Data Efficient Dense Cross-Lingual Information Retrieval

Luc Chen
lucchen@mit.edu

Yinan He
yinanhe@mit.edu

Alayna Nguyen
alaynang@mit.edu

Abstract

Cross-Lingual Information Retrieval (CIR) remains challenging due to limited annotated data and linguistic diversity, especially for low-resource languages. While dense retrieval models have significantly advanced retrieval performance, their reliance on large-scale training datasets hampers their effectiveness in multilingual settings. In this work, we propose two complementary strategies to improve data efficiency and robustness in CIR model fine-tuning. First, we introduce a paraphrase-based query augmentation pipeline leveraging large language models (LLMs) to enrich scarce training data, thereby promoting more robust and language-agnostic representations. Second, we present a weighted InfoNCE loss that emphasizes underrepresented languages, ensuring balanced optimization across heterogeneous linguistic inputs. Experiments on cross-lingual benchmark datasets demonstrate that our combined approaches yield substantial gains in retrieval quality, outperforming standard training protocols on small and imbalanced datasets. These results underscore the potential of targeted data augmentation and re-weighted objectives to build more inclusive and effective CIR systems, even under resource constraints.

1 Introduction

Cross-lingual information retrieval has garnered attention in recent years due to the significance of learning unified representation space across different languages. It aims to retrieve relevant documents from datasets using query from a different language. A successful CIR system allows users to access information across various languages, which is crucial in making content accessible to a broader audience, regardless of their native language. Information retrieval (IR) is also a pivotal part of Retrieval Augmented Generation, where suitable information is retrieved from an external knowledge base and provided to Large Language Models (LLMs) during text generation. CIR can expand the coverage of the knowledge base and handle cross-lingual queries.

Different languages vary significantly in terms of grammar, vocabulary, and cultural context, making it challenging to create shared semantic spaces that work across all languages. A major challenge of CIR is the high annotation cost for manually labeling paired data from a wide range of languages. This problem is exacerbated in low-resource languages, where it is more difficult to obtain unlabeled data and find skilled annotators.

An intuitive idea to tackle this problem is unsupervised learning, where no labes are needed. Inverse Cloze Task (Lee et al., 2019) obtains pseudo-queries from the context, and the context naturally serves as the target document for retrieval. Semi-supervised learning (Tarvainen and Valpola, 2017) leverages a small amounts of labeled data combined with a larger amount of unlabeled data. Self-supervised learning methods (Devlin, 2018) train models using pretext tasks derived from unlabeled data. While these methods achieve outstanding results without requiring large amounts of labeled, they depend heavily on substantial unlabeled data. This is not feasible in low-resource languages where data annotation and data collection are both expensive.

In this paper, we present two approaches to improve CIR model fine-tuning with limited data. We propose back-translation paraphrasing as strong query augmentation during CIR model fine-tuning with contrastive loss and a weighted InfoNCE loss for CIR with multiple query languages and low-resource languages.

2 Related Works

Traditional IR methods represent documents and queries as sparse features, which store the fre-

quency information of vocabularies present in the text. Many variations have been proposed to improve the feature quality. Notable ones include TF-IDF (Sparck Jones, 1972) and BM25 (Robertson et al., 2009). The high dimensional sparse features used in these methods are computationally expensive to handle and lacks semantic understanding, which causes over-reliance on near-exact match for document retrieval. LSA (Deerwester et al., 1990) was proposed to address these problems by using singular value decomposition for dimensionality reduction.

The developments in neural networks in natural language processing field prompted the use of neural network techniques in IR. Transformer-based architectures (Vaswani, 2017) has dominated the field since its inception. (Nogueira and Cho, 2019) uses concatenated query-document sequence (separated by [SEP] token) as the input to BERT model (Devlin, 2018). It encodes the query and document into a single embedding and predict the probability of the document being relevant.

