Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering

Arij Riabi[‡*] **Thomas Scialom**[⋆◇*] **Rachel Keraron**[⋆]
Benoît Sagot[‡] **Djamé Seddah**[‡] **Jacopo Staiano**[⋆]
[‡] Inria, Paris, France
[◇] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[⋆] reciTAL, Paris, France
{thomas,rachel,jacopo}@recital.ai
{arij.riabi,benoit.sagot,djame.seddah}@inria.fr

Abstract

Coupled with the availability of large scale datasets, deep learning architectures have enabled rapid progress on Question Answering tasks. However, most of those datasets are in English, and the performances of state-of-the-art multilingual models are significantly lower when evaluated on non-English data. Due to high data collection costs, it is not realistic to obtain annotated data for each language one desires to support.

We propose a method to improve Cross-lingual Question Answering performance without requiring additional annotated data, leveraging Question Generation models to produce synthetic samples in a cross-lingual fashion. We show that the proposed method allows to significantly outperform the baselines trained on English data only, establishing thus a new state-of-the-art on four multilingual datasets: MLQA, XQuAD, SQuAD-it and PIAF (fr).

1 Introduction

Question Answering is a fast-growing research field, aiming to improve the capabilities of machines to read and understand documents. Significant progress has recently been enabled by the use of large pre-trained language models (Devlin et al., 2019; Raffel et al., 2020), which reach human-level performances on several publicly available benchmarks, such as SQuAD (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2017).

Given that the majority of large scale Question Answering (QA) datasets are in English (Hermann et al., 2015; Rajpurkar et al., 2016; Choi et al., 2018), the development of QA systems targeting other languages is currently addressed via two cross-lingual QA datasets: XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020a), covering respectively 10 and 7 languages. Due to the cost of annotation, both are limited only to an evaluation set. They are comparable to the validation set of the original SQuAD (see more details in Section 3.3). In both datasets, each paragraph is paired with questions in various languages, allowing to evaluate models in a *cross-lingual* experimental scenario: the input context and the question can be in two different languages. This scenario has important practical applications, such as querying a set of documents in various languages.

Performing this cross-lingual task is complex and remains challenging for current models, assuming only English training data: transfer results are shown to rank behind training-language performance (Artetxe et al., 2020; Lewis et al., 2020a). In other words, multilingual models fine-tuned only on English data are found to perform significantly better on English than on other languages. Besides the almost simultaneous work of Shakeri et al. (2020), very few alternatives to such a simple zero-shot transfer method have been proposed so far.

In this paper, we propose to generate synthetic data in a cross-lingual fashion, borrowing the idea from monolingual QA research efforts (Duan et al., 2017). On English corpora, generating synthetic questions has shown to significantly improve the performance of QA models (Du et al., 2017; Golub et al., 2017; Du and Cardie, 2018; Alberti et al., 2019). However, the adaptation of this technique to cross-lingual QA is not straightforward: cross-lingual text generation is a challenging task *per se* which has not been yet extensively explored, in particular when no multilingual training data is available.

We explore two Question Generation scenarios: (i) requiring only SQuAD data; and (ii) using a translator tool to obtain translated versions of SQuAD. As expected, the method leveraging on a translator has shown to perform the best. Leveraging on such synthetic data, our best model obtains

[*]: equal contribution. The work of Arij Riabi was partly carried out while she was working at reciTAL.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7016–7030
November 7–11, 2021.

significant improvements on XQuAD and MLQA over the state-of-the-art for both Exact Match and F1 scores. In addition, we evaluate the QA models on languages not seen during training (even for the synthetic data) – using SQuAD-it (for Italian), PIAF (for French), and KorQUaD (for Korean) – reporting a new state-of-the-art for Italian and French, and observing significant improvements on Korean compared to zero-shot without augmentation. This indicates that the proposed method allows to capture better multilingual representations beyond the training languages. Our method paves the way toward multilingual QA domain adaptation, especially for under-resourced languages.

Our contributions can be summarized as follows:

- We present a data augmentation approach for Cross-Lingual Question Answering based on synthetic Question Generation;
- We report extensive experiments showing significant improvements on two multilingual evaluation datasets (XQuAD and MLQA);
- We additionally evaluate the proposed methodology on languages unseen during training, thus showing the potential benefits for QA on low-resource languages.

