Capstone Report: Cross-lingual Open-domain Question Answering

Abhishek Srivastava * **Aditya Srikanth Veerubhotla*** **Ameya Mahabaleshwarkar***
asrivas4 adityasv amahabal

Zhengbao Jiang (Mentor) **Graham Neubig (Advisor)**
zhengbaj gneubig

Abstract

Cross-lingual open-domain question answering is a challenging task that has typically been tackled using a reader and retriever setup, with these two components being optimized independently. In this paper, we present a novel approach that enables the sharing of information between the reader and retriever through a two-way knowledge distillation process. We use Fusion-in-Decoder Knowledge Distillation (FiD-KD) for distilling knowledge from reader to retriever, and propose two new methods for distilling knowledge from retriever to reader. In addition to our end-to-end modelling of the retriever-reader framework, we also explore methods for improving the individual reader and retriever components using self-training and cross-lingual adaptation. Through the experiments conducted in this work, we demonstrate that our end-to-end modelling of the retriever-reader framework achieves promising results and improves both components.

1 Introduction

Cross-Lingual Open-Domain Question-Answering (xODQA) is an important task because it has the potential to improve the accessibility and usability of information on the internet. By enabling users to pose questions in their own language and receive answers in the same language, this task can help overcome the language barrier that often prevents people from accessing valuable information and knowledge. Overall, the development of effective approaches for cross-lingual open-domain question-answering has the potential to greatly enhance the ability of people to access and use information from a wide range of sources.

While xODQA is similar to Open-Domain QA: given a question q, a QA system is expected to generate an answer a, leveraging a corpus C. The main difference between xODQA and ODQA lies in the construction of the Corpus, with xODQA having a multilingual corpus and ODQA enforcing a monolingual corpus. Prior work often breaks the task into two sub-problems: retrieval of relevant sources for a question which is done using a Retriever, and generating an answer from the retrieved sources performed by a Reader. Usually, these two components are trained independently and the outputs of the Retriever are input to the Reader. However, this weak coupling makes the errors of retriever severely affect the reader.

Recently, many approaches have been proposed that train the Retriever and Reader in an end-to-end manner, with the gradient flowing from the Reader to the Retriever, helping it with additional supervision. However, these focus more on the training of the Retriever, and no explicit signal is used for improving the Reader. In this work, we show that passing the query-evidence relevance scores output from the Retriever can help improve the Reader, and propose an iterative Forward-Backward algorithm for training the Retriever and Reader models in an E2E fashion. We show that our approach shows strong results over different baselines, and is a promising strategy for performing end-to-end training of ODQA systems.

2 Related Work

2.1 Optimizing retriever

2.1.1 Self-training with negatives

In dense retrieval, utilizing negatives in training is a popular self-training approach (Thakur et al., 2021). Most well-performing dense models utilize this type of iterative optimization and continue to improve over previous versions of the same model (Izacard and Grave, 2020a; Qu et al., 2020). The entire process involves retrieving top-k passages for each question from a version of the retriever and then tagging everything but the positive passages as negative examples. Afterwards, a new iteration

*Equal contribution

of the model is trained using the passages marked up previously (Sorokin et al., 2022).

2.1.2 Shared passage/query encoder

Earlier works in open-domain question answering and information retrieval utilize a bi-encoder network with separate passage and query encoders (Karpukhin et al., 2020). In contrast, more recent approaches have shown that having a shared encoder outperforms the bi-encoder approach with separate encoders (Sorokin et al., 2022; Lee et al., 2021).

2.2 Optimizing Reader

In an open-domain setup, after reducing the search space down from millions of documents to a few hundred (or less), the unordered set of retrieved documents can be utilized as a whole with a generative seq2seq reader model. There are two broad ways to utilize multiple documents for generating the answer. They differ primarily in where the passages are fused together in the network.

2.2.1 Fusion in Encoder

In this approach, the passages are concatenated together before encoding. Asai et al. (2021b) utilizes this approach where they retrieve top 15 passages using their retrieval module, and then feed the questions and concatenated passages into the generative reader model. Note that this approach can suffer from context fragmentation due to the input limit of 512 tokens in their encoder (with Transformers backend).