Self-supervised and unsupervised methods have gained popularity for pretraining models on massive datasets by leveraging unlabeled data to learn rich representations. (Huang et al., 2024) introduces an unsupervised CIR that uses the sequence likelihood estimation capabilities of cross-lingual language models to generate pseudo labels for dense retriever training. The framework consists of two stages to improve retriever performance iteratively. (Izacard et al., 2021) proposed the mContriever model, a retriever that specializes in multilingual and cross-lingual dense retrieval tasks trained using unsupervised contrastive learning. It uses a dual encoder architecture, where it separately encodes queries and documents into dense embeddings and computes their cosine similarity. It shows advantage in its scalability and ability to handle a wide variety of languages.

Strong data augmentation is a crucial part of self-supervised learning and unsupervised learning. (Bonifacio et al., 2022) makes use of the few-shot capabilities of LLMs to generate synthetic queries from documents for IR tasks. It shows that models fine-tuned on synthetic data can outperform strong baselines. However, documents alone are limited in low-resource languages and complete query generation from documents using LLM could be unstable as this would require LLM to fully understand the document.

3 Methods

In this section, we will describe two effective approaches to CIR model fine-tuning.

3.1 Data Augmentation with LLM

LLMs (Achiam et al., 2023; Dubey et al., 2024; Tiedemann and Thottingal, 2020) have displayed impressive language understanding abilities. We propose using this capability to expand low-resource language data using augmentation. There are several options to sentence augmentation using LLMs: summarization, elaboration, and rephrasing. Summarization condenses the information in a document, which reduces the precision of information. On the other hand, elaboration introduces additional information to the document, making it harder to train an encoder model that retains the core information. These challenges may cause inaccurate and misleading LLM outputs that can damage IR model performance. Rephrasing is less complex as it focuses on rearranging similar content in different words, making it a safer and more reliable augmentation strategy for low resource languages.

We focus on rephrasing queries using two techniques. The first is replacing keywords with synonyms which expands and deepens the model's understanding of a larger set of vocabularies. Another technique is modifying sentence structure (e.g. anastrophe). This introduces the model to diverse sentence structures during training, making it more robust toward complex sentences. Our data augmentation pipeline consists of the following steps:

1. **Translation**: Translate low resource language queries to English.

2. **Paraphrasing**: Paraphrase English text with prompts that encourage synonym substitution and sentence structure variation.

3. **Back-Translation**: Translate paraphrased English queries back to original low resource language.

4. **Dataset Integration**: Integrate the augmented queries and original corresponding document into dataset.

The augmentation process enhances the model's ability to handle syntactic and lexical variations across languages, improving its generalization to

diverse user queries. Figure 1 shows an example of a paraphrased query in a dataset which maps Chinese queries to English documents. This approach allows for scalability and can be extended to support low-resource languages.

```
Original Query (Chinese): 1869 年, 哪个国家发生了一起抢劫案?
Translated Query (English): In 1869, in which country was there a robbery?
Paraphrased Query (English): In which country was a robbery in 1869?
Paraphrased Query (Chinese): 1869年在哪个国家发生抢劫案?
```

Figure 1: An example of data augmentation process, showing the original query in Chinese, its English translation, a paraphrased query in English, and the back-translated paraphrased query in Chinese.

3.2 Weighted InfoNCE Loss

3.2.1 Review on InfoNCE Loss

The original mContriever relies on the contrastive InfoNCE loss (Oord et al., 2018). Given a batch of query-document pairs, InfoNCE aims to maximize the similarity between the query embedding and its corresponding positive document embedding while minimizing its similarity to other document embeddings in the batch, similar to the softmax-based cross entropy loss. Formally, let q_i and d_i^+ denote the query and its matching positive document, and d_j denotes other documents (including negatives), the InfoNCE loss is:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(q_i, d_i^+)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(q_i, d_j)/\tau\right)} \quad (1)$$

where τ is a temperature parameter and $\text{sim}(\cdot,\cdot)$ is the cosine similarity.

The InfoNCE loss is central to learning discriminative representations, ensuring that embeddings of semantically related pairs (e.g., queries and their relevant documents) are close, while embeddings of unrelated pairs remain distant.

