Zero-shot cross-lingual open domain question answering

Sumit Agarwal, Suraj Tripathi, Teruko Mitamura, Carolyn Penstein Rose
{sumita, surajt, teruko, cprose}@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University

Abstract

People speaking different kinds of languages search for information in a cross-lingual manner. They tend to ask questions in their language and expect the answer to be in the same language, despite the evidence lying in another language. In this paper, we present our approach for this task of cross-lingual open-domain question-answering. Our proposed method employs a passage reranker, the fusion-in-decoder technique for generation, and a wiki data entity-based post-processing system to tackle the inability to generate entities across all languages. Our end-2-end pipeline shows an improvement of 3 and 4.6 points on F1 and EM metrics respectively, when compared with the baseline CORA model on the XOR-TyDi dataset. We also evaluate the effectiveness of our proposed techniques in the zero-shot setting using the MKQA dataset and show an improvement of 5 points in F1 for high-resource and 3 points improvement for low-resource zero-shot languages. Our team, CMUmQA's submission in the MIA-Shared task ranked 1st in the constrained setup for the dev and 2nd in the test setting.

1 Introduction

Question Answering (QA), especially in English, is a popular research area in NLP with abundance of datasets like SQuAD (Rajpurkar et al., 2018), Natural Questions (Kwiatkowski et al., 2019) and different types of tasks including machine reading comprehension or extractive QA, cloze-completion and open-domain QA (Richardson et al., 2013; Chen et al., 2017). Open-domain QA is the task of answering natural language questions without any specified predefined context. It usually requires the system to first search for the relevant documents as the context w.r.t. a given question from either a local document repository or Wikipedia-like document collection, and then generate the answer.

Cross-lingual Open-Domain Question Answering is a challenging NLP task, where questions are given in a user's preferred language, and the system needs to find evidence in cross-lingual large-scale document collections, like Wikipedia, and return an answer in the user's preferred language, as indicated by their question. We work on this cross-lingual open-domain QA challenge as a part of the MIA Shared task. [1]

Recent advancements in Open-Domain QA, usually for English are made by following a Retriever-Reader architecture, where the retriever is aimed at retrieving relevant documents w.r.t. a given question, which can be modeled as a dense passage retriever trained on large-scale English QA datasets to fetch evidence passages (Karpukhin et al., 2020), while Reader aims at inferring the final answer from the retrieved documents, which is usually a neural MRC model (Chen et al., 2017) or a generative model (Izacard and Grave, 2021). Extending such approaches to a multilingual setting usually suffers from two major problems - 1) Answering questions from different language sources because the answer for low resource languages might lie in documents from high resource languages (Asai et al., 2021a), and Wikipedia which might fail in cases of same language retrieval. (Clark et al., 2020a). 2) Large-scale cross-lingual datasets are not available that supply passages in a diverse number of languages which can enable better training of cross-lingual retrievers.

One specific approach that has been followed for bringing multilingual QA close to English QA is that the non-English question is translated into English and the answer from the English QA system is translated back to the query language. These systems suffer from the problem of machine translation error propagating itself in the downstream question answering. And also, these systems aren't able to exploit the fact that for high resource languages like Spanish, and Chinese the evidence might lie in the target language itself which is eas-

[1] https://mia-workshop.github.io/shared_task.html

Proceedings of the Workshop on Multilingual Information Access (MIA), pages 91 - 99

ier than two-way translation.

In this paper, we aim to extend the task of cross-lingual question answering to tackle the following research questions - a) How can we adapt the retrieve-then-generate approaches for English open QA to cross-lingual QA that do not rely on machine translation? - b) How do multilingual QA models trained on a small set of languages perform in zero-shot settings?

We follow CORA (Asai et al., 2021b) which is a many-to-many multilingual QA model by following a four-stage pipeline for addressing cross-lingual QA. The DPR based on mBERT (mDPR) is a bi-encoder retriever that retrieves documents cross-lingually without relying on machine translation. XLM-RoBERTA which serves as a passage reranker is trained as a cross-encoder to capture the interactions between the question and the passage on the top k documents fetched by the mDPR retriever. The reranked documents are passed through a Fusion-in-Decoder based mT5 reader module which can effectively learn to collect evidence from multiple passages to arrive at the final answer. In some cases, the predicted answer is not in the target language as desired by the user because the generator is either not able to convert entities into the target language or evidence is directly extracted from a different language passage. Further, we use a postprocessing step to map entities from Wikidata to convert the answer into the target language.

