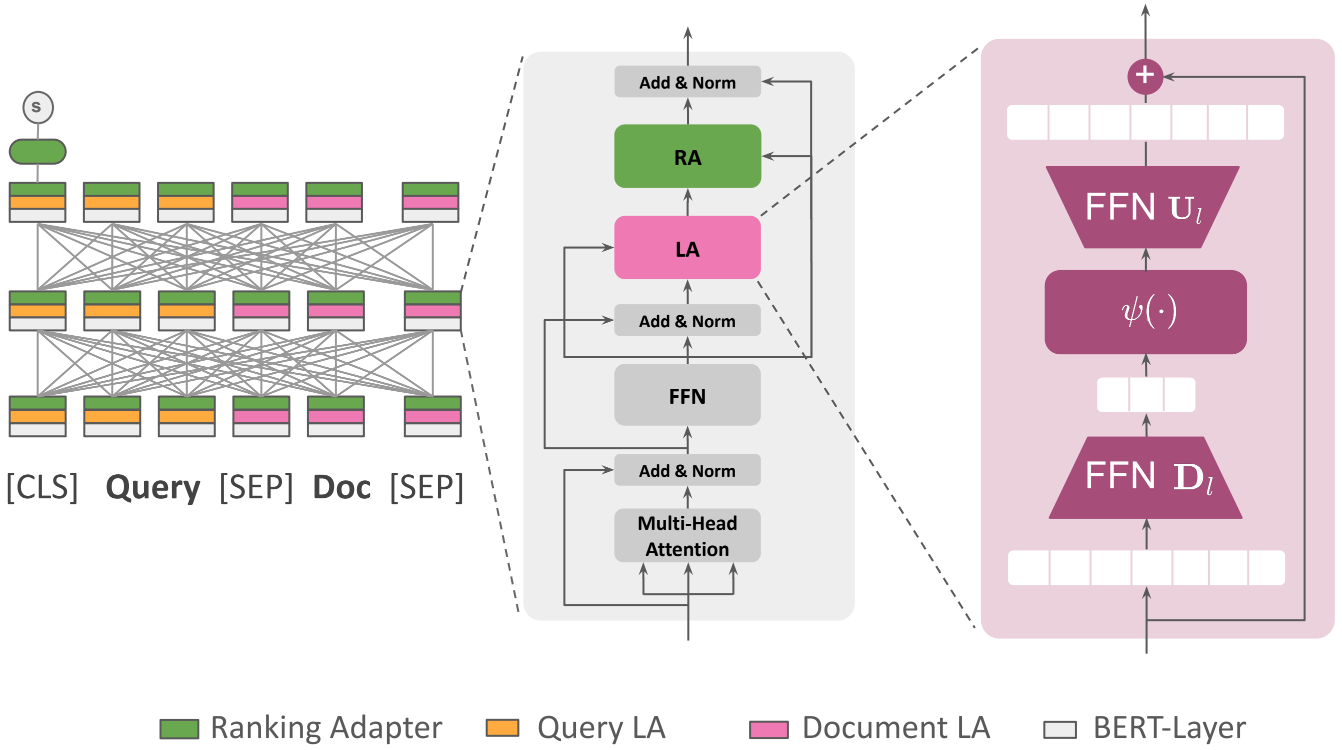
- **Bottleneck Adapters** proposed by (Houlsby et al., ICML'19).
- Instead of training the full model, inject and train adapter modules.
- **Parameter-Efficient** – Keep rest of the model frozen during training.
- **MAD-X** (Pfeiffer et al., EMNLP'19): Adapters encode task-specific kowledge, stacking adapters enables multi-task cross-lingual transfer.