2 Related Work

Question Answering (QA) QA is the task for which given a context and a question, a model has to find the answer. The interest for Question Answering goes back a long way: in a 1965 survey, Simmons (1965) reported fifteen implemented English language question-answering systems. More recently, with the rise of large scale datasets (Hermann et al., 2015), and large pre-trained models (Devlin et al., 2019), the performance drastically increased, approaching human-level performance on standard benchmarks – see for instance the SQuAD leader board.[1] More challenging evaluation benchmarks have recently been proposed: Dua et al. (2019) released the DROP dataset, for which the annotators were encouraged to provide adversarial questions; Burchell et al. (2020) released the MSQ dataset, consisting of multi-sentence questions.

However, all these works are focused on English. Another popular research direction focuses on the development of multilingual QA models. For this purpose, the first step has been to provide the community with multilingual evaluation sets: Artetxe et al. (2020) and Lewis et al. (2020a) concurrently proposed two different evaluation sets which are comparable to the SQuAD development set. Both reach the same conclusion: due to the lack of non-English training data, models do not achieve the same performance in Non-English languages than they do in English. To the best of our knowledge, no method has been proposed to fill this gap.

Question Generation (QG) QG can be seen as the dual task of QA: the input is composed of the *answer* and the *paragraph* containing it, and the model is trained to generate the *question*. Proposed by Rus et al. (2010), it has leveraged on the development of new QA datasets (Zhou et al., 2017; Scialom et al., 2019). Similar to QA, significant performance improvements have been obtained using pre-trained language models (Dong et al., 2019). Still, due to the lack of multilingual datasets, most previous works have been limited to monolingual text generation. We note the exceptions of Kumar et al. (2019) and Chi et al. (2020), who resorted to multilingual pre-training before fine-tuning on monolingual downstream NLG tasks. However, the quality of the generated questions is still found inferior to the corresponding English ones.

Question Generation for Question Answering Data augmentation via synthetic data generation is a well-known technique to improve models' accuracy and generalisation. It has found successful application in several areas, such as time series analysis (Forestier et al., 2017) and computer vision (Buslaev et al., 2020). In the context of QA, generating synthetic questions to complete a dataset has shown to improve QA performances (Duan et al., 2017; Alberti et al., 2019). So far, all these works have focused on English QA given the difficulty to generate questions in other languages without available data. This lack of data, and the difficulty to obtain some, constitutes the main motivation of our work and justifies exploring cost-effective approaches such as data augmentation via the generation of questions.

In a very recent work, almost simultaneous to our previously submitted version, Shakeri et al. (2020) address multilingual QA with a similar approach. However, we argue that their experimental protocol does not allow to totally answer the research question. We detail the differences in our discussion, Section 5.3.

[1] https://rajpurkar.github.io/SQuAD-explorer/

3 Data

3.1 English Training Data

SQuAD$_{en}$ The original SQuAD (Rajpurkar et al., 2016), which we refer as SQuAD$_{en}$ for clarity in this paper. It is one of the first, and among the most popular, large scale QA datasets. It contains about 100K question/paragraph/answer triplets in English, annotated via Mechanical Turk.[2]

QG datasets Any QA dataset can be reversed into a QG dataset, by switching the generation targets from the answers to the questions. In this paper, we use the *qg* subscript to specify when the dataset is used for QG (e.g. SQuAD$_{en;qg}$ indicates the English SQuAD data in QG format).

3.2 Synthetic Training Sets

SQuAD$_{trans}$ is a machine translated version of the SQuAD train set in the seven languages of MLQA, released by the authors together with their paper.

WikiScrap We collected 500 Wikipedia articles for all the languages present in MLQA. They are not paired with any question or answer. We use them as contexts to generate synthetic multilingual questions, as detailed in Section 4.2. Following the SQuAD$_{en}$ protocol, we used project Nayuki's code[3] to parse the top 10K Wikipedia pages according to the PageRank algorithm (Page et al., 1999). We then filtered out paragraphs with character length outside of a [500, 1500] interval. Articles with less than 5 paragraphs are discarded, since they tend to be less developed, in a lower quality or being only redirection pages. Out of the filtered articles, we randomly selected 500 per language.

3.3 Multilingual Evaluation Sets

XQuAD (Artetxe et al., 2020) is a human translation of the SQuAD$_{en}$ development set in 10 languages (Arabic, Chinese, German, Greek, Hindi, Russian, Spanish, Thai, Turkish, and Vietnamese), providing 1k QA pairs for each language.

