# Literature Review: Monolingual Information Retrieval in Many Languages

## General problem

The papers reviewed here are trying to solve is improving state-of-the art monolingual text retrieval across diverse languages. Recent advances in neural information retrieval such as Dense Passage Retrieval (Karpukhin et al., 2020) and ColBERT (Khattab and Zaharia, 2020) have been driven by large annotated English datasets such as Natural Questions (Kwiatkowski et al., 2019) and MS MARCO (Bajaj et al., 2018), and lead to advances in applications such as Open QA. The central question is how do we extend these techniques to other languages where such large public datasets don't exist? The papers here take different approaches to this problem, in particular evaluated on the Mr TyDi dataset.

## Article Summaries

### Zhang et al., 2021. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval

This paper introduces the Mr. TyDi dataset and sets main metrics and simple baselines, showing that even a poorly performing dense retriever can improve BM25 performance in a hybrid system.

The Mr. TyDi dataset is for benchmarking monolingual text retrieval in 11 typologically diverse languages. It builds on the TyDi QA (Clark et al., 2020) dataset, expanding it into a dataset for retrieval. Mr. TyDi was constructed by prompting annotators with the start of a Wikipedia article in a given lanugage, and getting them to write a question they were interested in elicited by the prompt. Then each question was searched on Wikipedia using Google, and the top-ranked article is annotated with which passage contains the best answer, or that the answer isn't contained in the article. In Mr. TyDi they start with the same raw Wikipedia dumps as Mr. TyDi, and collect all passages in each language and merge them with the passages in TyDi QA. Any questions with no answer are discarded, and the relevant passages in Mr. TyDi are the relevant passages in TyDi QA.

They set simple baselines using BM25, mDPR, and a hybrid between the two. The metrics they use are Reciprocal Rank@100, to measure the effectiveness of the ranking, and Recall@100 to put an upper bound on the effectiveness of a system that uses this as a retriever. For BM25 they considered both default parameters, and tuning the parameters on the development set, which had very similar results across all languages except Telugu, where it made a large improvement. Their mDPR model, Dense Passage Retrieval initialised on mBERT, was trained on the English Natural Questions dataset, and then evaluated on each language. The mDPR model performed worse than BM25 on all languages except for English. Their hybrid model consists of taking 1000 results from each of BM25 and mDPR, normalising the scores of each into the range [0,1], and then adding a weighted normalised mDPR score to the normalised BM25 score, where the weight is in [0,1] and maximised on MRR@100 on the dev set. They find the hybrid model is statistically significantly better (paired t-test, $p < 0.01$) on almost all of the languages for both MRR@100 and Recall@100. This is particularly interesting that even when the mDPR model is poor by itself in a hybrid it can increase the performance of the overall system both by increasing recall, or by improving the ranking.

### Wu et al., 2022. Unsupervised Context Aware Sentence Representation Pretraining for Multi-lingual Dense Retrieval

This paper presents a method to create aligned multi-lingual sentence embeddings using only monolingual corpora. The resulting embeddings are competitive on multi-lingual and cross-lingual retrieval tasks with methods using aligned bilingual data, such as LABSE (Feng et al., 2022).

Their approach is to train a model to contrastively predict a sentence nearby to a target sentence against random sentence from the same language, by a method called Contrastive Context Prediction (CCP). They start with XLM-R (Conneau et al., 2020) and further train it on the task to classify a sentence that occurred within a window of w words, from a random sentence in the same language. In particular they add a projection of two linear layers separated by a batch norm layer, followed by l2-normalisation and classify using softmax with temperature $\tau$. To get random examples they keep a memory bank of embeddings from previous batches. To prevent batch norm leaking information across batches they use an asymmetric batch norm, where the context and target sentences are alternately embedded with batch norm or running mean and variance across batches (and they show the necessity of this with an ablation study). This is trained on data extracted processed with CCNet (Wenzek, 2020) for 108 languages. The projection layers are then discarded and the CLS token of the final layer of the Roberta model is used as the embedding.