$$\text{Adapter}(h_l, r_l) = \mathbf{U}_l(\psi(D_l(h_l)) + r_l$$

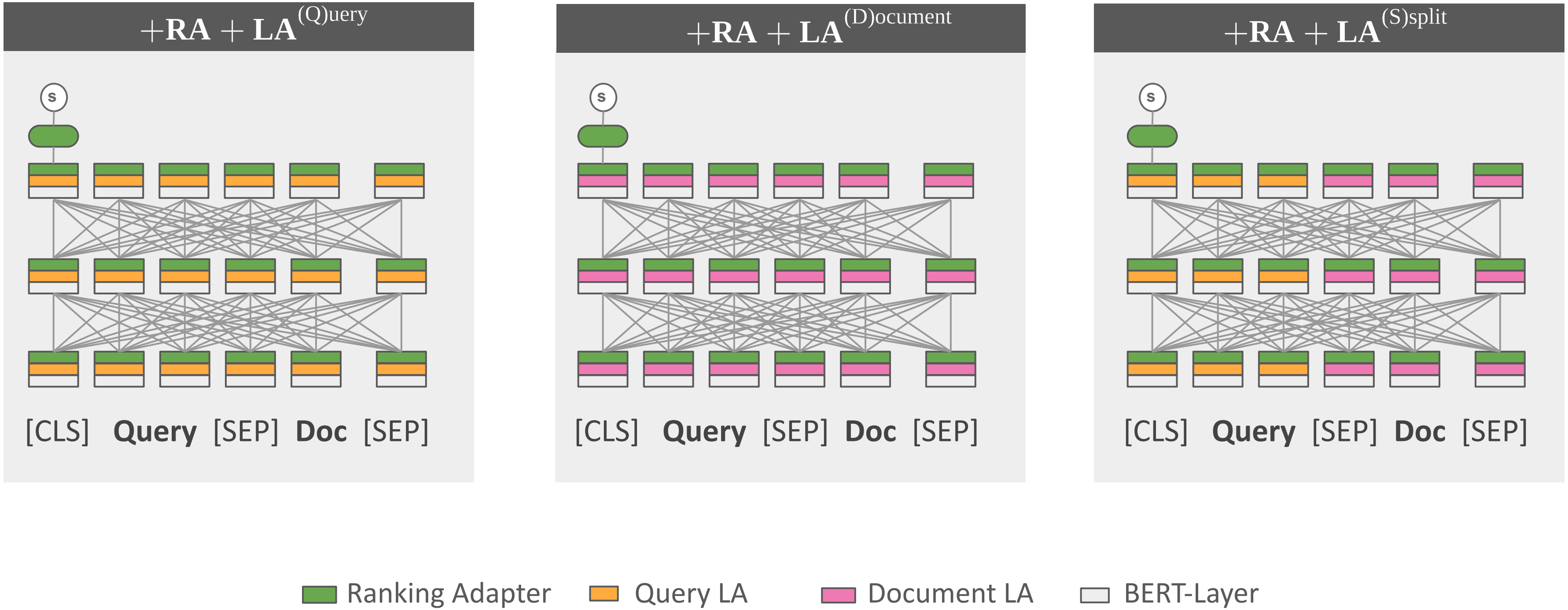
$$\text{Reduction Factor} = \frac{dim(h_l)}{dim(D_l(h_l))}$$

Image based on: *"Parameter-Efficient Transfer Learning for NLP"* Houlsby et al. (2019)

Adapters for CLIR



Composing Rerankers



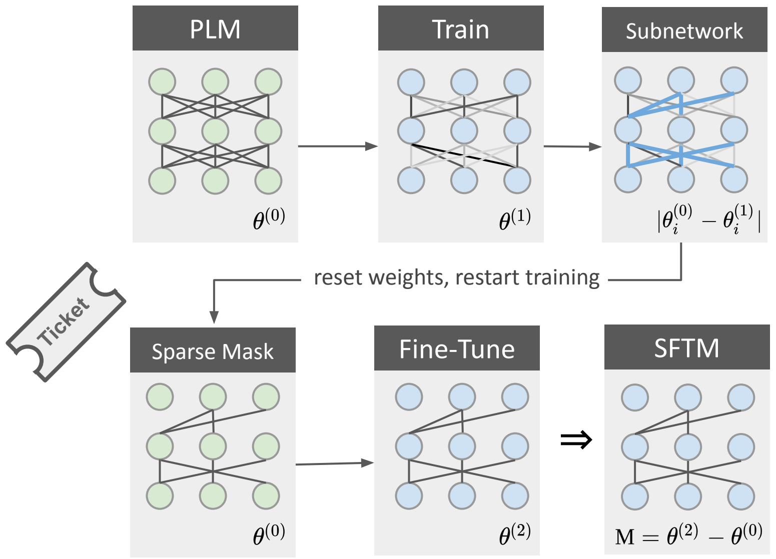
Sparse Fine-Tuning Masks (SFTMs)

Lottery Ticket Hypothesis

“A randomly-initialized, dense neural network contains a **subsetwork that is initialized such that** – when trained in isolation – **it can match the test accuracy of the original network** after training for at most the same number of iterations.”

