Boosting Zero-shot Cross-lingual Retrieval by Training on Artificially Code-Switched Data

Robert Litschko Ekaterina Artemova Barbara Plank
MaiNLP, Center for Information and Language Processing (CIS), LMU Munich, Germany
{robert.litschko, ekaterina.artemova, b.plank}@lmu.de

Abstract

Transferring information retrieval (IR) models from a high-resource language (typically English) to other languages in a zero-shot fashion has become a widely adopted approach. In this work, we show that the effectiveness of zero-shot rankers diminishes when queries and documents are present in different languages. Motivated by this, we propose to train ranking models on artificially code-switched data instead, which we generate by utilizing bilingual lexicons. To this end, we experiment with lexicons induced from (1) cross-lingual word embeddings and (2) parallel Wikipedia page titles. We use the mMARCO dataset to extensively evaluate reranking models on 36 language pairs spanning Monolingual IR (MoIR), Cross-lingual IR (CLIR), and Multilingual IR (MLIR). Our results show that code-switching can yield consistent and substantial gains of 5.1 MRR@10 in CLIR and 3.9 MRR@10 in MLIR, while maintaining stable performance in MoIR. Encouragingly, the gains are especially pronounced for distant languages (up to 2x absolute gain). We further show that our approach is robust towards the ratio of code-switched tokens and also extends to unseen languages. Our results demonstrate that training on code-switched data is a cheap and effective way of generalizing zero-shot rankers for cross-lingual and multilingual retrieval.

1 Introduction

Cross-lingual Information Retrieval (CLIR) is the task of retrieving relevant documents written in a language different from a query language. The large number of languages and limited amounts of training data pose a serious challenge for training ranking models. Previous work address this issue by using machine translation (MT), effectively casting CLIR into a noisy variant of monolingual retrieval (Li and Cheng, 2018; Shi et al., 2020, 2021; Moraes et al., 2021). MT systems are used to either train ranking models on translated training data (*translate train*), or by translating queries into the document language at retrieval time (*translate test*). However, CLIR approaches relying on MT systems are limited by their language coverage. Because training MT models is bounded by the availability of parallel data, it does not scale well to a large number of languages. Furthermore, using MT for IR has been shown to be prone to propagation of unwanted translation artifacts such as topic shifts, repetition, hallucinations and lexical ambiguity (Artetxe et al., 2020; Litschko et al., 2022a; Li et al., 2022). In this work, we propose a resource-lean MT alternative to bridge the language gap and propose to use *artificially code-switched* data.

We focus on zero-shot cross-encoder (CE) models for reranking (MacAvaney et al., 2020; Jiang et al., 2020). Our study is motivated by the observation that the performance of CEs diminishes when they are transferred into CLIR and MLIR as opposed to MoIR. We hypothesize that training on queries and documents from the same language leads to *monolingual overfitting* where the ranker learns features, such as exact keyword matches, which are useful in MoIR but do not transfer well to CLIR and MLIR setups due to the lack of lexical overlap (Litschko et al., 2022b). In fact, as shown by Roy et al. (2020) on bi-encoders, representations from zero-shot models are weakly aligned between languages, where models prefer non-relevant documents in the same language over relevant documents in a different language. To address this problem, we propose to use code-switching as an inductive bias to regularize monolingual overfitting in CEs.

Generation of synthetic code-switched data has served as a way to augment data in cross-lingual setups in a number of NLP tasks (Singh et al., 2019; Einolghozati et al., 2021; Tan and Joty, 2021). They utilize substitution techniques ranging from simplistic re-writing in the target script (Gautam et al., 2021), looking up bilingual lexicons (Tan and Joty,

Findings of the Association for Computational Linguistics: ACL 2023, pages 3096–3108
July 9-14, 2023

2021) to MT (Tarunesh et al., 2021). Previous work on improving zero-shot transfer for IR includes weak supervision (Shi et al., 2021), tuning the pivot language (Turc et al., 2021), multilingual query expansion (Blloshmi et al., 2021) and cross-lingual pre-training (Yang et al., 2020; Yu et al., 2021; Yang et al., 2022; Lee et al., 2023). To this end, code-switching is complementary to existing approaches. Our work is most similar to Shi et al. (2020), who use bilingual lexicons for full term-by-term translation to improve MoIR. Concurrent to our work, Huang et al. (2023) show that code-switching improves the retrieval performance on low-resource languages, however, their focus lies on CLIR with English documents. To the best of our knowledge, we are the first to systematically investigate (1) artificial code-switching to train CEs and (2) the interaction between MoIR, CLIR and MLIR.

