One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present **CORA**, a **C**ross-lingual **O**pen-**R**etrieval **A**nswer Generation model that can answer questions across many languages even when language-specific annotated data or knowledge sources are unavailable. We introduce a new dense passage retrieval algorithm that is trained to retrieve documents across languages for a question. Combined with a multilingual autoregressive generation model, CORA answers directly in the target language without any external translation or in-language retrieval modules as used in prior work. To train CORA, we extend available annotated data using an iterative training approach that improves retrieval for languages with limited resources. Our results show that CORA substantially outperforms the previous state of the art on multilingual open question answering benchmarks across 26 languages, 9 of which are unseen during training. Our analyses show the significance of cross-lingual retrieval and generation in many languages, particularly under low-resource settings. We will publicly release our code and model.

1 Introduction

Multilingual open QA is the task of answering a question from a large collection of multilingual documents. Most recent progress in open QA is made for English by building a pipeline consisting of a dense passage retriever trained on large-scale English QA datasets to find evidence passages in English (Lee et al., 2019; Karpukhin et al., 2020), followed by a reader to extracts the answer from retrieved passages. However, extending this approach to multilingual open QA poses new challenges. Answering many questions requires retrieving answers in knowledge sources of other languages than the original question since many languages have limited reference documents or the question inquires about concepts from other cultures (Asai et al., 2021; Lin et al., 2020). Yet, large-scale open QA training data are not available across languages.

To address these challenges, previous work in multilingual open QA (Ture and Boschee, 2016; Asai et al., 2021) translates questions into English, applies an English open QA system to answer in English, and then translates answers back to the target language. This pipeline approach suffers from error propagation of the machine translation component into the downstream QA, especially for low-resource languages. Moreover, it is not able to answer questions whose answers can be found in resources written in languages other than English or the target languages. For example, when a Japanese question asks about a Korean actor, it is useful to retrieve passages from Korean Wikipedia, not just English and Japanese. In this paper, we introduce a single QA model that answers questions in any language leveraging an iterative training and data expansion approach to enable cross-lingual retrieval and answer generation.

Our method (called CORA, Fig. 1) extends the *retrieve-then-generate* approach of English open QA with a single cross-lingual retriever and a generator that do not rely on language-specific or machine

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

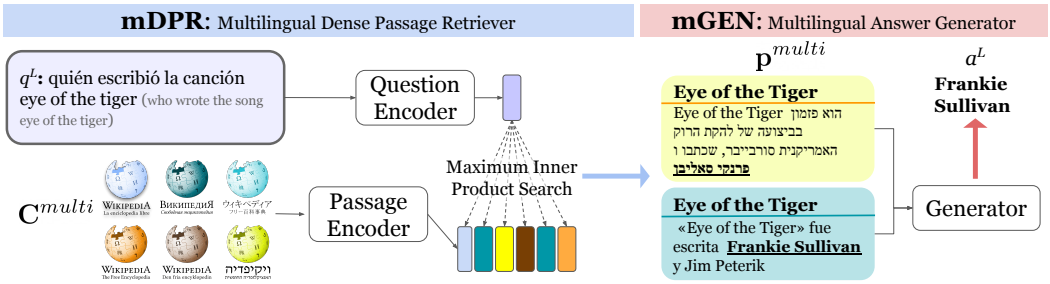

Figure 1: Overview of CORA using mDPR and mGEN.

translation modules. The multilingual retrieval module produces dense embeddings of a question and all multilingual passages, thereby retrieving passages across languages. The generation component is trained to output an answer in the target language conditioned on the retrieved multilingual passages. Those two components are iteratively trained to address limitations of the current multilingual QA datasets. In particular, each iteration proceeds over two stages by updating model parameters with available training data and expanding the training data cross-lingually by Wikipedia language links and model predictions. This approach does not require any additional human annotations or machine translation, and can be applied to many new languages with low resources.

Our experiments show that CORA advances the state of the art on two multilingual open QA datasets, XOR-TYDI QA (Asai et al., 2021) and MKQA (Longpre et al., 2020), across 26 typologically diverse languages; CORA achieves gains of 21.4 and 4.9 F1 points in XOR-TYDI QA and MKQA respectively, where MKQA data is not used for training. Moreover, CORA achieves F1 scores of roughly 30 over 9 languages on MKQA that have no training data or even reference Wikipedia documents, outperforming the state-of-the-art approach by 5.4 F1. Our controlled experiments and human analyses illustrate the impact of cross-lingual retrieval in improving QA performance. For example, we observe that CORA can find answers to 20% of the multilingual QA questions that are valid but are originally labeled as unanswerable, by using cross-lingual retrieval.

