ContrastiveMix: Overcoming Code-Mixing Dilemma in Cross-Lingual Transfer for Information Retrieval

Junggeun Do
Seoul National University
jg.do@snu.ac.kr

Jaeseong Lee
Seoul National University
tbvj5914@snu.ac.kr

Seung-won Hwang
Seoul National University
seungwonh@snu.ac.kr

Abstract

Multilingual pretrained language models (mPLMs) have been widely adopted in cross-lingual transfer, and code-mixing has demonstrated effectiveness across various tasks in the absence of target language data. Our contribution involves an in-depth investigation into the counterproductive nature of training mPLMs on code-mixed data for information retrieval (IR). Our finding is that while code-mixing demonstrates a positive effect in aligning representations across languages, it hampers the IR-specific objective of matching representations between queries and relevant passages. To balance between positive and negative effects, we introduce ContrastiveMix, which disentangles contrastive loss between these conflicting objectives, thereby enhancing zero-shot IR performance. Specifically, we leverage both English and code-mixed data and employ two contrastive loss functions, by adding an additional contrastive loss that aligns embeddings of English data with their code-mixed counterparts in the query encoder. Our proposed ContrastiveMix exhibits statistically significant outperformance compared to mDPR, particularly in scenarios involving lower linguistic similarity, where the conflict between goals is more pronounced. Our code is publicly available.[1]

1 Introduction

Multilingual pretrained language models (mPLMs) have been a key ingredient in cross-lingual transfer. Several studies (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2019; Feng et al., 2022a) have highlighted that the inherent multilinguality of these models facilitates knowledge transfer from high-resource to low-resource languages, even in data-scarce scenarios.

In cross-lingual transfer, code-mixing (Qin et al., 2021; Feng et al., 2022b) has been effective in both sentence-level and token-level zero-shot cross-lingual tasks. Code-mixing, employed in the absence of target language data, involves sentences containing words from multiple languages, aligns representations across different languages by training mBERT (Devlin et al., 2019) on code-mixed data.

However, code-mixing is less effective in the case of cross-lingual transfer for information retrieval (IR), compared to training solely on English data. In contrast to cross-lingual transfer for classification tasks, where the primary challenge is representation alignment, IR introduces the additional hurdle of representation matching between relevant query-passage pairs, where DPR (Karpukhin et al., 2020) is a de facto standard architecture. Our hypothesis is that code-mixing, while contributing to the shared goal of representation alignment, may adversely affect the IR-specific challenge of query-passage matching, by introducing noise into text embeddings.

Our contribution aims to strike a balance between the positive and negative effects of code-mixing on zero-shot IR performance, introducing ContrastiveMix as our proposed solution. Our distinction is disentangling the two objectives of **representation alignment** and **relevance matching** into two contrastive loss functions, involving both English and code-mixed data, to facilitate relevance matching. Specifically, we incorporate an additional contrastive loss, aligning embeddings of English data with their code-mixed counterparts, solely in the query encoder. With these two loss functions, our method aligns cross-lingual representations while mitigating interference caused by embedding noise during IR learning. As a result, ContrastiveMix effectively achieves both language alignment and relevance matching.

Experimental results demonstrate that ContrastiveMix outperforms mDPR, with statistically significant distinctions observed across eight di-

[1] https://github.com/DoJunggeun/contrastivemix

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 197–204
June 16-21, 2024

verse languages. In contrast, a naive code-mixing approach consistently falls short when compared to mDPR. These eight languages can be categorically divided into two groups based on linguistic similarities, one belonging to the Indo-European language family, the same as English (or the high similarity group), and the other constituting a low similarity group.

As the challenge of representation alignment is lower in the former group, consistently highlighted by the effectiveness of mPLMs in such group evidenced by Chi et al. (2020); Krishnan et al. (2021); Xu et al. (2022), mitigating the adverse effects on alignment becomes more apparent in the low similarity group. Our observations confirm that the impact of ContrastiveMix is notably more pronounced in the low similarity group, resulting in a widened performance gap of 1.29 MRR@100 when compared to mDPR.

Our contributions can be summarized as follows: (1) We demonstrate that training IR models on code-mixed data is not effective for cross-lingual transfer. (2) We propose a method to enhance zero-shot IR performance through code-mixing. (3) We analyze the effectiveness of code-mixing in IR through representation alignment.