Our contributions are as follows: (i) We show that training on artificially code-switched data improves zero-shot cross-lingual and multilingual rankers. (ii) We demonstrate its robustness towards the ratio of code-switched tokens and effectiveness in generalizing to unseen languages. (iii) We release our code and resources.[1]

2 Methodology

Reranking with Cross-Encoders. We follow the standard cross-encoder reranking approach (CE) proposed by Nogueira and Cho (2019), which formulates relevance prediction as a sequence pair (query-document pair) classification task. CEs are composed of an encoder model and a relevance prediction model. The encoder is a pre-trained language model (Devlin et al., 2019) that transforms the concatenated input `[CLS] Q [SEP] D [SEP]` into a joint query-document feature representation, from which the classification head predicts relevance. Finally, documents are reranked according to their predicted relevance. We argue that fine-tuning CEs on monolingual data biases the encoder towards encoding features that are only useful when the target setup is MoIR. To mitigate this bias, we propose to perturb the training data with code-switching, as described next.

Artificial Code-Switching. While previous work has studied code-switching (CS) as a natural phenomenon where speakers borrow words from other languages (e.g. anglicism) (Ganguly et al., 2016; Wang and Komlodi, 2018), we here refer to code-switching as a method to *artificially* modify monolingual training data. In the following we assume availability of English (EN–EN) training data. The goal is to improve the zero-shot transfer of ranking models into cross-lingual language pairs X–Y by training on code-switched data EN_X–EN_Y instead, which we obtain by exploiting bilingual lexicons similar to Tan and Joty (2021). We now describe two CS approaches based on lexicons: one derived from word embeddings and one from Wikipedia page titles (cf. Appendix A for examples).

Code-Switching with Word Embeddings. We rely on bilingual dictionaries $\mathcal{D}$ induced from cross-lingual word embeddings (Mikolov et al., 2013; Heyman et al., 2017) and compute for each EN term its nearest (cosine) cross-lingual neighbor. In order to generate EN_X–EN_Y we then use $\mathcal{D}_{\text{EN}\to\text{X}}$ and $\mathcal{D}_{\text{EN}\to\text{Y}}$ to code-switch query and document terms from EN into the languages X and Y, each with probability p. This approach, dubbed Bilingual CS (**`BL-CS`**), allows a ranker to learn inter-lingual semantics between EN, X and Y. In our second approach, Multilingual CS (**`ML-CS`**), we additionally sample for each term a different target language into which it gets translated; we refer to the pool of available languages as seen languages.

Code-Switching with Wikipedia Titles. Our third approach, **`Wiki-CS`**, follows (Lan et al., 2020; Fetahu et al., 2021) and uses bilingual lexicons derived from parallel Wikipedia page titles obtained from inter-language links. We first extract word n-grams from queries and documents with different sliding window of sizes $n \in \{1, 2, 3\}$. Longer n-gram are favored over shorter ones in order to account for multi-term expressions, which are commonly observed in named entities. In `Wiki CS` we create a single multilingual dataset where queries and documents from different training instances are code-switched into different languages.

3 Experimental Setup

Models and Dictionaries. We follow Bonifacio et al. (2021) and initialize rankers with the multilingual encoder mMiniLM provided by Reimers and Gurevych (2020). We report hyperparameters in Appendix C. For `BL-CS` and `ML-CS` we use multilingual MUSE embeddings[2] to induce bilingual

[1] https://github.com/MaiNLP/CodeSwitchCLIR

[2] https://github.com/facebookresearch/MUSE

	EN–EN	DE–DE	RU–RU	AR–AR	NL–NL	IT–IT	AVG	Δ_{ZS}
`Zero-shot`	35.0	25.9	23.8	**23.9**	27.2	26.9	25.5	-
`Fine-tuning`	35.0	30.3*	28.5*	27.2*	30.8*	30.9*	29.5	+4.0
`Zero-shot`$_{\text{Translate Test}}$	-	22.5*	18.2*	17.7*	24.7*	23.3*	21.3	-4.2
`ML-CS`$_{\text{Translate Test}}$	-	22.8*	18.6*	17.7*	24.7*	24.5*	21.7	-3.8
`BL-CS`	-	**26.0**	**25.5**	23.0	**27.5**	**27.2**	**25.8**	**+0.3**
`ML-CS`	34.0	25.9	24.7	21.3	27.2	26.9	25.2	-0.3
`Wiki-CS`	33.8*	25.6	24.1	20.5*	27.0	25.5*	24.5	-1.0

