Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin

David R. Cheriton School of Computer Science
University of Waterloo

Abstract

We present Mr. TYDI, a multi-lingual benchmark dataset for mono-lingual retrieval in eleven typologically diverse languages, designed to evaluate ranking with learned dense representations. The goal of this resource is to spur research in dense retrieval techniques in non-English languages, motivated by recent observations that existing techniques for representation learning perform poorly when applied to out-of-distribution data. As a starting point, we provide zero-shot baselines for this new dataset based on a multi-lingual adaptation of DPR that we call "mDPR". Experiments show that although the effectiveness of mDPR is much lower than BM25, dense representations nevertheless appear to provide valuable relevance signals, improving BM25 results in sparse–dense hybrids. In addition to analyses of our results, we also discuss future challenges and present a research agenda in multi-lingual dense retrieval. Mr. TYDI can be downloaded at `https://github.com/castorini/mr.tydi`.

1 Introduction

Retrieval approaches based on learned dense representations, typically derived from transformers, form an exciting new research direction that has received much attention of late. These dense retrieval techniques generally adopt a supervised approach to representation learning, where a labeled dataset is used to train two encoders—one for the queries and the other for texts from the corpus to be retrieved—whose output representation vectors are then compared with a simple comparison function such as inner product. Retrieval against a large text corpus is typically formulated as nearest neighbor search and efficiently executed using off-the-shelf libraries. In the literature, this is known as a "bi-encoder" design. Well-known examples include DPR (Karpukhin et al., 2020), ANCE (Xiong et al., 2021), and ColBERT (Khattab and Zaharia, 2020), but there is much recent work along these lines (Gao et al., 2020; Hofstätter et al., 2020; Hofstätter et al., 2021; Lin et al., 2021b), just to list a few papers.

Like all methods based on supervised machine learning, the effectiveness of the trained models on "out of distribution" (OOD) samples is an important issue since it concerns model robustness and generalizability. For dense retrieval, training data typically comprise (query, relevant passage) pairs, and in this context, OOD could mean that (1) the passage encoder is fed text from a different domain, genre, register, etc. than the training data, (2) the query encoder is fed queries that are different from the training queries, (3) the relationship between the inputs at inference time is different from the training samples (e.g., task variations), or (4) a combination of all of the above.

It is, in fact, already known that dense retrieval techniques generalize poorly across different corpora, queries, tasks, etc. Recently, Thakur et al. (2021) constructed a benchmark to specifically evaluate the zero-shot transfer capabilities of dense retrieval models by creating a framework that unifies over a dozen retrieval datasets spanning diverse domains. In a zero-shot setting, the authors found that BM25 remained the most effective overall. That is, dense retrieval techniques trained on one dataset can spectacularly fail on another dataset—exactly the out-of-distribution challenges we discussed above. In contrast, BM25, "just works" regardless of the corpus and queries, even though on "in distribution" samples, dense retrieval models are unequivocally more effective. Thus, learned dense representations are not as general and robust as BM25 representations.

This paper focuses on another aspect of the generalizability and robustness of learned representations for ranking: What if the encoders are applied to texts in languages different from the one they are trained in? Our focus is on *mono-*

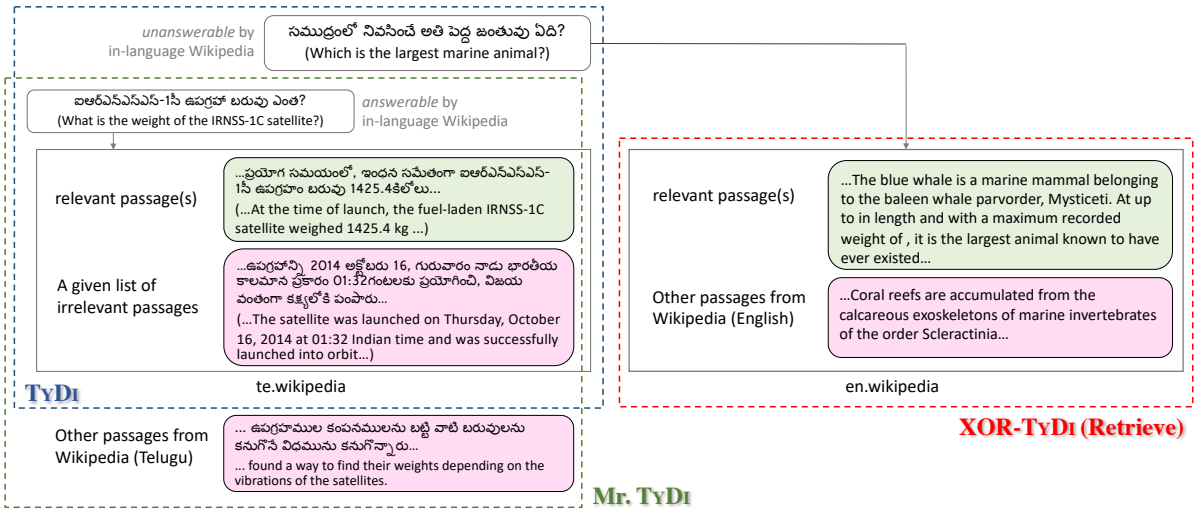

Figure 1: Comparison between TyDi, XOR-TyDi, and Mr. TyDi with an example in Telugu. The green blocks indicate relevant passages and the red blocks indicate non-relevant passages.

lingual retrieval in non-English languages (e.g., Bengali queries against Bengali documents) rather than *cross-lingual* retrieval, where documents and queries are in different languages (e.g., English queries against Arabic documents).

