IndicIRSuite: Multilingual Dataset and Neural Information Models for Indian Languages

Saiful haq[1,4], Ashutosh Sharma[3],
Omar Khattab[2], Niyati Chhaya[4], Pushpak Bhattacharyya[1]

[1]IIT Bombay, [2]Stanford University, [3]UIUC, [4]Hyprbots Systems Private Limited
Correspondence: saifulhaq@cse.iitb.ac.in

Abstract

In this paper, we introduce Neural Information Retrieval resources for 11 widely spoken Indian Languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu) from two major Indian language families (Indo-Aryan and Dravidian). These resources include (a) INDIC-MARCO, a multilingual version of the MS MARCO dataset in 11 Indian Languages created using Machine Translation, and (b) Indic-ColBERT, a collection of 11 distinct Monolingual Neural Information Retrieval models, each trained on one of the 11 languages in the INDIC-MARCO dataset. To the best of our knowledge, IndicIRSuite is the first attempt at building large-scale Neural Information Retrieval resources for a large number of Indian languages, and we hope that it will help accelerate research in Neural IR for Indian Languages. Experiments demonstrate that Indic-ColBERT achieves 47.47% improvement in the MRR@10 score averaged over the INDIC-MARCO baselines for all 11 Indian languages except Oriya, 12.26% improvement in the NDCG@10 score averaged over the MIRACL Bengali and Hindi Language baselines, and 20% improvement in the MRR@100 Score over the Mr. Tydi Bengali Language baseline. IndicIRSuite is available at github.com/saifulhaq95/IndicIRSuite.

1 Introduction

Information Retrieval (IR) models process user queries and search the document corpus to retrieve a ranked list of relevant documents ordered by a relevance score. Classical IR models, like BM25 (Robertson et al., 2009), retrieve documents that have lexical overlap with the query tokens. Recently, there has been a notable upsurge in adopting Neural IR models utilizing language models such as BERT (Devlin et al., 2018), which enable semantic matching of queries and documents. This shift has proven highly effective in retrieving and re-ranking documents. ColBERTv2 (Santhanam et al., 2021), one of the state-of-art neural IR models, has shown 18.5 points improvement in NDCG@10 Score over the BM25 model baseline on the MS MARCO dataset (Thakur et al., 2021).

The importance of dataset size outweighs domain-matching in training neural IR models (Zhang et al., 2022a). Due to the scarcity of large-scale domain-specific datasets, Neural IR models are first trained on the MS MARCO passage ranking dataset (Nguyen et al., 2016), and they are subsequently evaluated on domain-specific datasets in a zero-shot manner. MS MARCO dataset contains 39 million training triplets (q, +d, -d) where q is an actual query from the Bing search engine, +d is a human-labeled passage answering the query, and -d is sampled from unlabelled passages retrieved by the BM25 model. The MS MARCO dataset is in English, implying that neural IR models trained on it are effective only with English queries and passages.

Monolingual IR for non-English languages (Zhang et al., 2022b) (Zhang et al., 2021), Multilingual IR (Lawrie et al., 2023), and Cross-lingual IR (Lin et al., 2023; Sun and Duh, 2020) extend the English IR paradigm to support diverse languages. In Monolingual IR for non-English languages, the query and passages are in the same language, which is not English. In cross-lingual IR, the query is used to create a ranked list of documents such that each document is in the same language, which is different from the query language. In Multilingual IR, the query is used to create a ranked list of documents such that each document is in one of the several languages, which can be the same or different from the query language. In this work, we focus on Monolingual IR for non-English languages.

Monolingual IR for non-English languages involves training an encoder like mBERT (Devlin et al., 2018), on a large-scale general-domain

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 501–509

monolingual dataset for non-English languages to minimize the pairwise softmax cross-entropy loss. The trained models are subsequently finetuned or used in a zero-shot manner on small-scale domain-specific datasets. However, there is a notable lack of large-scale datasets like mMARCO (Bonifacio et al., 2021) for training monolingual neural IR models on many low-resource Indian languages. We introduce neural IR resources to address this scarcity and facilitate Monolingual neural IR across 11 Indian languages. Our contributions are:

- INDIC-MARCO, a multilingual dataset for training neural IR models in 11 Indian Languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu). For every language in INDIC-MARCO, there exists 8.8 Million passages, 1 Million queries, 39 million training triplets (query, relevant document, irrelevant document), and approximately one relevant document per query. To the best of our knowledge, this is the first large-scale dataset for training a neural IR system on 11 widely spoken Indian languages.

- Indic-ColBERT, a collection of 11 distinct Monolingual Neural Information Retrieval models, each trained on one of the 11 languages in the INDIC-MARCO dataset. Indic-ColBERT achieves 47.47% improvement in the MRR @10 score averaged over the INDIC-MARCO baseline for all 11 Indian languages except Oriya, 12.26% improvement in the NDCG @10 score averaged over the MIRACL Bengali and Hindi Language baselines, and 20% improvement in the MRR@100 Score over the Mr. Tydi Bengali Language baseline. To the best of our knowledge, this is the first effort for a neural IR dataset and models on 11 major Indian languages, thereby providing a benchmark for Indian language IR.