2.2.2 Fusion in Decoder `(FiD)`

To get around the input length limitation in the previous approach, Fusion in Decoder works relatively well (Izacard and Grave, 2020c). Given a set of retrieved passages, the FiD model first encodes each of them separately along with the question and then concatenates the representations. The concatenated sequence is then passed to the decoder for finally generating the answer. This approach has been widely adopted and used in a majority of works (Singh et al., 2021; Yu et al., 2021).

2.3 End-to-end approaches

As opposed to optimizing each component in a disjoint manner, there are approaches that enable joint optimization of the retriever and reader. Lewis et al. (2020) combine a pre-trained retriever with a pre-trained encoder-decoder (Reader/Generator) and fine-tune end-to-end. For a query, they use Maximum Inner Product Search (MIPS) to find the top-K most relevant documents of all documents. To make the final prediction, they treat the set of documents as a latent variable and marginalize over the encoder-decoder predictions given different documents.

Instead of jointly optimizing both components together, there are other approaches that enable communication between the reader and retriever to improve the performance. Izacard and Grave (2020a) propose to utilize the aggregated cross-attention scores from all the layers for each decoding step over each passage in `FID` as a relevance score for training the retriever. It is assumed that at the starting point, both retriever and reader perform sufficiently well as the scoring is done over the initial retrieved set which if bad, can deteriorate the performance in later steps. However, note that in this setup the distillation is one directional and the reader is always the teacher model with the retriever as student.

2.4 Cross-lingual Approaches

In cross-lingual settings, most recent works have adapted best performing approaches from the monolingual setup (majorly studied with English) and have seen promising results. An example if the `mFID` model that adapts `FID` for cross-lingual settings and uses a multilingual seq2seq model (such as mT5) while keeping the overall approach the same (Sorokin et al., 2022; Ri et al., 2022; Agarwal et al., 2022).

A majority of works also augment data through translation of original examples. Asai et al. (2021b) translate their entire training set into the number of languages in their setup. Additionally, they also translate all of Natural Questions set (English) (Kwiatkowski et al., 2019) to the languages. Similarly, a few other approaches have suggested similar augmentation techniques utilizing translations (Asai et al., 2022).

As a type of iterative data augmentation, Asai et al. (2021b) expands their training set by iteratively adding new cross-lingual examples through WikiData [1] anchored on answer entities. They find the corresponding Wikipedia article for an answer entity in different languages and then use a trained reader model to filter out paragraphs from that article that can generate the answer given the question.

[1] https://www.wikidata.org/w/api.php.

They show that this approach performs really well when done for a few iterations.

3 Experimental Setup

3.1 Task

To address the problem of cross-lingual open domain question answering, we can define a two-stage process, consisting of a retriever and a reader.

3.1.1 Retriever

In the retriever stage, we aim to find relevant passages from the evidence corpus that are likely to contain the answer to the given question. We can represent the question as a single entity q and the evidence corpus as a set of passages $P = p_1, p_2, ..., p_m$.

To measure the similarity between the question and each passage, we can use a loss function L_{sim} that measures the difference between the question and passage representations. We use in-batch negatives loss. Let P_{pos} be the set of passages that are known to be relevant (positive) to the given question, and let P_{neg} be the set of passages that are known to be irrelevant (negative) to the given question. The loss can then be defined as:

$$L_{sim}(\theta; q, P_{pos}, P_{neg}) = - \sum_{p \in P_{pos}} \log \frac{\exp(q_\theta \cdot p_\theta)}{\sum_{p' \in P_{pos} \cup P_{neg}} \exp(q_\theta \cdot p'_\theta)}$$