2 Related Work

2.1 Code-Mixing in NLP

Several studies (Zhang et al., 2019; Yang et al., 2020, 2021) have explored the use of code-mixed sentences to enhance representations across multiple languages, demonstrating success in various tasks such as machine translation and cross-lingual parsing. Qin et al. (2021) exhibited the capability of zero-shot cross-lingual transfer through multilingual code-mixing as data augmentation across several classification and sequence labeling tasks in 19 languages. Feng et al. (2022b) improved zero-shot performance in Part-of-Speech tagging and Named-Entity Recognition across 33 languages by considering token-level coherence based on the similarity between code-mixed sentences and English sentences.

2.2 Code-Mixing in Information Retrieval

Recent works (Huang et al., 2023; Litschko et al., 2023) utilized code-mixing for cross-lingual information retrieval in a document reranking manner. Litschko et al. (2023) employed bilingual and multilingual code-mixing, based on word embeddings and parallel Wikipedia titles, to improve cross-lingual retrieval between two languages and multilingual retrieval among multiple languages. However, they did not focus on cross-lingual transfer for monolingual IR.

3 Method

3.1 Preliminaries: Multilingual DPR

Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) is a widely used architecture using two BERT encoders to encode queries and passages. DPR is optimized to maximize the inner products of embeddings of a query and its relevant passage during training. At inference time, DPR retrieves passages with the most similar embedding to a given query through maximum inner product search.

The training objective involves comparing all pairs of queries and passages in a batch. With the query embedding denoted as q, the corresponding gold passage embedding as p_t, and passage embeddings as p_i (including $p_t, i = 1, 2, ..., N$), the loss function is defined as follows.

$$L_{ir}(q, p_i) = -\log\left(\frac{e^{q \cdot p_t}}{\sum_{i=1}^{N} e^{q \cdot p_i}}\right) \quad (1)$$

DPR initialized with multilingual BERT is referred to as multilingual DPR (mDPR). We use mDPR as baseline architecture.

3.2 NaiveMix: Naïve Code-Mixing

Naïve code-mixing approach, named NaiveMix, is to train models on code-mixed data, similar to previous studies (Qin et al., 2021; Huang et al., 2023; Litschko et al., 2023).

We conduct code-mixing through the following steps. First, at each training step, we select entire queries or passages with a probability of $\mathbf{r}_s$ in a batch. Second, we choose words to be replaced within each selected data with a probability of $\mathbf{r}_w$. Subsequently, we replace the selected words with equivalent terms in the target language.

3.3 Motivation: Mixed Effect of Code-Mixing

There are two conflicting objectives in achieving cross-lingual transfer for IR: aligning representations between source and target languages, and effectively matching relevant query-passage pairs. The first goal, language alignment, can be achieved through NaiveMix, since training mPLMs on code-mixed data typically contributes to language alignment, as indicated by Qin et al. (2021). However, NaiveMix hinders the additional goal of relevance matching due to the embedding differences between English and code-mixed text.

Further compounding this issue, embedding differences become larger for longer text representations, such as passages in IR. The average cosine similarities of mBERT embeddings between English data, in the validation set of Natural Questions (Kwiatkowski et al., 2019), and their code-mixed counterparts are 0.944 (SD = 0.036) for queries and 0.867 (SD = 0.257) for passages.

The lower similarity and substantial variation in code-mixed passages appear to pose a greater hindrance to relevance matching. Therefore, we apply code-mixing only to the query encoder to align language representations while avoiding interference with relevance matching.

3.4 Proposed: ContrastiveMix

To address the dilemma of two conflicting goals, we propose ContrastiveMix, which is designed to learn IR in English while transferring this knowledge to the target language through code-mixing. This approach involves training models on English data with an additional contrastive loss L_c, that aligns English query with its corresponding code-mixed query. By introducing this objective, we can separate the roles of English IR data and code-mixed data into IR learning and representation alignment, instead of directly training on code-mixed data.

Specifically, we implement this method as an in-batch contrastive loss. When the batch size is N, the embedding vectors of the English query and its code-mixed counterpart are denoted as q_s and q_t, respectively, and the embedding vectors of all code-mixed queries within the batch represented as q_j (including q_t, $j = 1, 2, ..., N$), the contrastive loss term is defined as follows:

$$L_c(q_s, q_t) = -\log\left(\frac{e^{q_s \cdot q_t}}{\sum_{j=1}^{N} e^{q_s \cdot q_j}}\right) \tag{2}$$

and the entire training objective is

$$L = L_{ir} + wL_c \tag{3}$$

where w is a hyperparameter for weighting the contrastive loss.

4 Experiment

4.1 Setting

Datasets. We train models on Natural Questions (Kwiatkowski et al., 2019) and evaluate their performance on Mr.TyDi (Zhang et al., 2021; Clark et al., 2020) and MIRACL (Zhang et al., 2023), multilingual datasets for monolingual information retrieval, containing 11 and 18 languages, respectively.