- Frankle & Carbin (ICLR'19)

Sparse Fine-Tuning Masks (Ansell et al., ACL'21)



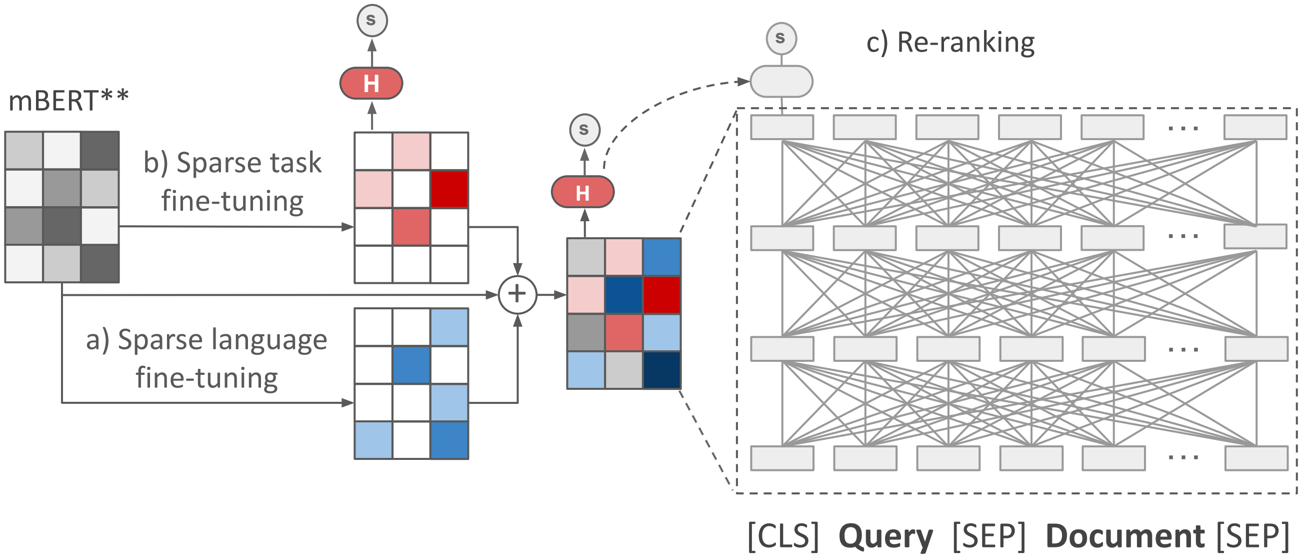
Training SFTMs

- **Find "Winning lottery ticket" –** Train pretrained LM (PLM) on a task and extract subnetwork with top k largest weight changes.
- **Sparse Fine-Tuning –** Reset weights and restart training, keeping all weights except for subnetwork frozen.
- SFTM is obtained as difference vector on subnetwork

Sparsity ≈ Reduction Factor (Adapters)