3.1.2 Reader

In the reader stage, we can use a fusion in decoder formulation, where we encode each passage in the evidence corpus P and concatenate the resulting representations together. The decoder then uses this concatenated representation to generate the answer to the given question. Let a be the sequence of tokens representing the answer to the question q. The reader can be trained to minimize the negative log-likelihood of the correct answer sequence given the question and evidence corpus as input:

$$L_{reader}(\theta; q, P, a) = - \log p_\theta(a|q, P)$$

3.2 Datasets

The XOR TyDi dataset (Asai et al., 2021a) is a multilingual open-retrieval QA resource that allows for cross-lingual answer retrieval. It is based on questions from the TyDi QA dataset (Clark et al., 2020) and includes three tasks involving finding documents in different languages using multilingual and English resources. The dataset consists of questions written by native speakers of 7 languages: Arabic, Bengali, Finnish, Japanese, Korean, Russian, and Telugu. The answers have been retrieved from a multilingual document collection.

Split	Multilingual			Cross-Lingual		
Lang	Train	Dev	Test	Train	Dev	Test
Arabic	15391	437	1133	2495	63	708
Bengali	2142	234	138	2246	265	425
English	60812	500	7251	-	-	-
Finnish	7291	389	1197	1965	111	615
Japanese	5158	369	869	2120	131	433
Korean	1576	212	507	2099	284	368
Russian	6956	393	1125	1826	107	568
Telugu	5043	407	712	1212	93	350

Table 1: Language-wise dataset statistics for our experiments. Multilingual refers to examples where both question and evidence are in the same language whereas cross-lingual refers to cases where they are in different languages.

The XOR-Retrieve task is a cross-lingual retrieval task where a question is written in the target language (e.g., Japanese), and a system is required to retrieve an English document that answers the question. The XOR-English Span task is a cross-lingual retrieval task where a question is written in the target language (e.g., Japanese), and a system is required to output a short answer in English. The XOR-Full task is a cross-lingual retrieval task where a question is written in the target language (e.g., Japanese), and a system is required to output a short answer in the target language.

Overall, the XOR TyDi dataset provides a valuable resource for researchers working on cross-lingual open-retrieval QA. It allows for the evaluation of systems on a variety of tasks involving cross-lingual retrieval and generation.

Table 1 consists of the language-wise dataset statistics for our setup. Due to compute issues and quick turnaround time, we subsample the original Wikipedia corpora to 5 million while maintaining the original global language proportion as much as possible during sampling. We present these proportion statistics in Table 2.

3.3 Evaluation

To evaluate the performance of our model, we will use the following metrics for the reader and retriever stages.

Lang	Original split	Subsampled
Arabic	1.3	0.21
Bengali	0.1	0.02
English	18	2.91
Finnish	0.9	0.15
Japanese	5.1	0.83
Korean	0.7	0.11
Russian	4.5	0.73
Telugu	0.3	0.05
Total	30.9	5.01

Table 2: Proportion of languages in the original and our subsampled corpus

3.3.1 Reader

For the reader stage, we will use exact match and f1-score to measure the accuracy of the generated answers.

Exact Match Exact match is a binary classification metric that measures the proportion of answers generated by the model that exactly match the ground truth answers in the test set. It is calculated as follows:

$$ExactMatch = \frac{1}{n}\sum_{i=1}^{n}[a_i = a_i^*]$$

where a_i is the predicted answer for the i-th question, a_i^* is the ground truth answer for the i-th question, and n is the total number of questions in the test set.

F1-score F1-score is a metric that combines precision and recall to measure the overall accuracy of the model. It is calculated as the harmonic mean of precision and recall. For the reader stage, we can calculate the precision, recall, and f1-score as follows:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F1Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

where tp is the number of tokens that are shared between the correct answer and the prediction, fp is the number of tokens that are in the prediction but not in the correct answer, and fn is the number of tokens that are in the correct answer but not in the prediction.

3.3.2 Retriever

For the retriever stage, we will use top-k accuracy with hasAnswer as positive, where hasAnswer refers to whether the retrieved passage contains the answer string or not. This will allow us to measure the accuracy of the retriever in identifying relevant passages from the evidence corpus.

Top-k accuracy To calculate top-k accuracy with hasAnswer as positive, we first assign a score of 0 or 1 to each question q based on whether any of the passages in the retrieved passage set P contains the answer substring. If any passage in P contains the answer, we assign a score of 1 to the question q.