We view this work as having three main contributions: First, we construct and share Mr. TyDi, a multi-lingual benchmark dataset for mono-lingual retrieval in eleven diverse languages, designed to evaluate ranking with learned dense representations. This dataset can be viewed as the "open-retrieval" condition of the TyDi multi-lingual question answering (QA) dataset (Clark et al., 2020), and "Mr" in Mr. TyDi stands for "multi-lingual retrieval". We describe the construction of this dataset and how it is different from existing resources. Second, we report zero-shot baselines for Mr. TyDi, including a dense retrieval method based on a multi-lingual version of DPR (Karpukhin et al., 2020) that we call "mDPR". Third, we present a number of initial findings about baseline results that highlight future challenges and begin to define a research agenda in multi-lingual dense retrieval. Most interestingly, we find that although the zero-shot effectiveness of mDPR is much worse than BM25, dense representations appear to provide valuable relevance signals, improving BM25 results in sparse–dense hybrids.

2 Background and Related Work

In presenting a new benchmark dataset, one important question to answer is: Why is a new resource needed? We begin by addressing this question. The introduction already lays out the intellectual motivation for our work. Thus, we focus here on explaining why existing datasets are not sufficient. The answer is summarized in Figure 1.

Mr. TyDi is constructed from TyDi (Clark et al., 2020), a question answering dataset covering eleven typologically diverse languages. For each language, the creators provided annotators with a prompt (the first 100 characters of a Wikipedia article), who were asked to write a question that cannot answered by the snippet. Then, for each question, annotators were given the top Wikipedia article returned by Google search and asked to label the relevance of each passage in the article as well as to identify a minimal answer span (if possible). Given this procedure, the answer to the question may or may not be found in the passages from the selected article. The answer passages are always in the same language as the questions. Note that the questions in different languages are not comparable as they are created independently rather than through translation.

The weakness of TyDi from our perspective is that it is essentially a machine reading comprehension task like SQuAD (Rajpurkar et al., 2016) because candidate passages are all included as part of the dataset (i.e., the passages are from the top Wikipedia article returned by Google search). Instead, we need a resource akin to what QA researchers call the "open-domain" or "open-retrieval" task, where the problem involves

retrieval from a much larger corpus (e.g., all of Wikipedia) (Chen et al., 2017). Thus, at a high level, Mr. TyDi can be viewed as an open-retrieval extension to TyDi.

Asai et al. (2021) created XOR-TyDi, a cross-lingual QA dataset built on TyDi by annotating answers in English Wikipedia for questions TyDi considered unanswerable in the original source (non-English) language. This was accomplished by randomly sampling 5,000 unanswerable (non-English) questions from TyDi, and then searching English Wikipedia articles for answers. Specifically, each non-English question was first translated into English; then, annotators were given the top-ranked English Wikipedia articles and asked to label passages containing the answer.

The XOR-TyDi dataset contains three overlapping tasks, but all of them are focused on the cross-lingual condition. Among the three tasks (XOR-EnglishSpan, XOR-Retrieve, and XOR-Full), XOR-Retrieve is most comparable to our work, but the retrieval target for the task is explicitly English Wikipedia articles rather than Wikipedia articles in the question's language. While the XOR-Full task requires systems to select answer spans from both English and target-language Wikipedia articles, the dataset does not provide ground truth for the intermediate retrieval step, thus it cannot be used for our evaluations.

The annotations of XOR-TyDi do not allow us to examine mono-lingual retrieval in non-English languages because the creators started with "unanswerable" non-English TyDi questions. Furthermore, since all answer passage annotations were performed on English Wikipedia, this doesn't help if we are interested in, for example, searching Finnish Wikipedia with Finnish questions.