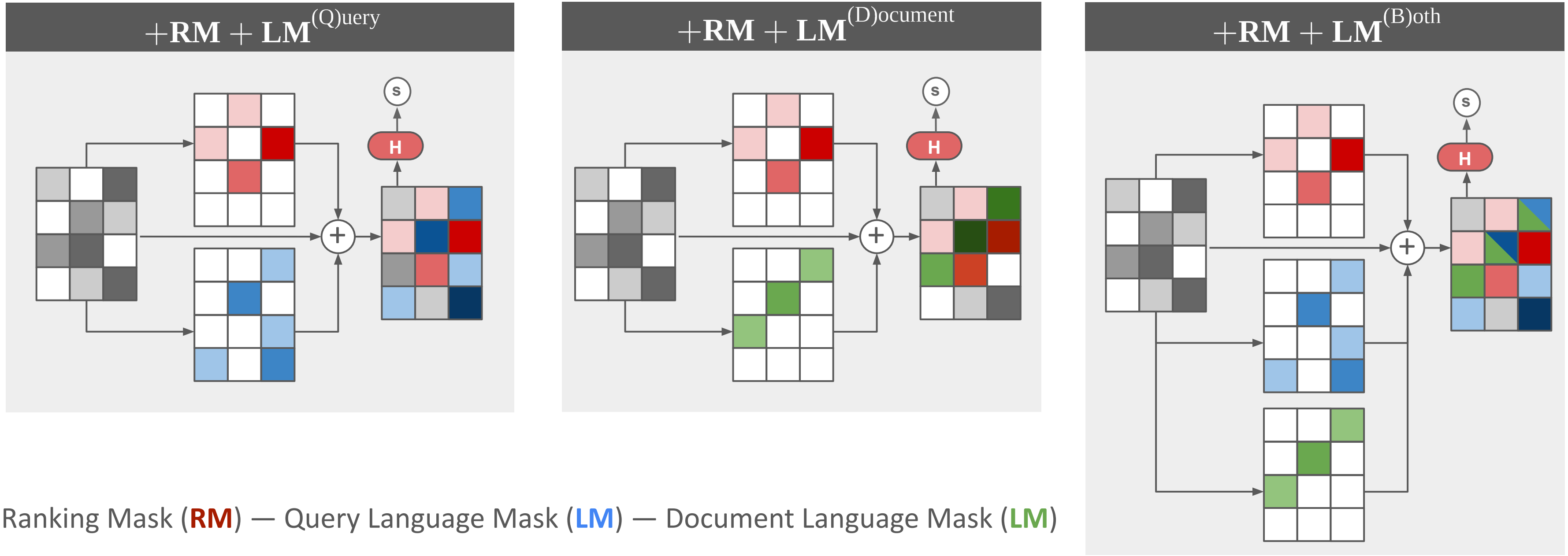
Illustration inspired by: *"The Lottery Ticket Hypothesis: A Survey"* Robert Tjarko Lange (2020)

Sparse Fine-Tuning Masks for CLIR*



*Following Ansell et al. (ACL'21), **(Devlin et al., NAACL'19)

Composing Rerankers



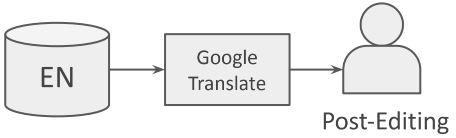
Ranking Mask (RM) — Query Language Mask (LM) — Document Language Mask (LM)

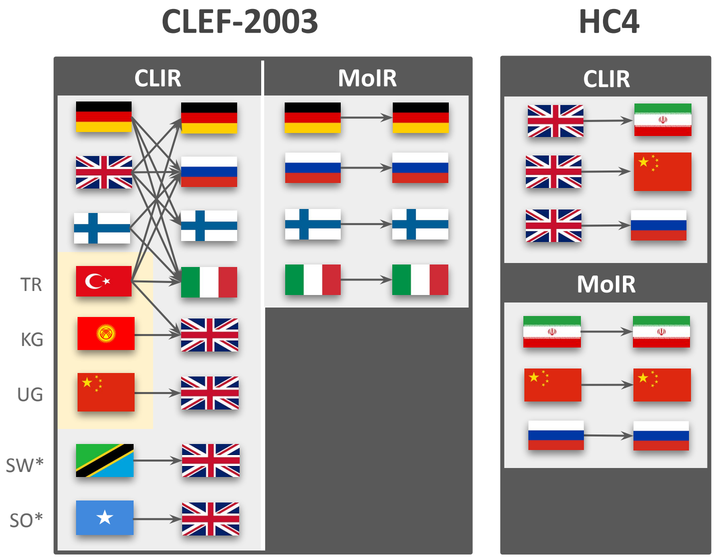
Outline

1. Introduction
2. Parameter-Efficient Reranking
3. **Experimental Setup**
    a. Evaluation Datasets
    b. Baselines
4. Results
5. Conclusion

Evaluation Datasets

- We **train** ranking models on MS-MARCO (Craswell et al., SIGIR'21).
- We **evaluate** ranking models on 29 language pairs from:
  - CLEF 2003 (Braschler, LNCS'03)
  - HC4 (Lawrie et al., ECIR'22)
- All models **rerank** the top-100 documents.
- In addition to existing CLEF languages **we release three new CLEF query languages**: Turkish (TR), Kyrgyz (KG), Uyghur (UG).