While the InfoNCE loss is effective for learning discriminative representations, it falls short when applied to multilingual scenarios, particularly for low-resource languages where data scarcity creates imbalances in representational capacity (Joshi et al., 2020). The model devotes proportionally less representational capacity to languages with fewer samples, limiting performance for those languages.

3.2.2 Weighted InfoNCE Loss

To solve the problems of InfoNCE loss, we propose modifications that acknowledge and compensate for data scarcity: weighted InfoNCE loss. The core motivation behind our proposed modifications is to re-balance the training signal.

The Weighted InfoNCE loss modifies the original contrastive objective to assign higher weights to underrepresented samples. Specifically, we increase the contribution of low-resource language pairs to the overall loss. This encourages the model to pay more attention to characteristics specific to low-resource languages.

$$\mathcal{L}_{\text{Weighted InfoNCE}} = \sum_{i=1}^{N}\frac{w_i \cdot \mathcal{L}_{\text{InfoNCE},i}}{\sum_{j=1}^{N} w_j} \quad (2)$$

where $w_i > 1$ for samples originating from low-resource languages and $w_i = 1$ for samples in high-resource languages. For example, we can set these static weights to $w_i = 2.0$ for low-resource samples and $w_i = 1.0$ otherwise. This targeted weighting ensures that the gradient updates give stronger preference to underrepresented languages, allowing the model to better account for their linguistic characteristics and ultimately improving retrieval performance in these settings.

4 Experiments

4.1 Dataset

For our experiments, we used the SWIM-IR Cross-Lingual dataset (Thakur et al., 2023). The SWIM-IR Cross-Lingual dataset is a synthetic dataset consisting of 28 million query-passage pairs across 33 languages, including low-resource languages like Swahili and Telugu. The passages are sampled from Wikipedia and paired with queries generated by PaLM-2 (Anil et al., 2023). The documents in SWIM-IR Cross-Lingual dataset are in english, while the queries are in other languages. The SWIM-IR Cross-Lingual dataset provides a robust foundation for training cross-lingual retrieval models, as it includes diverse linguistic and contextual variations in both queries and documents. The dataset's focus on relevance labels enables the evaluation of dense retrieval methods, such as those relying on contrastive learning with positive and negative pairs.

4.2 Evaluation Metrics

We use the standard recall at k (R@k) and Mean Reciprocal Rank (MRR) as evaluation metrics for our experiments. Recall measures the proportion of relevant documents retrieved within the top k

results. MRR measures the quality of a retrieval model using the rank of the first retrieved relevant document.

$$\text{R@k} = \frac{1}{N}\sum_{i=1}^{N} 1(\text{Retrieved}(q_i) \in \text{Top}_K(q_i)) \tag{3}$$

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\text{rank}_i} \tag{4}$$

where $1(\cdot)$ is the indicator function and N is the number of test queries.

4.3 Experiments on Data Augmentation with LLM

4.3.1 Implementation Details

We used the Chinese query mapping to English documents subset of the SWIM-IR Cross-Lingual dataset for fine-tuning with LLM data augmentation experiments. We selected Chinese as the primary language for our experiments because two members of our team are proficient in Chinese, allowing us to qualitatively assess the accuracy and linguistic quality of the paraphrased queries during our experiments. We used 5000 training pairs, 5000 validation pairs, and 5000 test pairs. For each of the 50000 Chinese query, we augmented it to form 5000 additional queries. Pairing the augmented Chinese queries with the corresponding English documents of the original queries, we obtain 5000 additional Chinese query-English document pairs. To augment a Chinese query, we used `opus-mt-zh-en` to translate Chinese query to English, `T5_Paraphrase_Paws` to rephrase the translated English query, and `opus-mt-zh-en` to translate the rephrased English query back to Chinese.

The dataset is tokenized using mContriever's multilingual tokenizer, with sequences standardized to 128 tokens through truncation or padding. This uniform preprocessing ensures consistency across languages, relying on model-driven adjustments rather than language-specific features.