MLQA (Lewis et al., 2020a) is an evaluation dataset in 7 languages (English, Arabic, Chinese, German, Hindi, and Spanish). The dataset is built from aligned Wikipedia sentences across at least two languages (full alignment between all languages being impossible), with the goal of providing natural rather than translated paragraphs. The QA pairs are manually annotated on the English sentences and then human translated on the aligned sentences. The dataset contains about 46k aligned QA pairs in total.

Language-specific benchmarks In addition to the two aforementioned multilingual evaluation corpora, we benchmark our models on three language-specific datasets for French, Italian and Korean, as detailed below. We choose these datasets since none of these languages are present in XQuAD or MLQA. Hence, they allow us to evaluate our models in a scenario where the target language is not available during training, even for the synthetic questions.

PIAF Keraron et al. (2020) provided an evaluation set in French following the SQuAD protocol, containing 3835 examples.

KorQuAD 1.0 the Korean Question Answering Dataset (Lim et al., 2019), a Korean dataset also built following the SQuAD protocol.

SQuAD-it Derived from SQuAD$_{en}$, it was obtained via semi-automatic translation to Italian (Croce et al., 2018).

4 Models

Recent works (Raffel et al., 2020; Lewis et al., 2019) have shown that classification tasks can be framed as a text-to-text problem, achieving state-of-the-art results on established benchmarks, such as GLUE (Wang et al., 2018). Accordingly, we employ the same architecture for both Question Answering and Generation tasks. This also allows fairer comparisons for our purposes, by removing differences between QA and QG architectures and their potential impact on the results obtained. In particular, we use a distilled version of XLM-R (Conneau et al., 2020): MiniLM-M (Wang et al., 2020) (see Section 4.3 for further details).

4.1 Baselines

QA$_{\text{No-synth}}$ Following previous works, we fine-tuned the multilingual models on SQuAD$_{en}$, and consider them as our baselines.

[2] Two versions of SQuAD have been released: v1.1, used in this work, and v2.0. The latter contains "unanswerable questions" in addition to those from v1.1. We use the former, since the multilingual evaluation datasets, MLQA and XQUAD, do not include unanswerable questions.

[3] https://www.nayuki.io/page/computing-wikipedias-internal-pageranks

English as Pivot Leveraging on translation models, we consider a second baseline method, which uses English as a pivot. First, both the question in language L_q and the paragraph in language L_p are translated into English. We then invoke the baseline model described above, $\text{QA}_{\text{No-synth}}$, to predict the answer. Finally, the predicted answer is translated back into the target language L_p. We used the google translate API.[4]

$\text{QA}_{\text{+SQuAD-trans}}$ the translated data $\text{SQuAD}_{\text{trans}}$ are used as additional training data to SQuAD_{en}, to train the QA model.

4.2 Question Generation Data Augmentation

In this work we consider data augmentation via generating synthetic questions, to improve the QA performance. Different training schemes for the question generator are possible, resulting in different quality of the synthetic data. Before this work, its impact on the final QA system remained unexplored in a multilingual context.

For all the following experiments, only the synthetic data changes. Given a specific set of synthetic data, we always follow the same two-stages protocol, similar to Alberti et al. (2019): we first train the QA model on the synthetic QA data, then on SQuAD_{en}. We also tried to train the QA model in one stage, with all the synthetic and human data shuffled together, but observed no improvements over the baseline.

We explored two different synthetic generation modes:

Synth the QG model is trained on $\text{SQuAD}_{en,qg}$ (i.e., English data only) and the synthetic data are generated on WikiScrap. Under this setup, the only annotated samples this model has access to are those from SQuAD-en.

Synth+trans the QG model is trained on $\text{SQuAD}_{trans,qg}$ in addition to $\text{SQuAD}_{en,qg}$. The questions can thus be in a different languages than the context. Hence, the model needs an indication about the language it is expected to generate the question in. To control the target language, we use a specific prompt per language, defining a special token `<LANG>`, which corresponds to the desired target language Y. Thus, the input is structured as `<LANG> <SEP> Answer <SEP> Context`, where `<LANG>` indicates to the model in what language the question should be generated, and `<SEP>` is a special token acting as a separator. These attributes offer flexibility on the target language. Similar techniques are used in the literature to control the style of the output (Keskar et al., 2019; Scialom et al., 2020; Chi et al., 2020).