We conduct our experiments on two multilingual open-domain QA datasets, XOR-TyDi QA (Asai et al., 2021a) and MKQA (Longpre et al., 2021) across 14 typologically diverse languages with CORA as the baseline. We also use English questions from NQ (Kwiatkowski et al., 2019) for training. Reranking the outputs from the retriever leads to consistent improvements across all languages in both XOR-TyDi and MKQA, even in zero-shot settings. We show that using a fusion-in-decoder based reader leads to 2.7 points improvement in EM and 0.6 points improvement in F1 for the XOR-TyDi dataset. Moreover, on applying Wikidata based postprocessing techniques we see a straight 4.6 points improvement in EM and 3 points in F1. We see that our proposed pipeline also helps in zero-shot settings for both high resource and low resource languages.

2 Datasets

In this work, we use the data corpus provided by the MIA Shared Task on Cross-lingual Open-Retrieval QA which consists of XOR-TyDi and MKQA corpus. The shared task also provides questions from the NQ corpus.

XOR-TyDi is the first corpus to combine information-seeking questions, and open-retrieval QA in the multilingual domain to enable cross-lingual answer retrieval. This dataset is an extension of the TyDi QA (Clark et al., 2020b) dataset and involves retrieving evidence passages from multilingual and English resources. This dataset consists of questions written by native speakers in 7 typologically diverse languages: Arabic, Bengali, Finnish, Japanese, Korean, Russian, and Telugu.

Language	#Train	#Dev	#Passages
En (English)	76.6k (n)	1.7k (m)	18M
Ar (Arabic)	18.4k (x)	3k (x,m)	1.3M
Bn (Bengali)	5k (x)	0.5k (x)	0.1M
Fi (Finnish)	9.7k (x)	2.7k (x,m)	0.9M
Ja (Japanese)	7.8k (x)	2.4k (x,m)	5.1M
Ko (Korean)	4.3k (x)	2.2k (x,m)	0.7M
Ru (Russian)	9.2k (x)	2.7k (x,m)	4.5M
Te (Telegu)	6.7k (x)	0.6k (x)	0.3M
Es (Spanish)	-	1.7k (m)	5.7M
Km (Khmer)	-	1.7k (m)	0.06M
Ms (Malay)	-	1.7k (m)	0.4M
Sv (Swedish)	-	1.7k (m)	4.6M
Tr (Turkish)	-	1.7k (m)	0.8M
Zh (Simplified Chinese)	-	1.7k (m)	3.4M
Total	137k	9115	45.86M

Table 1: Dataset Statistics showing the 14 diverse languages used in this task with the top 7 being seen and the bottom unseen. n, x and m denote the source of the dataset NQ, XOR-TyDi and MKQA respectively from which the examples are collected.

MKQA corpus was originally proposed in (Longpre et al., 2021) and consists of 10K question-answer pairs aligned across 26 typologically diverse languages (260K question-answer pairs in total). Answers for this corpus are heavily curated and obtained from language-independent data representation which makes this corpus ideal for evaluating across diverse languages and being independent of language-specific passages. MKQA corpus provided by the shared task is a filtered version that only includes questions with answer annotations and removes the "no answers" questions. For this task, 12 languages were collected from MKQA, six seen: Arabic, Finnish, Japanese, Korean, Russian, and six unseen(zero-shot): Spanish, Khmer, Malay, Swedish, Turkish, Simplified Chinese, each with

1.7k examples in the dev set.

NQ (Natural questions) is a factoid-based English question answering dataset with both short and long answers for each question from English Wikipedia. We focus on the subset which is given as a part of the training dataset in the shared task.

Table 1 shows the dataset statistics for the train and the dev across 7 different languages. We also experiment with MKQA corpus in the zero-shot setting with 6 languages.

3 Related Work

Open-domain question answering requires a model to answer questions without any pre-trained domain (Kwiatkowski et al., 2019). There have been some recent works to create a non-English QA corpus to analyze the model's effectiveness to transfer knowledge from the English language or other high-resource languages. Further, some works focus on generating loosely aligned data using translation or similar multilingual sources.