Table 1: MoIR: Monolingual results on mMARCO languages and averaged over all languages (excluding EN–EN) in terms of MRR@10. **Bold**: Best zero-shot performance for each language. Δ_{ZS}: Absolute difference to `Zero-shot`. Results significantly different from `Zero-shot` are marked with * (paired t-test, Bonferroni correction, $p < 0.05$).

	EN–DE	EN–IT	EN–AR	EN–RU	DE–IT	DE–NL	DE–RU	AR–IT	AR–RU	AVG	Δ_{ZS}
`Zero-shot`	24.0	23.0	14.0	18.3	15.0	19.7	12.9	7.7	7.1	15.7	-
`Fine-tuning`	29.7*	30.5*	26.5*	28.0*	26.9*	27.9*	25.5*	23.9*	22.7*	26.8	+11.1
`Zero-shot`$_{\text{Translate Test}}$	22.8	23.2	16.4	17.0	15.8	17.5	11.8	9.8	8.7	15.9	+0.2
`ML-CS`$_{\text{Translate Test}}$	24.9	24.6	17.9*	19.5	17.6	19.3*	14.3	12.2*	10.6*	17.9	+2.2
`BL-CS`	**26.9***	**27.3***	19.3*	22.8*	**20.4***	**22.8***	17.8*	**15.6***	14.1*	**20.8**	**+5.1**
`ML-CS`	26.5*	26.4*	18.1*	22.1*	19.8*	**22.8***	17.8*	15.3*	14.2*	20.3	+4.6
`Wiki-CS`	26.2*	26.4*	**19.4***	**22.9***	19.4*	22.4*	**18.3***	14.4*	14.1*	20.4	+4.7

Table 2: CLIR: Cross-lingual results on mMARCO in terms of MRR@10.

	Seen Languages					All Languages				
	X–EN	EN–X	X–X	AVG_{seen}	Δ_{seen}	X–EN	EN–X	X–X	AVG_{all}	Δ_{all}
`Zero-shot`	19.0	23.5	16.3	19.6	-	16.5	20.8	12.9	16.6	-
`Fine-tuning`	24.8*	26.4*	21.1*	24.1	+4.5	26.5*	26.5*	21.9*	25.0	+8.3
`ML-CS`	**24.2***	25.9*	**21.1***	**23.7**	+4.1	**21.6***	23.2*	17.0*	20.6	+3.9
`Wiki-CS`	23.6*	**26.0***	20.6*	23.4	+3.8	21.3*	**23.8***	**17.1***	**20.7**	**+4.0**

Table 3: MLIR: Multilingual results on mMARCO in terms of MRR@10. Left: Six seen languages for which we used bilingual lexicons to code-switch training data. Right: All fourteen languages included in mMARCO.

lexicons (Lample et al., 2018), which have been aligned with initial seed dictionaries of 5k word translation pairs. We set the translation probability $p = 0.5$. For `Wiki-CS`, we use the lexicons provided by the linguatools project.[3]

Baselines. To compare whether training on CS'ed data EN_X–EN_Y improves the transfer into CLIR setups, we include the zero-shot ranker trained on EN–EN as our main baseline (henceforth, `Zero-shot`). Our upper-bound reference, dubbed `Fine-tuning`, refers to ranking models that are directly trained on the target language pair X–Y, i.e. no zero-shot transfer. Following Roy et al. (2020), we adopt the *Translate Test* baseline and translate any test data into EN using using our bilingual lexicons induced from word embeddings. On this data we evaluate both the `Zero-shot` baseline (`Zero-shot`$_{\text{Translate Test}}$) and our `ML-CS` model (`ML-CS`$_{\text{Translate Test}}$).