To further align the embeddings of different languages there is an unsupervised affine calibration transformation between the embeddings of different languages. First for each language the embeddings are translated to have zero mean, and then scaled to have unit variance. Then an orthogonal rotation matrix is learned to align them using the adversarial method of Conneau et al., 2017.

They perform experiments on the model on cross-lingual sentence retrieval on Tatoeba (Artetxe and Schwenk, 2019) and zero-shot retrieval on XOR TyDi and Mr. TyDi. On Tatoeba they show that CCP with Calibration outperforms other monolingual models such as XLM-R and CRISS (Tran et al., 2020) and performs slightly worse than LABSE, but for pairs of non-English languages the performance is close to LABSE. For XOR TyDi and Mr. TyDi they look at the performance in the zero-shot setting of training on the English Natural

Questions dataset using Dense Passage Retrieval. CCP outperforms LABSE, XLM-R, and InfoXLM (Chi, et al. 2021) on Mr. TyDi, and is comparable to LABSE on XOR TyDi, outperforming mBERT, XLM-R, and InfoXLM.

### Izacard et al., 2022. Unsupervised Dense Information Retrieval with Contrastive Learning.

This paper shows that contrastive pre-training on which two randomly cropped sentences came from the same document significantly improves retrieval, including in a multilingual setting. The pre-training task involved for each passage of length up to 256 tokens picking two random spans containing 10-50% of the tokens, randomly deleting tokens with 10% probability, and training with contrastive InfoNCE loss against random negatives from previous batches. They show through ablation studies evaluated on BEIR that random spans is better than an Inverse Cloze Task (of detecting the middle of a span from its outside), that deletions improve downstream performance and more than random replacements, and that the performance generally increases with the batch size. To enable very large number of negatives (32768) they use MoCo (He, et al. 2020) instead of using in-batch negatives.

This approach leads to a large improvement on recall in Information Retrieval over BM25 and many strong neural systems. The Contriever model applies this contrastive pre-training on BERT base using 1 billion tokens with half the batches from CCNet, and half from Wikipedia. This model outperforms BM25 in Recall@100 on most BEIR datasets; although it is particularly weak on Touche-2020 which contains long documents, and Trec-COVID which contains COVID terms not in pre-training. When fine-tuned on MS MARCO it has higher recall@100 on most BEIR datasets than both BM25 and strong systems such as ColBERT and Splade v2 (Formal et al., 2021).

When trained independently on multilingual data this approach generates strong results in multilingual and cross-lingual retrieval. The mContreiver model applied the contrastive pre-training on mBERT on 29 languages (all the languages in Mr. TyDi and MKQA (Longpre et al., 2021) except zh-hk) uniformly sampling over languages. When fine-tuned on MS MARCO this has higher Recall@100 on both Mr. TyDi and MKQA than BM25 and mBERT or XLM-R fine-tuned on MS MARCO. They observed that pre-training only on the Mr. TyDi languages gave a performance improvement on Mr. TyDi.

### Zhang et al., 2022. Towards Best Practices for Training Multilingual Dense Retrieval Models.

This paper contributes recommendations for training multilingual dense retrieval models based on an extensive set of experiments with multilingual Dense Passage Retriever on Mr. TyDi. They find that in general fine-tuning on English MS MARCO before doing any other fine-tuning generally improves, or at least does not significantly degrade, results in all languages. MS MARCO is more significantly effective than the smaller Natural Questions dataset, despite Natural Questions being closer to the domain of Mr TyDi. For multilingual models fine tuning on either the specific lanugage or all available languages gives best results, but training on any other language from the dataset generally improves the result even if the script is different (with the exception of training on English after training on MS MARCO degraded performance in languages other than English). Monolingual models in the target language (when available; for Arabic, Finnish, Indonesian, and Korean) could sometimes outperform the multilingual models, but not by much. They also showed that even when a model is not pretrained in that lanugage fine tuning can lead to a limited increase in performance, with experiments in English DPR and AfriBERTa DPR. Finally hybrid solutions with BM25 consistently outperform the best dense retrieval system alone.