4.3 Implementation details

For all our experiments we use Multilingual MiniLM v1 (MiniLM-m) (Wang et al., 2020), a 12-layer with 384 hidden size architecture distilled from XLM-R Base multilingual (Conneau et al., 2020). With only 66M parameters, it is an order of magnitude smaller than state-of-the-art architectures such as BERT-large or XLM-large. We used the official Microsoft implementation.[5] For all the experiments –both QG and QA– we trained the model for 5 epochs, using the default hyperparameters. We used a single nVidia gtx2080ti with 11G RAM, and the training times amount to circa 4 and 2 hours for Question Generation and for Question Answering, respectively. To evaluate our models, we used the official MLQA evaluation scripts.[6] For reproducibility purposes, we make the code available.[7]

5 Results

5.1 Question Generation

We report examples of generated questions in Table 1.

Controlling the Target Language In the context of multilingual text generation, controlling the target language is not trivial.

When a QA model is trained only on English data, at inference, given a non-English paragraph, it predicts the answer in the input language, as one would expect, since it is an extractive process. Ideally, we would like to observe the same behavior for a Question Generation model trained only on English data (such as $Synth$), leveraging on the multilingual pre-training. Conversely to QA, QG is a language generation task. *Multilingual* generation is much more challenging, as the model's decoding ability plays a major role. When a QG model is fine-tuned only on English data (i.e SQuAD-en), its controllability of the target language suffers from catastrophic forgetting: the input language does not

[4] https://translate.google.com

[5] Publicly available at https://github.com/microsoft/unilm/tree/master/minilm.

[6] https://github.com/facebookresearch/MLQA/blob/master/mlqa_evaluation_v1.py

[7] https://anonymous.4open.science

Paragraph (EN) Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.
Answer Broncos
QG_{synth} What team did John Elway lead to victory at age 38?
$QG_{synth+trans}$ (target language = en) What team did John Elway lead to win in the Super Bowl?

Paragraph (ES) Peyton Manning se convirtió en el primer mariscal de campo de la historia en llevar a dos equipos diferentes a participar en múltiples Super Bowls. Ademas, es con 39 años, el mariscal de campo más longevo de la historia en jugar ese partido. El récord anterior estaba en manos de John Elway —mánager general y actual vicepresidente ejecutivo para operaciones futbolísticas de Denver— que condujo a los Broncos a la victoria en la Super Bowl XXXIII a los 38 años de edad.
Answer Broncos
QG_{synth} Where did Peyton Manning condujo?
$QG_{synth+trans}$ (target language = es) Qué equipo ganó el récord anterior? (*Which team won the previous record?*)
$QG_{synth+trans}$ (target language = en) What team did Menning win in the Super Bowl?

Paragraph (ZH) 培顿·曼宁成为史上首位带领两支不同球队多次进入超级碗的四分卫。他也以39 岁高龄参加超级碗而成为史上年龄最大的四分卫。过去的记录是由约翰·埃尔维保持的，他在38岁时带领野马队赢得第33 届超级碗，目前担任丹佛的橄榄球运营执行副总裁兼总经理
Answer 野马队
QG_{synth} What is the name for the name that the name is used?
$QG_{synth+trans}$ (target language = zh) 约翰·埃尔维在13岁时带领哪支球队赢得第33届超级碗? (*Which team did John Elvey lead to win the 33rd Super Bowl at the age of 13?*)
$QG_{synth+trans}$ (target language = en) What team won the 33th Super Bowl?

Table 1: Example of questions generated by the different models on an XQuAD's paragraph in different languages. For $QG_{synth+trans}$, we report the outputs given two target languages, the one of the context and English.

propagate to the generated text. While still relevant to the context, the synthetic questions are generated in English: for instance, in Table 1 we observe that the QG_{synth} model outputs English questions for the paragraphs in Chinese and Spanish. The same phenomenon was reported by Chi et al. (2020).

Cross-Lingual Training To overcome the aforementioned limitation on target language controllability (i.e. to enable the generation in other languages than English), multilingual data is needed. We can leverage on the translated versions of the dataset to add the required non-English examples. As detailed in Section 4.2, we simply use a specific prompt that corresponds to the target language (with N different prompts corresponding to the N languages present in the dataset). In Table 1, we show how $QG_{synth+trans}$ can generate questions in the same language as the input. These synthetic questions seem much more relevant, coherent and fluent, if compared to those produced by QG_{synth}: for the Spanish paragraph, the question is well formed and focused on the input answer; for Chinese (see bottom row of Table 1 for $QG_{synth+trans}$) is perfectly written.