[3]https://linguatools.org/wikipedia-parallel-titles

Datasets and Evaluation. We use use the publicly available multilingual mMARCO data set (Bonifacio et al., 2021), which includes fourteen different languages. We group those into six seen languages (EN, DE, RU, AR, NL, IT) and eight unseen languages (HI, ID, IT, JP, PT, ES, VT, FR) and construct a total of 36 language pairs.[4] Out of those, we construct setups where we have documents in different languages (EN–X), queries in different languages (X–EN), and both in different languages (X–X). Specifically, for each document ID (query ID) we sample the content from one of the available languages. For evaluation, we use the official evaluation metric MRR@10.[5] All models re-rank the top 1,000 passages provided for the passage re-ranking task. We report all results as averages over three random seeds.

[4]Due to computational limitations we don't exhaustively evaluate on all possible language pairs.

[5]We use the implementation provided by the `ir-measures` package (MacAvaney et al., 2022).

4 Results and Discussion

We observe that code-switching improves cross-lingual and multilingual re-ranking, while not impeding monolingual setups, as shown next.

Transfer into MoIR vs. CLIR. We first quantify the performance drop when transferring models trained on EN–EN to MoIR as opposed to CLIR and MLIR. Comparing `Zero-shot` results between different settings we find that the average MoIR performance of 25.5 MRR@10 (Table 1) is substantially higher than CLIR with 15.7 MRR@10 (Table 2) and MLIR with 16.6 MRR@10 (Table 3). The transfer performance greatly varies with the language proximity, in CLIR the drop is larger for setups involving typologically distant languages (AR–IT, AR–RU), to a lesser extent the same observation holds for MoIR (AR–AR, RU–RU). This is consistent with previous findings made in other syntactic and semantic NLP tasks (He et al., 2019; Lauscher et al., 2020). The performance gap to `Fine-tuning` on translated data is much smaller in MoIR (+4 MRR@10) than in CLIR (+11.1 MRR@10) and MLIR (+8.3 MRR@10). Our aim to is close this gap between zero-shot and full fine-tuning in a resource-lean way by training on code-switched queries and documents.

Code-Switching Results. Training on code-switched data consistently outperforms zero-shot models in CLIR and MLIR (Table 2 and Table 3). In AR–IT and AR–RU we see improvements from 7.7 and 7.1 MRR@10 up to 15.6 and 14.1 MRR@10, rendering our approach particularly effective for distant languages. Encouragingly, Table 1 shows that the differences between both of our CS approaches (`BL-CS` and `ML-CS`) versus `Zero-shot` is not statistically significant, showing that gains can be obtained without impairing MoIR performance. Table 2 shows that specializing one zero-shot model for multiple CLIR language pairs (`ML-CS`, `Wiki-CS`) performs almost on par with specializing one model for each language pair (`BL-CS`). The results of `Wiki-CS` are slightly worse in MoIR and on par with `ML-CS` on MLIR and CLIR.

Translate Test vs. Code-Switch Train. In MoIR (Table 1) both $\texttt{Zero-shot}_{\text{Translate Test}}$ and $\texttt{ML-CS}_{\text{Translate Test}}$ underperform compared to other approaches. This shows that zero-shot rankers work better on clean monolingual data in the target language than noisy monolingual data in English.

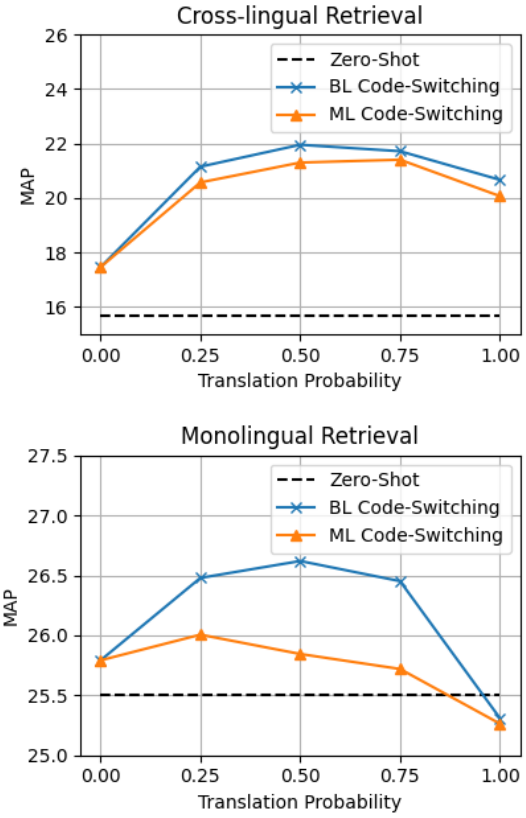

Figure 1: Retrieval performance in terms of mean average precision (MAP) for different translation probabilities, averaged across all language pairs.