Dictionary for Code-Mixing. We follow Qin et al. (2021) to use MUSE[2] (Lample et al., 2018) bilingual dictionaries to replace English words with their equivalents in target language. In cases where multiple target words are available in the MUSE dictionary, we randomly select one.

Sparse-Dense Hybrid Approach. Dense retrieval models can yield better performance by adopting the hybrid approach in most cases, with minimal additional computational costs. Therefore, we adopt the sparse–dense hybrid approach following Zhang et al. (2021), where the final retrieval score is computed by a linear combination of the BM25 score and the dense retrieval score. The weighting parameter is tuned in the range [0, 1] on the validation set, with a step size of 0.05.

Evaluation. We evaluate models using MRR@100 and Recall@100, as in Zhang et al. (2021), where each metric has been scaled within the [0, 100] range by multiplying by 100. MRR (Mean Reciprocal Ranking) assesses the model's ability to generate a high-quality ranking, while Recall provides an upper limit on overall effectiveness in an end-to-end scenario. Evaluation is conducted on all eight languages (Arabic, Thai, Japanese, Korean, Indonesian, Bengali, Finnish, Russian) common between Mr.TyDi and MUSE, and on all five languages (Chinese, Hindi, Persian, Spanish, French) not in Mr.TyDi and common between MIRACL and MUSE. Since the test set of MIRACL dataset is not publicly available, we used the validation set for evaluation.

4.2 Baselines

We compare ContrastiveMix with two baselines.

mDPR, trained on Natural Questions. We expect a certain degree of cross-lingual transfer due to multilinguality of mBERT.

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

Language Group	Low Similarity (target scenario)							High Similarity (ablation)		
Method	Ar	Th	Ja	Ko	Id	Fi	Avg	Ru	Bn	Avg
mDPR	47.49	46.80	35.35	36.38	48.73	37.58	42.06	43.61	**54.45**	**49.03**
NaiveMix	44.64*	44.43	32.98*	35.90	48.06	37.71	40.62	41.04*	50.26*	45.65
ContrastiveMix (ours)	**48.54***	**48.58***	**36.8***	**38.43***	**49.42***	**38.35***	**43.35**	**43.79**	54.11	48.95

Table 1: MRR@100 on the Mr.TyDi. Results significantly different ($p < 0.05$, paired t-test) from mDPR are starred.

Group	target	ablation				
Method	zh	fa	es	fr	hi	Avg
mDPR	44.93	48.09	63.12	41.65	**56.96**	52.46
NaiveMix	44.81	46.87*	62.97	40.69	55.05	51.40
Ours	**45.70**	**48.42**	**63.76**	**42.10**	56.51	**52.70**

Table 2: MRR@100 on the MIRACL.

Method	Query	Passage	MRR@100	Recall@100
mDPR			42.06	84.47
NaiveMix	✓	✓	40.62	84.34
		✓	39.28	82.51
	✓		41.89	84.62
ContrastiveMix	✓	✓	42.60	84.73
		✓	42.01	84.45
	✓		**43.35**	**84.91**

Table 3: Average performances in low-similarity group languages in Mr.TyDi.

NaiveMix, trained on code-mixed queries and passages. The hyperparameters for code-mixing are determined to $r_s = 0.2$ and $r_w = 0.5$ on the validation set.

4.3 Implementation Details

We trained models based on the mDPR architecture with separate query and passage encoders[3]. In all experiments, following Karpukhin et al. (2020), we trained models using the in-batch negative setting with a batch size of 128 and one additional negative passage per query, for up to 40 epochs with a learning rate of 10^{-5} using the Adam optimizer and linear scheduling with warm-up. We determined the loss weight parameter w through grid search on the validation set in the range of [0.001, 1]. We trained models based on Tevatron (Gao et al., 2023)[4] and evaluated them with Pyserini (Lin et al., 2021)[5].

4.4 Result

Table 1 and Table 2 present the results of experiments in terms of MRR@100. In Table 3, we provide macro-average performances of all method combinations in our target scenarios.

Effectiveness of ContrastiveMix. Table 1 shows the performance of ContrastiveMix in eight languages, which we categorized into high- and low-similarity groups, in terms of linguistic similarity to English. In the high similarity group (Bn, Ru), both belonging to the Indo-European language family, the objective of representation alignment is largely achieved in mDPR, as consistently observed in prior literature (Chi et al., 2020; Krishnan et al., 2021; Xu et al., 2022), making the dilemma between the two objectives less prominent (hence, we denote as ablation).