As mentioned some of the recent works in Question Answering (QA) aim to build systems that can work well with languages other than English. (Lewis et al., 2019) proposed MLQA which is a multi-way aligned extractive QA corpus. It consists of instances in 7 languages with each instance parallel between 4 languages on average. This work defines two tasks: the first one focuses on analyzing the model's ability to transfer by training and testing in different languages and the other task requires the model to retrieve passages in a different language than the question. One of the shortcomings of this corpus is that it contains context in the same language and therefore doesn't explicitly captures the cross-lingual aspect. This leads to a problem for a low-resource language question set as in real scenarios most of these questions have answers in a high resource language. (Liu et al., 2019) presents the XQA dataset to investigate cross-lingual OpenQA research. This corpus consists of the training set in English along with the development and test set in eight other languages. Their analysis of several baseline models indicates that the performance in a cross-lingual setting not only depends on the similarity of English and the target language but also on the complexity of the target language question set. Another work in the cross-lingual domain, XQuAD is proposed by (Artetxe et al., 2019) which is created by using a subset of SQuad v1.1 (Rajpurkar et al., 2018) corpus and translating them into ten other languages by professional translators. This paper also evaluates the hypothesis that multilingual models perform well due to the shared subword vocabulary and joint training across multiple languages and shows that monolingual representations can be adapted to produce similar performance without relying on a shared vocabulary or joint training.

Most previous works modeled cross-lingual QA as an extractive task which is mostly inspired by the datasets like XQuAD (Artetxe et al., 2019) which is a subset of SQuAD (Rajpurkar et al., 2018). The SQuAD dataset contains answer spans in the evidence passage to answer a given question. These answer spans were further used in the generation of the cross-lingual QA dataset, XQuAD, and therefore are more suitable to be modeled as an extractive task. More recently, there have been studies that work on generating answers from raw text. Works such as (Chi et al., 2019) (Kumar et al., 2019) study cross-lingual question generation. (Shakeri et al., 2020) proposed a method to generate multilingual question-answer pairs through the use of a single fine-tuned multilingual T5 generative model. Their work shows that these synthetic examples could be used to improve the performance of multilingual QA in the zero-shot setting on target languages. Previous works have also explored other variants of generative modeling but it was mostly limited to the domains where the model is expected to generate long answers. Recent work on FiD (Izacard and Grave, 2021) shows that generative approaches could achieve competitive results even in the cases where answers consist of a short text span. One of the widely used approaches for open-domain question answering named RAG (Lewis et al., 2020) makes use of the generative model approach. RAG model's reader module takes several retrieved passages from the retriever encoder simultaneously to generate the answer. Passage representations and their similarity score with the query are used to generate the final response in the reader module. Further, the RAG approach works efficiently at scale due to the independent processing of passages in the encoder module.

Bi-encoder retrievers are effective in bringing out relevant passages from a large index but sometimes reranking those passages is essential as the downstream reader can only see a limited number of them. (Fajcik et al., 2021) uses reranker

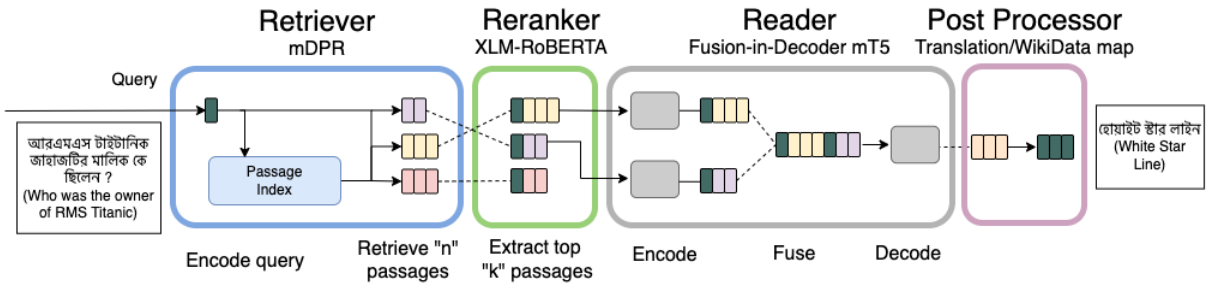

Figure 1: The proposed system architecture uses a mDPR based bi-encoder retriever which fetches the top 200 relevant passages from the passage index, followed by a XLM-R based cross-encoder for reranking. The top 50 passages are passed to FiD with mT5 to output the final response. The final response which is not in the target language is mapped using a Wikidata map to the target language.

after their retriever as a cross-encoder to improve the recall which proves effective in the end-to-end question answering pipeline. Incorporating Wikidata (Hu et al., 2021) in translation for sentences including named entities is common in literature because the pretrained multilingual reader is unable to translate these entities as it has not seen them during training.