2 Method

We define multilingual open QA as the task of answering a question q^L in a target language L given a collection of multilingual reference passages $\mathbf{C}^{multi}$, where evidence passages can be retrieved from *any language*. This definition extends the open QA setting proposed by Asai et al. (2021) as they assume the reference documents are either in English or the target language. These passages can come from Wikipedia articles that are not necessarily parallel over languages. We introduce CORA, which runs a *retrieve-then-generate* procedure to achieve this goal (Fig. 1).

2.1 CORA Inference

CORA directly retrieves multilingual passages for questions in any language and generates answers in the target language conditioned on those retrieved passages. More formally, CORA inference consists of two steps of (i) retrieving passages $\mathbf{p}^{multi}$ and (ii) generating the answer a^L from the retrieved passages.

$$\mathbf{p}^{multi} = \text{mDPR}(q^L, \mathbf{C}^{multi}), a^L = \text{mGEN}(q^L, \mathbf{P}^{multi}),$$

Multilingual Dense Passage Retriever (mDPR). mDPR extends Dense Passage Retriever (DPR; Karpukhin et al., 2020) to a multilingual setting. mDPR uses an iterative training approach to fine-tune a pre-trained multilingual language model (e.g., mBERT; Devlin et al., 2019) to encode passages and questions separately (§ 2.2). Once training is done, the representations for all passages from $\mathbf{C}^{multi}$ are computed offline and stored in local storage. At inference, mDPR independently obtains d-dimensional ($d = 768$) encodings of the question $\mathbf{e}_{q^L} = \text{mBERT}_q(q^L)$ and a passage $\mathbf{e}_{p^L} = \text{mBERT}_p(p)$, where a passage p is a fixed-length sequence of tokens from multilingual documents. It retrieves k passages with the k highest relevance scores to the question, where the

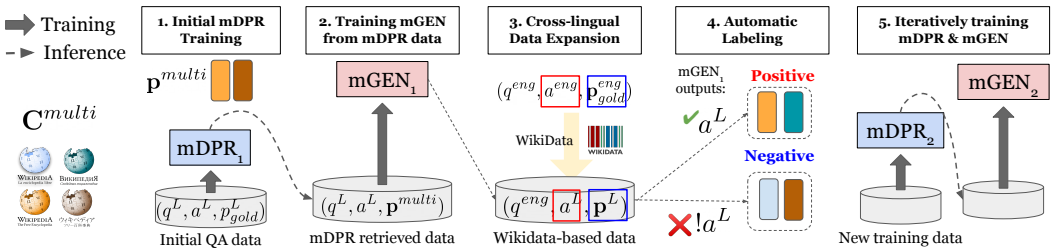

Figure 2: Overview of iterative training and data expansion.

relevance score between a passage p and a question q^L is estimated by the inner product of the encoding vectors, $\langle \mathbf{e}_{q^L}, \mathbf{e}_p \rangle$. The retrieved passages $\mathbf{p}^{multi}$ can be in any language in $\mathbf{C}^{multi}$.

Multilingual Answer Generator (mGEN). We use a multilingual sequence-to-sequence model (e.g., mT5; Xue et al., 2021) to generate answers in the target language token-by-token autoregressively given the retrieved multilingual passages $\mathbf{p}^{multi}$. We choose a generation approach because it can generate an answer in the target language L from passages across different languages.[1] Moreover, the generator can be adapted to new and unseen languages, some of which may have little or no translation training data. Specifically, the generator outputs the sequence probability for a^L as follows:

$$P(a^L | q^L, \mathbf{p}^{multi}) = \prod_{i}^{T} p(a^L_i | a^L_{<i}, q^L, \mathbf{p}^{multi}), \tag{1}$$

where a^L_i denotes the i-th token in the answer, and T is the length of the answer. We append a language tag to the question to indicate the target language. We can remove the dependency on $\mathbf{p}^{multi}$ and skip the retrieval step (i.e., *close-book* model), but this results in a performance drop (§ 4).

2.2 CORA Training

We introduce an iterative training approach that encourages cross-lingual retrieval and answer generation conditioned on multilingual passages (sketched in Fig. 2 and Alg. 1). Each iteration proceeds over two stages: **parameter updates** (§ 2.2.1) where mDPR and mGEN are trained given the current training data and **cross-lingual data expansion** (§ 2.2.2) where training data are automatically expanded by Wikipedia language links and model predictions. Our data expansion addresses the limitation of the current multilingual QA datasets.

Initial training data. Training data is initialized by combining multilingual QA (TyDi QA, Clark et al., 2020 and Xor-TyDi QA) and English open QA datasets (Natural Questions, Kwiatkowski et al., 2019). One training instance from these datasets comprises a question, a positive passage, and an answer. For each question, we use positive passages for the other questions in the training batch as negative passages (Karpukhin et al., 2020). Note that annotations in the multilingual QA datasets are limited: positive passages are taken either from English (Asai et al., 2021) or the question's language (Clark et al., 2020). Further, most of the non-English languages are not covered. Indeed, when we only train mDPR on this initial set, it learns to retrieve passages in the same languages or similar languages with irrelevant context.