2 Related work

The size of datasets holds greater importance than ensuring domain matching in the training of neural IR models (Zhang et al., 2022a). In terms of size and domain, mMARCO (Bonifacio et al., 2021) is the most similar to our work as it introduces a large-scale machine-translated version of MS MARCO in many languages, Hindi being the only Indian language. MIRACL (Zhang et al., 2022b) and Mr. Tydi (Zhang et al., 2021) also introduce datasets and models for Monolingual Neural IR in Hindi, Bengali, and Telugu.

FIRE[1] was the most active initiative from 2008 to 2012 for Multilingual IR in Indian languages. FIRE developed datasets for Multilingual IR in six Indian Languages (Bengali, Gujarati, Hindi, Marathi, Oriya, and Tamil). However, the size of these datasets is not large enough to train neural IR systems based on transformer models like mBERT (Devlin et al., 2018) and XLM (Lample and Conneau, 2019). In addition, the text in the FIRE dataset comes from newspaper articles (Palchowdhury et al., 2013), which is domain-specific; hence, the models trained on such datasets cannot generalize well to other domains. Due to the lack of large-scale datasets, Cross-lingual knowledge transfer via Distillation has become popular for neural IR in low-resource languages (Huang et al., 2023a) (Huang et al., 2023b).

The key distinction in our work from the earlier approaches is that we introduce monolingual datasets and neural IR models in 11 major Indian Languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu), that can also benefit Cross-lingual and Multilingual IR models from the cross-lingual transfer effects when trained on a large number of Indian Languages (Zhang et al., 2022a).

3 Datasets

3.1 INDIC-MARCO

We introduce the INDIC-MARCO dataset, a multilingual version of the MS MARCO dataset. We translate the queries and passages in the MS MARCO passage ranking dataset into 11 widely spoken Indian languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu) originating from two major language families (Indo-Aryan and Dravidian). The translation process utilizes the int-8 quantized version of the NLLB-1.3B-Distilled Model (Costa-jussà et al., 2022), available at CTranslate2[2] (Klein et al., 2020). We chose int-8 quantized version of NLLB-1.3B-Distilled Model for two reasons: (a) it has shown remarkable performance in terms of BLEU scores for many Indian languages as compared to IndicBART (Dabre et al., 2021)

[1] http://fire.irsi.res.in/fire/static/data

[2] https://forum.opennmt.net/t/nllb-200-with-ctranslate2/5090

and IndicTrans (Ramesh et al., 2022) (b) Quantization (Klein et al., 2020) enables faster inference with less computing power and little or no drop in translation quality. The machine translation process employs specific hyper-parameters: a beam width of 4, a maximum decoding sequence length of 200 tokens, a batch size of 64, and a batch type equal to 'examples'. Passages from the MS MARCO dataset are split into multiple sentences using the Moses SentenceSplitter[3], ensuring that each sentence serves as a translation unit in a batch of 64 sentences. In contrast, queries with an average length of 5.96 words (Thakur et al., 2021) are not sentence-split before translation. We also translate the MS MARCO Dev-Set(Small)[4] containing 6,390 queries (1.1 qrels/query) to obtain INDIC-MARCO Dev-set(Small). The translation process on an Nvidia A100 GPU with 80 GB VRAM takes approximately 1584 hours for passages in MS MARCO, 55 hours for queries in MS MARCO, and 1.5 hours for queries in MS MARCO Dev-Set(Small). Upon translation, the resulting INDIC-MARCO dataset comprises around 8.8 million passages, 530k queries, and 39 Million training triplets in 11 Indian languages. This dataset allows for training monolingual neural IR models for each language in the INDIC-MARCO dataset.

4 Models

4.1 Baselines

BM25 (Robertson et al., 2009) serves as a strong baseline as it performs better than many neural IR models on domain-specific datasets with exceptions (Thakur et al., 2021). It does not require any training. BM25 retrieves documents containing query tokens and assigns them a score for reranking based on the frequency of query tokens appearing in them and the document length. In this work, we use the BM25 implementation provided by Pyserini[5] with values for parameters k1=0.82 and b=0.68 for evaluation on INDIC-MARCO Dev-Set obtained after machine translation. We use Whitespace Analyzers to tokenize queries and documents during indexing and searching for all Indian languages except Hindi, Bengali, and Telugu, for which we use language-specific analyzers provided in Pyserini. BM25-tuned (BM25-T) presented in Mr. Tydi (Zhang et al., 2021) is optimized to maximize the MRR@100 score on the Mr. Tydi test-set using a grid search over the range [0.1, 0.6] for k1 and [0.1, 1] for b.

Multilingual Dense Passage Retriever (mDPR) is presented in both Mr. Tydi and MIRACL by replacing the BERT encoder in Dense Passage Retriever(DPR) (Karpukhin et al., 2020) with an mBERT encoder. In Mr. Tydi, mDPR is trained on English QA dataset (Kwiatkowski et al., 2019) and used in a zero-shot manner for indexing and retrieval of documents. In MIRACL, mDPR is trained on the MS MARCO dataset and used in a zero-shot manner for indexing and retrieving documents. Multilingual ColBERT (mCol) is introduced in MIRACL by replacing the BERT encoder in ColBERT (Santhanam et al., 2021) with an mBERT encoder. mCol is trained on the MS MARCO dataset and used in a zero-shot manner for indexing and retrieval of documents.

