MIA 2022 Shared Task: Evaluating Cross-lingual Open-Retrieval Question Answering for 16 Diverse Languages

Akari Asai[♣], **Shayne Longpre**[♠], **Jungo Kasai**[♣], **Chia-Hsuan Lee**[♣],
Rui Zhang[♦], **Junjie Hu**[♡], **Ikuya Yamada**[★◊], **Jonathan H. Clark**[‡], **Eunsol Choi**[†]
[♣]University of Washington [♠]Massachusetts Institute of Technology [♦]Penn State University
[♡]University of Wisconsin-Madison [★]Studio Ousia [◊]RIKEN
[‡]Google Research [†]The University of Texas at Austin
mia.nlp.workshop@gmail.com

Abstract

We present the results of the Workshop on Multilingual Information Access (MIA) 2022 Shared Task, evaluating cross-lingual open-retrieval question answering (QA) systems in 16 typologically diverse languages. In this task, we adapted two large-scale cross-lingual open-retrieval QA datasets in 14 typologically diverse languages, and newly annotated open-retrieval QA data in 2 underrepresented languages: Tagalog and Tamil. Four teams submitted their systems. The best system leveraging iteratively mined diverse negative examples and larger pretrained models achieves 32.2 F1, outperforming our baseline by 4.5 points. The second best system uses entity-aware contextualized representations for document retrieval, and significant improvements in Tamil (20.8 F1), whereas most of the other systems yield nearly zero scores. The official leaderboard[1] and baselines[2] models are publicly available.

1 Introduction

Open-retrieval[3] question answering (QA) is a task of answering questions in diverse domains given large-scale document collections such as Wikipedia (Chen and Yih, 2020). Despite the rapid progress in this area (Chen et al., 2017; Karpukhin et al., 2020; Lewis et al., 2020b), the systems have primarily been evaluated in English, yet open-retrieval QA in non-English languages has been understudied (Longpre et al., 2021; Asai et al., 2021a). Moreover, due to the task complexity, cross-lingual open-retrieval QA has unique challenges such as multi-step inference (retrieval and answer selection) and cross-lingual pattern matching (Lewis et al., 2020a; Schäuble and Sheridan, 1997), whereas other multilingual NLP tasks have their inputs specified at once (e.g. natural language inference) and typically only need to perform inference on one language at a time.

In this work, we introduce the MIA 2022 shared task on cross-lingual open-retrieval QA, which tests open-retrieval QA systems across typologically diverse languages. Compared to previous efforts on multilingual open-retrieval QA (Forner et al., 2008, 2010), this shared task covers a wider set of languages (i.e., 16 topologically diverse languages) and orders of magnitude more passages in retrieval targets (i.e., 40 million passages in total), and constitutes the first shared task for massive-scale cross-lingual open-retrieval QA. Four teams submitted systems, three of which significantly improve the baseline system based on a state-of-the-art multilingual open-retrieval QA system (Asai et al., 2021b).

Our analysis reveals that the system performance varies across languages even when the questions are parallel (as in one of our two settings), and several findings from the submitted systems shed light on the importance on entity-enhanced representations, leveraging more passages and data augmentation for future research in multilingual knowledge-intensive NLP. Our analysis suggests that (i) it is still challenging to retrieve passages cross-lingually, (ii) generating answers in the target language whose script differs from the script of evidence document is nontrivial, (iii) and potential answer overlaps in existing datasets may overestimate models' performance.

We formally introduce our task in Section 2, followed by data collection process for 16 languages in Section 3. We then introduce our baseline systems in Section 4 and the submitted systems. Section 5 presents our meta analysis of the systems performances, and we conclude by suggesting future improvements in this area.

[1] https://eval.ai/web/challenges/challenge-page/1638/leaderboard

[2] https://github.com/mia-workshop/MIA-Shared-Task-2022

[3] Also sometimes referred to as *open-domain* QA; we use *open-retrieval* as it is not ambiguous with the sense of "covering many domains."

2 Task Descriptions

We first formulate cross-lingual open-retrieval QA and introduce metrics used to evaluate systems' performance. We then present two submission tracks: constrained and unconstrained tracks.

2.1 Task Formulation

Cross-lingual open-retrieval QA is a challenging multilingual NLP task, where given questions written in a user's preferred language, a system needs to find evidence from large-scale document collections written in many different languages. The final answer needs to be in the user's preferred language which is indicated by their question, as in real-world applications. We follow the general definition of Asai et al. (2021b), where a system can retrieve evidence from documents in *any* languages, not limiting the retrieval target to certain languages as in Forner et al. (2008). For instance, a system needs to answer in Arabic to an Arabic question, but it can use evidence passages written in any language included in a large-document corpus such as English, German, Japanese and so on. In real-world applications, the issues of information asymmetry and information scarcity (Roy et al., 2022; Blasi et al., 2022; Asai et al., 2021a; Joshi et al., 2020) arise in many languages, hence the need to source answer contents from other languages—yet we often do not know *a priori* in which language the evidence can be found to answer a question.

2.2 Evaluation Metrics

Systems are evaluated using automatic metrics: token-level F1 and exact match (EM). Although EM is often used as the primary evaluation metric for English, the risk of surface-level mismatching (Min et al., 2020a) can be more pervasive in cross-lingual settings. Therefore, we use F1 as the primary metric and rank systems using the F1 scores. Evaluation is conducted using language-specific tokenization and evaluation scripts provided in the MIA shared task repository.[4] We use data from XOR-TyDi QA and MKQA (detailed in Section 3), and due to different characteristics these datasets have, we macro-average scores per language set on each dataset, and then macro-average those scores to produce an F1 score for XOR-TyDi QA and an F1 score for MKQA to compute the final scores for ranking.

