# The Impact of Incidental Multilingual Text on Cross-Lingual Transfer in Monolingual Retrieval

Andrew Liu*, Edward Xu*, Crystina Zhang, Jimmy Lin

David R. Cheriton School of Computer Science, University of Waterloo

**Abstract.** While great progress has been made in non-English monolingual passage retrieval in recent years, there has been little work exploring influential factors behind cross-lingual transfer capabilities in monolingual passage retrieval. In a retrieval corpus such as Wikipedia, incidental multilingual texts occur in forms including code-switching, translated name entities, and so on. In this work, we study how these naturally occurring multilingual texts impact the cross-lingual transfer of dense retrievers on monolingual passage retrieval. Results on 41 pairs of languages suggest that the cross-lingual transfer capacity of dense retrieval can be largely achieved with no incidental multilingual text, yet the decrease in effectiveness is correlated with the number of queries and documents containing incidental multilingual text. This suggests that cross-lingual transfer can be achieved by semantic understanding of the inputs alone and further enhanced by manually injecting more overlapping lexicons.[1]

## 1 Introduction

Text retrieval techniques have greatly evolved since the development of pretrained language models (pLM) [5, 16] on both English and multilingual benchmarks. Recent works show that current retrieval methods can be effectively adapted to multilingual scenarios simply by changing the backbone from the English pLM to a multilingual one [3, 10, 11, 21–23]. This is an example of the transfer effect, where a model trained on one task or domain shows capability in a different but related task or domain due to the model having learned representations generalizable across these tasks. Meanwhile, many works have studied the effect of multilingual pLMs on NLP tasks by investigating how transfer is related to cross-lingual traits such as shared subwords, shared linguistic features, and many other factors [14, 19, 8, 1, 4]. However, there are few works exploring influential factors behind the cross-lingual transfer effectiveness in multilingual text retrieval [22, 6], where *cross-lingual transfer* refers to the scenario where a retriever is applied to a target language $L_t$ when it is only trained on a different source language $L_s$, and *its effectiveness* is quantified by standard IR metrics (e.g., nDCG@10) on the test collection in the target language.

Incidental multilingual text (IMT), a natural occurrence of multilingual text in the same sentence or paragraph and possibly as a result of code-switching, cross-lingual reference, quoting, etc., is prevalent in retrieval corpora like Wikipedia [4]: See an

*Equal contribution

[1] https://github.com/Andrwyl/IMT-in-monolingual-IR

馕坑（维吾爾語：تونۇر 拉丁维文：tonur 西里尔维文：тонур）是用于烤制馕的功能等同于烤炉的盐土制火灶。整体如同一只倒扣在地上的碗，底大口小。还被用于烤制烤包子、馕坑烤肉、烤全羊等广受欢迎的新疆特色食物。

Translation:

*The naan pit is a salt-and-earth fire stove used for baking naan, which has the same function as an oven. The whole thing is like a bowl turned upside down on the ground, with a large bottom and a small mouth. It is also used to bake popular Xinjiang specialty foods such as baked buns, Nang pit barbecue, and roasted whole lamb.*

Fig. 1: An example of incidental multilingual text (IMT) in a passage in Chinese extracted from Wikipedia. The non-Chinese texts are highlighted in red and will be removed in the token removal procedure to be described in Section 2.

example in Figure 1. Previous work observes a large number of shared tokens amongst retrieval corpora in different languages due to incidental multilingual text [22], which we reproduced on MIRACL [23], a large-scale multilingual retrieval dataset (Table 1). Given the large number of shared tokens it produces, it is intuitive to assume that they serve as major anchors of cross-lingual transfer. In this work, we study the impact of incidental multilingual text on cross-lingual transfer in monolingual passage retrieval. Specifically, we remove all tokens that are in a language script different from the script of the training or evaluation data (Details in Section 2). We then compare the impact of this action on cross-lingual transfer effectiveness.

We find that the effectiveness of cross-lingual transfer is mostly preserved without incidental multilingual text: removing IMT tokens shows insignificant differences in the transfer results in 38 of 41 pairs of languages, where the decrease in effectiveness in most pairs of languages is less than 2% and overall less than 10%. This suggests that models are able to perform cross-lingual transfer via semantic understanding of the input sentences or passages. However, the decrease in effectiveness strongly correlates with the number of queries and positive passages that contained incidental multilingual text, implying that injecting more overlapping lexicon may add value to cross-lingual transfer on top of pure semantic understanding.