Meanwhile, in our target group for cross-lingual transfer into the low similarity group (Ar, Th, Ko, Ja, Id, Fi), ContrastiveMix significantly outperforms mDPR and NaiveMix in terms of MRR@100 across all languages. Results on the MIRACL dataset shown in Table 2 also exhibit similar trends to those explained. Moreover, unlike NaiveMix, there was no significant performance degradation in terms of Recall@100 in any case, as shown in Table 4 in Appendix A.

Analysis of Representation Alignment. Models trained with code-mixed queries show better performance compared to others, as shown in Table 3. This is attributed to better representation alignment between languages, evident in UMAP visualization (McInnes et al., 2018).

As shown in Figure 1, the query encoder of ContrastiveMix clearly demonstrates a significant overlap in representations between English (green dots) and the target language (red dots). In contrast, as presented in Figure 2 in Appendix A, encoders of mDPR reveal distinct language clusters, aligning with findings from prior studies (Krishnan et al., 2021; Xu et al., 2022). NaiveMix shows an overlap in the query encoder, but not in the passage encoder

[3]initialized with bert-base-multilingual-cased
[4]https://github.com/texttron/tevatron
[5]https://github.com/castorini/pyserini

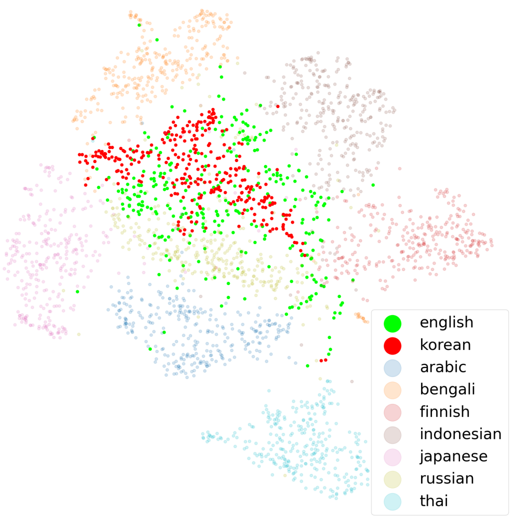

Figure 1: Visualization of the representations of different languages, in the validation set of Mr. TyDi, from query encoder of ContrastiveMix model trained with Korean as the target language.

as in mDPR, as shown in Figure 3 in Appendix A. These results indicate that code-mixed queries contribute to improving language alignment, but code-mixed passages do not, which justifies the design decision of ContrastiveMix: applying contrastive loss for language alignment only to the query encoder.

NaiveMix models show a performance drop despite better language alignment, due to disrupted relevance matching. Their passage-only version performs the worst, as code-mixed passages fail in language alignment. In contrast, even passage-only ContrastiveMix performs comparably to mDPR, indicating the contribution of relevance matching. Query-only ContrastiveMix, our proposed method, shows the best performance, further benefiting from improved language alignment.

5 Conclusion

This paper identifies and overcomes a dilemma in using code-mixing for cross-lingual transfer in IR. Specifically, we add contrastive loss, designed to align the embeddings of English sentences with their code-mixed counterparts, as a key component of the training objective. Our approach experimentally demonstrated better performance than mDPR, with statistical significance.

Limitation

Although it is necessary to carefully consider context to ensure appropriate replacements when dealing with polysemy, we randomly selected from among multiple candidates in the MUSE dictionary during code-mixing. Additionally, we did not consider token-level coherence in our approach, unlike Feng et al. (2022b).

Acknowledgements

This research was partially supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was also partially supported by IITP grant funded by MSIT (No.2022-0-00077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data). We would like to thank Google's TPU Research Cloud (TRC) program for providing access to Cloud TPUs, and the anonymous reviewers for their reviews. Hwang is a corresponding author.

References

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

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*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022a. 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.

Yukun Feng, Feng Li, and Philipp Koehn. 2022b. Toward the limitation of code-switching in cross-lingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5966–5971, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. 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*, SIGIR '23, page 3120–3124, New York, NY, USA. Association for Computing Machinery.

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.

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.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Robert Litschko, Ekaterina Artemova, and Barbara Plank. 2023. Boosting zero-shot cross-lingual retrieval by training on artificially code-switched data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3096–3108, Toronto, Canada. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2022. Cross-linguistic syntactic difference in multilingual BERT: How good is it and how does it affect transfer? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8073–8092, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021. Multilingual agreement for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 233–239, Online. Association for Computational Linguistics.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-lingual dependency parsing using code-mixed TreeBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 997–1006, Hong Kong, China. Association for Computational Linguistics.

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, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A Appendix

Table 4 presents the experimental results in terms of Recall@100. Figure 2 and Figure 3 show the visualization of representations of different languages in the validation set of Mr. TyDi.