2.3 Tracks

For the shared task, we defined two tracks based on the resource used to train systems: *constrained* and *unconstrained* settings. Systems trained only on the official training data qualify for the constrained track, while systems trained with additional data sources participate in the unconstrained track.

Constrained Track. To qualify as a constrained track submission, participants are required to use the official training corpus, which consists of examples pooled from XOR-TyDi QA and Natural Questions (Kwiatkowski et al., 2019). See more data collection details in Section 3. No other QA data may be used for training. We allow participants to use off-the-shelf tools for linguistic annotations (e.g. POS taggers, syntactic parsers), as well as any publicly available unlabeled data and models derived from these (e.g. word vectors, pre-trained language models). In the constrained setup, participants may not use external blackbox APIs such as Google Search API and Google Translate API for inference, as those models are often trained on additional data, but they are permitted to use them for offline data augmentation or training.

Unconstrained track. Any model submissions using APIs or training data beyond the scope of the constrained track are considered for the *unconstrained* setting. Participants are required to report the details of their additional resources used for training, for transparency. For instance, a submission might use publicly available QA datasets, such as CMRC 2018 (Cui et al., 2019) and FQuAD (d'Hoffschmidt et al., 2020), to create larger-scale training data.

3 Shared Task Data

The MIA shared task data is derived from two large-scale multilingual evaluation sets: XOR-TyDi QA (Asai et al., 2021a) and MKQA (Longpre et al., 2021). We first discuss the source datasets, and then discuss how the target languages are selected, and how the data is split into training and evaluation sets. Table 1 shows the included languages, their language groups, the size of training, development and test data, and the number of Wikipedia passages available in each language.

[4] For non-spacing languages (i.e., Japanese, Khmer, and Chinese), we use off-the-shelf tokenizers including Mecab, khmernltk and jieba to tokenize both predictions and ground-truth answers.

	Language Family		# of examples			# Wiki. passages
Language	Family	Branch	Train	Development	Test	
Arabic (ar)	Afro-Asiatic	Semitic	18,402	3,145	5,590	1,304,828
Bengali (bn)	Indo-European	Indo-Iranian	5,007	2,248	5,203	179,936
English (en)	Indo-European	Germanic	76,635	1,758	5,000	18,003,200
Spanish (es)	Indo-European	Italic	0	1,758	5,000	5,738,484
Finnish (fi)	Uralic	Finnic	9,762	2,732	1,368	886,595
Japanese (ja)	Japonic	Japonic	7,815	2,451	6,056	5,116,905
Khmer (km)	Austroasiatic	Khmer	0	1,758	5,000	63,037
Korean (ko)	Koreanic	Han	4,319	2,231	6,048	638,864
Malay (ms)	Austronesian	Malayo-Poly.	0	1,758	5,000	397,396
Russian (ru)	Indo-European	Balto-Slavic	9,290	2,776	6,910	4,545,635
Swedish (sv)	Indo-Europea	Germanic	0	1,758	5,000	4,525,695
Chinese (zh)	Sino-Tibetan	Sinitic	0	1,758	5,000	3,394,943
Telugu (te)	Dravidian	South-Central	6,759	2,322	6,873	274,230
Surprise Languages						
Tagalog (tl)	Austronesian	Malayo-Poly.	0	0	350	–
Tamil (ta)	Dravidian	Southern	0	0	350	–

Table 1: List of the languages, their families and amount of data available in the MIA shared task data. The last two languages are surprise languages hidden from the participants.

3.1 Source Datasets

XOR-TyDi QA (Asai et al., 2021a) is a cross-lingual open-retrieval QA dataset covering 7 languages built upon TyDi QA (Clark et al., 2020). Asai et al. (2021a) collect answers for questions in TyDi QA that are *unanswerable* using the same-language Wikipedia. As the questions are inherited from TyDi QA, they are written by native speakers to better reflect their own interests and linguistic phenomena, and they are not parallel across languages. We use data for the XOR-full setting, where some questions can be answered based on the target language's Wikipedia (monolingual) while others require evidence only presented in English Wikipedia (cross-lingual). We use all of the 7 languages covered by XOR-TyDi QA: Arabic (ar), Bengali (bn), Finnish (fi), Japanese (ja), Korean (ko), Russian (ru), Telugu (te).

MKQA (Longpre et al., 2021) comprises the largest set of languages and dialects (26) for open-retrieval QA, spanning 14 language families. There are 10k question and answer pairs per language. The questions are human-translated from English Natural Questions (Kwiatkowski et al., 2019) and the answers are re-annotated for higher quality – chosen independently of any web pages or document corpora. From MKQA, we sample the 6,758 parallel examples which are answerable. We select 12 of the 26 languages to lower the computational barrier: Arabic (ar), English (en), Spanish (es), Finnish (fi), Japanese (ja), Khmer (km), Korean (ko), Malay (ms), Russian (ru), Swedish (sv), Turkish (tr), and traditional Chinese (zh-cn).

3.2 Language Selection

We select a subset of languages from each resource (i) to cover a wide range of languages and typological features with a sufficient scale, and (ii) to compare participating model performance between questions that are translated from English and ones that are naturally generated by native speakers. The natively-written questions from XOR-TyDi QA allow measuring systems' quality on questions that are likely to serve information need expressed by speakers of each language, whereas the human-translated questions of MKQA allow measuring the performance on the target script and language, holding constant the question content. For this reason, we include 5 languages present in both XOR-TyDi QA and MKQA to compare the gap between cultural and linguistic model generalization: Arabic, Finnish, Japanese, Korean, and Russian.