### Bonifacio, et al. 2022. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset.

This paper introduces the mMARCO dataset, a machine translated version of the MS MARCO dataset, and shows it can improve multilingual and cross-lingual retrieval. They translate the English MS MARCO dataset into 13 languages; Spanish, Portuguese, French, Italian, German, Dutch, Russian, Chinese, Japanese, Vietnamese, Arabic, Hindi, and Indonesian. They do this using both OPUS-MT (Tiedemann and Thottingal, 2020), and Google Translate, but find that OPUS-MT translations have lower BLEU on Tatoeba, which correlates with worse retrieval with BM25, and so they focus on Google Translate. They find mColBERT retriever, and mT5 and mMiniLM cross-encoder re-rankers outperform BM25 in MRR@10 when trained in a subset of the languages, and evaluated on the others. On mMARCO they show mT5 and then Portuguese and English fine-tuning gives better MRR@10 than a Portuguese T5 fine-tuned on any combination of English and Portuguese, but adding the other languages to fine-tuning leads to a slight reduction in performance. On Mr TyDi they show fine-tuning a mT5 cross-encoder re-ranker on mMARCO leads to slightly higher MRR@10 than training on English MS MARCO (0.551 vs 0.532).

## Compare and contrast

These papers all address how to train a multilingual text retriever given only English labelled data for a similar task, either Natural Questions or MS MARCO. Both Zhang et al., 2021 and Zhang et al., 2022 show that a hybrid system with mDPR and BM25 is often significantly stronger than either system (even where mDPR itself is weak). Izacard et al., 2022 and Wu et al., 2022 both show that contrastive pre-training on nearby sentences or passages in all the languages significantly improves performance in all languages after fine-tuning on an English labelled dataset. Bonifacio, et al. 2022 takes the unique approach of using machine translation on MS MARCO, and show that it improves the performance of a mT5 re-ranker.

Even though the papers took different approaches, there were many similar observations about multilingual data and models. Both Wu et al., 2022, and Zhang et al., 2022 all noted that contrastive training is most effective when negatives are sampled from the same language, otherwise it degenerates into a lanugage detection task. Izacard et al., 2022 and Bonifacio, et al. 2022 both noted that pre-training in additional languages could result in worse results, however Zhang et al., 2022 observed that the best results in each language came from fine-tuning across all languages.

The table below summarises the Recall@100 for the retrieval models from the different papers. This metric is a chosen as a strong indicator of the retrieval quality; MRR@100 could be improved by adding a cross-encoder to re-rank the results. Note the difference in Recall@100 in mDPR on Natural Questions between Zhang et al., 2021 and Zhang et al., 2022; this difference could be due to a difference in model training and hyperparameter selection, which shows we need to be careful when comparing results across papers.

| Paper | Model | Train | Ar | Bn | En | Fi | Id | Ja | Ko | Ru | Sw | Te | Th | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Izacard et al., 2022 | mContriever | MS MARCO | 88.7 | 91.4 | 77.2 | 88.1 | 89.8 | 81.7 | 78.2 | 83.8 | 91.4 | 96.6 | 90.5 | 87.0 |
| Bonifacio, et al. 2022 | mColBERT | mMARCO | 85.9 | 91.8 | 78.6 | 82.6 | 91.1 | 70.9 | 72.9 | 86.1 | 80.8 | 96.9 | - | - |
| Wu et al., 2022 | CCP | Natural Questions | 82.0 | 88.3 | 80.1 | 78.7 | 87.5 | 80.0 | 73.2 | 77.2 | 75.1 | 88.8 | 88.9 | 81.8 |
| Zhang et al., 2021 | BM25 + mDPR | Natural Questions | 86.3 | 93.7 | 69.6 | 78.8 | 88.7 | 77.8 | 70.6 | 76.0 | 78.6 | 82.7 | 87.5 | 80.9 |
| Wu et al., 2022 | InfoXLM | Natural Questions | 80.6 | 86.0 | 80.4 | 74.9 | 86.9 | 78.8 | 71.7 | 76.7 | 72.4 | 86.7 | 87.4 | 80.2 |
| Wu et al., 2022 | XLM-R | Natural Questions | 81.3 | 84.2 | 77.6 | 78.2 | 88.6 | 78.5 | 72.7 | 77.4 | 63.3 | 87.5 | 88.2 | 79.8 |
| Izacard et al., 2022 | XLM-R | MS MARCO | 79.9 | 84.2 | 73.1 | 81.6 | 87.4 | 70.9 | 71.1 | 74.1 | 73.9 | 91.2 | 89.5 | 79.7 |
| Izacard et al., 2022 | mBERT | MS MARCO | 81.1 | 88.7 | 77.8 | 74.2 | 81.0 | 76.1 | 66.7 | 77.6 | 74.1 | 89.5 | 57.8 | 76.8 |
| Wu et al., 2022 | LABSE | Natural Questions | 76.2 | 91.0 | 78.3 | 76.0 | 85.2 | 66.9 | 64.4 | 74.4 | 75.0 | 88.9 | 83.4 | 78.2 |
| Izacard et al., 2022 | mContriever | - | 82.0 | 89.6 | 48.8 | 79.6 | 81.4 | 72.8 | 66.2 | 68.5 | 88.7 | 80.8 | 90.3 | 77.2 |
| Zhang et al., 2021 | BM25 (tuned) | - | 80.0 | 87.4 | 55.1 | 72.5 | 84.6 | 65.6 | 79.7 | 66.0 | 76.4 | 81.3 | 85.3 | 75.8 |