In CLIR, where *Translate Test* bridges the language gap between X and Y, we observe slight improvements of +0.2 and +2.2 MRR@10 (Table 2). However, in both MoIR and CLIR *Translate Test* consistently falls behind code-switching at training time.

Multilingual Retrieval and Unseen Languages. Here we compare how code-switching fares against `Zero-shot` on languages to which neither model has been exposed to at training time. Table 3 shows the gains remain virtually unchanged when moving from six seen (+4.1 MRR@10 / +3.8 MRR@10) to fourteen languages including eight unseen languages (+3.9 MRR@10 / +4.0 MRR@10). Results in Appendix B confirm that this holds for unseen languages on the query, document and both sides, suggesting that the best pivot language for zero-shot transfer (Turc et al., 2021) may not be monolingual but a code-switched language. On seen languages `ML-CS` is close to MT (`Fine-tuning`).

Ablation: Translation Probability. The translation probability p allows us to control the ratio of code-switched tokens to original tokens, with $p = 0.0$ we default back to the `Zero-shot` base-

	EN–X	X–EN	X–X
No Code Switching (Zero-Shot)			
No overlap	12.2	11.0	7.4
Some overlap	29.7	22.4	19.9
Significant overlap	44.6	36.4	45.5
All queries	23.5	19.0	16.3
Multilingual Code Switching (ML-CS)			
No overlap	15.5 (+3.3)	17.8 (+6.8)	13.0 (+5.6)
Some overlap	31.7 (+2.0)	27.2 (+4.8)	25.3 (+5.4)
Significant overlap	44.7 (+0.2)	37.8 (+1.4)	45.1 (-0.5)
All queries	25.9 (+2.4)	24.2 (+5.3)	21.1 (+4.8)

Table 4: MLIR results on seen languages (MRR@10) broken down into queries that share no common tokens (no overlap), between one and three tokens (some overlap) and more than three tokens (significant overlap) with their relevant documents. Gains of ML-CS are shown in brackets. EN–X has 3,116 queries with no overlap, 3,095 with some overlap and 769 with significant overlap. X–EN has 3,147 queries with no overlap, 2,972 with some overlap and 861 with significant overlap. X–X has 3,671 queries with no overlap, 2,502 with some overlap and 807 with significant overlap.

line, with $p = 1.0$ we attempt to code-switch every token. [6] Figure 1 (top) shows that code-switching a smaller portion of tokens is already beneficial for the zero-shot transfer into CLIR. The gains are robust towards different values for p. The best results are achieved with $p = 0.5$ and $p = 0.75$ for BL-CS and ML-CS, respectively. Figure 1 (bottom) shows that the absolute differences to Zero-shot are much smaller in MoIR.

Monolingual Overfitting. Exact matches between query and document keywords is a strong relevance signal in MoIR, but does not transfer well to CLIR and MLIR due to mismatching vocabularies. Training zero-shot rankers on monolingual data biases rankers towards learning features that cannot be exploited at test time. Code-Switching reduces this bias by replacing exact matches with translation pairs,[7] steering model training towards learning interlingual semantics instead. To investigate this, we group queries by their average token overlap with their relevant documents and evaluate each group separately on MLIR.[8] The results are shown in Table 4. Unsurprisingly, rankers work best when there is significant overlap between query and document tokens. However, the performance gains resulting from training on code-switched data (ML-CS) are most pronounced for queries with some token overlap (up to +5.4 MRR@10) and no token overlap (up to +6.8 MRR@10). On the other hand, the gains are much lower for queries with more than three overlapping tokens and range from -0.5 to +1.4 MRR@10. This supports our hypothesis that code-switching indeed regularizes monolingual overfitting.

5 Conclusion

We propose a simple and effective method to improve zero-shot rankers: training on artificially code-switched data. We empirically test our approach on 36 language pairs, spanning monolingual, cross-lingual, and multilingual setups. Our method outperforms zero-shot models trained only monolingually and provides a resource-lean alternative to MT for CLIR. In MLIR our approach can match MT performance while relying only on bilingual dictionaries. To the best of our knowledge, this work is the first to propose artificial code-switched training data for cross-lingual and multilingual IR.