Surprise languages. In addition, we newly annotated data in Tagalog (tl) and Tamil (ta), where little work studies open-retrieval QA (Liu et al., 2019). For each language, we sample 350 MKQA English examples, where the answer entities have an Wikipedia article in the target language. The 350 questions are all translated using Gengo's human translation,[5] but the answers are automatically translated using Wikidata. This annotation results in 350 well-formed examples in Tagalog (tl) and Tamil (ta). Surprise languages are released two

[5] https://gengo.com/

weeks before the system submission deadline to test systems' ability to perform zero-shot transfer (Hu et al., 2020) to unseen languages that are substantially different from the languages they are trained on. Except for one system, all of the submissions directly apply their systems to the new languages without any training or adding new target languages' Wikipedia.

3.3 Data Statistics

Table 1 presents the list of the languages and statistics of the train, development and test set data in each target language.

Training data. Our training data consists of Natural Questions (Kwiatkowski et al., 2019) for English and XOR-TyDi QA for the other languages in the shared task.[6] In the constrained track (Section 2.3) only this data source is permitted for providing QA supervision, though other tools are permissible for data augmentation.

Evaluation data. Our evaluation sets span 16 languages: 7 from XOR-TyDi QA and 12 from MKQA with an overlap of five languages and two surprise languages newly annotated for this shared task following MKQA annotation schema. We found that the original XOR-TyDi QA validation and test splits have different proportions of the in-language and cross-lingual questions, resulting in large performance gaps between dev and test subsets as reported by Asai et al. (2021b). We re-split XOR-TyDi QA so that the validation and test sets have similar ratios of the two question types of in-language and cross-lingual questions. In-language questions are answerable from Wikipedia in the question's language, and are often easier to answer while the other category requires cross-lingual retrieval between the target language and English, and are more challenging. Further, we add aliases that can be retrieved via the Wikimedia API to the gold answers, following MKQA, thereby avoiding penalizing models for generating correct answers with surface-level differences. For MKQA we split the answerable examples into a validation set of 1,758 questions and a test set of 5,000 question. We add the newly annotated data for the surprise languages (Tamil and Tagalog) to the test set only.

3.4 Limitations

False negatives in evaluations. First, because the original source questions and answers are from TyDi QA or Natural Questions, their answers are annotated based on a single Wikipedia article in English or the question language. MKQA answers are re-labeled by English speakers without any Wikipedia or web corpus, but small portion of the answers can be geographically incorrect for that regions of the languages the data is translated into (e.g., when the first harry potter movie was released?). As we generalize the task setting to *cross-lingual open retrieval*, there are inconsistent contents across articles in different languages leading to many possible answers. However, because we only have one answer, this can penalize correct answers (Palta et al., 2022). It is a common issue that open-retrieval QA datasets do not comprehensively cover all valid answers (Min et al., 2020a; Asai and Choi, 2021), and this can be more prevalent in multilingual settings due to transliteration of entities or diverse ways to express numeric in some languages (Al-Onaizan and Knight, 2002).

English American-centric biases. Second, the MKQA questions as well as the new data annotated for this shared task are translated from English. This annotation scheme enables us to scale up to many typologically diverse languages, but the resulting questions are likely to be Western- or specifically American-centric, rather than reflecting native speakers' interests and unique linguistic phenomena (Clark et al., 2020). We try to reduce such English-centric bias by only using the questions whose answer entities are also included in Tamil or Tagalog Wikipedia, though this constrains the distribution to simple factoid questions. We also found that in some languages, MKQA answers have high overlap with their English counterparts.

4 Baseline Models

We use a state-of-the-art open-retrieval QA model as our baseline. We open source the code, trained checkpoints, training data, and intermediate/final prediction results.[7]

4.1 Modeling

Our baseline model is based on CORA (Asai et al., 2021b), which has two components: mDPR for

[6] See the training data linked at https://github.com/mia-workshop/MIA-Shared-Task-2022#training-data

[7] https://github.com/mia-workshop/MIA-Shared-Task-2022

document retrieval and mGEN for answer generation. Both mDPR and mGEN are based on multilingual pretrained models to process data written in many different languages without relying on external translation modules.

Given a question q^L written in a language L, mDPR $\mathcal{R}$ retrieves top N passages: $\mathbf{P} = p_1, \ldots, p_N = \mathcal{R}(q^L)$. mDPR includes all of the target languages' Wikipedias as its retrieval target, except for the two surprise languages. mGEN $\mathcal{G}$ takes as input q and $\mathbf{P}$ and generates an answer a^L in the target language: $a^L = \mathcal{G}(q, \mathbf{P})$. mDPR is a multilingual extension of DPR (Karpukhin et al., 2020), which employs a dual-encoder architecture based on BERT (Devlin et al., 2019) and retrieves top passages based on the dot-product similarities between encoded representations. During training, mDPR optimizes the loss function as the negative log likelihood of the positive passages. mGEN simply concatenates the question and a set of top K passages, and the fine-tuned multilingual encoder-decoder model generates a final answer in the target language. Unlike some prior work in English conducting end-to-end training of the retriever and reader (Lewis et al., 2020c; Guu et al., 2020), we train mDPR and mGEN independently. Note that during mGEN training, we use the passages retrieved by the trained mDPR, as in Izacard and Grave (2021a).

4.2 Training and Hyperparameters

We use the official training data for training. We also leverage the long answer annotations in the Natural Questions dataset and the gold paragraph annotations of XOR-TyDi QA to create mDPR training data, released at the shared task repository.[8] After training mDPR, we run it on the shared task training data questions to obtain top passages, and then use those retrieved passages to train the mGEN model: mGEN is trained to generate the gold answer given an input query and top retrieved passages.

mDPR uses multilingual BERT-base uncased (Devlin et al., 2019), and mGEN is fine-tuned from mT5-base (Xue et al., 2021). For mDPR, we use the same hyperparameters as in DPR (Karpukhin et al., 2020), and train it for 30 epochs, and take the last checkpoint. For mGEN, we follow Asai et al. (2021b) hyperparameters.