| Paper | Model | Train | Ar | Bn | En | Fi | Id | Ja | Ko | Ru | Sw | Te | Th | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zhang et al., 2022 | mDPR | MS MARCO | 79.9 | 82.0 | 75.8 | 69.3 | 75.8 | 73.8 | 64.5 | 72.8 | 68.6 | 79.7 | 64.8 | 73.4 |
| Zhang et al., 2021 | BM25 (default) | - | 79.3 | 86.9 | 53.7 | 71.9 | 84.3 | 64.5 | 61.9 | 64.8 | 76.4 | 75.8 | 85.3 | 73.2 |
| Wu et al., 2022 | mBERT | Natural Questions | 69.5 | 71.2 | 74.9 | 64.5 | 73.9 | 66.2 | 56.5 | 67.4 | 53.7 | 43.3 | 52.9 | 63.1 |
| Zhang et al., 2022 | mDPR | Natural Questions | 65.0 | 77.9 | 67.8 | 56.8 | 68.5 | 58.4 | 53.3 | 64.7 | 52.8 | 36.6 | 51.5 | 59.4 |
| Zhang et al., 2021 | mDPR | Natural Questions | 62.0 | 67.1 | 47.5 | 37.5 | 46.6 | 53.5 | 49.0 | 49.8 | 26.4 | 35.2 | 45.5 | 47.3 |

Some of the papers also investigate the related tasks of cross-lingual retrieval and fine-tuning on Mr TyDi data. Both Wu et al., 2022 and Izacard et al., 2022 show that contrastive sentence level pre-training improves the related task of cross-lingual retrieval; asking a question in one language and returning a result from another language. Izacard et al., 2022 and Zhang et al., 2022 show that training on Mr. TyDi data leads to very large improvements, and both note that pre-training with MS MARCO helps boost the performance. The Recall@100 from these two papers is listed below.

| Paper | Model | Train | Ar | Bn | En | Fi | Id | Ja | Ko | Ru | Sw | Te | Th | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Izacard et al., 2022 | mContriever | MS MARCO; Mr. TyDi | 94.0 | 98.6 | 92.2 | 92.7 | 94.5 | 88.8 | 88.9 | 92.4 | 93.7 | 98.9 | 95.2 | 93.6 |
| Zhang et al., 2022 | BM25 + mDPR | MS MARCO; Mr. TyDi | 93.2 | 94.6 | 85.7 | 90.9 | 94.8 | 88.3 | 85.3 | 89.8 | 90.3 | 98.2 | 94.6 | 91.6 |
| Zhang et al., 2022 | mDPR | MS MARCO; Mr. TyDi | 90.0 | 95.5 | 84.1 | 85.6 | 86.0 | 81.3 | 78.5 | 84.3 | 87.6 | 96.6 | 88.3 | 87.1 |