Limitations

This paper does not utilize any major linguistic theories of code-switching, such as (Belazi et al., 1994; Myers-Scotton, 1997; Poplack, 2013). Our approach to generating code-switched texts replaces words with their synonyms in target languages, looked up in a bilingual lexicon. Furthermore, we do not make any special efforts to resolve word sense or part-of-speech ambiguity. To this end, the resulting sentences may appear implausible and incoherent.

Acknowledgements

We thank the members of the MaiNLP research group as well as the anonymous reviewers for their feedback on earlier drafts of this paper. This research is in parts supported by European Research Council (ERC) Consolidator Grant DIALECT 101043235.

[6] Due to out-of-vocabulary tokens the percentage of translated tokens is slightly lower: 23% for $p = 0.25$, 45% for $p = 0.5$, 68% for $p = 0.75$ and 92% for $p = 1.0$. In Wiki CS 90% of queries and documents contain at least one translated n-gram, leading to 20% of translated tokens overall.

[7] We analyzed a sample of 1M positive training instances and found a total of 4,409,974 overlapping tokens before and 3,039,750 overlapping tokens after code-switching (ML-CS, $p = 0.5$), a reduction rate of ~31%.

[8] We use the model's SentencePiece tokenizer (Kudo and Richardson, 2018) and ignore the special tokens <s>, </s>, <pad>, <unk> and <mask>.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

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*.

Hedi M Belazi, Edward J Rubin, and Almeida Jacqueline Toribio. 1994. Code Switching and X-bar Theory: The Functional Head Constraint. *Linguistic inquiry*, pages 221–237.

Rexhina Blloshmi, Tommaso Pasini, Niccolò Campolungo, Somnath Banerjee, Roberto Navigli, and Gabriella Pasi. 2021. IR like a SIR: Sense-enhanced Information Retrieval for Multiple Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1041, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Arash Einolghozati, Abhinav Arora, Lorena Sainz-Maza Lecanda, Anuj Kumar, and Sonal Gupta. 2021. El volumen louder por favor: Code-switching in task-oriented semantic parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1009–1021, Online. Association for Computational Linguistics.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer enhanced named entity recognition for code-mixed web queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Debasis Ganguly, Ayan Bandyopadhyay, Mandar Mitra, and Gareth J. F. Jones. 2016. Retrievability of code mixed microblogs. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 973–976. ACM.

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. CoMeT: Towards code-mixed translation using parallel monolingual sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 47–55, Online. Association for Computational Linguistics.

Junxian He, Zhisong Zhang, Taylor Berg-Kirkpatrick, and Graham Neubig. 2019. Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3223, Florence, Italy. Association for Computational Linguistics.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1085–1095, Valencia, Spain. Association for Computational Linguistics.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1048–1056, New York, NY, USA. Association for Computing Machinery.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 26–31, Marseille, France. European Language Resources Association.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An empirical study of pre-trained transformers for Arabic information extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4727–4734, Online. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Jaeseong Lee, Dohyeon Lee, Jongho Kim, and Seungwon Hwang. 2023. C2lir: Continual cross-lingual transfer for low-resource information retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, pages 466–474. Springer.

Bo Li and Ping Cheng. 2018. Learning neural representation for CLIR with adversarial framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1861–1870, Brussels, Belgium. Association for Computational Linguistics.

Wing Yan Li, Julie Weeds, and David Weir. 2022. MuSeCLIR: A multiple senses and cross-lingual information retrieval dataset. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1128–1135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Robert Litschko, Ivan Vulić, and Goran Glavaš. 2022a. Parameter-efficient neural reranking for cross-lingual and multilingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1071–1082, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022b. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2022. Streamlining evaluation with ir-measures. In *European Conference on Information Retrieval*, pages 305–310. Springer.