4.2 Indic-ColBERT

Indic-ColBERT (iCol) is based on ColBERT (Khattab and Zaharia, 2020) for training and ColBERTv2 (Santhanam et al., 2021) for compression and inference. There are some distinctions: it uses mBERT as query-document encoder, and is trained on INDIC-MARCO. Model architecture comprises (a) a query encoder, (b) a document encoder, and (c) max-sim function (same as ColBERTv2). Given a query with q tokens and a document with d tokens, the Query encoder outputs q fix-sized token embeddings, and the document encoder outputs d fix-sized token embeddings. The maximum input sequence length for the query, q_{max}, and, for the document, d_{max}, is set before giving them to the respective encoders. If q is less than q_{max}, we append $q_{max} - q$ [MASK] tokens to the input query, and if q is greater than q_{max}, q is truncated to q_{max}. If d is less than d_{max}, then d is neither truncated nor padded. If d is greater than d_{max}, d is truncated to d_{max}. The max-sim function is used to obtain the relevance score of a document for a query using the encoded representations.

5 Experiment Setup

We train 11 distinct Indic-ColBERT (iCol) models separately for 50k iterations with a batch size of 128 on the first 6.4 million training triplets from the INDIC-MARCO dataset to optimize the pairwise softmax cross entropy loss function, where each

[3]https://pypi.org/project/mosestokenizer/
[4]https://ir-datasets.com/MS MARCO-passage.html
[5]https://github.com/castorini/pyserini

Language	MRR@10			Recall@1000		
	BM25	mCol	**iCol**	BM25	mCol	**iCol**
Assamese	0.078	0.095	**0.176**	0.449	0.503	**0.698**
Bengali	0.112	0.159	**0.221**	0.622	0.691	**0.788**
Gujarati	0.100	0.141	**0.232**	0.539	0.653	**0.805**
Hindi	0.125	0.171	**0.223**	0.678	0.729	**0.772**
Kannada	0.089	0.156	**0.219**	0.520	0.691	**0.787**
Malayalam	0.076	0.124	**0.198**	0.442	0.603	**0.742**
Marathi	0.085	0.143	**0.207**	0.476	0.655	**0.750**
Oriya	**0.086**	0.002	0.002	**0.484**	0.022	0.016
Punjabi	0.113	0.134	**0.211**	0.603	0.637	**0.766**
Tamil	0.088	0.144	**0.202**	0.495	0.661	**0.756**
Telugu	0.1007	0.144	**0.206**	0.569	0.648	**0.749**

Table 1: Results on INDIC-MARCO Dev-Set(Small). mColBERT (mCol) is trained on MS MARCO dataset (Nguyen et al., 2016). Indic-ColBERT are 11 distinct monolingual neural IR models trained on INDIC-MARCO.

Language	Mr. Tydi test-set					MIRACL Dev-set			
	BM25	BM25-T	mDPR	mCol	**iCol**	BM25	mDPR	mCol	**iCol**
Bengali	0.418	0.413	0.258	0.414	**0.501**	0.508	0.443	0.546	**0.606**
Hindi	-	-	-	-	-	0.458	0.383	0.470	**0.483**
Telugu	0.343	**0.424**	0.106	0.314	0.393	**0.494**	0.356	0.462	0.479

Table 2: Results on Mr. Tydi test-set (MRR@100) and MIRACL Dev-set (NDCG@10): For Mr. Tydi test-set, we use official BM25, BM25-tuned (BM25-T) and mDPR model scores (Zhang et al., 2021); mCol (mColBERT trained on MS MARCO), and iCol (Indic-ColBERT trained on INDIC-MARCO) are tested in a zero-shot manner. For the MIRACL dev-set, we use official BM25, mDPR, and mCol(mColBERT) model scores (Zhang et al., 2022b); iCol (Indic-ColBERT trained on INDIC-MARCO) is tested in a zero-shot manner.

triplet contains a query, a relevant passage and an irrelevant passage in one of the 11 languages on which the model is trained. The mBERT encoder is finetuned from the official "bert-base-multilingual-uncased" checkpoint, and the remaining parameters are trained from scratch.

6 Results

Indic-ColBERT (iCol) outperforms baseline models (BM25, BM25-T, mDPR, mCol) by 20%, in MRR@100 Score and on Mr. Tydi test-set (Refer Table 2) for Bengali Language. For Telugu, Indic-ColBERT (iCol) outperforms 3 (BM25, mDPR, mCol) out of 4 baselines in terms of MRR@100 scores. Indic-ColBERT (iCol) outperforms baseline models (BM25, mDPR, mCol) by 19.29% in Bengali and 5.4% in Hindi, in NDCG@10 Score on MIRACL dev-set(Refer Table 2). For Telugu, Indic-ColBERT (iCol) outperforms 2 (mDPR, mCol) out of 3 baselines in terms of NDCG@10 scores. Indic-ColBERT (iCol) outperforms baseline models (BM25, mCol) by 47.47% in MRR@10 Score on INDIC-MARCO Dev-Set(Small) (Refer Table 1) averaged over all 11 Indian languages (excluding Oriya).

We do not see any improvements for Oriya because mBERT used in Indic-ColBERT is not pretrained on Oriya and Assamese. Assamese demonstrates a 125% MRR@10 improvement over the BM25 baseline, attributed to its linguistic similarity with Bengali (indicated by the mColBERT model outperforming BM25 by 21% in MRR@10 Score) and the high-quality data in INDIC-MARCO, further enhancing the MRR@10 score by 104%, making INDIC-MARCO a significant contributor to the advancement for a low-resource language like Assamese which mBERT does not support.