## 2 Incidental Multilingual Text (IMT)

*IMT Token Removal.* The same token removal procedure applies for both training data in the source language ($L_s$) and evaluation data in the target languages ($L_t$). Given training data in $L_s$, we remove words detected to not be in the same script as $L_s$. Specifically, we use an off-the-shelf language detection library `langdetect`.[2] We pre-process the training data for the language $L_s$ by first tokenizing the data using the mBERT tokenizer, then removing all tokens that `langdetect` finds to be in a script other than the script of $L_s$. For example, if $L_s$ is Russian, our process will remove all *non-Cyrillic* tokens from training passages. The remaining tokens, which are in the same script as $L_s$, form new training data used for fine-tuning mDPR. Similarly, given evaluation corpus and queries in $L_t$, we remove tokens that are detected to be in a different script from $L_t$. Punctuation and numbers are always kept in this process.

*How IMT Tokens Contribute to Overlapping Tokens between Corpora in Two Languages.* As mentioned in the Introduction, the corpora of two languages exhibit a large number

[2] `https://pypi.org/project/langdetect`

Table 1: Statistics of the removed IMT tokens and the overlapping tokens across languages, where **blue** highlights the highest values and **orange** highlights the lowest values. In all tables, $L_s$ indicates the source language and $L_t$ indicates the target language.

(a) Overlap percentage of unique tokens between languages $L_s$ and $L_t$ in MIRACL [23]. The value at entry {column $L_s$, row $L_t$} is computed as $\frac{\text{No. } \textit{unique} \text{ shared tokens between } L_s \text{ and } L_t}{\text{No. } \textit{unique} \text{ tokens in } L_s}$.

| | fa | hi | bn | ru | fr | es | fi | id | sw | te | ar | th | zh | ja |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **fa** | 100.0% | 50.2% | 52.9% | 33.2% | 35.7% | 31.7% | 34.8% | 32.7% | 43.4% | 48.5% | 48.5% | 42.6% | 39.1% | 37.4% |
| **hi** | 44.1% | 100.0% | 51.4% | 30.6% | 32.8% | 28.8% | 32.0% | 30.0% | 40.4% | 57.1% | 40.0% | 41.8% | 37.9% | 36.0% |
| **bn** | 41.6% | 46.1% | 100.0% | 26.1% | 28.6% | 25.3% | 27.8% | 26.7% | 34.9% | 45.1% | 34.9% | 35.3% | 31.5% | 30.8% |

(b) Identical to Table 1a yet considering token frequencies.

| | fa | hi | bn | ru | fr | es | fi | id | sw | te | ar | th | zh | ja |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **fa** | 100.0% | 24.9% | 5.3% | 36.8% | 66.5% | 62.4% | 62.4% | 58.1% | 71.9% | 14.1% | 84.9% | 3.8% | 23.9% | 25.9% |
| **hi** | 36.9% | 100.0% | 59.5% | 23.6% | 62.8% | 61.3% | 58.1% | 54.0% | 70.1% | 38.4% | 47.1% | 55.3% | 28.1% | 21.7% |
| **bn** | 47.0% | 60.7% | 100.0% | 24.2% | 65.1% | 62.5% | 56.9% | 52.3% | 67.9% | 38.9% | 63.0% | 24.2% | 32.6% | 23.4% |

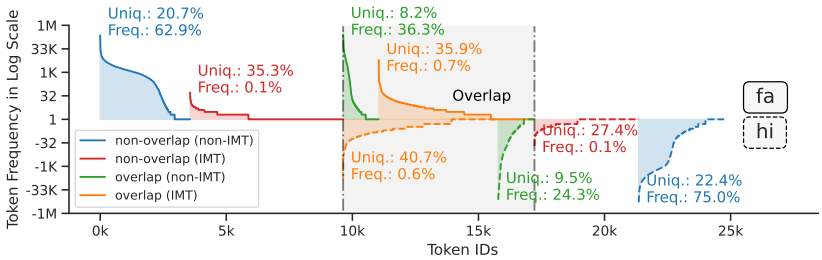


Fig. 2: An example of overlapping and IMT tokens among `fa` and `hi`. Each point at x-axis indicates a unique token. Y-axis show token frequencies in log scale, where positive bar indicates frequencies in `fa`, and negative bar indicates frequencies in `hi`. Colors indicates whether the tokens overlap with the other language and whether it is IMT.

of unique overlapping tokens (Table 1a) with overall high frequencies (Table 1b). We are now ready to have a deeper look on the role of IMT tokens in these statistics. This is illustrated in Fig 2 using the `fa`–`hi` language pair as an example, where the grey block in the middle represents all overlapping tokens between the two corpora. First, IMT-tokens primarily account for a large number of unique overlapping tokens, as shown by the projection of the **orange** area onto the x-axis. Second, although IMT tokens are not highly frequent in the source language, they are usually frequent terms in the target languages (the **green** area on the opposite side of the y-axis). These patterns illustrate the potential of IMT tokens as anchors of cross-lingual transfer.