In Table 2 we report the BLEU4 scores for $QG_{synth+trans}$ grouped by the language of the question. As expected, the score is maximized on the diagonal (same languages for the context and the question). Still, most of these scores are lower on non-English languages. It is interesting to note BLEU4 correlates with the QA scores: 0.51 Pearson coefficient. The reasons are two folds: 1) QA and QG share the same Language Model, which might struggle for the same languages; 2) the better the QG, the better the synthetic data, therefore the better the QA performs. We discuss further in Section 5.3 how this impacts the QA performance.

q/c	en	es	de	ar	hi	vi	zh
en	14.5	8.9	7.2	5.9	6.5	8.4	6.0
es	9.0	10.	6.6	4.2	5.9	6.3	4.6
de	6.2	4.8	6.3	3.1	3.7	5.0	3.2
ar	2.8	2.2	2.4	3.3	2.0	2.3	2.1
hi	7.9	6.7	6.6	5.8	8.3	6.6	5.2
vi	9.1	7.3	7.2	6.0	6.5	12.3	6.1
zh	9.2	8.0	7.8	6.1	7.2	8.0	15.0

Table 2: BLEU-4 scores on MLQA test for $QG_{synth+trans}$. Columns show context language, rows show question language.

In addition to BLEU, we also report the QA F1 scores for different QA models when applied on the generated questions in the supplementary material. Yet, we warn the reader that these results should be taken with caution: evaluating NLG is known to be an open research problem; BLEU is known to

suffer from important limitations (Novikova et al., 2017), which might be accentuated in a multilingual context (Lee et al., 2020). For this reason, we conducted a manual qualitative analysis on a small number of samples. Note that the annotators need to have a professional level in the language of the generated question to evaluate its fluency, and to be bilingual, when evaluating its relevance w.r.t. input context in our cross language scenario. This is a significant challenge to conduct a large scale evaluation.

So far, our results (see at the end of Supplementary Material) for Arabic and German show an overall good quality in the questions: only one question for Arabic was genuinely missing the point while for German there were 2 lexical questionable choices that invalidate the question (out of 10 samples for both languages so far). This indicates that Arabic questions could actually be better than what their low BLEU score shows. Arabic has a very different morphological structure that could explain such low BLEU (Bouamor et al., 2014). This emphasizes the limitation of the current automatic metrics in a multilingual context.

5.2 Question Answering

We report the main results of our experiments on XQuAD and MLQA in Table 3. The scores correspond to the average over all the different possible combination of languages (*de-de*, *de-ar*, etc.).

English as Pivot Using English as a pivot does not lead to good results. This may be due to the evaluation metrics, which are based on n-grams similarity. For extractive QA, F1 and EM metrics measure the overlap between the predicted answer and the ground truth. Therefore, meaningful answers worded differently are penalized, a situation that is likely to occur because of the back-translation mechanism. This makes automatic evaluation challenging for this setup, as metrics suffer from similar difficulties as those observed for text generation (Sulem et al., 2018). As an additional downside, this model requires multiple translations *at inference time*. For these reasons, we decided not to explore this approach further.

Synthetic without translation (+synth) Compared to the MiniLM baseline, we observe a small performance increase for MiniLM$_{\text{+synth}}$ (Exact Match increases from 29.5 to 33.1 on XQuAD and from 26.0 to 27.5 on MLQA).

During the self-supervised pre-training stage, the model was exposed to multilingual inputs. Yet, for a given input, the target language was always consistent, preventing the model to be exposed to such a cross-lingual scenario. The synthetic inputs are composed of questions in English (see examples in Table 1) while the contexts can be in any languages. Therefore, the QA model is exposed for the first time to a cross-lingual scenario. We hypothesise that such a cross-lingual ability is not innate for a default multilingual model: exposing a model to this scenario allows to develop this ability and contributes to improve its performance.

Synthetic with translation (+synth-trans) : For MiniLM$_{\text{+synth-trans}}$, we obtain a much larger improvement over its baselines, MiniLM, compared to MiniLM$_{\text{+synth}}$, on both MLQA and XQuAD. Also, it outperforms MiniLM$_{\text{+SQuADtrans}}$, indicating the benefit of our proposed approach. This supports the intuition developed in the previous paragraph: independently of the multilingual capacity of the model, a cross-lingual ability is developed when the two inputs components are not exclusively written in the same language. In Section 5.3, we discuss this phenomenon more in depth.