Another point of comparison worth discussing is MKQA (Longpre et al., 2020), which comprises 10k question–answer pairs aligned across 26 typologically diverse languages. Questions are paired with exact answers in the different languages, and evaluation is conducted in the open-retrieval setting by matching those answers in retrieved text. There are two main differences between our work and MKQA: First, MKQA was created via translation in order to achieve cross-lingual alignment. In addition to the possible translation artifacts that such a process might introduce, which Clark et al. (2020) discussed at length, we argue that forced alignment creates non-natural questions, for the simple reason that speakers of different languages are likely to be interested in different topics. This is different from the "geographically dependent" questions that MKQA tries to avoid. Take the question "*who starred in the movie bridge over the river kwai*" as an example: While it does not involve any geographical preference, the question is probably less likely to be asked in, say, Swahili, compared to in English or Thai. Second, the builders of MKQA explicitly made the decision to create "retrieval-independent answer annotations" that are linked to Wikidata entities and a few other value types. This decision, we feel, restricts the range of natural language questions that are covered. The cross-lingual aspect of the dataset appears to be primarily limited to entity translations, which likely do not cover a wide range of linguistic phenomena (which is the reason that we are interested in typologically diverse languages to begin with). Thus, we believe that Mr. TyDi fills a gap in the evaluation space that is currently not occupied.

Many multi-lingual (both mono-lingual and cross-lingual) information retrieval and question answering datasets have been constructed over the past decades, via community-wide evaluations at TREC, FIRE, CLEF, and NCTIR. These test collections are typically built on newswire articles, although some evaluations use Wikipedia and scientific texts. While no doubt useful for evaluation, these test collections usually comprise only a small number of queries (at most a few dozen) with relevance judgments, which are insufficient to fine-tune dense retrieval models. Furthermore, whereas TyDi at least draws from comparable corpora (i.e., Wikipedia articles), these test collections are built on corpora from much more diverse sources. This makes it difficult to generalize across different languages. For these reasons, the above-mentioned IR and QA test collections are not suitable for tackling the research questions we are interested in.

3 Mr. TyDi

Having justified the need for a new benchmark dataset, this section describes the construction of Mr. TyDi, which can be best described as an open-retrieval extension to TyDi.

Corpus The formulation of any text ranking problem begins with a corpus $\mathcal{C} = \{d_i\}$ comprising the units of text to be retrieved. As the starting point, we used exactly the same raw Wikipedia dumps as TyDi.

		Train		Dev		Test		Corpus Size
		# Q	# J	# Q	# J	# Q	# J	
Arabic	(Ar)	12,377	12,377	3,115	3,115	1,081	1,257	2,106,586
Bengali	(Bn)	1,713	1,719	440	443	111	130	304,059
English	(En)	3,547	3,547	878	878	744	935	32,907,100
Finnish	(Fi)	6,561	6,561	1,738	1,738	1,254	1,451	1,908,757
Indonesian	(Id)	4,902	4,902	1,224	1,224	829	961	1,469,399
Japanese	(Ja)	3,697	3,697	928	928	720	923	7,000,027
Korean	(Ko)	1,295	1,317	303	307	421	492	1,496,126
Russian	(Ru)	5,366	5,366	1,375	1,375	995	1,168	9,597,504
Swahili	(Sw)	2,072	2,401	526	623	670	743	136,689
Telugu	(Te)	3,880	3,880	983	983	646	664	548,224
Thai	(Th)	3,319	3,360	807	817	1,190	1,368	568,855
Total		48,729	49,127	12,317	12,431	8,661	10,092	58,043,326

Table 1: Descriptive statistics for Mr. TYDI: the number of questions (# Q), judgments (# J), and the number of passages (Corpus Size) in each language.

Relevance annotations in TYDI are provided at the passage level (in the passage selection task), and thus we kept the same level of granularity in our corpus preparation. For articles covered by TYDI (identified by the article titles), we retained the original passages. For articles that are not covered by TYDI, we prepared passages using WikiExtractor[1] based on natural discourse units (e.g., two consecutive newlines in the wiki markup). Unfortunately, Clark et al. (2020) did not precisely document their passage segmentation method, but based on manual examination of our results, the generated passages appear qualitatively similar to the TYDI passages.

The result of the corpus preparation process is, for each language, a collection of passages from the Wikipedia articles in that language. To form the final passages that comprise the basic unit of retrieval, we prepend the title of the Wikipedia article to each passage. This creates retrieval units that can be more readily understood in isolation.[2]

Task While Mr. TYDI is adapted from a QA dataset, our task is mono-lingual *ad hoc* retrieval. That is, given a question in language L, the task is to retrieve a ranked list of passages from $\mathcal{C}_L$, the Wikipedia collection in the same language (prepared in the manner described above), where the retrieved passages are ranked according to their relevance to the given question.

Our assumption here is a standard "retriever–reader" framework (Chen et al., 2017) or a multi-stage ranking architecture (Lin et al., 2020), where we focus on the retriever (what IR researchers call candidate generation or first-stage retrieval). For end-to-end question answering, the output of the retriever would be fed to a reader for answer extraction. This focus on retrieval allows us to explore the research questions outlined in the introduction, and this formulation is consistent with previous work in dense retrieval, e.g., Karpukhin et al. (2020).