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

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Guilherme Moraes, Luiz Henrique Bonifácio, Leandro Rodrigues de Souza, Rodrigo Nogueira, and Roberto Lotufo. 2021. A cost-benefit analysis of cross-lingual transfer methods. *arXiv preprint arXiv:2105.06813*.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical Structure in Code-Switching*. Oxford University Press.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Shana Poplack. 2013. "sometimes i'll start a sentence in spanish y termino en español": Toward a typology of code-switching. *Linguistics*, 51(s1):11–14.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Peng Shi, He Bai, and Jimmy Lin. 2020. Cross-lingual training of neural models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773, Online. Association for Computational Linguistics.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual training of dense retrievers for document retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 251–253, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: Cross-lingual Data Augmentation for Natural Language Inference and Question Answering. *arXiv preprint arXiv:1905.11471*.

Samson Tan and Shafiq Joty. 2021. Code-mixing on sesame street: Dawn of the adversarial polyglots. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3596–3616, Online. Association for Computational Linguistics.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3154–3169, Online. Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Jieyu Wang and Anita Komlodi. 2018. Switching languages in online searching: A qualitative study of web users' code-switching search behaviors. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, page 201–210, New York, NY, USA. Association for Computing Machinery.

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 the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W. Oard. 2022. C3: Continued pretraining with contrastive weak supervision for cross language ad-hoc retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2507–2512, New York, NY, USA. Association for Computing Machinery.

Jian Yang, Shuming Ma, Dongdong Zhang, ShuangZhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9386–9393.

Puxuan Yu, Hongliang Fei, and Ping Li. 2021. Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference 2021*, WWW '21, page 1029–1039, New York, NY, USA. Association for Computing Machinery.

A Code-Switching Examples

Approach	Query	Document
`Zero-Shot`	What is an affinity credit card program?	Use your PayPal Plus credit card to deposit funds. If you have a PayPal Plus credit card, you are able to instantly transfer money from it to your account. This is a credit card offered by PayPal for which you must qualify.
`Fine-tuning`	Was ist ein Affinity-Kreditkartenprogramm?	Используйте свою кредитную карту PayPal Plus для внесения средств. Если у вас есть кредитная карта PayPal Plus, вы можете мгновенно переводить деньги с нее на свой счет. Это кредитная карта, предлагаемая PayPal, на которую вы должны претендовать.
`BL-CS`	Denn is einem affinity credit card programms?	Использовать your PayPal плюс кредита билет попытаться депозиты funds. если you have a PayPal плюс credit билет, скажите are able to instantly переход денег from it попытаться ваши account. This is a credit билет offered by paypal for причём you может qualify.
`ML-CS`	What is это affinità credit card program?	Use jouw PayPal Plus credit geheugenkaarten to depositi funds. إذا you хотя ein الائتمان aggiunta credit card, you are попытаться quindi sofort transfer geld from questo إلى deine account. Это является a кредита card offerto by paypal voor which you devono للتأهل
`Wiki-CS`	What is an affinity Kreditkarte program?	Use your PayPal Plus carta di credito to deposit funds. If you have a PayPal Plus carta di credito, you are able to instantly transfer denaro from it to your account. This is a carta di credito offered by PayPal for which you mosto qualify.

Table 5: Different Code-Switching strategies on a single training instance for the target language pair DE–RU (Query ID: 711253, Document ID: 867890, label: 0). **`Zero-shot:`** Train a single zero-shot ranker on the original EN–EN MS MARCO instances (Bajaj et al., 2016). **`Fine-tuning:`** Fine-tune ranker directly on DE–RU, we use translations (Google Translate) provided by the mMARCO dataset (Bonifacio et al., 2021). **Bilingual Code-Switching (`BL-CS`)**: Translate randomly selected EN query tokens into DE and randomly selected EN document tokens into RU, each token is translated with probability $p = 0.5$; **Multilingual Code-Switching (`ML-CS`)**: Same as `BL-CS` but additionally sample for each token its target language uniformly at random. **`Wiki-CS`**: Translate n-grams extracted with a sliding window. Tokens within a single query/document are code-switched with a single language; across training instances languages are randomly mixed. We use the following "seen languages": English, German, Russian, Italian, Dutch, Arabic.