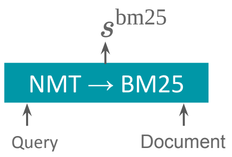


*We use Swahili (SW) and Somali (SO) query translations from Bonab et al. (CIKM'19)

Baselines

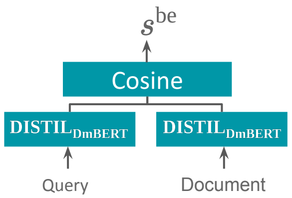
NMT→BM25 (PR)

- **Query Translation** with SOTA NMT models (Fan et al., JML'21; Mirzakhalov, EMNLP'21).
- Retrieve documents with BM25 (**Monolingual Lexical Retrieval**).



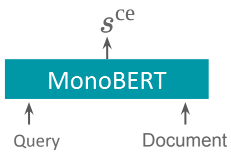
$\text{DISTIL}_{\text{DmBERT}}$ (PR)

- Multilingual encoder trained with **Knowledge DISTILation** (Reimers et al., EMNLP'20).
- **Bi-Encoder**: Encode query and document independendly, compute relevance score as cosine similarity between representations.



MonoBERT

- **Full fine-tuning**: mBERT-based ranking model (Nogueira et al., arxiv'19) trained on MS-MARCO.
- Zero-shot reranking with **Cross-Encoders** (MacAvaney et al., ECIR'20).



PR = Preranker

Outline

1. Introduction
2. Parameter-Efficient Reranking
3. Experimental Setup
4. **Results**
    a. CLIR results on CLEF 2003
    b. CLIR results on extended CLEF and HC4
    c. Impact of NMT on CLIR
5. Conclusion

CLIR Results on CLEF 2003

	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$_{DmBERT}$ (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
Adapters	+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
SFTMs	+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

Mean Average Precision (MAP)

Preranker (PR): Bi-Encoder

- **MonoBERT** trained on MS-MARCO (EN-EN) improves preranker results on all languages pairs involving English*, mixed results on other language pairs.
- **Ensembling** preranker and reranker (average rank) improves retrieval results.
- Both **Adapters**** and **SFTMs**** improve over baselines while training with fewer parameters.

*Except for EN-RU **Adaptes and SFTMs have reduction factors of 16 and 2, respectively (see paper for ablation).

CLIR Results on extended CLEF and HC4

Model	CLEF 2003 SW–EN	SO–EN	KG–EN	UG–EN	HC4 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
$+\text{RA} + \text{LA}^D$	**.407**	**.166**	.305	.155	.259	.189	.234	.245	**.228**
$+\text{RM} + \text{LM}^D$	.389	.161	**.311**	**.165**	**.267**	**.196**	**.241**	**.247**	.225

Mean Average Precision (MAP)

Preranker (PR): NMT+BM25*

- NMT casts CLIR into a noisy variant of MoIR. Both **Adapters** and **SFTMs** improve over baselines.
- Results on **low-resource** and **distant languages** generally lower than results on high-resource languages.
- But: Gains are less pronounced when preranker/MonoBERT results are low.

We found low results to be related to NMT quality!

* **DISTIL**$_{\textbf{DmBERT}}$ does not support languages used here.

Impact of NMT on CLIR

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**.

- **Topic shifts**: sports vs. business
- **“Hallucinations”**: queries consisting of unrelated text and repetitions*
- **Copy source words**: *Nugleerka* (Nuclear), *Jarmalka* (Germany)
- Slight **lexical** and **Semantic variations**: *flooding* vs. *floods*, *holland* vs. *netherlands*

*Filtering out queries that contain more than two repetitions improves from 0.157 to 0.280 MAP (MonoBERT)

Outline

1. Introduction
2. Parameter-Efficient Reranking
3. Experimental Setup
4. Results
5. **Conclusion**

Conclusion

In this work we …

… introduced **modular** and **parameter-efficient** neural rerankers for effective cross-lingual transfer.

… demonstrate the effectiveness of Adapters and SFTMs for Cross-Lingual IR.

… released **three new CLEF query languages** to encourage research on low-resource CLIR.

More results in our paper:

- **Monolingual IR (MoIR)** results on high- and low-resource languages.
- **Parameter Efficiency** – Ablation of different levels of sparsity (reduction factor).
- **AdapterDrop** (Rücklé et al., EMNLP'21) – Speed vs. effectiveness, drop Adapters in lower layers.

github.com/rlitschk/ModularCLIR   @ robert.litschko@uni-mannheim.de