Questions and Judgments To prepare the questions, we started with all questions provided by TYDI and removed those without any answer passages or whose answer passages are all empty. We consider all non-empty annotated answer passages from TYDI as relevant to the corresponding question in Mr. TYDI. We adopt the development set of TYDI as our test set, since the original test data are not public. A new development set was created by randomly sampling 20% of questions from the original training set. We observed that some of the questions in TYDI are shared between the training and development set (but labeled with different answer passages). In these cases, we retained the duplicate questions only in the training set. Descriptive statistics for Mr. TYDI are shown in Table 1, where languages are identified by their two letter ISO-639 language codes.

In summary, relevant passages in Mr. TYDI are imputed from TYDI. Since Clark et al. (2020) only asked annotators to assess the top-ranked article for each question, there are likely relevant passages that have not been identified. Following standard assumptions in information retrieval, unjudged passages are considered non-relevant. Thus, it is likely

[1] https://github.com/attardi/wikiextractor

[2] Mr. TYDI v1.1 contains these article titles, whereas v1.0 did not. Results reported in this paper are with v1.1; for differences, please refer to the earlier version of our paper posted on arXiv.

that ranking models will retrieve false negatives, i.e., passages that are relevant, but would not be properly rewarded.

In other words, our judgments are far from exhaustive. This might be a cause for concern, but is a generally accepted practice in IR research due to the challenges of gathering complete judgments. The widely used MS MARCO datasets (Bajaj et al., 2018), for example, share this characteristic of having "sparse judgments". No claim is made about the exhaustiveness of the annotations, as both Mr. TyDi and MS MARCO provide only about one good answer per question. From a methodological perspective, findings based on MS MARCO "sparse judgments" are largely consistent with results from more expensive evaluation efforts (to gather more complete judgments), such as the TREC Deep Learning Tracks (Craswell et al., 2020, 2021). We expect a similar parallel here: more exhaustive judgments will change the absolute scores, but will likely not affect the findings qualitatively.

Metrics We evaluate results in terms of reciprocal rank and recall at a depth k of 100 hits. The first metric quantifies the ability of a model to generate a good ranking, while the second metric provides an upper bound on end-to-end effectiveness (e.g., when retrieval results are fed to a reader for answer extraction). The setting of $k = 100$ is consistent with work in the QA literature.

4 Baselines

We provide a few "obvious" baselines for Mr. TyDi as a starting point for future research:

BM25 We report results with bag-of-words BM25, a strong traditional IR baseline, with the implementation provided by Pyserini (Yang et al., 2017; Lin et al., 2021a), which is built on the open-source Lucene search library. Lucene provides language-specific analyzers for nine of the eleven languages in Mr. TyDi; for these languages, we used the Lucene implementations. For Telugu (Te) and Swahili (Sw), since Lucene does not provide any language-specific implementations, we simply used its whitespace analyzer. We report BM25 scores on two conditions, with default and tuned k_1 and b parameters; the default settings are $k_1 = 0.9, b = 0.4$. Tuning was performed on the development set, on a per-language basis, via grid search on $k_1 \in [0.1, 1.6]$ and $b \in [0.1, 1.0]$, with step size 0.1, optimizing MRR@100.

mDPR Dense passage retriever (DPR) by Karpukhin et al. (2020) is a well-known bi-encoder model for open-domain QA that we adapt to mono-lingual retrieval in non-English languages by simply replacing BERT with multi-lingual BERT (mBERT),[3] but otherwise keeping all other aspects of the training procedure identical. This adaptation, which we call mDPR, was trained on the English QA dataset Natural Questions (Kwiatkowski et al., 2019) using Facebook's open-source codebase.

Our retrieval experiments with mDPR can be characterized as zero shot: We applied the same mBERT document encoder to convert passages from all eleven languages into dense vectors; similarly, we applied the same mBERT question encoder to all questions. Retrieval in each language was performed using Facebook's Faiss library for nearest neighbor search (Johnson et al., 2017); we used the FlatIP indexes. Experiments were conducted using the same codebase as the DPR replication experiments of Ma et al. (2021), with the Pyserini toolkit (Lin et al., 2021a) .

Our choice of zero-shot mDPR as a baseline deserves some discussion. At a high level, we are interested in the generalizability of dense retrieval techniques in out-of-distribution settings (in this case, primarily different languages). Operationally, our experimental setup captures the scenario where the model does not benefit from any exposure to the target task, even (question, relevant passage) pairs in the English portion of Mr. TyDi. This makes the comparison "fair" to BM25, which is similarly not provided any labeled data from the target task (in the case with default parameters).

Sparse–Dense Hybrid Our hybrid technique combines the scores of sparse (BM25) and dense (mDPR) retrieval results. The final fusion score of each document is calculated by $s_{\text{sparse}} + \alpha \cdot s_{\text{dense}}$, where s_{sparse} and s_{dense} represent the scores from sparse and dense retrieval, respectively. This strategy is similar to the one described by Ma et al. (2021). We take 1000 hits from mDPR and 1000 hits from BM25 and normalize the scores from each into $[0, 1]$ since the range of the two types of scores otherwise are quite different. If one hit isn't found in the other, the normalized score for that hit is set to zero. The weight α was tuned in $[0, 1]$ with a simple line search on the development set by optimizing MRR@100 with step size 0.01.