4 Baseline

We use CORA (Asai et al., 2021b) as our baseline model which is a unified multilingual open QA model for many languages. CORA model is a combination of two models: Multilingual Dense Passage Retriever (mDPR) and Multilingual Answer Generation (mGEN).

mDPR extends Dense Passage Retriever (DPR) to a multilingual setting and uses an iterative training approach to fine-tune a pre-trained multilingual model (mBERT) to encode passages and questions separately. mGEN uses a multilingual sequence-to-sequence model (mT5) to generate answers in the target language token-by-token given the retrieved multi-lingual passages. The generation approach is used as it can generate an answer in the target language from passages across different languages.

5 Methodology

We employ the widely used Retrieve-then-generate, figure 1, architecture for open domain question answering. To tackle the challenge of passing a limited number of passages to the generator, we use a reranker. We also use a postprocessor to convert named entities into the query language.

5.1 Bi-encoder Retriever

Following the baseline (Asai et al., 2021b), we used the mDPR model trained on hard negatives mined using BM-25 and adversarial examples mined using hard negatives from 1st iteration of the mDPR model.

5.2 Cross-encoder Reranker

The multilingual retriever is trained as a bi-encoder where the questions and passages are encoded separately and compared during the inference time. Therefore, there is no cross-attention captured between the question and passage which is essential for cross-language retrieval. Cross-encoder can't be used for the retrieval because it isn't computationally feasible to compare the query with all the passages (which are in millions). Further, we can only pass a limited number of k passages (less than 50) to the reader model. Therefore, reranking is very essential in this scenario. Hence, we take the top k (here 200) passages fetched by the retriever and train a XLM-RoBERTA (Conneau et al., 2020) model as a cross-encoder by scoring the positive passages higher. We followed (Qu et al., 2020) to train the reranker by using a cross-entropy loss on the [CLS] token output. The negative passages for the reranker are mined using the finetuned baseline mDPR model.

5.3 Fusion-in-Decoder Reader

Given the recent popularity of generative reader models in English Open-domain QA (Izacard and Grave, 2021) (Lewis et al., 2020), we used a FiD model as a reader model with mT5 (Xue et al., 2021) which encodes the top-50 reranked/retrieved passages one-by-one and concatenates them before passing it to the decoder. FiD is useful because

Model	Target Language L_i Ar	Bn	Fi	Ja	Ko	Ru	Te	F1	EM
Baseline (CORA)	51.3	28.7	44.3	43.2	29.8	41.3	44.1	39.8	30.3
mDPR + mFiD	53.5	26.3	46.4	42.4	28.3	42.6	43.0	40.4	33.0
mDPR + mFiD + P	54.4	**31.7**	**46.7**	42.9	33.3	43.1	45.2	42.5	34.6
mDPR + mFiD + RR	55.1	28.0	46.2	43.2	30.6	42.8	44.5	41.5	34.0
mDPR + mFiD + RR + P	**55.6**	30.4	46.3	**43.7**	**34.7**	**43.2**	**45.8**	**42.8**	**34.9**
Baseline (Test)	49.7	**34.0**	39.5	39.7	25.6	41.0	36.2	37.9	-
mDPR + mFiD + RR + P (Test)	**55.1**	30.6	**41.3**	**42.4**	**28.8**	**42.6**	**40.8**	**40.2**	-

Table 2: XOR-TyDi dev and test set performance across 7 different languages for different ablations of our components. We tried settings where we removed the RR(reranker) and P(Postprocessing).