2.2.1 Parameter Updates

mDPR updates. (line 3 in Alg. 1). Let $\mathcal{D} = \{\langle q^L_i, p^+_i, p^-_{i,1}, \cdots, p^-_{i,n} \rangle\}_{i=1}^{m}$ be m training instances. Each instance consists of a question q^L_i, a passage that answers the question (positive passage) p^+_i, and n passages that do not answer the question (negative passages) $p^-_{i,j}$. mDPR is updated by minimizing the negative log likelihood of positive passages:

$$\mathcal{L}_{\text{mdpr}} = -\log \frac{\exp(\langle \mathbf{e}_{q^L_i}, \mathbf{e}_{p^+_i} \rangle)}{\exp(\langle \mathbf{e}_{q^L_i}, \mathbf{e}_{p^+_i} \rangle) + \sum_{j=1}^{n} \exp(\langle \mathbf{e}_{q^L_i}, \mathbf{e}_{p^-_{i,j}} \rangle)}. \tag{2}$$

[1] An alternative approach of answer extraction requires translation for all language pairs (Asai et al., 2021).

Algorithm 1: Iterative training that automatically mines training data.

Data: Input QA pairs: (q^L, a^L)

1 initialize training data $\mathbf{B}^1 = (q^{\mathbf{L}}, a^{\mathbf{L}}, p_{gold}), \mathbf{L} = \{\text{Eng}, L\}$;

2 **while** $t < T$ **do**

3 $\quad \Theta^t_{mDPR} \leftarrow Train(\theta^{t-1}_{mDPR}, \mathbf{B}^t)$ /* Train mDPR */

4 $\quad \mathbf{p}^{multi} \leftarrow \text{mDPR}(q^{\mathbf{L}}, \text{embedding}(\mathbf{C}^{multi}))$ /* Retrieve passages */

5 $\quad \theta^t_{mGEN} \leftarrow Train(\theta^{t-1}_{mGEN}, (q^{\mathbf{L}}, a^{\mathbf{L}}, \mathbf{p}^{multi}))$ /* Train mGEN */

6 $\quad$ For $\mathbf{L}$ == Eng, $\mathbf{p}^{multi}$+ = LangLink$(q^{\mathbf{L}}, \mathbf{C}^{multi})$) /* Mine data using Wikidata */

7 $\quad$ **For** $p_i \in \mathbf{p}^{multi}$: **if** mGEN$(q^{\mathbf{L}}, p_i)$ == a^L **then** $positives.add(p_i)$ **else** $negatives.add(p_i)$

8 $\quad \mathbf{B}^{t+1}$ += $(q^{\mathbf{L}}, a^{\mathbf{L}}, positives, negatives)$ /* Add new training data */

9 $\quad t \leftarrow t + 1$

10 **end**

mGEN updates (lines 4-5 in Alg. 1). After updating mDPR, we use mDPR to retrieve top k passages $\mathbf{p}^{multi}$ for each q^L. Given these pairs of the question and the retrieved passages $(q^L, \mathbf{p}^{multi})$ as input, mGEN is trained to generate answer a^L autoregressively (Eq. (1)) and minimize the cross-entropy loss. To train the model to generate in languages not covered by the original datasets, we translate a^L to other languages using Wikipedia language links and create new synthetic mGEN training data.[2]

2.2.2 Cross-lingual Data Expansion

After the parameter updates, we mine new training data using mDPR and Wikipedia language links and label the new data by mGEN predictions. This step is skipped in the final iteration.

Expansion by trained mDPR and language links (line 6 in Alg. 1). Trained mDPR can discover positive passages in another language that are not covered by initial training data. At each iteration, we use retrieved passages $\mathbf{p}^{multi}$ for q^L (lines 4 in Alg. 1) as a source of new positive and negative passages. This enables expanding data between language pairs not in the original data. To add more languages, we use language links between Wikipedia articles and Wikidata entities.

Wikipedia maintains article-level language links that connect articles on the same entity over languages. We use these links to expand training data from the English QA dataset of Natural Questions. Denote a training instance by $(q^{Eng}, a^{Eng}, p_{gold})$. We first translate the English answer a^{Eng} to a target language a^L by language links. We use language links again to look up the English Wikipedia article that the gold passage p_{gold} comes from. We then find articles in non-English languages in the reference documents that correspond to this article. We add passages from these articles to $\mathbf{p}^{multi}$ as positive passage candidates, which are passed to mGEN to evaluate if they lead to a^L or not.