*Sentences Coherence and Information Completeness.* As our experiments remove tokens from the sentence, it potentially creates incoherent text that potentially hurts retrieval results. We thus investigate whether the introduced incoherence would disrupt the general understanding of the passage and lead to undesired external variables in the experiments. We curate five levels of information loss and paragraph comprehensibility after the token removal as shown under the "Initial Order" in Fig. 3, and then ask GPT-4 to categorize

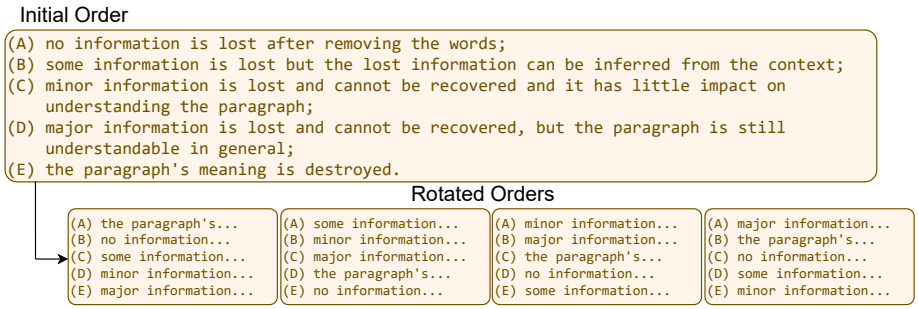


Fig. 3: The five levels of information loss and paragraph comprehensibility after the token removal, from mild to severe.

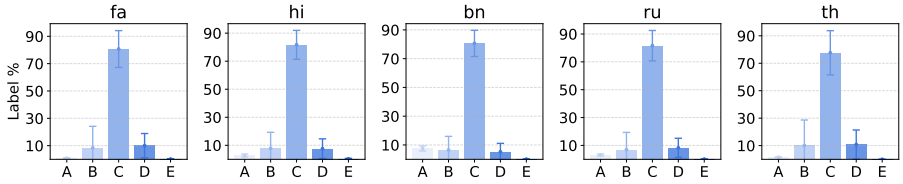


Fig. 4: Percentage of "information loss" categories labeled by GPT-4. The heights of the bars show the average percentage, and the error bars show the standard deviation. The distributions on the rest of the languages are similar.

a sampled set of documents into one of the five levels. To affirm its capability on this task, authors first annotate 20 documents per language, which have been translated into English via Google Translate, and then compared GPT-4 results with human-annotated results. Note that the human annotation process is finished before GPT-4, i.e., authors have not been exposed to the GPT-4 results before annotations. We find that on the sample set, GPT-4 results are largely aligned with human annotations.[3] We thus scaled up the examination using GPT-4 to 200 documents per language.

To mitigate the effect of positional bias of LLMs for multiple-choice questions [17], we generate 5 rounds of annotations, each with a different rotation of contents under (A–E) options as illustrated under "Rotated Orders" in Fig. 3. Fig. 4 shows aggregated results of all rotation variations, where the rotated orders have been translated back to the (A–E) options in the initial order. While certain variation is observed, the model recognized most of the token removal as C (*minor lost*), next as B (*recoverable*) and D (*major lost*), and almost no instances are categorized as E (*meaning destroyed*). We believe this shows that while token removal slightly breaks coherence, it does not cause fundamental detriments to the underlying meaning of documents.

[3]Among the examined examples, GPT-4 has identical assessments with humans in 65% of them, and shows stricter assessment (e.g., human assesses it as A but GPT-4 assesses it as B) in 95% of the examples, which indicates that results in Figure 2 approximate the upper-bound of the information loss during the token removal process.