B Results on Unseen Languages

	Unseen QL		Unseen DL			Unseen Both					
	FR–EN	ID–NL	EN–PT	DE–VT	IT–ZH	ES–FR	FR–PT	ID–VT	PT–ZH	AVG	Δ_{ZS}
Zero-shot	18.3	13.7	23.2	10.9	9.4	19.0	18.7	11.8	9.6	15.0	-
Fine-tuning	30.0*	27.2*	30.8*	24.8*	25.0*	29.0*	29.0*	25.8*	25.4*	27.4	+12.2
Multilingual CS	**21.4***	**18.3***	25.9*	**15.5***	14.8*	**22.7***	**21.9***	**16.4***	14.7*	**19.1**	**+4.1**
Wiki CS	21.0*	17.2*	**26.2***	15.4*	**15.0***	21.9*	20.5*	15.3*	**14.8***	18.6	+3.4

Table 6: CLIR results on unseen mMARCO languages in terms of MRR@10. **Bold**: Best zero-shot model for each language pair. Δ_{ZS}: Absolute difference to the zero-shot baseline. Results significantly different from the zero-shot baseline are marked with * (paired t-test, Bonferroni correction, $p < 0.05$). Results include unseen query languages (QL), unseen document languages (DL) and unseen languages on both sides.

	FR–FR	ID–ID	ES–ES	PT–PT	ZH–ZH	VT–VT	AVG	Δ_{ZS}
Zero-shot	**27.2**	**26.8**	**28.2**	**27.9**	**24.8**	22.8	26.3	-
Fine-tuning	30.5*	30.6*	31.5*	31.2*	29.1*	28.6*	30.3	+4.0
Multilingual CS	26.4	26.7	27.6	27.3	22.3	**23.1***	25.6	-0.7
Wiki CS	25.8*	25.5*	27.1*	26.5*	22.2*	21.8*	24.8	-1.8

Table 7: MoIR: Monolingual results on unseen mMARCO languages in terms of MRR@10.

C Hyperparameters, Datasets and Infrastructure

Hyperparameter	Value
Maximum sequence length	512
Learning rate	2e-5
Training steps	200,000
Batch size	64
Warm-up steps (linear)	5,000
Positive-to-negative ratio	1:4
Optimizer	AdamW (Loshchilov and Hutter, 2019)
Encoder Model	`nreimers/mMiniLMv2-L6-H384-distilled-from-XLMR-Large`
Encoder Parameters	106,993,920

Table 8: Hyperparameter values for re-ranking models. Following Reimers and Gurevych (2020) we extract negative samples from training triplets provided by MS MARCO (Bajaj et al., 2016). In the passage re-ranking task we re-rank for 6980 queries 1,000 passages respectively (qrels.dev.small). We construct 36 different language pairs from the mMARCO dataset (Bonifacio et al., 2021).

Setup	
GPU	NVIDIA A100 (80 GB)
Avg. Training Duration (per model)	13 h
Avg. Test (per language pair)	2 h

Table 9: Computational environment. We use `Huggingface` to train our models (Wolf et al., 2020), NLTK for tokenization, `ir-measures` for evaluating MRR@10 (MacAvaney et al., 2022) and `SciPy` for significance testing.

D Bilingual Lexicon Sizes

Language	MUSE vocabulary	Parallel Wikipedia titles
Arabic	132,480	432,359
German	200,000	1,113,422
Italian	200,000	999,243
Dutch	200,000	822,563
Russian	200,000	906,750

Table 10: Size of bilingual lexicons. Two lexicons are used to substitute the words in English with their respective cross-lingual synonyms: (i) multilingual word embeddings provided by MUSE (Lample et al., 2018), (ii) Wikipedia page titles obtained from inter-language links, provided by linguatools project.[9] The Wikipedia-based lexicons are several times larger that the MUSE vocabulary.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations (after Conclusion)

☑ A2. Did you discuss any potential risks of your work?
We train small distilled models (mentioned in Section 3 and Appendix C) to reduce environmental impact.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, Section 1 Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2

☑ B1. Did you cite the creators of artifacts you used?
Section 2, Section 3 and Appendix C.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We use Wikipedia as a resource for creating bilingual lexicons (Wiki-CS), which is openly licensed under "Wikipedia: Creative Commons Attribution-ShareAlike license."
MS-MARCO and mMARCO are standard benchmarks that have been released for non-commercial research purposes. (https://microsoft.github.io/msmarco/)
MUSE is distributed under "Attribution-NonCommercial 4.0 International" license. (https://github.com/facebookresea

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 1

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We make use of existing research data and do not release new corpora.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3, Appendix D

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix C

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3, Appendix C

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3, Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix C

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.