4.3 Pre-processing Knowledge Corpus.

Following DPR and mDPR, we split each article into 100-token chunks based on whitespace. For non-spacing languages (e.g., Japanese, Thai), we tokenize the articles using off-the-shelf tokenizers (i.e., MeCab for Japanese[9] and Thai NLP for Thai[10]). We exclude passages with less than 20 tokens. Total numbers of passages for each language are listed in Table 1.

5 Shared Task Submissions

Four teams submitted their final systems to our EvalAI (Yadav et al., 2019) leaderboard,[11] three of which significantly outperformed the original baseline described in Section 4. We summarize the submitted systems here and refer readers to their system description paper for details. All of the systems are constrained submissions, and we did not receive any submission to unconstrained track.

mLUKE+FiD. Tu and Padmanabhan (2022) adapt the retrieve-then-read baseline system with several improvements, including (a) using an mLUKE encoder (Ri et al., 2022) for dense retrieval, (b) combining sparse and dense retrieval, (c) using a fusion-in-decoder reader (Izacard and Grave, 2021b), and (d) leveraging Wikipedia links to augment the training data with additional target language labels.

For retrieval, Tu and Padmanabhan (2022) use the 2019/02/01 Wikipedia snapshot as their document corpora, matching the baseline. They include the Wikipedia snapshots for Tamil and Tagalog to evaluate on the surprise languages. Their sparse retriever searches the monolingual corpora only, while their dense retriever searches all corpora.

CMUmQA. Agarwal et al. (2022) build a four-stage pipeline for a retrieve-then-read approach, based on the CORA open-retrieval system (Asai et al., 2021b) that searches evidence documents in any language for target questions (many-to-many QA; Asai et al., 2021b), without relying on translation. They first apply an mBERT-based DPR retrieval model, followed by a reranker (Qu et al., 2021) with XLM-RoBERTA (Conneau et al., 2020). While it is computationally intractable to use for

[8] https://github.com/mia-workshop/MIA-Shared-Task-2022#training-data

[9] https://taku910.github.io/mecab/.

[10] https://github.com/PyThaiNLP/pythainlp.

[11] https://eval.ai/web/challenges/challenge-page/1638/leaderboard

System	Macro F1			Language F1						
	Total	XOR	MKQA	Arabic	Bengali	Finnish	Japanese	Korean	Russian	Telugu
(a) Texttron	**32.02**	**45.50**	18.54	**56.37**	**42.43**	**43.13**	**44.71**	**34.37**	**47.79**	**49.72**
(b) mLUKE-FID	31.61	40.93	22.29	45.33	30.48	41.01	43.45	31.21	42.62	42.40
(c) CMUmQA	31.53	40.20	**22.87**	55.06	30.56	41.25	42.44	28.76	42.56	40.75
(d) ZusammenQA	27.00	37.95	16.04	49.66	33.99	39.54	39.72	25.59	40.98	36.16
(e) Baseline	27.55	37.95	17.14	51.66	31.96	38.68	40.89	25.35	39.87	37.26

Table 2: Final results on the XOR-TyDi QA subsets of the MIA 2022 shared task. The grayed entry indicates the baseline system.

sys	Language F1													
	ar	en	es	fi	ko	ma	ja	km	ru	sv	tr	zh	tm	ta
(a)	13.62	33.24	28.98	25.26	13.07	29.04	23.11	3.96	20.11	29.75	**28.15**	11.30	0.00	0.00
(b)	12.67	39.63	30.85	25.22	12.81	29.09	20.49	2.36	18.82	29.62	26.16	**22.60**	**20.75**	20.95
(c)	**13.94**	**42.58**	**32.11**	**26.75**	**14.59**	**31.13**	**22.72**	**8.71**	**22.36**	**31.48**	26.59	18.00	2.74	**26.42**
(d)	8.73	35.32	25.54	20.42	6.78	24.10	14.27	6.06	12.01	25.97	20.27	13.95	0.00	11.14
(e)	9.52	36.34	27.23	22.70	7.68	25.11	15.89	6.00	14.60	26.69	21.66	13.78	0.00	12.78

Table 3: Final results on the MKQA subsets of the MIA 2022 shared task. The grayed entry indicates the baseline.

retrieval, the reranker has the advantage of encoding a question and a passage together, rather than independently. An mT5-based fusion-in-decoder is then applied to generate an answer. As the final step of their pipeline, Wikidata is used to translate English entities in the answer into the target language, if any.

ZusammenQA. Hung et al. (2022) follow the retrieve-then-read system, but with the expansion of several components, along with training methods and data augmentation. Their retriever ensembles supervised models (mDPR and mDPR with a MixCSE loss; Wang et al., 2022) along with unsupervised sparse (Oracle BM-25) and unsupervised dense models (DISTIL, LaBSE, MiniLM, MPNet).

The reader system is based on mGEN, but with domain adaptation by continued masked language modeling on the document corpora, to better adapt to Wikipedia and the target languages. The training data is augmented using Dugan et al. (2022) that generates question-answer pairs from raw document corpora and translates them into multiple languages.

Texttron. This submission also follows the retrieve-then-read structure: the retrieval model performs dense passage retrieval with XLM-RoBERTa Large (Conneau et al., 2020), and the reading model uses mt5 large. The retrieval text is split into paragraphs (as opposed to 100-word text segments) extracted by the WikiExtractor package. The retrieval model is trained on a combination of three types of custom training data: target-to-target (both the query and retrieved paragraphs are in the target language), target-to-English (the query is in the target language and the retrieval paragraphs are in English), and English-to-English (both the query and retrieved paragraphs are in English). These data are created based on BM25 retrieval and query translation.