7 Ablation Study

In this section, we perform ablation study with three different machine translation models and two different document splitting schemes. We compare the NDCG@10 scores of Indic-ColBERT models trained on machine translated MS-MARCO data using NLLB-600M, NLLB-1.3B and IndicTrans2. As shown in Table 4, the impact of translation quality

Language	Mr. Tydi test-set					MIRACL Dev-set			
	BM25	BM25-T	mDPR	mCol	**iCol**	BM25	mDPR	mCol	**iCol**
Bengali	0.869	**0.874**	0.671	0.846	0.864	0.909	0.819	**0.913**	0.894
Hindi	-	-	-	-	-	0.868	0.776	**0.884**	0.811
Telugu	0.758	**0.813**	0.352	0.589	0.688	**0.831**	0.762	0.830	0.768

Table 3: Results on Mr. Tydi test-set (Recall@100) and MIRACL Dev-set (Recal@100): For Mr. Tydi test-set, we use official BM25, BM25-tuned (BM25-T) and mDPR model scores (Zhang et al., 2021); mCol (mColBERT trained on MS MARCO), and iCol (Indic-ColBERT trained on INDIC-MARCO) are tested in a zero-shot manner. For the MIRACL dev-set, we use official BM25, mDPR, and mCol(mColBERT) model scores (Zhang et al., 2022b); iCol (Indic-ColBERT trained on INDIC-MARCO) is tested in a zero-shot manner.

Language	Translation Model + Splitting Scheme			
	NLLB-600M	NLLB-1.3B		IndicTrans2
	Moses	Moses	Full-Stop	Moses
Bengali	0.592	0.606	**0.614**	0.602
Hindi	0.464	0.483	0.493	**0.497**
Telugu	**0.523**	0.479	0.475	0.469

Table 4: Results on MIRACL Dev-Set(NDCG@10).

on retrieval effectiveness follows a different trend for each language. In terms of chrF++ score, IndicTrans2 performs better than NLLB-1.3B which performs better than NLLB-600M on Flores-200 devtest (Gala et al., 2023) (Costa-jussà et al., 2022). For Telugu, we observe a negative correlation between translation quality and retrieval effectiveness, where the Indic-Colbert trained on data translated using NLLB-600M model, which has the lowest chrF++ score among the three machine translation models, gives the best retrieval effectiveness. For Hindi, we observe a positive correlation between the translation quality and retrieval effectiveness. For Bengali, we don't observe any correlation between translation quality and retrieval effectiveness.

Each document in MS-MARCO dataset is first split into sentences, each sentence is translated by the machine translation model and finally the translated sentences are merged back into the document. We experimented with two different document splitting schemes. We compare the NDCG@10 scores for Indic-ColBERT models trained on machine translated MS-MARCO dataset using NLLB-1.3B model on sentences obtained from Moses Splitting and Full-stop Splitting schemes. As shown in Table 4, we can observe "NLLB-1.3B + Full-Stop Splitting" outperforms "NLLB-1.3B + Moses Splitting" for Hindi and Bengali Languages.

8 Summary, conclusion, and future work

We present IndicIRSuite, featuring INDIC-MARCO, a multilingual neural IR dataset in 11 Indian languages, and Indic-ColBERT, comprising 11 monolingual neural IR models based on ColBERTv2. Our results demonstrate performance enhancements over baselines in Mr. Tydi, MIRACL, and INDIC-MARCO, particularly benefiting low-resource languages like Assamese. INDIC-MARCO proves valuable for such languages, not supported by models like mBERT but linguistically akin to Bengali. We also perform an ablation to find the impact of translation quality and sentence splitting on retrieval effectiveness. Future work includes expanding IndicIRSuite to Multilingual and Crosslingual IR.

Limitations

The primary limitation of our study is the absence of a comprehensive comparison of the trained IR models across out-of-domain datasets beyond MIRACL and Mr. Tydi. It is imperative to delve deeper into the translation quality, specifically assessing whether it exhibits pronounced "translationese." A more exhaustive examination is warranted, particularly in cases where the proposed models, such as Indic-ColBERT, demonstrate subpar performance compared to baseline models, as observed in the instance where Indic-ColBERT lags behind the BM25 Baseline for the Telugu Language in Mr. Tydi test-set and MIRACL Dev-set.

Ethics Statement

We want to emphasize our commitment to upholding ethical practices throughout this work. This work publishes a large-scale machine-translated dataset for neural information retrieval in 11 Indian languages - Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu. MS MARCO passage ranking Dataset in the English language used as a Source dataset for translation is publicly available, and no annotators were employed for data collection. We have cited the datasets and relevant works used in this study.

References

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023a. Cross-lingual knowledge transfer via distillation for multilingual information retrieval.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023b. 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*. ACM.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

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.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The opennmt neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109.

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. *arXiv preprint arXiv:1901.07291*.

Dawn Lawrie, Eugene Yang, Douglas W Oard, and James Mayfield. 2023. Neural approaches to multilingual information retrieval. In *European Conference on Information Retrieval*, pages 521–536. Springer.

Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, Nandan Thakur, Jheng-Hong Yang, et al. 2023. Simple yet effective neural ranking and reranking baselines for cross-lingual information retrieval. *arXiv preprint arXiv:2304.01019*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Sauparna Palchowdhury, Prasenjit Majumder, Dipasree Pal, Ayan Bandyopadhyay, and Mandar Mitra. 2013. Overview of fire 2011. In *Multilingual Information Access in South Asian Languages: Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*, pages 1–12. Springer.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

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

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Shuo Sun and Kevin Duh. 2020. Clirmatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. *arXiv preprint arXiv:2108.08787*.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022a. Towards best practices for training multilingual dense retrieval models.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022b. Making a miracl: Multilingual information retrieval across a continuum of languages. *arXiv preprint arXiv:2210.09984*.

A Examples

Snapshots from the INDIC-MARCO dataset are shown in Figure 1, Figure 2 and Figure 3.

Language	INDIC-MARCO Dataset
Assamese	বিজ্ঞানীসকলৰ মাজত যোগাযোগৰ উপস্থিতি মানহাননেটনৰ প্ৰকল্পৰ সফলতাৰ বাবে সমানেই গুৰুত্বপূৰ্ণ আছিল যেনে বৈজ্ঞানিক বুদ্ধিমত্তাৰ বাবে।.পৰমাণু বিজ্ঞানী আৰু অভিযন্তাসকলৰ আশ্চৰ্যজনক সাফল্যৰ ওপৰত কেৱল মাত্ৰ এটা মেঘহৈহে ৰৈ আছে যে তেওঁলোকৰ সফলতাৰ অৰ্থ কি আছিল; শত সহস্ৰ নিৰীহ লোকৰ জীৱন ধ্বংস কৰা হৈছিল.
Bengali	বিজ্ঞানী মনুষ্যাদের মধ্যে যোগাযোগের উপস্থিতি মানহাটেন প্রকল্পের সাফল্যের জন্য সমান গুরুত্বপূর্ণ ছিল যেমন বৈজ্ঞানিক বুদ্ধিমত্তা ছিল.পরমাণু গবেষক ও প্রকৌশলীদের আশ্চর্যজনক সাফল্যের উপর নির্ভর করছে একমাত্র মেঘ, তাদের সাফল্যের অর্থ কী ছিল; শত শত হাজার নির্দোষ মানুষের জীবন ধ্বংস হয়ে গেছে.
Gujarati	વૈજ્ઞાનિક મનમાં સંચારની હાજરી મેનહટન પ્રોજેક્ટની સફળતા માટે એટલી જ મહત્વપૂર્ણ હતી જેટલી વૈજ્ઞાનિક બુદ્ધિ હતી.અણુ સંશોધકો અને ઇજનેરોની પ્રભાવશાળી સિદ્ધિઓ પર માત્ર એક જ વાદળ છે કે તેમની સફળતાનો ખરેખર અર્થ શું હતો; સેંકડો હજારો નિર્દોષ જીવન ભૂંસી નાખવામાં આવ્યા હતા.