Language Group	Low Similarity							High Similarity		
Method	Ar	Th	Ja	Ko	Id	Fi	Avg	Ru	Bn	Avg
mDPR	87.29	**87.81**	81.40	74.63	90.68	84.99	84.47	**81.68**	**92.79**	**87.24**
NaiveMix	86.47	87.43	78.86*	74.78	**91.91***	**86.57***	84.34	79.93	90.54*	85.24
ContrastiveMix (ours)	**87.45**	87.77	**82.23***	**75.22**	90.99	85.80*	**84.91**	81.32	91.80	86.56

Table 4: Recall@100 on the Mr.TyDi.

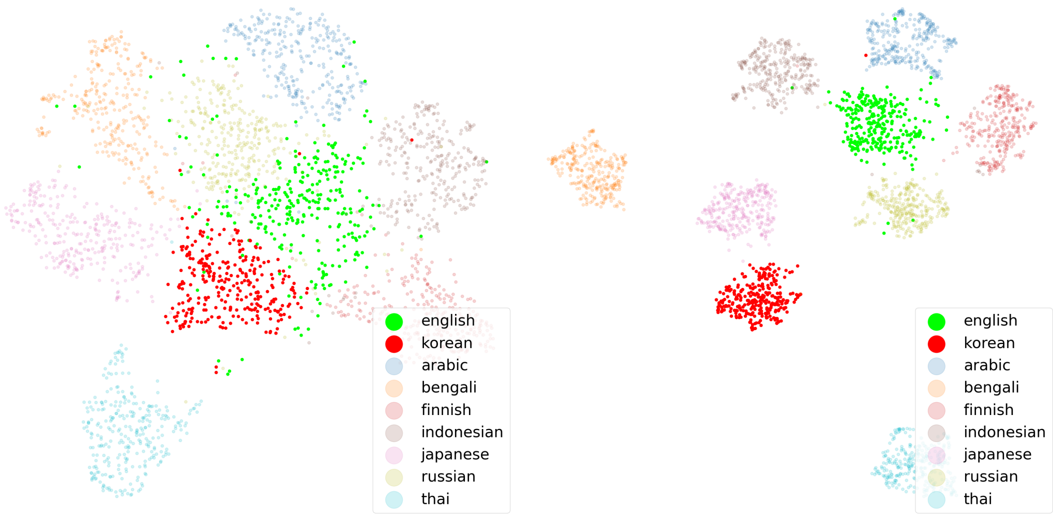

Figure 2: Visualization of the representations from query encoder (left) and passage encoder (right) of mDPR model trained with Korean as the target language.

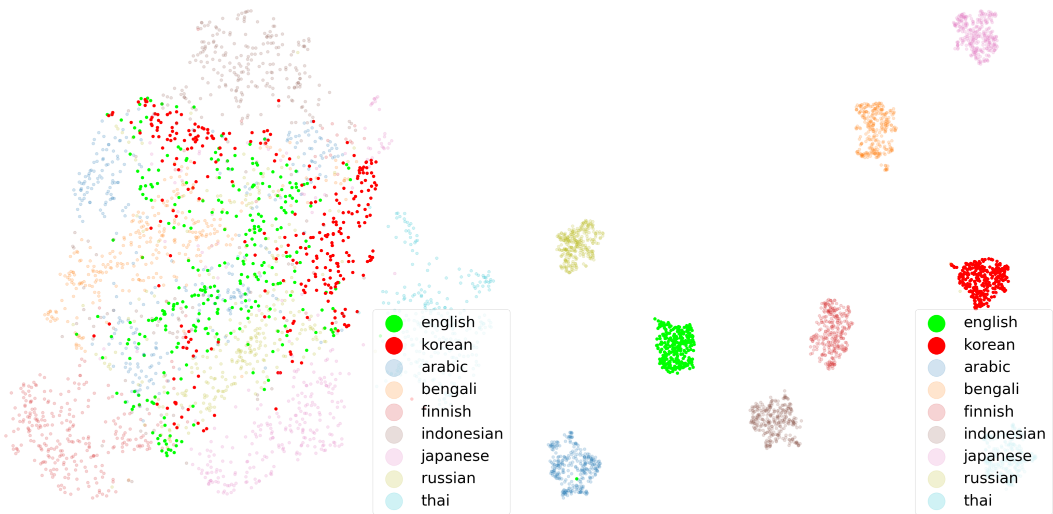

Figure 3: Visualization of the representations from query encoder (left) and passage encoder (right) of NaiveMix model trained with Korean as the target language.