The top-k accuracy is then calculated as the proportion of questions with a score of 1, as follows:

$$TopKAccuracy = \frac{\sum_{i=1}^{n}[hasAnswer_i = 1]}{n}$$

where $hasAnswer_i$ is the score for the i-th question, and n is the total number of questions in the test set.

This metric allows us to evaluate the performance of the retriever in identifying relevant passages from the evidence corpus. We can use this metric to compare the performance of our model to other models on the same task.

3.4 Baseline

XLM-R (Cross-lingual Language Representation) (Conneau et al., 2019) is a large-scale pre-trained multilingual language model developed by Facebook AI. It is trained on a diverse and extensive dataset of parallel and non-parallel texts from a variety of languages, and is designed to capture cross-lingual syntactic and semantic information. It can be fine-tuned for various natural language processing tasks, including question answering.

mT5 (Multilingual Text-to-Text Transformer) (Xue et al., 2020) is a pre-trained multilingual language model developed by Google Research. It is trained on a large corpus of text from multiple languages, and is designed to generate human-like text in a variety of languages. It can be fine-tuned for various natural language generation tasks, including question answering.

We will use XLM-R as the retriever in our baseline model, and mT5 as the reader (with Fusion in Decoder setup). We make these choices based on past work as well as our own experiments that found XLM-R to be better than mBERT (Devlin

et al., 2018) and mT5 better than mBART (Liu et al., 2020).

4 Method

The following sections describe the experiments and methodology for optimizing the $mDPR$ retriever and $mFiD$ reader individually, as well as an end-to-end framework go iteratively optimizing these two components of the cross-lingual open-domain question-answering pipeline.

4.1 Optimizing retriever

4.1.1 Self-training

Self-training approaches have been shown to be effective for improving the performance of the retriever component in open-domain question-answering. These approaches involve using the output of the retriever to automatically generate additional training data, which can then be used to fine-tune the retriever (Qu et al., 2021; Izacard and Grave, 2020b). Inspired by the work of Sorokin et al. (2022), we iteratively train the retriever in following steps:

1. Perform retrieval on the corpus using the $mDPR_i$ model for queries in the initial train set
2. From the set of top-k retrieved passages for each query, select the gold passages (P^+) which are known apriori, and keep the rest as hard negative examples (P^-)
3. Using the set of positive passages (P^+) and hard negative passages (P^-) identified in the previous step, train $mDPR_{i+1}$ the next iteration of model
4. Repeat steps 1 - 3 for a pre-defined number of iterations

For the initial iteration when there is no trained model available, we train the $mDPR_0$ model using gold passages.

4.1.2 Entity augmentation

Hu et al. (2021) show in their work that entity-based denoising pre-training helps improve entity translation accuracy within sentences for machine translation models. Building upon this idea, we perform an entity-based augmentation on the training set. Our method leverages pre-trained Named Entity Recognition models for identifying entity spans in passages. These spans are mapped to their translations in different languages using `WikiMap`. The original entites in the passages are then replaced with these translations. This augmented data is then used for training the $mDPR$ model.

4.1.3 Cross-lingual data augmentation

Cross-lingual data augmentation has been used as a method for adapting NLP models to a setting where different segments of the input text can be expected to be in different languages (Singh et al., 2019). Within the context of this work, a passage (p) in language L_p and a query (q) in language L_q are termed as a cross-lingual pair if $L_p \neq L_q$. Adhering to this condition, we dynamically convert 50% of the passage-query pairs in the training set into a cross-lingual pair. This is achieved by translating the query from language L_q to L'_q, such that $L_p \neq L'_q$ is ensured. The target language is uniformly sampled from the set of languages present in the corpus. The dynamic conversion of training examples to cross-lingual examples, and uniform sampling of target languages ensures that with enough training steps, the model is able to see the same passage-query pair in different cross-lingual language settings, thus avoiding bias towards any particular language. The logic behind this process is defined in algorithm 1.