[3] Specifically, the `bert-base-multilingual-cased` model provided by HuggingFace (Wolf et al., 2020).

	Ar	Bn	En	Fi	Id	Ja	Ko	Ru	Sw	Te	Th	Avg
BM25 (default)	0.368	0.418	0.140	0.284	0.376	0.211	0.285	0.313	0.389	0.343	0.401	0.321
BM25 (tuned)	0.367	0.413	0.151	0.288	0.382	0.217	0.281	0.329	0.396	0.424	0.417	0.333
mDPR	0.260	0.258	0.162	0.113	0.146	0.181	0.219	0.185	0.073	0.106	0.135	0.167
hybrid	0.491†	0.535†	0.284†	0.365†	0.455†	0.355†	0.362†	0.427†	0.405	0.420	0.492†	0.417

(a) MRR@100

	Ar	Bn	En	Fi	Id	Ja	Ko	Ru	Sw	Te	Th	Avg
BM25 (default)	0.793	0.869	0.537	0.719	0.843	0.645	0.619	0.648	0.764	0.758	0.853	0.732
BM25 (tuned)	0.800	0.874	0.551	0.725	0.846	0.656	0.797	0.660	0.764	0.813	0.853	0.758
mDPR	0.620	0.671	0.475	0.375	0.466	0.535	0.490	0.498	0.264	0.352	0.455	0.473
hybrid	0.863†	0.937	0.696†	0.788†	0.887†	0.778†	0.706†	0.760†	0.786	0.827	0.875†	0.809

(b) Recall@100

Table 2: Results of BM25 (with default and tuned parameters), mDPR, and the sparse–dense hybrid on the test set of Mr. TYDI. The symbol † indicates significant improvements over BM25 (tuned) (paired t-test, $p < 0.01$).

5 Results and Analysis

We performed experiments on Mr. TYDI v1.1, where each passage contains the title of the Wikipedia article and the passage text. Table 2 reports results on the test set across all eleven languages; mean reciprocal rank (MRR) in the top table and recall in the bottom table, both at a cutoff of 100 hits; the final column reports the average across all languages. The rows report BM25 results (default and tuned), followed by results of mDPR and the sparse–dense hybrid. For the hybrid method, statistically significant improvements over tuned BM25 are denoted with the symbol † based on paired t-tests ($p < 0.01$).

5.1 High-Level Findings

By comparing scores in each column, we observe that the absolute effectiveness of the techniques varies greatly across languages. Absolute scores are difficult to compare because both the questions and the underlying corpora are different. However, three high-level findings emerge:

First, we find that tuning BM25 parameters yields at most minor improvements for most languages, both in terms of MRR@100 and recall except for Telugu (cases where scores decrease slightly can be explained by noise in the training/test splits). This is a bit of a surprise, as parameter tuning usually yields larger overall gains, e.g., in the MS MARCO collections (Bajaj et al., 2018). Regardless, tuned BM25 serves as a competitive baseline for the remainder of our experiments.

Second, we notice that mDPR underperforms BM25 across all languages except for English. That is, in a zero-shot setting, retrieval using learned dense representations from mDPR (fine-tuned with NQ) is a lot worse than just retrieval using BM25-based representations. Clearly, mDPR is far less robust in cross-lingual generalizations. Even within the same language, mDPR seems to be sensitive to characteristics of the training data. Effectiveness on the English portion of Mr. TYDI is only slightly better than BM25, likely arising from the fact that we are applying an NQ-trained model on "out-of-distribution" questions.

Based on Karpukhin et al. (2020) and the experiments by Ma et al. (2021), we would have expected mDPR to beat BM25 for in-distribution training and inference. Since NQ is also based on Wikipedia, corpus differences are less likely an issue; these results suggest that questions in TYDI and NQ are qualitatively different.

Third, despite the fact that mDPR effectiveness is quite a bit worse than BM25, the MRR@100 of the sparse–dense hybrid is significantly higher than tuned BM25 for nine of the eleven languages (the exceptions are Swahili and Telugu). Rephrased differently, this means that although mDPR by itself is a poor dense retrieval model in a zero-shot setting, it nevertheless contributes valuable relevance signals that are able to improve over tuned BM25. On average, the hybrid results are around eight and five points absolute higher than tuned BM25 in terms of MRR@100 and recall, respectively.