We fine-tuned the last encoder layer of the mContriever model for 3 epochs, using a batch size of 128 on 1 A100 GPU. We experimented with fine-tuning using InfoNCE loss and triplet loss. We used a learning rate of 3×10^{-5} when using InfoNCE loss and 1×10^{-5} when using triplet loss. For all experiments, we used the AdamW optimizer with 0.01 weight decay.

4.3.2 Results

We report our test performances in Table 1. "Raw" refers to the mContriever model tested without fine-tuning, "InfoNCE" and "Triplet" refers to fine-tuning with InfoNCE loss and triplet loss without the additional 5000 LLM augmented pairs, and "InfoNCE + Aug" and "Triplet + Aug" refers to fine-tuning with InfoNCE loss and triplet loss with the additional 5000 LLM augmented pairs.

Our results show fine-tuning with additional LLM augmented data consistently improves the CIR model performance, where "InfoNCE + Aug" outperforms "InfoNCE" and "Triplet + Aug" outperforms "Triplet", with "InfoNCE + Aug" having the highest performance across all 4 evaluation metrics used.

4.4 Experiments on Weighted InfoNCE Loss

4.4.1 Implementation Details

To construct a balanced yet resource-sensitive training set, we first define a fixed training size for high-resource languages and proportionally scale subsets for low-resource languages. Specifically, we sample 1,500 training examples per relatively high-resource language, such as Chinese, Spanish, and French, to ensure robust coverage of their linguistic variability. This size represents a typical upper limit for languages with abundant data, enabling the model to capture nuanced semantic and syntactic features effectively.

For low-resource languages, we experimented with Bengali, Swahili, Telugu, and Thai, where we deliberately choose a smaller subset size, set to $0.3 \times 1,500 = 450$ examples. This reduction acknowledges that generally low-resource settings offer substantially fewer annotated samples. The chosen ratio of 0.3 serves two primary purposes. First, it simulates a realistic data scarcity scenario, forcing the model to generalize from a relatively smaller pool of examples. Second, it maintains a controlled proportion of low-resource data relative to the high-resource data, preventing the model from disproportionately focusing on languages with sparse training signals while still ensuring that these underrepresented languages contribute meaningfully to the overall training objective.

Our choice of sampling 30% of high-resource data for sampling low-resource languages is moti-

Metric	Raw	InfoNCE	InfoNCE + Aug	Triplet	Triplet + Aug
MRR	0.3677	0.5440	**0.5720**	0.4788	0.5316
R@1	0.2938	0.4532	**0.4832**	0.3902	0.4418
R@5	0.4444	0.6484	**0.6700**	0.5750	0.6314
R@10	0.5096	0.7100	**0.7398**	0.6450	0.6986

Table 1: Performance metric comparison across different fine-tuning methods. Fine-tuning with InfoNCE loss and LLM augmented queries outperforms all other configurations.

vated by the findings in (Conneau, 2019), which highlighted the importance of balancing language representation in multilingual training. This parameter ensures that low-resource languages contribute effectively to the model's learning process without disproportionately skewing the overall training dynamics.

A custom test set of 1,000 examples is reserved for Bengali. By creating an evaluation set of a fixed, reasonably sized subset, we ensure that we can fairly assess the impact of our proposed loss functions and optimization strategies on retrieval quality.

After preparing the data, we shuffle and split the combined training portion into a training set (90%) and a validation set (10%) for all languages. This ensures that tuning hyperparameters and evaluating early stopping criteria can be done on a portion of the data not directly used to optimize the model. Furthermore, the validation set provides insight into how adjustments to weighting strategies and alignment losses transfer to unseen data without prematurely exposing the model to the held-out test sets.

For our experiments with Weighted InfoNCE loss, we set the low-resource weight for the InfoNCE term to 2.0, while high-resource samples retain a weight of 1.0. This choice is motivated by practical considerations rather than extensive hyperparameter tuning. Due to limited computational resources and time constraints, it was not feasible to exhaustively search for the optimal weighting factor. Instead, we relied on informed reasoning and a small number of preliminary tests to identify a weight that is sufficiently large to amplify the low-resource signal, yet not so large as to destabilize training or overshadow high-resource performance.