5.3 Discussion

Cross Lingual Generalisation To explore the models' effectiveness in dealing with cross-lingual inputs, we report in Figure 1 the performance for our MiniLM$_{\text{+synth-trans}}$ setup, varying the number of samples and the languages present in the synthetic data. The abscissa x corresponds to the progressively increasing number of synthetic samples used; at $x = 0$, it corresponds to the MiniLM$_{\text{+trans}}$ baseline, where the model has access only to the original English data from SQuAD$_{en}$. We explore two sampling strategies for the synthetic examples:

0. *All Languages* corresponds to sampling the examples from any of the different languages.
0. Conversely, for *Not All Languages*, we progressively added the different languages: for $x = 50K$, all the 50K synthetic data are on a unique language input, $L1$. Then for $x = 100K$, the synthetic data are from either $L1$, or an additional language $L2$; finally, for $x = 250K$, all MLQA languages are present.

In Figure 1, we observe that the performance for *All Languages* increases largely at the beginning, then remains mostly stable. Conversely, we note a gradual improvement for *Not All Languages*, as

	#Params	Trans.	XQuAD	MLQA
MiniLM (Wang et al., 2020)	66M	No	42.2 / 29.5	38.4 / 26.0
XLM (Hu et al., 2020)[8]	340M	No	68.5/52.8	**65.4** / 47.9
English as Pivot	66M	Yes	46.2 / 30.9	36.1 / 23.0
$MiniLM_{+synth}$	66M	No	44.8 / 33.1	39.8 / 27.5
$MiniLM_{+SQuAD-trans}$	66M	Yes	55.0 / 40.7	49.5 / 35.3
$MiniLM_{+synth-trans}$	66M	Yes	63.3 / 49.1	56.1 / 41.4
$MiniLM_{+SQuAD-trans+synth-trans}$	66M	Yes	62.5 / 48.6	55.0 / 40.4
$XLM-R_{+synth-trans}$	340M	Yes	**74.3 / 59.2**	**65.3 / 49.2**

Table 3: Results (F1 / EM) of the different QA models on XQuAD and MLQA. XLM corresponds to the large version.

	#Params	PIAF (fr)	KorQuAD	SQuAD-it
MiniLM (Wang et al., 2020)	66M	58.9 / 34.3	53.3 / 40.5	72.0 / 57.7
mBert (Devlin et al., 2019)	110M	64.4 / 42.5	-	74.1 / 62.5
CamemBERT (Martin et al., 2020)	340M	68.9 / -	N/A	N/A
$MiniLM_{+synth}$	66M	58.6 / 34.5	52.1 / 39.0	71.3 / 58.0
$MiniLM_{+synth-trans}$	66M	63.9 / 40.6	60.0 / 48.8	74.5 / 62.0
$XLM-R_{+synth-trans}$	340M	**72.1 / 47.1**	**63.0 / 52.8**	**80.4 / 67.6**

Table 4: Zero-shot results (F1 / EM) on PIAF, KorQuAD and SQuAD-it for our different QA models, compared to various baselines. For mBert on SQuAD-it, we report the score from Croce et al. (2018). Note that CamemBERT is a French version of RoBERTa, an architecture widely outperforming BERT.

more languages are made available during training. This shows that when all the languages are present in the synthetic data, the model immediately develops cross-lingual abilities.

However, it appears that even with only one language pair present, the model is able to develop a cross-lingual ability that brings benefits on other languages: of Figure 2, we can see that most of the improvement is happening given only one cross-lingual language pair (i.e. English and Spanish).

Unseen Languages To measure the benefit of our approach on unseen languages (i.e. not present in the synthetic data from MLQA/XQuAD), we test our models on three QA evaluation sets: PIAF (fr), KorQuAD and SQuAD-it (see Section 3.3). The results are consistent with the previous experiments on MLQA and XQuAD. Our $MiniLM_{+synth-trans}$ model outperforms its baseline by more than 4 Exact Match points, while $XLM-R_{+synth-trans}$ obtains a new state-of-the-art. Notably, our multilingual $XLM-R_{+synth-trans}$ outperforms CamemBERT on PIAF, even if the latter is a pure monolingual, in-domain language model.