Because absolute scores vary widely across languages, it is helpful to normalize the effectiveness of tuned BM25 to 1.0 and scale the effectiveness of mDPR and the hybrid approach appropriately; this is shown in Figure 2 (left) for MRR@100. As

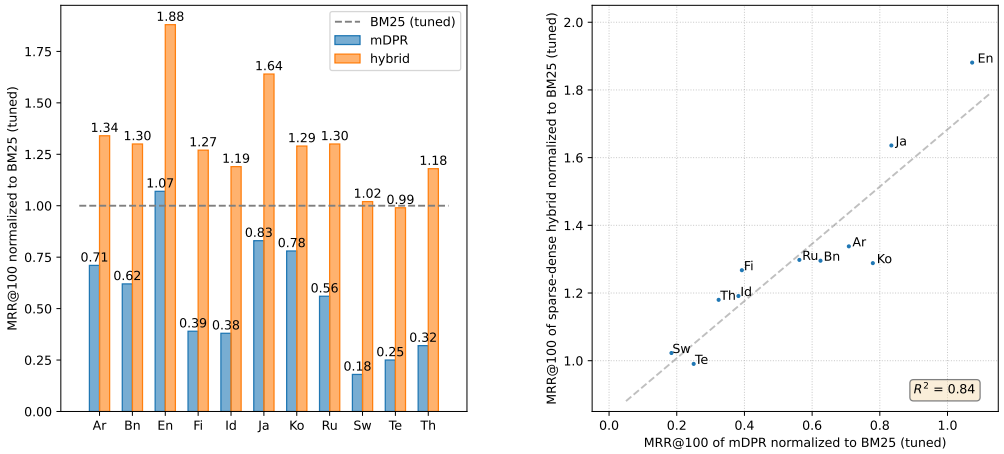

Figure 2: MRR@100 of mDPR and the sparse–dense hybrid normalized with respect to BM25 for each language (left); corresponding pairs for each language plotted as a scatter plot (right).

an example, from the leftmost bars, we see that the MRR@100 of mDPR in Arabic is 71% of BM25 but the sparse–dense hybrid improves over BM25 by 34% (stat. sig). We additionally plot the relation between the normalized effectiveness of mDPR and the hybrid approach in Figure 2 (right).

This plot shows a clear positive (linear) correlation, that is, better mDPR (relative) effectiveness translates into bigger improvements over BM25 in the sparse–dense hybrid. What is surprising, though, is that this relationship seems to hold even if the mDPR results are poor. For example, in Thai, the MRR@100 of mDPR is only 32% of BM25 (tuned), yet the hybrid yields a statistically significant 18% relative gain in the hybrid approach. However, there appear to be limits to our simple linear combination of relevance signals: for both Swahili and Telugu, the hybrid approach does not outperform tuned BM25, likely because mDPR effectiveness is too poor.

5.2 Components of Effectiveness

To provide a more in-depth analysis, we attempt to untangle effectiveness into two separate components: (1) retrieving a relevant passage and (2) placing the relevant passages into top ranks. The recall figures in Table 2 already quantify the first component, but MRR@100 alone does not tell the complete story for the second component, since the metric averages a bunch of zeros for questions where relevant passages do not appear in the top-100 hits. It could be the case, for example, that mDPR provides a good ranking for those queries where it retrieves a relevant result in the top 100.

The results of such an analysis, comparing BM25 (tuned) and mDPR, are shown in Figure 3 for all languages (ordered alphabetically). Each plot consists of a histogram and a line graph. The histogram captures the distribution of the ranks (binned by ten) where the relevant passage appears for each question.[4] Questions for which no relevant passage was found in the top-100 hits are tallied in the rightmost bar ("Not Found"). Thus, all questions are either in the rightmost bar (not found in the top-100 hits) or in one of the top-100 bins; these are exactly the components of recall, so the histograms are a more fine-grained way to visualize recall.

The superimposed line graphs in each plot show the ratio of the number of questions falling in each bin to the total number of questions in all top-100 bins (that is, we remove the "Not Found" bin and renormalize). These plots answer the following question: Given that the relevant passage appeared in the top-100 hits, how well did the model perform at ranking it? In other words, we have isolated the ranking ability of the model.

Looking only at the line graphs, these results tell us that for Arabic, Japanese, and Korean, BM25 and mDPR are comparable when we focus only on ranking—that is, given that the relevant passage appears in the top-100 hits. In other words, MRR@100 differences for these languages come mostly from the fact that mDPR misses many relevant passages that BM25 finds (i.e., exhibits lower

[4]If the question has multiple retrieved relevant passages, we only consider the smallest rank among them (i.e., the highest ranked relevant passage).