Automatic labeling by mGEN predictions (lines 7-8 in Alg. 1). A passage p_i from $\mathbf{p}^{multi}$ may not always provide sufficient information to answer the question q^L. We take instances generated from the two expansion methods described above, and run mGEN on each passage to predict an answer for the question. If the answer matches the correct answer a^L, then the passage p_i is labeled as a *positive passage*; otherwise we label the input passage as a *negative passage*. We add these new positive and negative passages to the training data, and in the next iteration, mDPR is trained on this expanded training set (§ 2.2.1).

3 Experiments

We evaluate CORA on two multilingual open QA datasets across 28 typologically diverse languages.[3] CORA achieves state-of-the-art performance across 26 languages, and greatly outperforms previous approaches that use language-specific models such as question translation.

[2]This data augmentation is only done after the third epoch of initial training to prevent the model from overfitting to synthetic data. See the appendix for more detail.

[3]A full list of the language family and script types are in the appendix.

3.1 Datasets and Knowledge Sources

Multilingual open QA datasets differ in covered languages, annotation schemes, and target application scenarios. We evaluate F1 and EM scores over the questions with answer annotations from two datasets, following the common evaluation practice in open QA (Lee et al., 2019).

XOR-TYDI QA. XOR-TYDI QA (Asai et al., 2021) is a multilingual open QA dataset consisting of 7 languages, where questions are originally from TYDI QA (Clark et al., 2020) and posed by information-seeking native speakers. The answers are annotated by extracting spans from Wikipedia in the same language of the question (*in-language data*) or by translating English spans extracted from English Wikipedia to the target language (*cross-lingual data*).

MKQA. MKQA (Longpre et al., 2020) is an evaluation dataset created by translating 10k Natural Questions (Kwiatkowski et al., 2019) to 25 target languages. The parallel data enables us to compare the models' performance across typologically diverse languages. This dataset does not have any training data, and only five languages are covered in the XOR-TYDI QA training data.

Collection of multilingual documents. We use the Wikipedia dumps of 13 diverse languages from all XOR-TYDI QA languages and a subset of MKQA languages, leading to about 31.5M passages.[4] Several MKQA languages are not included so that we can evaluate over unseen languages and make indexing and inference computationally feasible. We chose the 13 languages to cover languages with a large number of Wikipedia articles and a variety of both Latin and non-Latin scripts.

Training-evaluation language splits. CORA is trained on XOR-TYDI QA; mDPR uses gold passages from languages in XOR-TYDI QA and also synthetically mined data from 13 languages (as explained in § 2.2.2). The mGEN model is trained on XOR-TYDI QA and synthetically generated data for 17 languages (§ 2.2.1). We call languages on which we train CORA with annotated data or synthetically generated data *seen* languages, while call others *unseen* languages. MKQA is only used for evaluation, and 9 of its languages are never included during training.

3.2 Baselines and Experimental Setting

We compare CORA with the following strong baselines adopted from Asai et al. (2021).

Translate-test (MT + DPR). As used in most previous work (e.g., Asai et al., 2021), this method translates the question to English, extracts an answer in English using DPR, and then translates the answer back to the target language. The translation models are obtained from MarianMT (Junczys-Dowmunt et al., 2018) and trained on the OPUS-MT dataset (Tiedemann, 2012).

Monolingual baseline (BM 25). This baseline retrieves passages in the target language and extracts the answer from the retrieved passages. Due to the lack of training data in most of the target languages, we use a BM25-based lexical Pyserini implementation (Lin et al., 2021). We then feed the retrieved documents to a multilingual QA model to extract final answers.

MT+Mono. This baseline combines results from the translate-test method (MT+DPR) and the monolingual method (BM25) to enable retrieving passages in both English and the target language. Following Asai et al. (2021), we prioritize predictions from the monolingual pipeline if they are over a certain threshold; otherwise we output predictions from the translate-test method.[5]

Closed-book baseline. This model uses an mT5-base sequence-to-sequence model that takes a question as input and generates answers in the target language without any retrieval component at inference time (Roberts et al., 2020). This baseline assesses the models' ability of memorizing and retrieving knowledge from its parameters without using the knowledge from reference documents.

CORA details. For all experiments, we use a single retriever (mDPR) and a generator (mGEN) that use the same passage embeddings. mDPR uses multilingual-BERT-base-uncased,[6] and the generator uses mT5-base for fine-tuning. We first fine-tune mDPR using gold passages from Natural Question, and then further fine-tune it using XOR-TYDI QA and TYDI QA's gold passage data. We exclude

[4]Downloaded from `https://archive.org/details/wikimediadownloads?and%5B%5D=year%3A%222019%22`.

[5]For the languages not supported by Pyserini, we always output translate-test' predictions.

[6]XLM-RoBERTa (Conneau et al., 2020) did not improve our results.