Kannada	ವೈಜ್ಞಾನಿಕ ಮನಸ್ಸುಗಳ ನಡುವೆ ಸಂವಹನವು ಮ್ಯಾನ್ಹ್ಯಾಟನ್ ಯೋಜನೆಯ ಯಶಸ್ಸಿಗೆ ಅಷ್ಟೇ ಮುಖ್ಯವಾಗಿತ್ತು..ಪರಮಾಣು ಸಂಶೋಧಕರು ಮತ್ತು ಎಂಜಿನಿಯರ್ಗಳ ಪ್ರಭಾವಶಾಲಿ ಸಾಧನೆಗಳ ಮೇಲೆ ಅಂಟಿಕೊಂಡಿರುವ ಏಕೈಕ ಮೋಡವೆಂದರೆ ಅವರ ಯಶಸ್ಸಿನ ಅರ್ಥ; ನೂರಾರು ಸಾವಿರ ಮುಗ್ಧ ಜೀವಗಳನ್ನು ಅಳಿಸಿಹಾಕಲಾಗಿದೆ.
Malayalam	ശാസ്ത്രീയ മനസ്സിനിടയില് ആശയവിനിമയത്തിന്റെ സാന്നിധ്യം മാൻഹാന് ട് പ്രോജക്റ്റിന്റെ വിജയത്തിന് സയന്റിഫിക് ബുദ്ധി പോലെ തന്നെ പ്രധാനപ്പെട്ടതായിരുന്നു.ആണവ ഗവേഷകരുടെയും എൻജിനീയർമാരുടെയും ശ്രദ്ധേയമായ നേട്ടങ്ങൾക്കു മേൽ തൂങ്ങുന്ന ഒരേയൊരു മേഘം അവരുടെ വിജയം ശരിക്കും അർത്ഥമാക്കുന്നത്; നൂറുകണക്കിന് നിരപരാധികളായ ജീവൻ മായ്ച്ചത്.
Marathi	वैज्ञानिक मनामध्ये संवाद असणे हे मान्हॅटन प्रकल्पाच्या यशासाठी तितकेच महत्त्वाचे होते, जितके वैज्ञानिक बुद्धिमत्ता होते..अणु संशोधक आणि अभियंत्यांच्या प्रभावी कामगिरीवर लटकलेला एकमेव ढग म्हणजे त्यांचा यशस्वी होण्याचा अर्थ; शेकडो हजार निर्दोष जीव नष्ट.
Oriva	ବୈଜ୍ଞାନିକମାନଙ୍କ ମଧ୍ୟରେ ଯୋଗାଯୋଗର ଉପସ୍ଥିତି ମଧ୍ୟ Manhattan ପ୍ରକଳ୍ପର ସଫଳତା ପାଇଁ ସମାନ ଭାବରେ ଗୁରୁତ୍ୱପୂର୍ଣ୍ଣ ଥିଲା ।.ପରମାଣୁ ଗବେଷକ ଓ ଇଞ୍ଜିନିୟରଙ୍କ ସଫଳତା ଉପରେ କେବଳ ଗୋଟିଏ ମେଘ ରହିଛି, ଯାହା ସେମାନଙ୍କ ସଫଳତାକୁ ପ୍ରକୃତରେ ବୁଝାଇଥାଏ । ଶହ ଶହ ନିରୀହ ପ୍ରାଣ ହରାଇଲା.- ନୁଁ ଜାଣେ ।
Punjabi	ਵਿਗਿਆਨਕ ਦਿਮਾਗਾਂ ਵਿਚ ਸੰਚਾਰ ਦੀ ਮੌਜੂਦਗੀ ਮੈਨਹੱਟਨ ਪ੍ਰੋਜੈਕਟ ਦੀ ਸਫਲਤਾ ਲਈ ਉਨੀ ਹੀ ਮਹੱਤਵਪੂਰਨ ਸੀ ਜਿੰਨੀ ਵਿਗਿਆਨਕ ਬੁੱਧੀ.ਪ੍ਰਮਾਣੂ ਖੋਜਕਰਤਾਵਾਂ ਅਤੇ ਇੰਜੀਨੀਅਰਾਂ ਦੀ ਪ੍ਰਭਾਵਸ਼ਾਲੀ ਪ੍ਰਾਪਤੀ ਉੱਤੇ ਸਿਰਫ ਇੱਕ ਹੀ ਬੱਦਲ ਲਟਕਦਾ ਹੈ ਕਿ ਉਨ੍ਹਾਂ ਦੀ ਸਫਲਤਾ ਦਾ ਅਸਲ ਅਰਥ ਕੀ ਸੀ; ਸੈਂਕੜੇ ਹਜ਼ਾਰਾਂ ਨਿਰਦੋਸ਼ ਜਾਨਾਂ ਮਿਟਾ ਦਿੱਤੀਆਂ ਗਈਆਂ.
Tamil	அறிவியல் மனதில் தகவல்தொடர்பு இருப்பது மன்ஹாட்டன் திட்டத்தின் வெற்றிக்கு சமமாக முக்கியமானது, அறிவியல் அறிவு.அணுசக்தி ஆராய்ச்சியாளர்கள் மற்றும் பொறியாளர்களின் அற்புதமான சாதனைகள் மீது மட்டுமே மேகம் தொங்குகிறது, அவர்களின் வெற்றி உண்மையில் என்ன அர்த்தம்; நூற்றுக்கணக்கான ஆயிரக்கணக்கான அப்பாவி உயிர்கள் அழிக்கப்பட்டன.
Telugu	శాస్త్రీయ మనస్సులలో కమ్యూనికేషన్ ఉనికి మాన్హాటన్ ప్రాజెక్టు విజయానికి శాస్త్రీయ మేధస్సు వలె ముఖ్యమైనది.అణు పరిశోధకులు, ఇంజనీర్లు సాధించిన అద్భుతమైన విజయాలు వాటి విజయానికి మాత్రమే ప్రతీకగా ఉన్నాయి. వందల వేల మంది అమాయక ప్రాణులు నాశనం అయ్యాయి.