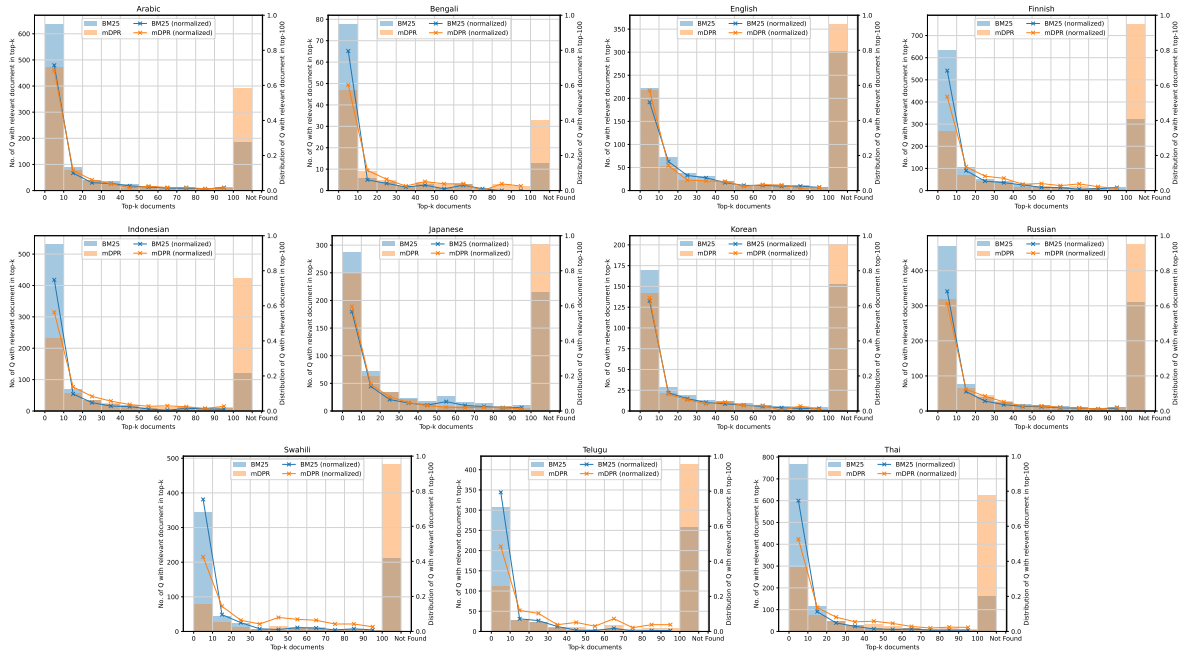

Figure 3: Analysis of recall and ranking effectiveness comparing BM25 (tuned) and mDPR. In each plot, the histogram shows the distribution of relevant passages; lines plot the distribution of relevant passages normalized to only questions where a relevant passage appears in the top-100 hits.

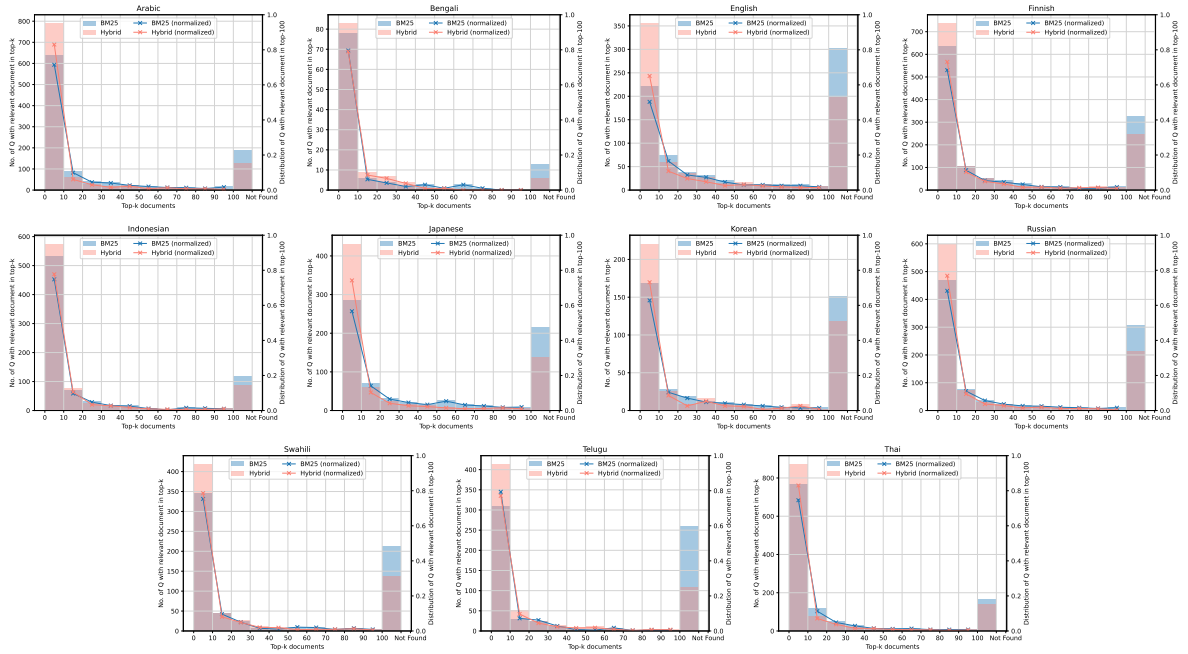

Figure 4: Analysis of recall and ranking effectiveness comparing BM25 (tuned) and the sparse–dense hybrid. Each plot is constructed in a similar manner as the plots in Figure 3.

recall). For the other languages, BM25 appears to exhibit *both* better recall and better ranking. Consider Swahili, for example, BM25 places many more relevant passages in the top 10 and also has far fewer questions where no correct answer appears in the top-100 hits. Thus, this analysis isolates the different failure modes of mDPR (dense retrieval) relative to BM25 (sparse retrieval).