Models	Target Language L_i F1							Macro Average		
	Ar	**Bn**	**Fi**	**Ja**	**Ko**	**Ru**	**Te**	**F1**	**EM**	**BLEU**
CORA	**54.8**	**35.5**	**41.3**	**39.2**	**22.3**	**44.8**	**42.5**	**40.1**	**31.1**	**28.9**
state-of-the-art	31.5	19.0	18.3	8.8	20.1	19.8	13.6	18.7	12.1	16.8
MT+Mono	25.1	12.7	20.4	12.9	10.5	15.7	0.8	14.0	10.5	11.4
MT+DPR	7.6	5.9	16.2	9.0	5.3	5.5	0.8	7.2	3.3	6.3
BM25	31.1	21.9	21.4	12.4	12.1	17.7	–	–	–	–
Closed-book	14.9	10.0	11.4	22.2	9.4	18.1	10.4	13.8	9.6	7.4

Table 1: Performance on XOR-FULL (test data F1 scores and macro-averaged F1, EM and BLEU scores). The previous state-of-the-art results are from the baseline that combines Google Custom Search in the target language and Google Translate + English DPR for cross-lingual retrieval (Asai et al., 2021). Pyserini do not support Telugu.

Setting	–	Included in XOR-TYDI QA						Seen by mDPR and mGEN			
	Avg. over all L.	**En**	**Ar**	**Fi**	**Ja**	**Ko**	**Ru**	**Es**	**Sv**	**He**	**Th**
CORA	22.0	39.8	13.1	**28.3**	**16.6**	**8.9**	**22.0**	**33.4**	**31.6**	**15.3**	**9.1**
MT+Mono	14.1	19.3	6.9	17.5	9.0	7.0	10.6	21.3	20.0	8.9	8.3
MT+DPR	17.1	**43.3**	**16.0**	21.7	9.6	5.7	17.6	28.4	19.7	8.9	6.9
BM25	–	19.4	5.9	9.9	9.1	6.9	8.1	14.7	10.9	–	4.9
Closed	4.5	8.0	4.6	3.6	6.5	3.8	4.1	6.6	4.8	3.8	2.1

Table 2: F1 scores on MKQA seen languages over 6.7k questions with short answer annotations.

the training questions in Natural Questions and TYDI QA used to create the MKQA or XOR-TYDI QA evaluation set. We run two iterations of CORA training. All hyperparameters are in the appendix.

4 Results and Analysis

4.1 QA Results

XOR-TYDI QA. Table 1 reports the scores of CORA and the baselines in XOR-TYDI QA. CORA, which only uses a single retriever and a single generator, outperforms the baselines and the previous state-of-the-art model on XOR-TYDI QA by a large margin across all 7 languages. CORA achieves a gain of 21.4 macro-averaged F1 points over the previous state-of-the-art method, which uses the black-box APIs of Google Translate and Google Custom Search.

MKQA. Tables 2 and 3 report the F1 scores of CORA and the baselines on MKQA under *seen* and *unseen* settings.[7] Longpre et al. (2020) only report results in a simplified setting with gold passages given in advance, where retrieval is unnecessary. CORA significantly outperforms the baselines in all languages by a large margin except for Arabic and English. CORA yields larger improvements from the translate-test baseline in the languages that are distant from English and with limited training data such as Malay (MS; 29.2 vs. 12.6) and Hebrew (He; 15.3 vs. 8.9). The performance drop of the translate-test model from English to other languages indicates the error propagation from the translation process. BM25 performs very poorly in some low-resource languages such as Thai because of the lack of answer content in the target languages' Wikipedia. MT+Mono underpeforms the MT+DPR baseline in MKQA since it is challenging to rerank answers from two separate methods with uncaliberated confidence scores. In contrast, CORA uses a single retrieval model to retrieve passages across languages, achieving around 30 F1 in a majority of the 26 languages.

Ablations: Impact of CORA components. We compare CORA with the following three variants to study the impact of different components on XOR-TYDI QA and a subset of MKQA (~5% of the evaluation set). (i) **mDPR$_1$ + mGEN$_1$** only trains CORA using labeled annotated data and

[7] Following previous work in open QA, but different from the official script of MKQA (Longpre et al., 2020), we disregard the questions labeled as "no answer" to ensure that a model cannot increase aggregate F1 and EM scores simply by predicting "no answer." As shown in our human analysis, it is difficult to prove an answer does not exist in the millions of multilingual documents even if the annotation says so.