Texttron also used multiple stages of training and negative sample mining to tune their final dense retriever with hard negatives: a combination of BM25 and examples from the previous iteration of retrieval that had low token overlap with the gold answers. No system description was available.

6 Main Results

Tables 2 and 3 show final results on XOR-TyDi QA and MKQA subsets, respectively.

Macro performance. Texttron, mLUKE + mFiD, and CMUmQA significantly improve the baseline performance. Among the constraint submissions, Texttron yields the best performance. While several systems achieve higher than 40 average F1 on XOR-TyDi QA, only two systems achieve higher than 20 average F1 on MKQA, demonstrating how difficult it is to build a system that performs well in many languages without language-specific supervision. Texttron significantly outperforms other baselines on XOR-TyDi

QA while CMUmQA shows the best MKQA performance among the submitted systems.

Language-wise performance. The performance varies across different languages. Among XOR-TyDi QA, all of the systems struggle in Korean and Bengali, while in Arabic, Japanese and Russian, they generally show relatively high F1 scores.

On MKQA, where all of the questions are parallel, the performance still significantly differs across languages. Almost all of the systems report lower than 10 F1 in Khmer and Tamil, which are less represented in existing pretraining corpora (Xue et al., 2021) and use their own script systems—with the notable exception of mLUKE + FiD, which achieves 20.8 F1 on Tamil. mLUKE + FiD achieves substantially better performance than other systems in Tamil. This is partially because they also include the Tamil Wikipedia passages for passage retrieval, while other systems, including the baseline, do not. As discussed in Asai et al. (2021b), all systems show lower scores in the languages that are distant from English and use non-Latin scripts (e.g., Cyrillic for Russian, Hangul for Korean).

7 Analysis

We provide further analysis on the submitted systems. In Section 7.1 we provide a brief summary of the findings from the submitted system descriptions. Section 7.2 provides performance comparison over answer-type, and answer overlap with English or training data. We then analyze the degree of answer agreements among the submitted systems to understand which questions remain challenging in Section 7.3. We further conduct manual error analysis in five languages in Section 7.4.

7.1 Summary of Findings

In this section, we highlight several effective techniques from the submitted systems. Overall, a surprisingly wide range of complementary, and potentially additive, methods all reported strong benefits, including: (i) larger and longer pre-trained models for retrieving and reading, (ii) a reranking step with fusion-in-decoder multi-passage cross-encodings, (iii) iterative dense retrieval tuning with progressively harder negative example mining, (iv) using entity-aware retrieval encodings, (v) combining dense and sparse retrievers, (vi) data augmentation, and (vii) leveraging Wikidata answer post-processing for language localization. We discuss some of these below.

These findings highlight various techniques migrating the performances in English retrieval systems. And most of all, they emphasize that cross-lingual retrieval still poses the major bottleneck to the end-to-end task, while large multilingual fusion-in-decoder reader systems can operate well when given sufficient evidence. These findings suggest multilingual retrieval is the most important avenue for future research, especially on questions not easily answered by English Wikipedia. Moreover, retrieving evidence cross-lingually is keys for other knowledge intensive NLP tasks such as fact verification (Thorne et al., 2018) and knowledge-grounded dialogues (Dinan et al., 2019) beyond open-retrieval QA.

Entity representations. Using entity-aware representations for the passage retriever's encoders gives a large performance improvement; As shown in analysis by Team Utah (Tu and Padmanabhan, 2022), replacing mBERT encoders in DPR with mLUKE improves by 1.22 F1 on XOR macro-average and 1.85 MKQA macro F1. We hypothesize that the mLUKE may capture better cross-lingual entity alignment than mBERT as it leverages inter-language links in Wikipedia during pretraining. This sheds light on the potential effectiveness of multilingual entity contextualized representations for cross-lingual passage representations, which is an under-explored direction.

Combining dense and sparse retrievers & hard negatives. Texttron and Team Utah combine both BM25 and mDPR, while ZusammenQA explore a diverse set of unsupervised and supervised retrieval approaches including BM25 and LaBSE (Feng et al., 2022). Team Utah shows that combining BM25 with mDPR helps, while ZusammenQA shows that only using BM25 gives significantly lower scores than the original baseline (Hung et al., 2022), as BM25 does not have cross-lingual phrase matching capabilities. Texttron iteratively trained their dense retriever, mining increasingly hard negative examples using BM25 and query translation, filtered using simple heuristics.

Fusion-in-Decoder and passage reranking. Team Utah and CMUmQA demonstrate that Fusion-in-Decoder architectures outperform simply concatenating passages as in mGEN (Fusion-in-Encoder). While Fusion-in-Encoder simply concatenates retrieved passages in a retrieved order, Fusion-in-Decoder encodes each of the retrieved passages independently and then concatenate them. This may help the model to pay more attentions to the passages that are ranked lower by the retriever but indeed provides evidence to answer. Recent work in open domain QA also demonstrates that the Fusion-in-Decoder architecture is more competitive than prior systems that simply concatenate passages (Fajcik et al., 2021; Asai et al., 2022).

Team Utah show increasing the number of passages improves performance, while CMUmQA show that cross-encoder reranking is particularly beneficial for Fusion-in-Decoder.

Data augmentation. ZusammenQA introduces data augmentation using Google Translate to translate the training data into target languages. **AUG-QA** translates question-answer pairs into target languages, while **AUG-QAP** translates question, answer and the original training data passages into the target languages. They found that the AUG-QAP and AUG-QA both improve performance from their direct counterpart without data augmentation.