Model	Seen Target Language L_i Ar	En	Fi	Ja	Ko	Ru	Zero Shot Target Language L_i Es	Km	Ms	Sv	Tr	Zh	F1	EM
Baseline (CORA)	8.77	27.9	23.3	15.2	8.3	14.0	24.9	5.7	22.6	24.1	20.6	13.1	17.4	13.5
mDPR + mFiD	8.8	39.7	25.2	14.3	6.3	13.3	29.7	7.7	**30.1**	28.6	25.7	9.8	19.9	16.0
mDPR + mFiD + P	**14.5**	39.7	25.1	20.6	13.6	22.6	**30.2**	7.8	29.4	28.2	25.4	15.1	22.7	17.2
mDPR + mFiD + RR	9.3	40.6	26.2	14.9	6.5	14.6	29.5	8.3	29.9	**29.9**	**26.7**	10.6	20.6	16.5
mDPR + mFiD + RR + P	14.2	**40.6**	**26.1**	**21.5**	**14.8**	**22.7**	29.8	**8.3**	29.3	29.6	26.5	**16.2**	**23.3**	**17.8**
Basline (Test)	9.5	36.3	22.7	7.7	15.9	14.6	27.2	6.0	25.1	26.7	21.7	13.8	17.1	-
mDPR + mFiD + RR + P (Test)	**13.9**	**42.6**	**26.8**	**14.6**	**22.7**	**22.4**	**32.1**	**8.7**	**31.1**	**31.5**	**26.6**	**18.0**	**22.9**	-

Table 3: MKQA dev and test set performance across 12 different languages for different ablations of our components. There were 6 languages which were in a zero-shot setting. We note from the dataset statistics presented in 1, Es, Sv, Zh are high-resource whereas others are low resource languages.

it also learns to rerank the documents to collect the evidence from the documents. We followed the baseline to use mT5 as the underlying cross-lingual language model. This mT5 with FiD model, which we call mFID, is trained on the training data that is given, which is a mixture of Natural Questions and XOR-TyDi. This mFiD acts as a cross-lingual fusion reader without the necessity to translate the passages/free-text answers from which the evidence is collected. To add extra supervision during training, we always pass the gold passage which has the answer along with the other passages.

5.4 Answer Post Processing with Wikidata

Figure 2 shows the discrepancy between the language of the predicted output and the language of the question. This is because if the evidence is collected from a different language passage, the answer is not translated by the model in cases where there are entities in the answer. After all, the model hasn't seen those entities while training. In such cases, we require post-processing to convert entities in other languages to the answer language. We have only tackled the English case because we have seen that most of the time the model is not able to translate the English entity. We collected en-xx Wikidata [2] maps for the languages in our dev set to convert those English predicted answers.

Question	Predicted Answer	Gold Answer
ఆక్సిజన్ చిత్ర కథానాయకుడు ఎవరు? (Who is the protagonist of the film 'Oxygen'?)	Anu Emmanuel	అను ఇమ్మాన్యుయేల్
ভুটানের সবচেয়ে জনবহুল শহর কোনটি ? (Which is the most populous city of Bhutan?)	Thimpu	থিম্ফু
세상에서 가장 넓은 산은 무엇인가? (Which is the widest mountain in the world?)	Mount Everest	에베레스트산

Figure 2: The figure shows the predicted answer by the model which is in English (usually entities) which the model can't translate and hence requires post processing to convert to the final language.

6 Results & Discussion

Table 2 shows the performance of various components of our pipeline and compares it with the baseline on the XOR-TyDi dev and test set and Table 3 shows the overall performance on the MKQA dev set. MKQA dev set has 6 languages that have no training data. We do not add any training data using data augmentation for these languages as we want to evaluate improvement in models in a

[2] https://www.Wikidata.org/wiki/Wikidata:Main_Page

Model	R@50 Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg
Baseline (mDPR)	67.8	52.9	**60.6**	16.6	40.2	60.9	50.4	53.5
XLM-R Reranker	**68.9**	**56.5**	60.3	**19**	**44.6**	**61.2**	**53.2**	**55.1**

Table 4: Reranking performance for R@50 across 7 different languages in the XOR-TyDI dev set.

Model	Seen R@50 Ar	En	Fi	Ja	Ko	Ru	Zero Shot R@50 Es	Km	Ms	Sv	Tr	Zh	Avg
Baseline (mDPR)	25.5	70.5	55.5	13.6	20.7	35.9	62.2	16.1	59.6	63.0	56.4	12.5	41.0
XLM-R Reranker	**27.7**	**73.5**	**58.5**	**15.9**	**23.3**	**38.9**	**64.7**	**18.6**	**62.3**	**65.2**	**59.0**	**15.4**	**43.6**

Table 5: Reranking performance for R@50 across 12 different languages in the MKQA dev set. 6 languages were zero shot (not seen in the training corpus).

zero-shot setting that is comparable to real-world scenarios.