# Future work

These papers give a foundation for approaches to monolingual retrieval across many languages, but there are open questions on analysing the impact of these techniques, on whether other techniques from adjacent domains are effective, and on extending this to open question answering. Each paper introduces a set of modelling choices that leads to some improvement, but when comparing papers its hard to understand why one performs better than another. This gives an opportunity to do detailed ablations to understand what techniques actually contribute to improvement. The ideas in these papers of hybrid sparse-dense systems, contrastive pretraining, and translation only scratch the surface of ideas from English information retrieval and from other multilingual tasks, which could potentially lead to fruitful results in this context. Another way to extend this is to move the focus from information retrieval to open question answering; these systems are getting very high recall, what is the best way to integrate them into an open question answering system across multiple languages?

These papers show different techniques for zero-shot retrieval in many languages, but it's hard to tell which of the changes make the most impact. The table showing Recall@100 in a zero-shot setting shows many gaps because different authors made different modelling choices it's hard to understand the impact of a single change. For example mMARCO was only used to train mColBERT; how much gain does it get over mColBERT trained with English MS MARCO, or how much improvement would mDPR or mContriever get from training on mMARCO? Hybrid models with BM25 drastically improve mDPR which indicates they are complementary; is mContriever also highly complementary to BM25 or are the improvements over mDPR because it returns more documents that BM25 would? Even when there is a direct comparison details of the training can have a large impact which can lead to uncertainty in comparing across papers; Zhang et al., 2021 and Zhang et al., 2022 have quite different results for mDPR trained on Natural Questions. Does mContriever perform better than CCP because the contrastive modelling approach is better, or is it because of the base model (mBERT vs XLM-R), or the contrastive pre-training corpus, or training on MS MARCO rather than Natural Questions? The performance of XLM-R on Natural Questions in Wu et al., 2022 is very similar to that of XLM-R on MS MARCO in Izacard et al., 2022; is the difference in dataset really smaller for this model than mBERT or mDPR, or is it due to different hyperparameter selection?

The approaches in these papers are all effective at improving multilingual retrieval, but there are other ideas from English retrieval or other multilingual settings that could also be effective. For example Lassance, 2023 proposes an approach to multilingual SPLADE for fine tuning on the data, but is it effective in a zero-shot setting with mMARCO? Another direction is training on automatically generated queries (Oguz et al., 2022), can this be extended to a multilingual setting? Wu et al., 2022 mentioned aligning embeddings but didn't explore its impact on transferring English retrieval to other languages; this approach has been effective in cross-lingual retrieval (Huang et al., 2023) and for multi-lingual semantic textual similarity (Reimers and Gurevych, 2020), can it improve monolingual retrieval in diverse languages?

# References

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Transactions of the Association for Computational Linguistics, 7:597–610.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs].

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset. arXiv:2108.13897 [cs].

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. Transactions of the Association for Computational Linguistics, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. arXiv:2109.10086 [cs].

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In pages 9729–9738.

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, pages 1048–1056, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118 [cs].

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 39–48, New York, NY, USA. Association for Computing Machinery.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. Transactions of the Association of Computational Linguistics.

Carlos Lassance. 2023. Extending English IR methods to multi-lingual IR. arXiv:2302.14723 [cs].

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. Transactions of the Association for Computational Linguistics, 9:1389–1406.

Barlas Oguz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Scott Yih, Sonal Gupta, and Yashar Mehdad. 2022. Domain-matched Pre-training Tasks for Dense Retrieval. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1524–1534, Seattle, United States. Association for Computational Linguistics.

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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual Retrieval for Iterative Self-Supervised Training. In Advances in Neural Information Processing Systems, volume 33, pages 2207–2219. Curran Associates, Inc.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 4003–4012, Marseille, France. European Language Resources Association.

Ning Wu, Yaobo Liang, Houxing Ren, Linjun Shou, Nan Duan, Ming Gong, and Daxin Jiang. 2022. Unsupervised Context Aware Sentence Representation Pretraining for Multi-lingual Dense Retrieval. arXiv:2206.03281 [cs].

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. In Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022. Towards Best Practices for Training Multilingual Dense Retrieval Models. arXiv:2204.02363 [cs].