Setting	Seen in mGEN							Unseen								
	Da	**De**	**Fr**	**It**	**Nl**	**Pl**	**Pt**	**Hu**	**Vi**	**Ms**	**Km**	**No**	**Tr**	**cn**	**hk**	**tw**
CORA	**29.7**	**29.9**	**31.5**	**28.7**	**32.3**	**26.1**	**28.6**	**21.1**	**21.8**	**29.2**	**5.4**	**27.9**	**22.8**	**5.8**	**8.1**	**6.2**
MT+Mono	19.3	21.6	21.9	20.9	21.5	24.6	19.9	16.5	15.1	12.6	1.2	17.4	16.6	5.4	3.8	5.9
MT+DPR	26.2	25.9	21.9	25.1	28.3	24.6	24.7	15.7	15.1	12.6	1.2	18.3	18.2	3.3	3.8	3.8
BM25	9.5	12.5	–	13.6	12.8	–	13.4	7.4	–	–	–	9.4	8.8	2.8	–	3.3
Closed	4.7	5.6	5.8	5.3	5.5	4.0	4.4	5.5	5.9	5.3	1.9	4.1	3.8	2.6	2.3	2.4

."cn" denotes "'zh-cn' (Chinse, simplified), "hk" denotes "zh-hk" (Chinese, Hong Kong) and "tw" denotes "zh-tw" (Chinese, traditional).

Table 3: F1 scores on MKQA in languages unseen by mDPR and not included in $\mathbf{C}^{multi}$

Setting	XOR-TYDI QA				MKQA					
	Avg. F1	Ar	Ja	Te	Avg. F1	Fi	Hu	Es	Th	Ms
CORA	**31.4**	**42.6**	**33.4**	**26.1**	**23.2**	**30.3**	**21.5**	**34.4**	**7.1**	**31.8**
(i) mDPR_1 + mGEN_1	27.9	36.2	29.8	21.1	17.3	23.1	19.4	28.5	5.7	22.4
(ii) DPR (trained NQ)+mGEN	24.3	30.7	29.2	19.0	17.9	20.1	16.2	29.4	5.5	22.2
(iii) CORA, $\mathbf{C}^{multi}$={Eng.}	18.1	19.8	22.3	14.8	18.4	22.7	16.0	31.1	5.1	27.3

Table 4: Ablation studies on XOR-TYDI QA development set and a subset of MKQA (350 questions for each of 26 languages).

measures the impact of the iterative training. (ii) **DPR (trained NQ) + mGEN** replaces mDPR with a multilingual BERT-based DPR trained on English data from Natural Questions (NQ) and encodes all passages in $\mathbf{C}^{multi}$. This configuration assesses the impact of cross-lingual training data. (iii) **CORA, $\mathbf{C}^{multi}$={Eng.}** only retrieves from English during inference. This variant evaluates if English reference documents suffice to answer multilingual questions.

Results in Table 4 show performance drops in all three variants. This leads to the following conclusions: (i) iterative learning and data mining process is useful, (ii) mDPR trained with cross-lingual data outperforms DPR with multi-lingual BERT, and (iii) languages other than English are important in answering multilingual questions.

4.2 Analysis

	Setting	**Seen**					**Unseen**					
	Lang	Es	Fi	Ja	Ru	Th	Pt	Ms	Tr	Zh-Cn	Zh-Hk	Km
	Script	Latn		Jpan	Cyrl	Thai	Latn			Hant		Khmr
mDPR	R^{L}@10	54.0	49.1	21.1	40.3	17.7	47.4	46.8	40.3	10.9	16.3	13.1
	R^{multi}@10	62.6	55.1	42.0	54.0	31.1	59.7	58.6	53.1	43.1	44.6	23.1
DPR(NQ)	R^{L}@10	52.3	46.0	24.6	36.0	12.6	45.7	48.8	32.0	9.1	14.0	13.4
	R^{multi}@10	63.1	53.1	32.9	49.1	29.4	56.8	58.0	44.0	36.3	39.4	23.4

Table 5: Retrieval recall performance on MKQA as the percentage of the questions where at least one paragraph includes answer string in the target language R^{L}@10, or in any language R^{multi}@10. The results are on the same subset as in the ablations.

Retrieval performance and relationship to the final QA performance. We evaluate CORA's retrieval performance on MKQA using two recall metrics that measure the percentage of questions with at least one passage among the top 10 that includes a string in an answer set in the target language (R^{L}@10) or in the union of answer sets from all languages that is available in MKQA (R^{multi}@10).

Table 5 reports retrieval results for mDPR and multilingual-bert based DPR trained on NQ (**DPR (NQ)**, which is equivalent to (ii) from ablations). We observe that mDPR performs well in Indo-European languages with Latin script, even when the language is unseen. Interestingly, there is a significant performance gap between R^{L}@10 and R^{multi}@10 in languages with non-Latin script (e.g., Japanese, Russian, Chinese), which can indicate that our model often retrieves relevant documents from other languages with Latin scripts such as English or Spanish in those languages as well. Our mDPR outperforms **DPR (NQ)** by a large margin in unseen languages with limited resources, as

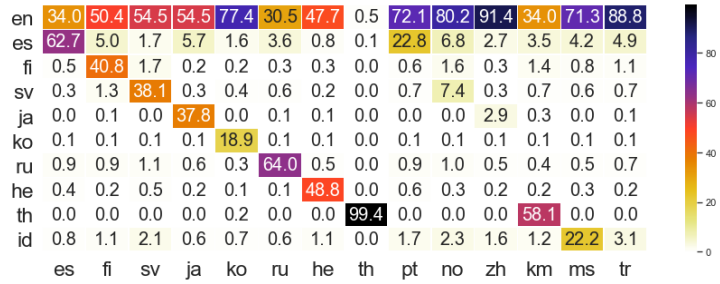

Figure 3: Breakdown of the languages of retrieved reference passages for sampled MKQA questions (%). The x and y axes indicate target (question) and target languages respectively.