Wikipedia answer localization. CMUmQA and others used Wikidata entity maps to localize answers to the correct target script following Longpre et al. (2021). This process was particularly effective for localizing short answers into a target language from English due to the overwhelming English bias of retrieval and generative systems finetuned on English. As a result, CMUmQA obtains the best MKQA performance among the submitted systems.

7.2 Performance Comparison

In this section, we group questions based on several factors (e.g., answer types) and compare the models' performance across different sub-groups.

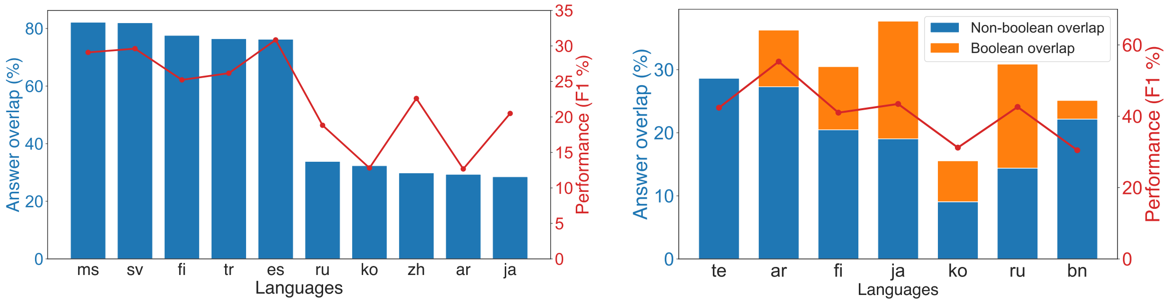
Answer types. MKQA provides answer categories for each question. We analyze the per-category model performance to understand what types of questions remain challenging. The original MKQA source data except for the unanswerable subsets has the following answer type distributions: Entity (42%), Date (12%), Number (5%), Number with Unit (4%), Short Phrase (3%), Boolean (yes, no; 1%), Unanswerable (14%), and Long Answers (13%). The Unanswerable and Long Answers categories are excluded from the MIA 2022 shared task evaluation data.

We present the percentage of the questions where any of the submitted system predictions match the annotated gold answers in English, Spanish, Japanese and Chinese in Table 4. In all of the languages, the systems show relatively higher exact matching rate in Entity types questions except for Chinese and Japanese. In those languages, many of the entity names are written in their own script systems (e.g., Chinese characters, katakana), which is challenging to be generated from the evidence passages written in other languages; it is known to be challenging to translate an entity name from one language to another using different script systems (Wang et al., 2017). In English and Spanish, the systems show significantly higher accuracy on entity and date than in Japanese or Chinese, while the systems struggle in Boolean questions. XOR-TyDi QA Japanese subset shows higher percentage of boolean questions than other subsets, which potentially helps the systems in Japanese and Chinese MKQA boolean questions. All of the systems show significantly lower performance in short phrase questions, indicating the difficulty of generating phrase length answers beyond simple factoid questions with entity or date answers.

	en	es	ja	zh
Number with units	7.77	3.56	1.94	3.88
Entity	**58.18**	**53.19**	34.42	15.75
Number	27.07	29.83	21.27	25.70
Date	28.14	28.49	6.10	11.37
Short phrases	8.60	7.81	5.08	5.08
Binary	32.99	31.96	**79.38**	**75.25**

Table 4: The percentage of the exact match per answer types in English (en), Spanish (es), Japanese (ja) and Chinese (zh).

Answer overlaps with English. We analyze performances across languages by examining the relationship between the final performance and the number of the questions whose answers are the same as English answers. Figure 1a shows mLUKE + FiD, which is the second best system in our shared task, and answer overlap with the English subsets for each MKQA language except for Khmer and two surprise languages. We observe a clear correlation between the answer overlap and final performance among those languages. The model performs well on the languages where many answers are the same as English answers. Finnish, on the other hand, shows relatively lower performance compared to other languages with high answer overlap (i.e., Malay, Swedish, Spanish). Among the languages with low answer overlap, on the Japanese



(a) MKQA performance vs. answer overlap with English answers.

(b) XOR-TyDi QA performance vs. answer overlap between train and test sets.

Figure 1: Performance vs. answer overlap between train and test sets.

and Chinese sets, the system shows relatively high F1 scores compared to the other languages with lower than 40% overlap (i.e., Russian, Korean, Arabic). This is likely because Chinese and Japanese show higher accuracy on Boolean type questions than other languages as discussed above.

Answer overlap with training data. Prior work shows that the high overlap between train and test data can result in the overestimated performance of the systems (Lewis et al., 2021). In XOR-TyDi QA, the questions are annotated by native speakers of the target languages, so the percentage of the train-test overlap can vary across languages. We calculate the percentage of the answers for the test data questions that also appear as gold answers in XOR-TyDi QA training data. We then check whether the degree of the answer overlap between the train and test sets correlate with the final XOR-TyDi QA test performance.

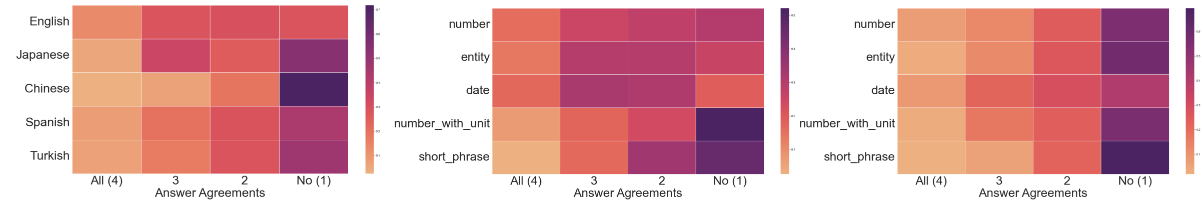
Figure 1b shows the performance and train-test overlap percentage. Although we can see the percentage of overlap between train and test data varies across languages, it is not particularly correlated with the final performance. For instance, Bengali actually shows relatively high overlap between train and test data (over 25% answer overlap), but the performance is much lower than Telugu, whose answer overlap ratio is close to that of Bengali. We also found that the percentage of the Boolean questions (yes, no) significantly differs across languages: in Japanese, around 10% of the questions are Boolean questions, while in Telugu, almost no questions are Boolean. The original TyDi QA data is annotated by different groups of annotators for each language, and thus such question distributions can differ (Clark et al., 2020).