## 3 Experimental Setup

*Model.* This work uses dense passage retriever (DPR) [9], one of the earliest dense retrieval models [9, 12, 20]. We chose the model for the simplicity of the architecture and the training process. DPR encodes the query and passage independently into vector representations $E_Q(q)$ and $E_P(p)$, and then measures their similarity by inner product. During training, encoders are optimized with NCE loss [15]. Following previous works [2, 21, 22], we adopt mDPR, which initializes DPR model with multilingual BERT (mBERT) to support retrieval in multiple non-English languages.

*Data.* All experiments on this work are based on MIRACL [23], a large-scale multilingual retrieval dataset that provides training and evaluation data for typologically diverse languages. Data is released under the Apache-2.0 License. For simplicity, we refer to each language using its ISO-2 code.[4]

*Training and Inference.* All our experiments are based on Tevatron [7], a flexible framework for the training and inference of common retrieval methods. During training, we start from the mBERT checkpoint provided by HuggingFace [18], then finetune it on two versions of the MIRACL training data in each language $L$: the official MIRACL training data[5] (baselines) and the one with non-$L$-script tokens removed as described in Section 2. Similarly, we adopt two versions of evaluation collections given language $L$: the official MIRACL collection[6] and the collection with non-$L$-script tokens removed. We evaluate the models using the official metric nDCG@10 of MIRACL. We use the same hyper-parameters as previous works [22]: All experiments are fine-tuned for 40 epochs with a batch size of 64 and a learning rate of 1e-5. We limit the query lengths to 32 tokens and the passage lengths to 256 tokens on both training and inference steps.

*Language Selection.* Models are evaluated on three languages: `fa`, `hi`, and `bn`, and trained extensively in 14 languages from MIRACL. These three evaluation languages form a language group that contains both similar and diverse linguistic features: On the one hand, they belong to the same language family, have a substantial amount of loan words from each other, and have the same word order (Subject–object–verb; SOV). On the other hand, they are all written in different scripts, which makes direct token sharing impossible. They also have different gender systems and morphological typologies (Persian is agglutinative, while the other two are fusional).

## 4 Results and Analysis

Table 2 compares zero-shot results before and after removing incidental multilingual text as mentioned in Section 2. We fine-tune models on fourteen source languages individually and evaluate on three target languages. Under each source–target language pair, we report

[4]Language names to their ISO-2 code: Persian (`fa`), Hindi (`hi`), Bengali (`bn`), Russian (`ru`), French (`fr`), Spanish (`es`), Finnish (`fi`), Indonesian (`id`), Swahili (`sw`), Telugu (`te`), Arabic (`ar`), Thai (`th`), Chinese (`zh`), Japanese (`ja`).

[5]`https://huggingface.co/miracl/miracl`

[6]`https://huggingface.co/miracl/miracl-corpus`

Table 2: nDCG@10 on the development set of Persian (fa), Hindi (hi), Bengali (bn) from MIRACL. **Column**: source languages $L_s$; **Rows**: target languages $L_t$. Under each target language, the first two rows are trained on official MIRACL training data (*w/ IMT*) and data with non-self script tokens removed (*w/o IMT*). The third row (%) is the relative effectiveness of *w/o IMT* compared to *w/ IMT*, highlighted in **blue**, where higher saturation indicates higher performance.[7] Significantly different results are marked with $^\dagger$ ($p < 0.01$ with paired t-tests).

| | w/ IMT? | fa | hi | bn | ru | fr | es | fi | id | sw | te | ar | th | zh | ja |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fa | ✓ | 0.474 | 0.318 | 0.349 | 0.390 | 0.283 | 0.380 | 0.383 | 0.337 | 0.303 | 0.371 | 0.433 | 0.410 | 0.312 | 0.393 |
| | ✗ | – | 0.318 | 0.322$^\dagger$ | 0.372 | 0.287 | 0.376 | 0.383 | 0.334 | 0.307 | 0.369 | – | 0.373$^\dagger$ | 0.310 | 0.375 |
| | % | – | 100.1% | 92.5% | 95.3% | 101.7% | 98.8% | 99.9% | 99.1% | 101.3% | 99.4% | – | 91.0% | 99.3% | 95.5% |
| hi | ✓ | 0.343 | 0.385 | 0.343 | 0.335 | 0.218 | 0.297 | 0.308 | 0.299 | 0.273 | 0.327 | 0.363 | 0.310 | 0.244 | 0.339 |
| | ✗ | 0.331 | – | 0.340 | 0.315 | 0.221 | 0.293 | 0.309 | 0.307 | 0.253 | 0.308 | 0.342 | 0.296 | 0.252 | 0.313 |
| | % | 96.5% | – | 99.3% | 94.1% | 101.6% | 98.8% | 100.6% | 102.7% | 92.8% | 94.4% | 94.3% | 95.4% | 103.3% | 92.1% |
| bn | ✓ | 0.441 | 0.359 | 0.597 | 0.502 | 0.263 | 0.369 | 0.453 | 0.432 | 0.362 | 0.497 | 0.498 | 0.523 | 0.337 | 0.482 |
| | ✗ | 0.396 | 0.380 | – | 0.458$^\dagger$ | 0.274 | 0.365 | 0.466 | 0.430 | 0.364 | 0.521 | 0.495 | 0.493 | 0.328 | 0.488 |
| | % | 89.8% | 105.8% | – | 91.2% | 104.0% | 99.0% | 102.8% | 99.6% | 100.5% | 104.8% | 99.3% | 94.2% | 97.4% | 101.3% |