Other training settings such as tokenizer, learning rate, optimizer, follows the same settings in section 4.2.

In summary, the chosen subset sizes and proportions reflect a strategic compromise. They impose a challenging yet realistic scenario for low-resource languages like Bengali, while still providing sufficient data from high-resource languages to enable robust multilingual retrieval. Our goal is to demonstrate that the proposed loss adjustments can improve low-resource performance without sacrificing general quality, all under controlled experimental conditions that approximate real-world data imbalances.

4.4.2 Results

To evaluate the effectiveness of our proposed weighted InfoNCE loss, we compare the baseline mContriever model fine-tuned using only the standard InfoNCE loss with our proposed Weighted InfoNCE loss.

We evaluate on a dedicated test set of 1,000 examples for Bengali, Swahili, Telugu, and Thai. Our results are reported in Table 2.

We observe the following key trends. First, the weighted InfoNCE loss demonstrates consistent improvements over the baseline InfoNCE loss across all metrics (MRR, R@1, R@5, R@10). This validates the hypothesis that emphasizing low-resource languages via weighting leads to better retrieval performance. For Bengali, weighted loss improves MRR from 0.6507 to 0.6683 and R@1 from 0.5530 to 0.5860. These gains suggest that Bengali benefits from additional weighting, likely due to its morphological complexity and moderate representation in the dataset. However, the improvements are modest, possibly because Bengali queries are inherently ambiguous or noisy. Swahili achieves the highest MRR of 0.8886 with InfoNCE loss and remains competitive (0.8845) under weighted loss. This strong baseline performance could be attributed to Swahili's simpler grammar and well-represented, high-quality data in the dataset. The minor drop under weighted loss suggests that Swahili may not benefit significantly from additional weighting, as it is already performing near its optimal capacity.

For Telugu, weighted loss marginally improves MRR from 0.6625 to 0.6658 and R@10 from

Language	Loss Type	MRR	R@1	R@5	R@10
Bengali (bn)	InfoNCE Loss	0.6507	0.5530	0.7630	0.8130
	Weighted InfoNCE Loss	**0.6683**	**0.5860**	**0.7660**	**0.8350**
Swahili (sw)	InfoNCE Loss	**0.8886**	**0.8500**	0.9410	0.9580
	Weighted InfoNCE Loss	0.8845	0.8420	**0.9430**	**0.9600**
Telugu (te)	InfoNCE Loss	0.6625	0.5700	0.7620	0.8310
	Weighted InfoNCE Loss	**0.6658**	**0.5710**	**0.7670**	**0.8420**
Thai (th)	InfoNCE Loss	0.7681	0.6970	0.8510	0.8980
	Weighted InfoNCE Loss	**0.8420**	**0.7870**	**0.9110**	**0.9510**

Table 2: Performance metric comparison for low-resource languages using InfoNCE loss and weighted InfoNCE loss.

0.8310 to 0.8420. These small gains indicate that Telugu may be limited by dataset quality or unique linguistic challenges, such as its rich morphology and complex syntax. The relatively modest improvement suggests that further optimization or augmentation may be required for Telugu. Thai shows substantial improvement with weighted loss, where MRR increases from 0.7681 to 0.8420 and R@1 rises significantly from 0.6970 to 0.7870. This result indicates that the model benefits greatly from the weighted approach, likely due to Thai's underrepresentation and tokenization challenges (e.g., lack of word boundaries). Emphasizing Thai in the loss function compensates for these challenges effectively.

The variability across languages highlights the interplay between linguistic complexity, dataset quality, and representation. While Swahili and Thai benefit significantly due to their simpler linguistic structure or effective weighting, Telugu and Bengali exhibit more modest gains due to their inherent challenges or noisier datasets. These results demonstrate that the weighted InfoNCE loss is an effective strategy for enhancing low-resource performance, although its impact varies by language. The findings suggest that further customization, such as dynamic weighting or language-specific preprocessing, could yield even greater gains in future work.