XOR-TyDi QA vs. MKQA. Arabic, Japanese, Korean, and Finnish are included both in MKQA and XOR-TyDi QA, but their performance on the two subsets significantly differ; In general, the XOR-TyDi QA F1 scores are much higher than MKQA (e.g., Japanese: 44.71 vs. 23.11). We hypothesize that this happens because we do not have training data for MKQA and all MKQA questions tend to require cross-lingual retrieval as the questions are translated from English and answers are American-centric. In contrast, half of the questions in XOR-TyDi QA are from TyDi QA, and the answers are grounded to their own languages' Wikipedia. Cross-lingual retrieval is generally more challenging than monolingual retrieval (Zhang et al., 2021). In addition, all of the XOR-TyDi QA cross-lingual questions are labeled "unanswerable" in TyDi QA, and can be more difficult to answer than its monolingual counterparts.

To further test this hypothesis, we evaluate the submitted systems' performance on XOR-TyDi QA's cross-lingual and monolingual subsets in Table 5. We can clearly see that all of the baseline's performance deteriorates on the cross-lingual subsets, while they show high F1 scores across languages on the monolingual subsets.

7.3 Prediction Agreement

We analyze how often all of the systems agree on the same answers on the MKQA test data in five languages. In particular, we compare all of the four system predictions on the English, Japanese, Chinese, Spanish and Turkish subsets of the MKQA test data, and check the prediction agreements based on the number of the unique predictions among the union of the predictions. We can see that in English and Spanish, the agreement is high (e.g., in 40% of the questions, all or three of the



(a) MKQA Answer agreement.

(b) Per-category agreement (En).

(c) Per-category agreement (Ja).

Figure 2: Answer agreements of the four submitted systems.

	Avg.		Arabic		Bengali		Finnish		Japanese		Korean		Russian		Telugu	
Sys.	cl	m	cl	m	cl	m	cl	m	cl	m	cl	m	cl	m	cl	m
(a)	27.2	58.8	28.3	64.8	29.4	63.3	30.3	51.3	29.5	55.9	22.2	53.2	24.9	55.8	25.9	67.6
(b)	21.4	54.2	22.2	65.2	21.7	44.6	24.1	51.7	27.4	55.2	19.5	49.3	19.1	50.8	16.0	62.2
(c)	20.3	52.5	21.5	64.6	20.5	42.3	25.5	50.9	26.3	53.1	17.6	44.9	16.0	51.2	14.7	60.3
(d)	19.5	49.7	22.4	59.0	19.0	51.5	23.8	46.9	25.2	50.2	15.3	38.8	16.9	47.0	14.0	54.2

Table 5: Final results on MIA 2022 Shared Task XOR-TyDi QA cross-lingual ("cl") / monolingual subsets ("m"). Systems (a), (b), (c) and (d) are Texttron, mLUKE-FID, CMUmQA, and ZusammenQA, respectively.

four systems agree on the same answers), while the agreement is lower in other languages, particularly in Japanese and Chinese.

To understand the phenomena, we breakdown the prediction agreement statistics in English and Japanese into different answer categories. Figure 2b and Figure 2c show per-category prediction agreements in English and Japanese, respectively. While in English, systems show high agreements in date, entity and number type questions, in Japanese, the agreement rate is lower across category, potentially because of their diverse formats of number and dates, as well as the transliteration of the entity names.

7.4 Error Analysis

We conduct a set of error analysis in five languages (i.e., English, Japanese, Korean, Chinese and Telugu) on randomly sampled 30 questions, where none of the submission systems' predictions exactly match any of the ground truth answers.

Error types. We classify the errors into following categories: (i) incorrect predictions, (ii) answers are semantically correct in different languages (incorrect languages), (iii) incorrect gold answers, (iv) semantically-equivalent predictions in the target language but are penalized because gold answers do not cover all of the potential gold answers (not comprehensive gold answers), (v) questions are open-ended or ambiguous (e.g., entity ambiguity), (vi) questions' granularity is unclear (unclear question granularity; e.g., year v.s. month, kilometers v.s. meters), (vii) questions are highly subjective (e.g., who is the best singer ever), (viii) temporal or geographical dependency in questions.

The first two error types, (i) and (ii), reveal the limitations of models. The error type (iii) and (iv) are considered answer annotation errors (Min et al., 2020a; Asai and Choi, 2021). The last four error types (v), (vi), (vii) and (viii) requires some specifications or context (Zhang and Choi, 2021; Min et al., 2020b).

Error analysis schema. We recruit native speakers of the five target languages and ask them to classify the errors into the aforementioned categories. We present the predictions of all of the systems as well as the intermediate retrieval results of the mLUKE + FiD.