For XOR-TyDi, we see that overall we achieve 3 and 4.6 points improvement in F1 and EM respectively, whereas for MKQA we achieve a 5.9 points F1 improvement. We see that adding FiD to the baseline mDPR leads to a significant increase in F1 for languages like Ar, Fi, and Ru. Further, adding postprocessing to the FiD output leads to a significant increase in both F1 and EM and the model outperforms the baseline for all the languages. Reranking is crucial for FiD because it sees only 50 documents as compared to the 100 that the baseline uses. Reranking the top 200 documents retrieved helps the resulting FiD which shows a consistent improvement over the baseline except for Bn and Ko for which the model usually outputs lots of entities in English that require post-processing. We get the best results for applying postprocessing (P) on the outputs by the reranker(RR) + FiD model. For zero-shot settings in the MKQA dataset, we also see consistent 5 points of F1 improvement over the baseline due to the combined effect of the FiD, reranker, and the postprocessor. For high-resource zero-shot languages Es, Sv, Zh we observe around 5 points improvement over baseline whereas for Km which is an extremely low-resource language we still show around 3 points improvement. We now provide detailed results and analysis for each component of our pipeline.

6.1 Reranker

Table 4 captures the reranker performance over the baseline mDPR model. Applying the reranker to the baseline mDPR gives a consistent improvement in R@50 for all the languages leading to about 1.6 points improvement. This essentially is because of two reasons - cross interactions between question and passage and the fact that XLM-RoBERTA is a better language model than mBERT. These reranked passages also help in the downstream FiD model (See Table 2) and lead to a 1 point F1 and EM improvement over the model which didn't receive the reranked passages (mDPR + mFiD). We also think that this model lacks in performance over the baseline for Bn because the number of training examples for Bn is very low as compared to other languages. For the MKQA dataset, in table 5, we see a significant increase over the mDPR model across all the languages for R@50. We see a 2.6 improvement in R@50 over the baseline. For zero-shot languages, we also see a significant increase in recall showing the effectiveness of the XLM-R model for unseen languages. This increased recall further helps in the downstream reader improvements as well.

6.2 Reader

The FiD with mT5 (mFiD) reader performs better than the normal mT5 as can be seen by the 3 points EM improvement over the baseline in Table 2, although FiD just uses 50 documents as compared to the 100 documents used by the baseline. The fusion-in-decoder approach also is an effective reranker by itself in searching for evidence to arrive at the final answer. Table 3 shows the final performance of the model on the MKQA dataset as well. We see that mFiD with reranking has 3 points improvement over the baseline and also shows great improvements for unseen languages like Es, Ms, and Tr. This further corroborates the effectiveness of the reranker and the mT5 based FiD for unseen

Question	Lang	Baseline Prediction	Our Model Answer	Gold Answer	Category
Как называется национальный костюм австрийских женщин? (What is the national costume of Australian women called?)	Ru	Баварский национальный костюм (Bavarian National Costume)	Дирндль (Drindl)	Дирндль (Drindl)	Rerank
من كان أول خليفة للدولة الأموية؟ (What is the oldest university in Netherlands?)	Ar	جامعة أوتريخت (Utrecht University)	جامعة لايدن (Leiden University)	جامعة لايدن (Leiden University)	FiD
스웨덴에서 가장 인구밀도가 높은 도시는 어디인가? (Which is the most densely populated city in Sweden?)	Ko	런던 (London)	스톡홀름 (Stockholm)	스톡홀름 (Stockholm)	Post Process
সমুদ্রপৃষ্ঠ থেকে মিরিকের গড় উচ্চতা কত ? (What is the average height of Mirik from sea level?)	Bn	১৪৯৫ মিটার (1495 m)	১৪৯৫ মিটার (1495 m)	1495 মি (1495 m)	Dataset Flaw

Figure 3: The figure shows 1 example each for our component improvement and also points to a flaw in the dataset. Here category indicates the model component which led to the correct prediction compared to the baseline model.

languages. Also, it is worth noting that for En language in MKQA, simply adding FiD improves the performance by 12 points, showing the advantage of the fusion-in-decoder technique.

6.3 Postprocessor

Converting entities using Wikidata maps in English to target languages is useful, especially for the high-resource languages, in both XOR-TyDI and MKQA datasets because the answers are expected in the question language and the reader models (mFiD) can't translate named entities. Table 2 shows that for languages Bn, and Ko the improvement of post-processing is the maximum because the predicted answers of these languages are usually in English, which when converted to their entities boost's the F1 and EM. In the zero-shot setting, there is a huge improvement of 6 points in Zh because of the same reason.