The same analysis comparing BM25 (tuned) and the sparse–dense hybrid is shown in Figure 4. These plots reveal how the hybrid is improving the BM25 results. We see that gains in Bengali, Indonesian, Swahili, and Telugu come mostly from higher recall. That is, ranking capabilities are roughly comparable (the line plots largely overlap) but the hybrid approach has fewer queries where the relevant passage does not appear in the top-100 hits. For Thai, the gain comes from better ranking, while recall is just a small bit better (the "Not Found" bars are pretty close). For the other languages, hybrid improves both recall and ranking.

6 Future Work

Mr. TyDi provides a resource to begin exploring mono-lingual *ad hoc* retrieval with both dense and sparse retrieval techniques. In this paper, we have focused primarily on zero-shot baselines. Although zero-shot dense retrieval (mDPR) does not appear to be effective by itself, relevance signals from the model do appear to be complementary to sparse retrieval (bag-of-words BM25). We have identified *how* they are complementary (better recall vs. better ranking), but the behavior varies across languages, and we do not yet have an explanation for *why*; for example, do the typological characteristics of the language play a role?

For our experiments, we have decided to focus on zero-shot effectiveness because it serves as the natural baseline of any technique that tries more sophisticated approaches. Thus, the baselines here are foundational to any future work. We have explicitly decided not to report any language-specific fine-tuning results here, although preliminary experiments suggest that such techniques do bring about benefits. We have not yet systematically explored the broad design space of what Lin et al. (2020) calls "multi-step fine-tuning strategies", paralleling the explorations of Shi et al. (2020) in the context of transformer-based reranking models. There are many possible variations, for example, how many languages to use, what order to sequence data from different languages, possible data augmentation using machine translation, complementary data from other tasks, etc. There are a number of experiments that will allow us to tease apart the effects of language versus other aspects of training data distribution (e.g., NQ vs. TyDi). Exploration of this vast design space is the focus of our immediate future work.

In addition, we believe that our dataset can provide a probe to examine the nature of multi-lingual transformer models. Our experimental results show that absolute effectiveness varies quite a bit across languages. Some of these variations may be due to the nature of the queries, the size of the corpora, etc. However, we hypothesize that inherent properties of the transformer model play important roles as well, e.g., the size of the pretraining corpus in each language, typological and other innate characteristics of the languages, etc. We hope that Mr. TyDi can help us untangle some of these issues.

7 Conclusion

In this work, we introduce Mr. TyDi, a multi-lingual benchmark dataset for mono-lingual retrieval in eleven typologically diverse languages, built on TyDi. We describe zero-shot experiments using BM25, mDPR, and a sparse–dense hybrid. The experimental results are not surprising: as is already known from complementary experiments, dense retrieval techniques do not generalize well to out-of-distribution input. However, we find that even poor dense retrieval results provide valuable relevance signals in a sparse–dense hybrid.

Of course, this is only the starting point. With Mr. TyDi, we now have a resource to explore our motivating research questions regarding the behavior of dense retrieval models when fed "out of distribution" data, and from there, devise techniques to increase the robustness and generalizability of our techniques. The potential broader impact of this work is more equitable distribution of information access capabilities across diverse languages of the world—to help non-English speakers access relevant information in their own languages.

Acknowledgements

This research was supported in part by the Canada First Research Excellence Fund and the Natural Sciences and Engineering Research Council (NSERC) of Canada; computational resources were provided by Compute Ontario and Compute Canada.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268v3*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *arXiv:2102.07662*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv:2003.07820*.

Luyu Gao, Zhuyun Dai, Zhen Fan, and Jamie Callan. 2020. Complementing lexical retrieval with semantic residual embedding. *arXiv:2004.13969*.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv:2010.02666*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 113–122.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv:1702.08734*.

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.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, pages 39–48.

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.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. 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.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: BERT and beyond. *arXiv:2010.06467*.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *arXiv:2007.15207*.

Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. 2021. A replication study of dense passage retriever. *arXiv:2104.05740*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas.

Peng Shi, He Bai, and Jimmy Lin. 2020. Cross-lingual training of neural models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773.

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:2104.08663*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of Lucene for information retrieval research. In *Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 1253–1256, Tokyo, Japan.