Error analysis results. Table 6 provides the error analysis result. Besides modeling errors, we found that the original annotations themselves exhibit some issues, which underestimates models' performance. Across languages, annotators found non-negligible proportion of the errors happen as the original gold answers do not cover all of the possible answer aliases or the answer granularity is unclear. For instance, an English question asks "what is the temperature at the center of earth" and the gold answer is 6000 °C. Several systems answer in Fahrenheit or Kelvin, and got zero F1 score. Several questions are also temporal or geographical dependent such as "who was the last person appointed to the u.s. supreme court" or クリミナル・マインドの新シーズンが公開されるのはいつか (when is the next season of Criminal Minds will

	English	Arabic	Japanese	Korean	Chinese
(i) incorrect predictions	12	9	23	16	12
(ii) incorrect languages	0	2	3	0	2
(iii) incorrect gold answers	2	4	5	1	0
(iv) not comprehensive gold answers	10	1	7	5	6
(v) ambiguous question	3	7	6	15	5
(vi) unclear question granularity	3	2	1	2	0
(vii) subjective question	0	0	0	0	0
(viii) temporal or geographical dependency in questions	4	4	1	4	5

Table 6: Error analysis on sampled questions where all of the submissions unanimously fail to predict the correct answers. We show the percentage of the errors in each category.

be released?). Although situation-grounded QA has been recently studied (Zhang and Choi, 2021), there's little work that analyzes this phenomena in multilingual settings, where the particularly geographical dependence can be even more prevalent. Question ambiguity is also common in multilingual QA.

8 Conclusion and Discussions

We have presented the MIA 2022 Shared Task on cross-lingual open-retrieval QA systems in 16 typologically diverse languages, many of which are unseen during training. Several submissions improved significantly over our baseline based on a state-of-the-art cross-lingual open-retrieval QA system and investigated a wide range of techniques. Those results shed light on the effectiveness of several techniques in this challenging task, such as entity-enhanced representations, sparse-dense retrieval, and better interactions between passages. We further conducted detailed performance analysis on different subsets of the datasets, such as languages, answer types, the necessity of cross-lingual retrieval as well as detailed error analysis. We also suggest several bottlenecks in the area.

Acknowledgements

We would like to acknowledgments and Noah A. Smith for serving as our steering committee. We are grateful to Google for providing funding for our workshop. We thank GENGO translators to translate questions into Tamil and Tagalog. we thank the EvalAI team, particularly Ram Ramrakhya, for their help with hosting the shared task submission site. We thank Maraim Masoud for her help in error analysis.

References

Sumit Agarwal, Suraj Tripathi, Teruko Mitamura, and Carolyn Penstein Rose. 2022. Zero-shot cross-lingual open domain question answering. In *Proc. of MIA*.

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proc. of ACL*.

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *Proc. of ACL*.

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-guided generation for knowledge-intensive nlp tasks. In *In Proc. of NAACL*.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *Proc. of NAACL*.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. In *Proc. of NeurIPS*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proc. of ACL*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proc. of ACL*.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proc. of ACL: Tutorial Abstracts*.

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. *TACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. of ACL*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proc. of EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of EMNLP*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proc. of ICLR*.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan, and Chris Callison-Burch. 2022. A feasibility study of answer-unaware question generation for education. In *Findings of ACL*.

Martin Fajcik, Martin Docekal, Karel Ondrej, and Pavel Smrz. 2021. R2-D2: A modular baseline for open-domain question answering. In *Findings of EMNLP*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proc. of ACL*.

Pamela Forner, Danilo Giampiccolo, Bernardo Magnini, Anselmo Peñas, Álvaro Rodrigo, and Richard Sutcliffe. 2010. Evaluating multilingual question answering systems at CLEF. In *Proc. of LREC*.

Pamela Forner, Anselmo Peñas, Eneko Agirre, Iñaki Alegria, Corina Forăscu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe, and Erik Tjong Kim Sang. 2008. Overview of the clef 2008 multilingual question answering track. In *Proc. of CLEF*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proc. of ICML*.

Chia-Chien Hung, Tommaso Green, Robert Litschko, Tornike Tsereteli, Sotaro Takeshita, Marco Bombieri, Goran Glavaš, and Simone Paolo Ponzetto1. 2022. ZusammenQA: Data augmentation with specialized models for cross-lingual open-retrieval question answering system. In *Proc. of MIA*.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *Proc. of ICLR*.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proc. of EACL*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proc. of ACL*.

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 *Proc. of EMNLP*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *TACL*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. MLQA: Evaluating cross-lingual extractive question answering. In *Proc. of ACL*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proc. of NeurIPS*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020c. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. of NeurIPS*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proc. of EACL*.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *Proc. of ACL*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *TACL*.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei

Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen-tau Yih. 2020a. NeurIPS 2020 efficientQA competition: Systems, analyses and lessons learned. In *Proc. of NeurIPS 2020 Competition and Demonstration Track*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020b. AmbigQA: Answering ambiguous open-domain questions. In *Proc. of EMNLP*.

Shramay Palta, Haozhe An, Yifan Yang, Shuaiyi Huang, and Maharshi Gor. 2022. Investigating information inconsistency in multilingual open-domain question answering.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proc. of NAACL*.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proc. of ACL*.

Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. 2022. Information asymmetry in wikipedia across different languages: A statistical analysis. *JASIST*.

Peter Schäuble and Páraic Sheridan. 1997. Cross-language information retrieval (CLIR) track overview. In *Proc. of TREC*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proc. of NAACL*.

Zhucheng Tu and Sarguna Janani Padmanabhan. 2022. MIA 2022 shared task submission: Leveraging entity representations, dense-sparse hybrids, and fusion-in-decoder for cross-lingual question. In *Proc. of MIA*.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. SNCSE: Contrastive learning for unsupervised sentence embedding with soft negative samples.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for WMT17. In *Proc. of WMT*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL*.

Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. 2019. EvalAI: Towards better evaluation systems for ai agents.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proc. of EMNLP*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proc. of MRL*.