7 Error and Qualitative Analysis

7.1 Qualitative analysis

We present a qualitative analysis, in figure 3, of our model that highlights the component-wise improvement. For the first example with the category "Rerank", it implies that the original mDPR retrieval top-50 docs didn't have the ground truth passage but due to the reranker module, we were able to move ground truth passage into the top-50 passages and finally generate the correct answer. The second example indicates the improvement of the reader module due to the use of the FiD technique. In this example, both baseline and reranker retrieval output had the ground truth passage but only our reader module (FiD) can generate the correct response. For the 3rd example, our reader module generates an answer in the English language but our post-processing module can identify this English entity from the Wikidata mapping and convert it to the source language as expected by this task. For the last example, we try to highlight that both the models can generate the correct response but F1 comes 0 for both of them due to the limitation of the dataset (could have provided multiple answers) and the evaluation metric used for this task.

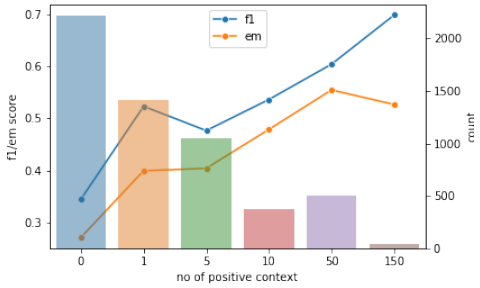

Figure 4: The graph shows the performance of our best model with respect to number of positive contexts.

7.2 # Positive context analysis

We look at the reranked results which are used by our best model to see the effect of the number of positive passages (passages containing the right answer) on the F1 and EM metrics. Figure 4 shows that with an increased number of positive contexts (greater than 10), it's easier for the FiD model to collect evidence and arrive at the final answer, which indicates that the retriever is the bottleneck. If the retriever is good enough to pull up multiple positive contexts having the correct answer, the FiD model will perform better in those cases.

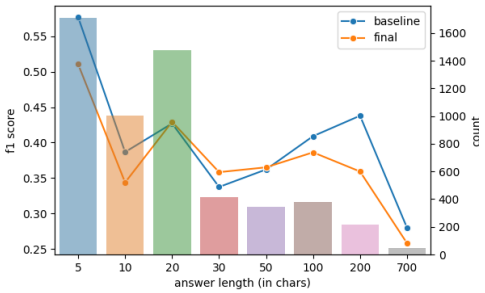

Figure 5: The graph shows a comparison of our model with the baseline model across correct answer length (in chars).

7.3 Answer length analysis

Figure 5 shows that our best model performs better than the baseline model for questions that have short answers whereas for long answers the baseline model outperforms our model. The FiD mT5 model might have learned some bias to truncate at short answers and it fails to emit long answers, which the normal mT5 does better.

8 Conclusion & Future Work

We introduced a modular end-to-end system with a retriever, reranker, reader, and postprocessor for cross-lingual question answering and showed improvements in both normal and zero-shot settings. These cross-lingual models consist of a large number of parameters and are very resource-intensive. The retriever model takes around 1 hr/epoch whereas the fusion-in-decoder model takes 8 hr/epoch on A6000 GPUs. For future work, we think it would be interesting to try sparse retrieval methods (Formal et al., 2021) in cross-lingual settings and also try incorporating more knowledge from Wikidata based entities in the pipeline.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

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

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

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

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xianling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training. *CoRR*, abs/1909.10481.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *TACL*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020b. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *CoRR*, abs/2003.05002.

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*.

Martin Fajcik, Martin Docekal, Karel Ondrej, and Pavel Smrz. 2021. R2-d2: A modular baseline for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 854–870.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2021. Deep: Denoising entity pre-training for neural machine translation.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 874–880. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. *CoRR*, abs/1906.02525.

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

Patrick Lewis, Ethan Perez, Aleksandra 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. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: evaluating cross-lingual extractive question answering. *CoRR*, abs/1910.07475.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

S. Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *CoRR*, abs/2010.08191.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Siamak Shakeri, Noah Constant, Mihir Sanjay Kale, and Linting Xue. 2020. Multilingual synthetic question and answer generation for cross-lingual reading comprehension. *CoRR*, abs/2010.12008.

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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.