5 Future Experimentation

To further enhance the adaptability and robustness of our multilingual retrieval system, we propose several future directions. One key focus is Dynamic Weight Optimization, where instead of relying on static weights (e.g., weight = 2), we experiment with dynamically learning weights during training. This approach would allow the model to adapt better to the dramatic differences between low-resource languages by assigning higher weights to underrepresented languages or those contributing higher losses (Piñeiro-Martín et al., 2024). For instance, languages with lower representation in the dataset could be weighted inversely proportional to their prevalence, ensuring they are not overshadowed during training. This idea aligns with our findings, where static weights improved performance for low-resource languages, but further dynamic adjustments could optimize results.

Additionally, we propose expanding evaluation on Low-Resource languages by leveraging multilingual datasets with diverse linguistic structures. Incorporating zero-shot and few-shot learning techniques could provide a powerful way to address data scarcity in these settings. Another critical direction is evaluation across domains, where we assess the model's adaptability in specialized fields such as biomedicine or law. These domains pose unique challenges due to their specialized vocabulary and sentence structures, offering an opportunity to evaluate the retrieval model's real-world applicability. Lastly, we aim to explore more robust paraphrasing models to improve data augmentation during training. Advanced models capable of preserving critical proper nouns and domain-specific terminology in queries could mitigate current challenges where paraphrasing inadvertently alters key information, potentially degrading retrieval performance.

5.1 Contrastive Learning using Hard Negative Pairs

In addition to the standard document encoder and query encoder training using the similarity score, we can perform supervised contrastive learning using hard negative samples. For every query at every

training iteration, we can identify the top k negative documents that have the highest similarity score. These are hard negative samples that the model failed to identify as dissimilar to the query and can be used for contrastive learning for targeted learning. Document-document pairs with non-overlapping queries and query-query pairs with non-overlapping documents but with high similarity scores can be used for document encoder and query encoder contrastive learning.

6 Impact Statement

This work contributes to advancing CIR by addressing challenges in low-resource language settings and exploring data-efficient methods by incorporating data augmentation techniques and novel loss functions. Improved CIR systems can enable diverse user communities to access multilingual datasets, fostering inclusive knowledge-sharing. The scalability of our approach provides practical solutions for expanding retrieval systems to underserved languages. This research can also be expanded to other domains such as healthcare or legal texts, which are less focused on in current CIR research.

However, our work also presents several ethical considerations. Augmented data with paraphrasing and translation may inadvertently introduce biases or inaccuracies, which could propagate errors in downstream applications. For instance, paraphrasing models may fail to preserve proper nouns or domain-specific terminology, leading to semantic drift. Additionally, CIR systems can reinforce systemic biases if training data is unbalanced or reflective of sociolinguistic inequalities. Future research should carefully evaluate the fairness and reliability of retrieval outputs, particularly in sensitive domains, to ensure equitable access to information. We recognize the need for interdisciplinary efforts to address these challenges and responsibly develop technologies for diverse global users who speak different languages.

7 Supplemental Information

Our code is available in this Google Drive folder: https://drive.google.com/drive/folders/1jIE0PlCTcwxVa-UGcMNwg2OVq92c3xKi?usp=sharing.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2387–2392.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chao-Wei Huang, Chen-An Li, Tsu-Yuan Hsu, Chen-Yu Hsu, and Yun-Nung Chen. 2024. Unsupervised multilingual dense retrieval via generative pseudo labeling. *arXiv preprint arXiv:2403.03516*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, María del Carmen López-Pérez, and Georg Rehm. 2024. Weighted cross-entropy for low-resource languages in multilingual speech recognition. *arXiv preprint arXiv:2409.16954*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Wieting, Jimmy Lin, and Daniel Cer. 2023. Leveraging llms for synthesizing training data across many languages in multilingual dense retrieval. *CoRR*, abs/2311.05800.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.