three scores: (1) *w/ IMT*: using the official training and evaluation data, which contains IMT text; (2) *w/o IMT*: tokens in a non-self script are removed from training and evaluation data as mentioned in Section 2; (3) *Percentage*, the relative percentage of *w/o IMT* scores compared to *w/ IMT*. Entries under *Percentage* are highlighted in **blue**, where higher saturation indicates higher effectiveness and the white background indicating relative effectiveness at 90%.

*Cross-lingual Transfer Capacity is Mostly Preserved without Incidental Multilingual Text.* As most visibly shown by the percentage row, all source–target language pairs achieve over 90% relative effectiveness after shared tokens are removed, where most of them have minimal effectiveness loss (< 2%) or even outperform the "w/ IMT" baseline. Among the 41 pairs, only 3 pairs show a significant drop in transfer effectiveness after token removal. This indicates that the cross-lingual transfer capacity of dense retrievers can be effectively achieved via the semantic understanding of the input sentences alone, with no overlapping tokens between the training and evaluation data.

*Correlation with Affected Queries and Positive Passages.* While the overall impact on cross-lingual transfer is less than 10%, the decrease in effectiveness still varies across different language pairs. We find this could be partially explained by the number of *affected* queries and positive passages in training data,[8] where *affected* queries or passages are the ones containing incidental multilingual text to be removed. Fig. 5 shows how the relative effectiveness (y-axis) varies with the percentage of affected queries (x-axis; Fig. 5 left) and positive passages (x-axis; Fig. 5 right). Both show a strong negative correlation with $r < -0.5$, indicating that a higher percentage of affected

[7] "w/o IMT" score is not computed for pair fa–ar since they are in the same script.

[8] We also examine (1) the affected queries and positive passages in the evaluation data, and (2) the total number of removed tokens (Table 1), yet no conclusive observation has been shown regarding these two traits.

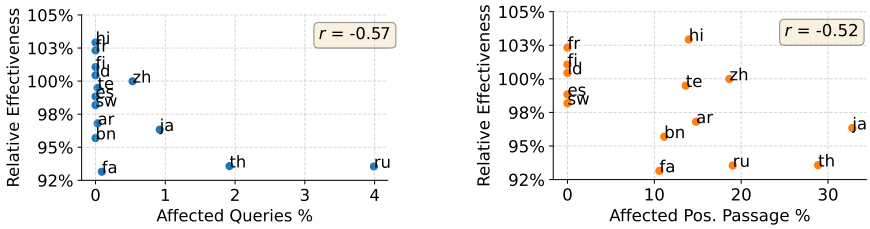


Fig. 5: Averaged relative effectiveness on `fa`, `hi`, and `bn` regarding the percentage of affected queries (left) or affected positive passages (right) in the training set, where "affected" means some tokens in the queries or passages are in non-self-language and thus have been removed. $r$: Pearson Correlation Coefficient.

queries or positive passages correlate with a larger decrease in cross-lingual transfer results. This suggests that while cross-lingual transfer is possible without any overlapping tokens between the training and evaluation data, manually injecting more overlapping vocabulary can possibly enhance the cross-lingual transfer.

## 5 Related Work

Many papers have studied the impact of shared tokens with inconsistent results. It has been reported that more shared tokens lead to better cross-lingual transfer results on the NER task [14] and other NLP tasks, including POS and dependency parsing [19]. On the other hand, it has also been found that shared tokens do not play a significant role in cross-lingual transfer in entailment and NER [8] and that multilingual language models could be pretrained without shared vocabulary [1]. All of the above works are conducted on multilingual pLM pretraining or NLP tasks. On retrieval, previous works [13] have found that code-switching data could assist in cross-lingual retrieval, yet it remains unclear how shared tokens might influence language transfer in monolingual retrieval.