	Ja	Es
retrieval errors	28	48
different lang	18	0
incorrect answer	22	36
annotation error	22	12
invalid question	10	4

Table 6: Error categories (%) on 50 errors sampled from Japanese (Ja) and Spanish (Es) data.

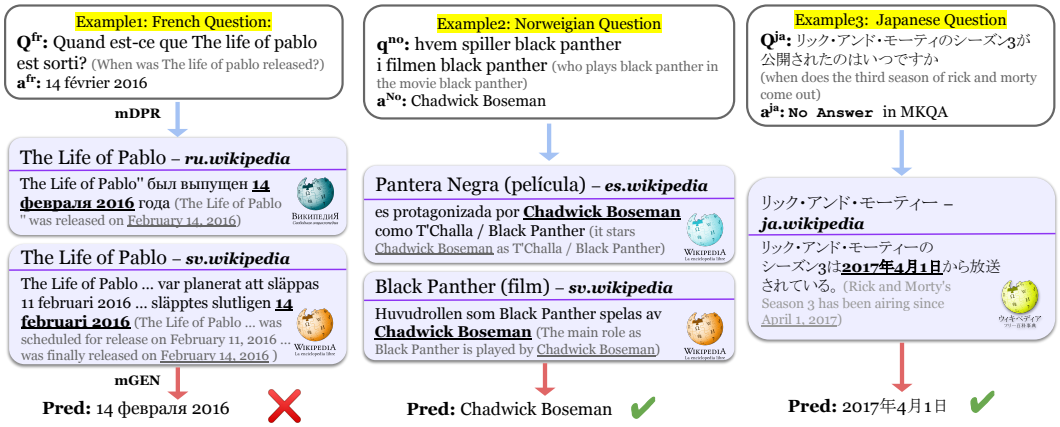

Figure 4: Cross-lingual retrieval and generation examples for thee MKQA questions.

shown in Table 3. Nevertheless, we still see low performance in Khmer or Thai even with R^{multi}@10 metric. We also observe that passage and query embeddings for those languages are far from other languages, which can be further studied in a future work. We provide two-dimensional encoded document representations in Appendix.

Visualizing cross-lingual retrieval. Fig. 3 breaks down retrieved reference languages for each target language. Our multilingual retrieval model often retrieves documents from the target language (if its reference documents are available), English, or its typologically similar languages. For example, CORA often retrieves Spanish passages for Portuguese questions and Japanese passages for Chinese questions; Japanese and Chinese overlap in script.

To further evaluate this, we conduct a controlled experiment: we remove Spanish, Swedish and Indonesian documents' embeddings and evaluate CORA on relevant languages Danish, Portuguese and Malay. We observe performance drops of 1.0 F1 in Danish, 0.6 F1 in Portuguese, and 3.4 F1 in Malay. This illustrates that cross-lingual retrieval from closer languages in fact helps CORA to achieve better performance in those languages.

Error analysis and qualitative examples. Table 6 analyzes errors made by CORA by manually inspecting 50 Japanese and Spanish wrong predictions from MKQA. We observe six error categories: (a) retrieval error, (b) generation language error (different lang), (c) incorrect answer generation (incorrect answer), (d) answer annotation errors (e.g., a correct alias isn't covered by gold answers, or Wikipedia information is inconsistent with English.), and (e) ambiguous or ill-formed questions such as "who won X this year" (invalid question). The table shows that in both Japanese and Spanish, the retrieval errors are dominant. In Japanese, CORA often generates the correct answers in English, not in Japanese, which hurts its final performance.

Fig. 4 shows some qualitative examples. The first example shows an error in (b): mGEN is generating an answer in Russian, not in French though the answer itself is correct. This type of errors happen especially when retrieved passages are in other languages than the target and English.

Human evaluation on cross-lingual retrieval results. To observe how cross-lingual retrieval between distant languages is actually helping, we sample 25 Norwegian questions for which Spanish passages are included among the top 10 retrieved results. As seen in Fig. 3, CORA retrieves Spanish passages for 6.8% of the Norwegian questions. A bilingual Spanish native speaker judges if the retrieved Spanish paragraphs actually answer the given Norwegian questions. We found that in 96% of the cases, the retrieved Spanish paragraphs are relevant in answering the question. One such example is presented in Fig. 4, second example.