Algorithm 1 Dynamic cross-lingual data augmentation

Require: passage (p) in language L_p, query (q) in language L_q, corresponding query translations Q, and set of languages in the corpus L
  if $L_p \neq L_q$ **then**
    `return` p, q
  else
    `sample` $L'_q \epsilon\ L \mid L'_q \neq L_p$
    `select` $q' \epsilon\ Q \mid q'$ is in language L'_q
    `return` p, q'
  end if

The translations for the queries were obtained from the `NLLB 1.2B`[2] model. For English queries in the training set, translations were obtained in all non-English languages present in the corpus, and for non-English queries, corresponding English translations were used. The $mDPR$ model is then trained on this augmented training set.

[2] https://huggingface.co/docs/transformers/model_doc/nllb

4.2 Optimizing reader

4.2.1 Cross-lingual regularization of decoder output distributions

Liang et al. (2021) employ bidirectional KL-divergence minimization as a form of regularization between sub-models resulting from dropout, with the aim of bringing the output distributions of these sub-models for the same input closer to each other. It can be hypothesized that to make the reader model language invariant, and thus better in cross-lingual settings, the output distributions of the model obtained with a set of passages and the same input query, but in different languages should be similar. Thus, we adapt this idea to the reader model using a multi-task setting, where we also minimize an additional jensen-shannon divergence (JSD) along with the generation loss. The jensen-shannon divergence is computed between the output distributions over T answer (a) tokens obtained by - (1) passing a query (q) and set of passages (P) through the $mFiD$ reader model, and (2) passing a translated version of the same query (q') with the same set of passages (P). The additional loss to be minimized is represented by the following equation:

$$\sum_{t=1}^{T} JSD(P(a_t|q', P), P(a_t|q, P))$$

Calculating the the JSD between these two output distributions requires the answer tokens to be same in both cases even if the queries are in different languages. This contradicts the nature of the task which requires the answer to be in the same language as that of the query. To deal with this, we additionally experiment with appending of language ID to the query to guide the reader model to decode the answer in a specific language as indicated by the language ID, irrespective of the query language.

4.2.2 Cross-lingual regularization of decoder cross-attention

Along the lines of cross-lingual regularization of decoder output distributions, we hypothesize that ensuring similarity between the $log - softmax$ of aggregate decoder cross-attention scores over passages (P), for a query (q) and a translated version of the same query (q') for generating an answer (a), would also induce cross-linguality in the reader model. To achieve this, we experiment with minimizing an additional JSD as defined by the following equation:

$$JSD(\texttt{xAttn}(q', P), \texttt{xAttn}(q, P))$$

Here, $\texttt{xAttn}(.)$ indicates the function to obtain the $log - softmax$ of aggregated $mFiD$ decoder cross-attention scores for each passage in the set of passages (P).

4.2.3 Cross-lingual data augmentation

Cross-lingual data augmentation for the reader follows the same logic as described ins Section 4.1.3. An added condition is observed for the reader according the nature of task, where the answer is also translated from language L_a into language L'_a such that $L'_a = L'_q$, where L'_q is the resulting language of the query after cross-lingual augmentation.

4.3 End-to-end modelling

To optimize the retriever and reader in an end-to-end manner, we propose an iterative, two-way knowledge distillation framework where the retriever and reader alternately act as the teacher and student model. In what we call as the `forward` iteration of the framework, the retriever model is the teacher, where as the reader model acts as the teacher in the `backward` iteration.

4.3.1 `Forward` (Knowledge distillation from retriever to reader)

The retriever scores obtained during the retrieval of passages (P) for a given query (q) are indicative of the relative importance of these passages for answering the given query. In traditional retriever-reader frameworks, this relevance information is not passed on to the reader, which could effectively benefit from it while identifying the relevant source of information for answering the query. To enable this channel of communication, and to allow distillation of knowledge from the retriever to the reader we propose two methods:

- **Minimizing KL-divergence between retriever (Ret^r) scores and reader cross-attention (`xAttn`) scores**: To implicitly encourage the reader model to pay more attention to passages that are assigned relatively higher scores by the retriever, the reader model is trained in a multi-task manner where the generation loss is minimized along with the KL-divergence between the reader scores

and the decoder cross-attention scores aggregated over all tokens of a passage, over all attention heads for all decoder layers. Following equations describe this loss calculation in this method:

$$L_1 = -\log p(a|q, P)$$

$$L_2 = KL(\texttt{xAttn}(q, P), Ret^r(q, P))$$

$$Loss = L_1 + \lambda * L_2$$

The additional loss is multiplied by a factor λ which can be tuned. For the results presented in the following sections, a λ value of 0.01 is used.