Figure 1: INDIC-MARCO translations for the MS-MARCO document "The presence of communication amid scientific minds was equally important to the success of the Manhattan Project as scientific intellect was. The only cloud hanging over the impressive achievement of the atomic researchers and engineers is what their success truly meant; hundreds of thousands of innocent lives obliterated"

Language	INDIC-MARCO Dataset
Assamese	মানহাটেন প্ৰজেক্টৰ প্ৰজেক্টৰ লক্ষ্য আছিল পৰমাণু বোমা সম্ভৱ হৈ উঠিবনে নাই সেয়া পৰীক্ষা কৰা.এই প্ৰকল্পৰ সফলতাই পৃথিৱীখনক চিৰকালৰ বাবে সলনি কৰিব। এই প্ৰকল্পই এই ক্ষমতাশালী বস্তুটো মানুহৰ দ্বাৰা সৃষ্টি কৰিব পৰা যাব।.
Bengali	মানহাটেন প্রকল্পের উপর লেখা - মানহাটেন প্রকল্প মানহাটেন প্রকল্পটি ছিল পরমাণু বোমা তৈরি করা সম্ভব কিনা তা পরীক্ষা করা.এই প্রকল্পের সাফল্য চিরকালের জন্য পৃথিবীকে চিরকালের জন্য পরিবর্তন করবে, এটা জানানো হবে যে এত শক্তিশালী কিছু মানুষের তৈরি হতে পারে।.
Gujarati	મેનહટન પ્રોજેક્ટ પરનો પ્રયોગ - મેનહટન પ્રોજેક્ટ મેનહટન પ્રોજેક્ટ એ જોવાનું હતું કે શું અણુ બોમ્બ શક્ય બનાવવી.આ પ્રોજેક્ટની સફળતા વિશ્વને કાયમ માટે બદલી નાખશે. તે જાણીને કે આ શક્તિશાળી કંઈક માનવસર્જિત થઈ શકે છે..
Kannada	ಮ್ಯಾನ್ಹ್ಯಾಟನ್ ಯೋಜನೆ - ಮ್ಯಾನ್ಹ್ಯಾಟನ್ ಯೋಜನೆ ಪರಮಾಣು ಬಾಂಬ್ ಸಾಧ್ಯವೇ ಎಂಬುದನ್ನು ನೋಡಲು ಮ್ಯಾನ್ಹ್ಯಾಟನ್ ಯೋಜನೆ.ಈ ಯೋಜನೆಯ ಯಶಸ್ಸು ಶಾಶ್ವತವಾಗಿ ಜಗತ್ತನ್ನು ಶಾಶ್ವತವಾಗಿ ಬದಲಾಯಿಸುತ್ತದೆ. ಈ ಶಕ್ತಿಶಾಲಿ ವಸ್ತುಗಳನ್ನು ಮನುಷ್ಯನಿಂದ ಮಾಡಬಹುದೆಂದು ತಿಳಿಯುವಂತೆ ಮಾಡುತ್ತದೆ..
Malayalam	മാൻഹട്ടൻ പദ്ധതി - മാൻഹട്ടൻ പദ്ധതി ആറ്റോമിക് ബോംബ് സാധ്യമാക്കുമോ എന്ന് പരിശോധിക്കുക.ഈ പദ്ധതിയുടെ വിജയം ലോകത്തെ എന്നെന്നേക്കുമായി മാറ്റിമറിക്കും. മനുഷ്യന് സൃഷ്ടിച്ചതാകാം..
Marathi	मॅनहॅटन प्रकल्प - मॅनहॅटन प्रकल्प मॅनहॅटन प्रकल्प हा अणुबॉम्ब शक्य आहे का हे पाहण्यासाठी होता.या प्रकल्पाचा यशस्वी होणे हे जग कायमचे बदलू शकते. हे सर्व लोकांना कळेल की, इतक्या शक्तिशाली गोष्टी माणसाच्या हातून बनवता येतात..
Oriva	Manhattan Project ଉପରେ ପ୍ରବନ୍ଧ - Manhattan Project Manhattan Project ଏକ ପରମାଣୁ ବୋମା ସମ୍ଭବ କି ନୁହେଁ ତାହା ଦେଖିବା ପାଇଁ Manhattan Project.ଏହି ପ୍ରକଳ୍ପର ସଫଳତା ଦୁନିଆକୁ ସବୁଦିନ ପାଇଁ ବଦଳାଇଦେବ । ଏହା ପ୍ରାୟ ଏହା ଜଣା ପଡିବ ଯେ ମଣିଷ ତିଆରି କରିପାରେ ।.- ନୁଁ ଜାଣେ ।
Punjabi	ਮੈਨਹੱਟਨ ਪ੍ਰੋਜੈਕਟ'ਤੇ ਲੇਖ - ਮੈਨਹੱਟਨ ਪ੍ਰੋਜੈਕਟ ਮੈਨਹੱਟਨ ਪ੍ਰੋਜੈਕਟ ਇਹ ਦੇਖਣਾ ਸੀ ਕਿ ਕੀ ਪ੍ਰਮਾਣੂ ਬੰਬ ਸੰਭਵ ਬਣਾਉਣਾ ਹੈ.ਇਸ ਪ੍ਰਾਜੈਕਟ ਦੀ ਸਫਲਤਾ ਦੁਨੀਆਂ ਨੂੰ ਹਮੇਸ਼ਾ ਲਈ ਬਦਲ ਦੇਵੇਗੀ। ਇਹ ਜਾਣ ਕੇ ਕਿ ਅਜਿਹਾ ਕੁਝ ਸ਼ਕਤੀਸ਼ਾਲੀ ਇਨਸਾਨ ਦੁਆਰਾ ਬਣਾਇਆ ਜਾ ਸਕਦਾ ਹੈ।.
Tamil	மன்ஹாட்டன் திட்டம் - மன்ஹாட்டன் திட்டம் அணு குண்டு சாத்தியமா என்பதைப் பார்க்கும் நோக்கம்.இந்த திட்டத்தின் வெற்றி உலகை மாற்றிவிடும். மனிதனால் உருவாக்கக்கூடிய சக்தி வாய்ந்த ஒன்று..
Telugu	మాన్హాటన్ ప్రాజెక్టు - మాన్హాటన్ ప్రాజెక్టు అణు బాంబు సాధ్యమా అని చూడాలని.ఈ ప్రాజెక్టు విజయవంతం కావడం వల్ల ప్రపంచం ఎప్పటికీ మారుతుంది. ఈ శక్తివంతమైన వస్తువు మానవుడు తయారు చేయగలదని తెలియజేస్తుంది..

Figure 2: INDIC-MARCO translations for the MS-MARCO document "The Manhattan Project and its atomic bomb helped bring an end to World War II. Its legacy of peaceful uses of atomic energy continues to have an impact on history and science."