The closest work is Do et al. [6], which proposes to enhance cross-lingual transfer in monolingual retrieval by leveraging *manual* code-mixing, while we focus on examining the impact of *naturally occurring* multilingual text in retrieval corpora.

## 6 Conclusion

In this work, we examined the impact of incidental multilingual text on the cross-lingual transfer effectiveness in monolingual passage retrieval when using dense models (mDPR). Extensive results on 41 languages pairs show that, while the incidental multilingual text results in a large number of shared tokens between the training and evaluation corpus, they don't significantly contribute to the cross-lingual transfer effect and that the transfer is possible without any overlapping tokens in the training and evaluation data. However, the observed effectiveness drop is strongly correlated with the number of affected queries and positive documents in the training data. This hints that while transfer is possible through pure semantic understanding, manually injecting more overlapping vocabulary could possibly enhance the cross-lingual transfer. We hope this work can shed light on how cross-lingual transfer occurs naturally and how it can be further improved.

## References


1. Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4623–4637. Association for Computational Linguistics, Online (Jul 2020)
2. Asai, A., Yu, X., Kasai, J., Hajishirzi, H.: One question answering model for many languages with cross-lingual dense passage retrieval. In: Advances in Neural Information Processing Systems. vol. 34, pp. 7547–7560 (2021)
3. Bonifacio, L.H., Jeronymo, V., Abonizio, H.Q., Campiotti, I., Fadaee, M., , Lotufo, R., Nogueira, R.: mMARCO: A multilingual version of ms marco passage ranking dataset. ArXiv **abs/2108.13897** (2021)
4. Briakou, E., Cherry, C., Foster, G.: Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9432–9452. Association for Computational Linguistics, Toronto, Canada (Jul 2023)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Minneapolis, Minnesota (Jun 2019)
6. Do, J., Lee, J., Hwang, S.w.: ContrastiveMix: Overcoming code-mixing dilemma in cross-lingual transfer for information retrieval. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). pp. 197–204. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024)
7. Gao, L., Ma, X., Lin, J., Callan, J.: Tevatron: An efficient and flexible toolkit for neural retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3120–3124. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023)
8. K, K., Wang, Z., Mayhew, S., Roth, D.: Cross-lingual ability of multilingual bert: An empirical study. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=HJeT3yrtDr
9. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (Nov 2020)
10. Lawrie, D.J., Mayfield, J., Oard, D.W., Yang, E.: HC4: A new suite of test collections for ad hoc clir. ArXiv **abs/2201.09992** (2022), https://arxiv.org/pdf/2201.09992.pdf
11. Lawrie, D.J., Yang, E., Oard, D.W., Mayfield, J.: Neural approaches to multilingual information retrieval. In: European Conference on Information Retrieval (2022), https://arxiv.org/abs/2209.01335
12. Lin, S.C., Yang, J.H., Lin, J.: In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In: Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). pp. 163–173. Online (Aug 2021)
13. Litschko, R., Artemova, E., Plank, B.: Boosting zero-shot cross-lingual retrieval by training on artificially code-switched data. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 3096–3108. Association for Computational Linguistics, Toronto, Canada (Jul 2023)

14. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual BERT? In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4996–5001. Association for Computational Linguistics, Florence, Italy (Jul 2019)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:231591445
16. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1) (jan 2020), https://jmlr.org/papers/volume21/20-074/20-074.pdf
17. Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., Plank, B.: "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics ACL 2024. pp. 7407–7416. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting (Aug 2024)
18. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Online (Oct 2020)
19. Wu, S., Dredze, M.: Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 833–844. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
20. Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Bennett, P.N., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021) (2021)
21. Zhang, X., Ma, X., Shi, P., Lin, J.: Mr. TyDi: A multi-lingual benchmark for dense retrieval. In: Proceedings of the 1st Workshop on Multilingual Representation Learning. pp. 127–137. Punta Cana, Dominican Republic (Nov 2021)
22. Zhang, X., Ogueji, K., Ma, X., Lin, J.: Toward best practices for training multilingual dense retrieval models. ACM Trans. Inf. Syst. **42**(2) (sep 2023)
23. Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., Lin, J.: MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. Transactions of the Association for Computational Linguistics **11**, 1114–1131 (09 2023)