- **Reader cross-attention offset with retriever scores**: In this method, the $log - softmax$ of retriever scores for each passage are multiplied with a factor λ (0.01 for experiments in this work) and then added to the attention scores in all attention heads of all decoder layers in the reader. This is a more explicit way of guiding the reader cross-attention towards passages that are considered important by the $mFiD$ retriever for the given query, as indicated by the retriever scores.

4.3.2 `Backward` (Knowledge distillation from reader to retriever)

Due to the inductive bias of the $mFiD$ reader, the model learns to pay more attention to passages that contain relevant information for answering the given query. Through this process, the $mFiD$ model has an implicit re-ranking capability which is expressed through it's cross-attention scores. To distil this knowledge from the reader to the retriever, we make use of the `FiD-KD` methodology proposed by Izacard and Grave (2020b). We treat the aggregated cross attention scores as soft labels for training and $mDPR$ model which minimizes the KL-divergence between the retriever scores and the $softmax$ of aggregated reader cross-attention scores.

5 Results and Analysis

Table 3 shows the results of our experiments for the Retriever. It can be observed that self-training is a strong technique for improving performance, with the Dev accuracy improving over each iteration. While the improvement in Test accuracy is low between $mDPR_0$ and $mDPR_1$, this can be explained by the lack of cross-lingual passages retrieved from the previous iteration, which increases in between the first iteration and the second. When comparing $mDPR_1$ + Entity Aug. and $mDPR_2$, we see an improvement in test performance, which indicates that the model is able to learn cross-lingual representations of entities, which improve the test performance. Stronger results are obtained when explicit cross-lingual data is provided, leading to an improvement from 0.501 to 0.507 between $mDPR_1$ + Entity Aug. and $mDPR_1$ + xLingual Data Aug. However, our experiments suggest that the best training technique is to leverage the fine-grained relevance information output from the Reader model, providing an improvement of 0.107 in accuracy over $mDPR_2$.

The results of our experiments on the reader are presented in Table 4. We can see that there is a large difference between the Dev dataset and the Test dataset scores, indicating a large domain shift between the two sets. This is because a larger fraction of cross-lingual examples are present as a part of the test set over the dev set. Interestingly, we observe that the performance in FiD drops in the dev and test sets when cross-lingual data is added. Upon further analysis, we observed that the performance in cross-lingual test examples improves, but a greater decrease in metrics is observed for the multilingual examples. This happens due to the fact that the training data is enforced to have mostly cross-lingual data, but the dev and test data have mostly multilingual data. This demands a strategy to perform augmentation better.

The results also show that the results from the Output regurlarization experiments show a drastic drop in performance as compared to the baseline. We believe this is due to the strong regularization effects of the JSD loss, and further experimentation is needed to control its effects. We also observe a slight drop in performance in the case of $mFiD$ + xLingual $xAttn$ regularization. The evidence from the cross-lingual experiments with the biases in the datasets seem to suggest that we need better strategies to mine cross-lingual examples for training, and better inductive biases for conditioning on the evidences.