Language	INDIC-MARCO Dataset
Assamese	মানহাটেন প্ৰজেক্ট আৰু ইয়াৰ পাৰমাণৰিক বোমা দ্বিতীয় বিশ্বযুদ্ধৰ অন্ত পেলাবলৈ সহায় কৰিছিল.পাৰমাণৰিক শক্তিৰ শান্তিপূৰ্ণ ব্যৱহাৰৰ ঐতিহাই ইতিহাস আৰু বিজ্ঞানত প্ৰভাৱ পেলাই আহিছে.
Bengali	মানহাটেন প্রকল্প এবং এর পরমাণু বোমা দ্বিতীয় বিশ্বযুদ্ধের সমাপ্তিতে সাহায্য করেছিল.পারমাণবিক শক্তির শান্তিপূর্ণ ব্যবহারের ঐতিহ্য ইতিহাস ও বিজ্ঞানকে প্রভাবিত করে চলেছে.
Gujarati	મેનહટન પ્રોજેક્ટ અને તેના પરમાણુ બોમ્બથી બીજા વિશ્વયુદ્ધનો અંત આવ્યો.અણુ ઊર્જાના શાંતિપૂર્ણ ઉપયોગની તેની વારસો ઇતિહાસ અને વિજ્ઞાન પર અસર કરતી રહે છે.
Kannada	ಮ್ಯಾನ್ಹ್ಯಾಟನ್ ಯೋಜನೆ ಮತ್ತು ಅದರ ಪರಮಾಣು ಬಾಂಬ್ ವಿಶ್ವ ಸಮರ II ರ ಅಂತ್ಯಕ್ಕೆ ನೆರವಾಯಿತು.ಪರಮಾಣು ಶಕ್ತಿಯ ಶಾಂತಿಯುತ ಬಳಕೆಗೆ ಅದರ ಪರಂಪರೆ ಇತಿಹಾಸ ಮತ್ತು ವಿಜ್ಞಾನದ ಮೇಲೆ ಪ್ರಭಾವ ಬೀರುತ್ತಿದೆ.
Malayalam	മാൻഹട്ടൻ പദ്ധതിയും ആറ്റോമിക് ബോംബും രണ്ടാം ലോകമഹായുദ്ധത്തിന് അന്ത്യം കുറിക്കാൻ സഹായിച്ചു.ആണവോർജ്ജത്തിന്റെ സമാധാനപരമായ ഉപയോഗം ചരിത്രത്തിലും ശാസ്ത്രത്തിലും സ്വാധീനം ചെലുത്തുന്നത് തുടരുന്നു.
Marathi	मॅनहॅटन प्रकल्प आणि त्याच्या अणुबॉम्बने दुसऱया महायुद्धाचा अंत केला.अणुऊर्जेच्या शांततापूर्ण वापराचा वारसा इतिहास आणि विज्ञानावर प्रभाव पाडत आहे.
Oriya	ମାନ୍ହାଟନ୍ ପ୍ରକଳ୍ପ ଏବଂ ଏହାର ପରମାଣୁ ବୋମା ଦ୍ୱିତୀୟ ବିଶ୍ୱଯୁଦ୍ଧର ଅନ୍ତ ଘଟାଇଥିଲା.ପରମାଣୁ ଶକ୍ତିର ଶାନ୍ତିପୂର୍ଣ୍ଣ ବ୍ୟବହାରର ପରମ୍ପରା ଇତିହାସ ଓ ବିଜ୍ଞାନ ଉପରେ ପ୍ରଭାବ ପକାଇବ.- ମୁଁ ଜାଣେ ।
Punjabi	ਮੈਨਹੱਟਨ ਪ੍ਰੋਜੈਕਟ ਅਤੇ ਇਸ ਦੇ ਪ੍ਰਮਾਣੂ ਬੰਬ ਨੇ ਦੂਜੇ ਵਿਸ਼ਵ ਯੁੱਧ ਦਾ ਅੰਤ ਕਰਨ ਵਿੱਚ ਮਦਦ ਕੀਤੀ.ਪ੍ਰਮਾਣੂ ਊਰਜਾ ਦੀ ਸ਼ਾਂਤੀਪੂਰਨ ਵਰਤੋਂ ਦੀ ਇਸ ਦੀ ਵਿਰਾਸਤ ਦਾ ਇਤਿਹਾਸ ਅਤੇ ਵਿਗਿਆਨ ਉੱਤੇ ਅਸਰ ਜਾਰੀ ਹੈ.
Tamil	மன்ஹாட்டன் திட்டமும் அதன் அணு குண்டுகளும் இரண்டாம் உலகப் போருக்கு முடிவை ஏற்படுத்த உதவியது.அணுசக்தி அமைதியான முறையில் பயன்படுத்தப்படுவது வரலாறு மற்றும் அறிவியலில் தொடர்ந்து தாக்கத்தை ஏற்படுத்துகிறது.
Telugu	మ్యాన్హాటన్ ప్రాజెక్ట్ మరియు దాని అణు బాంబు రెండవ ప్రపంచ యుద్ధం ముగియడానికి సహాయపడ్డాయి.అణు శక్తి యొక్క శాంతియుత వినియోగం యొక్క దాని వారసత్వం చరిత్ర మరియు శాస్త్రంపై ప్రభావం చూపుతూనే ఉంది.

Figure 3: INDIC-MARCO translations for the MS-MARCO document "Essay on The Manhattan Project - The Manhattan Project The Manhattan Project was to see if making an atomic bomb possible. The success of this project would forever change the world forever making it known that something this powerful can be manmade."