Human analysis on unanswerable questions. CORA retrieves passages from a larger multilingual document collection, which may further improve the answer coverage. MKQA includes the questions that are marked as unanswerable by English native speakers given English knowledge sources. We sample 400 unanswerable Japanese questions whose top one retrieved passage is from non-English Wikipedia articles. Among these 329 questions are ill formed (also discussed in Asai and Choi, 2021). For 17 out of the 71 remaining questions, CORA's answer is correct. This finding indicates the significance of cross-lingual retrieval and potential room for improvement in annotating multilingual open QA datasets. The third example in Fig. 4 shows one of these examples.

5 Related Work and Broader Impacts

English and non-English open QA. Despite the rapid progress in open QA (Chen et al., 2017; Karpukhin et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021), most prior work has been exclusively on English. Several prior attempts to build multilingual open QA often rely on additional MT or language-specific retrieval models (Asai et al., 2021; Ture and Boschee, 2016).

Many-languages-one models. Recently, several work introduced single multilingual models for many languages by using pre-trained multilingual models such as mBERT or mT5 in several NLP tasks (e.g., entity linking: Botha et al., 2020; De Cao et al., 2021; semantic role labeling: Mulcaire et al., 2019; Lyu et al., 2019; Fei et al., 2020). This work conducts the first large-scale study of a unified multilingual open QA model across many languages, and we show state-of-the-art performance in 26 typologically diverse languages.

Synthetic data creation for multilingual machine reading comprehension. Several recent work augment multilingual machine reading comprehension training data by generating new question and answer pairs from randomly sampled non-English Wikipedia paragraphs using a sequence-to-sequence model (Riabi et al., 2020; Shakeri et al., 2020). This work focuses on multilingual open QA, which involve cross-lingual retrieval going beyond extracting answers from multilingual passages. To this end, we augment existing training data by using trained models and language links, which can be further improved by applying similar question generation methods.

Societal impacts. We will make our code publicly available. CORA can perform open QA in unseen languages and could benefit the society in building QA systems for low-resource languages such as African languages, hence enabling research in that direction. By removing the necessity of using external blackbox APIs, we can examine and address wrong answers because of its prediction error or misinformation that exists in Wikipedia. This would help us to mitigate the potential negative impact by CORA outputting a wrong answer when it used by people who seek information.

6 Conclusion

To address the information needs from many non-English speakers, a QA system has to conduct cross-lingual passage retrieval. This work presents CORA, a unified multilingual open QA model that retrieves multilingual passages and generates answers in the target languages, conditioned on the retrieved passages. CORA does not require language-specific translation or retrieval components and can even answer questions in unseen new languages. We conduct extensive experiments on two multilingual open QA datasets across 28 languages, 26 of which CORA advances the state of the art, outperforming competitive models by up to 20 F1 points. Our extensive analysis and manual evaluation reveal that CORA effectively retrieves semantically relevant passages beyond language boundaries, and can even find answers to the questions that are previously considered unanswerable due to lack of sufficient evidence in annotation languages.

References

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

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

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *EMNLP*.

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

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 *ACL*.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2021. Multilingual autoregressive entity linking. *arXiv preprint arXiv:2103.12528*.

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

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *ACL*.

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

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *ACL (System Demonstrations)*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*.

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. *TACL*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

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

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *arXiv preprint arXiv:2102.10073*.

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

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

Chunchuan Lyu, Shay B. Cohen, and Ivan Titov. 2019. Semantic role labeling with iterative structure refinement. In *EMNLP-IJCNLP*.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *NAACL-HLT*.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2020. Synthetic data augmentation for zero-shot cross-lingual question answering. *arXiv preprint arXiv:2010.12643*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Siamak Shakeri, Noah Constant, Mihir Sanjay Kale, and Linting Xue. 2020. Zero-shot multilingual synthetic question and answer generation for cross-lingual reading comprehension. *arXiv preprint arXiv:2010.12008*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*.

Ferhan Ture and Elizabeth Boschee. 2016. Learning to translate for multilingual question answering. In *EMNLP*.

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 *NAACL*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section § 4.
   (b) Did you describe the limitations of your work? [Yes] See Section § 4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section § 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will submit our code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All hyperparameters will be included in appendix. Data splits are from the original datasets.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] That information will be included in Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section § 3.

(b) Did you mention the license of the assets? [Yes] Both two datasets are under MIT license and we will add that information in Appendix.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We will submit our code as well as our human annotated results from § 4.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section § 4

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] As the dataset used in this work is originally created for general purpose QA and we believe that those datasets contain few examples that risk our participants. We will describe that in details in Appendix.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We will include the instructions in Appendix.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] As the dataset used in this work is originally created for general purpose QA and we believe that those datasets contain few examples that risk our participants. We will describe that in details in Appendix.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]