Finally, the results show that our proposed technique of utilizing the Retriever scores as an additional signal to the reader is a promising direction. We observe improvement in all the metrics, with a large improvement in the test sets. However, the re-

Retriever Approach	Dev. Acc. @k=50	Test Acc. @k=50
$mDPR_0$	0.59	0.4
$mDPR_1$	0.665	0.403
$mDPR_2$	0.745	0.494
$mDPR_1$ + xLingual Data Aug.	0.746	0.507
$mDPR_1$ + Entity Aug.	0.74	0.501
$mDPR_1$ + `FiD-KD`	**0.749**	**0.51**

Table 3: `mDPR` retriever results for experiments described in Section 4

Reader Approach	Dev. EM	Dev. F1	Test EM	Test F1
$mFiD$	49.4	58.249	23.919	29.351
$mFiD$ + xLingual Data Aug.	45.3	51.119	22.116	28.163
$mFiD$ + xLingual Output Regularization	14	21.562	5.731	10.874
$mFiD$ + xLingual Output Regularization w/ Lang ID	13.4	22.682	5.405	10.645
$mFiD$ + xLingual `xAttn` regularization	48.9	55.782	21.071	26.803
$mFiD$ + Ret^r Scores KL-Divergence	**52.1**	**60.745**	**26.84**	**32.536**
$mFiD$ + Ret^r Scores Attention Offset	**52.1**	**60.68**	**26.748**	**32.758**

Table 4: `mFiD` reader results for experiments described in Section 4

sults from the Attention Offset and KL-Divergence experiments are similar to each other, requiring further further experimentation across datasets to identify the better technique.

6 Conclusion

Cross-lingual open-domain question-answering (xODQA) is an important task that, if performed effectively, has the potential to improve access to information across a diverse set of demographics. In this work, we explore multiple strategies to improve the standard reader-retriever framework for performing xODQA. We follow two lines of methodology - (1) strengthening the reader and retriever individually, and (2) enabling two-way distillation of knowledge between the reader and retriever which, to the best of our knowledge, was only explored from reader to retriever in previous literature.

Our results demonstrate that self-training, cross-lingual data augmentation, and entity augmentation techniques can significantly improve the retriever component. The same however does not hold true for the reader, where these techniques result in an over all decrease in performance. We hypothesize that this decrease in performance arises due to the difference in proportion of cross-linguality between the training and test datasets. Exploring methods for regularizing the effect of these techniques so that overall performance still holds is, hence, an important direction for future research.

As observed in the results, the two-way knowledge distillation framework helps improve both, the reader and the retriever. The respective inductive biases of these two components encode information that can be useful for the other component in performing it's task. The presented `Froward-Backward` optimization framework is capable of successfully leverage this, as is evident from the results. Similar knowledge distillation techniques have been shown to improve performance over multiple iterations. Thus, scaling the `Froward-Backward` optimization loop to multiple iterations is a direction that can be explored in future work.

References

Sumit Agarwal, Suraj Tripathi, Teruko Mitamura, and Carolyn Penstein Rose. 2022. Zero-shot cross-lingual open domain question answering. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 91–99, Seattle, USA. Association for Computational Linguistics.

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

Akari Asai, Shayne Longpre, Jungo Kasai, Chia-Hsuan Lee, Rui Zhang, Junjie Hu, Ikuya Yamada, Jonathan H. Clark, and Eunsol Choi. 2022. Mia 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages.

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

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

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

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2021. DEEP: denoising entity pre-training for neural machine translation. *CoRR*, abs/2111.07393.

Gautier Izacard and Edouard Grave. 2020a. Distilling knowledge from reader to retriever for question answering.

Gautier Izacard and Edouard Grave. 2020b. Distilling knowledge from reader to retriever for question answering. *CoRR*, abs/2012.04584.

Gautier Izacard and Edouard Grave. 2020c. Leveraging passage retrieval with generative models for open domain question answering.

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, Online. Association for Computational Linguistics.

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*.

Haejun Lee, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher D. Manning, and Kyoung-Gu Woo. 2021. You only need one model for open-domain question answering.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *CoRR*, abs/2106.14448.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for 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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems*, volume 34, pages 25968–25981. Curran Associates, Inc.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. *CoRR*, abs/1905.11471.

Nikita Sorokin, Dmitry Abulkhanov, Irina Piontkovskaya, and Valentin Malykh. 2022. Ask me anything in your native language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 395–406, Seattle,

United States. Association for Computational Linguistics.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2021. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. *CoRR*, abs/2110.04330.