On the correlation between BLEU4 and QA scores To measure the impact of the quality of the generated questions on the QA performance, we computed the Pearson correlation between the BLEU4 and the QA scores. The coefficient is equal to 0.65 ($p < .001$). When we observe the correlations grouping the samples w.r.t. their language question (i.e. the rows in Table 2), we obtain: *en 0.94; es 0.84; de 0.46; ar 0.36; hi 0.33; vi 0.73; zh 0.92*. We observe stronger correlation for languages with higher BLEU scores (i.e en & zh), and lower for the Arab that had the lowest BLEU, indicating an impact on the final QA score in par to the quality of the synthetic questions.

Differences with Shakeri et al. (2020) A very recent work has addressed multilingual QA with a very similar approach. However, we note a major difference in our respective experiments regarding the choice for the QA and QG models. Shakeri et al. (2020) choose mBert for QA and T5-m for QG. We would like to emphasize that because T5-m significantly outperforms mBert it is not clear where the improvement comes from: is it due to the proposed approach, or simply from a distillation effect from T5-m to mBert? In our case, we

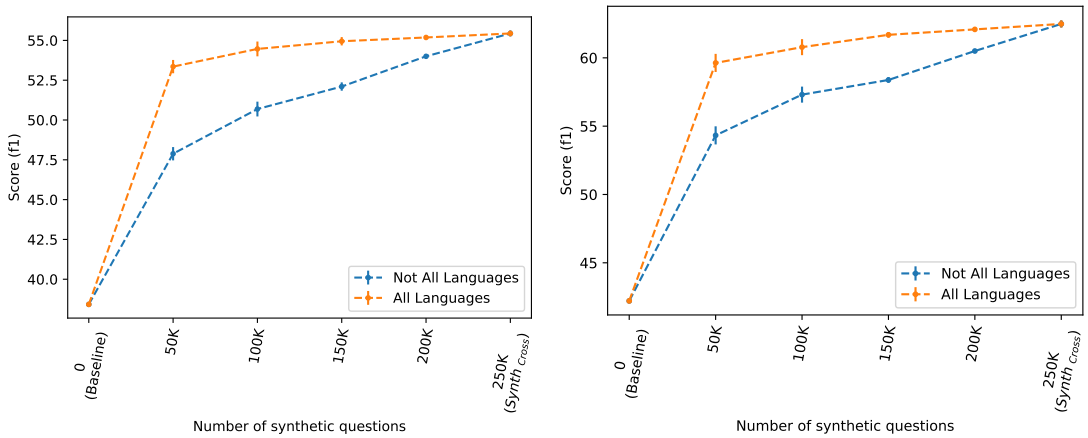

Figure 1: Left: F1 score on MLQA, for models with different number of synthetic data in two setups: for *All Languages*, the synthetic questions are sampled among all the five languages in MLQA; for *Not All Languages*, the synthetic questions are sampled progressively from only one language, two, . . . , to all five for the last point, which corresponds to *All Languages*. We report the standard deviation over five different permutations of the language ordering. Note that, as expected, the more the synthetic data, the lower the variance in the results. Right: same as on the left, but evaluated on XQuAD.

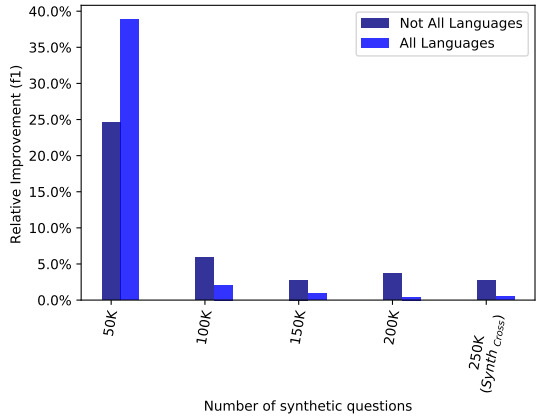

Figure 2: The relative variation in performance for the models in Figure 1.

deliberately used MiniLM *for both QA and QG*: this allows a fairer investigation about the benefits of the proposed approach.

Hidden distillation effect The relative improvement for our best synthetic configuration *+synth-trans*, over the baseline, is above 60% EM for MiniLM (from 29.5 to 49.5 on XQuAD and from 26.0 to 41.4 on MLQA). Significantly higher than that observed for XLM-R (+11.7% on XQuAD and +2.71% on MLQA), it indicates that XLM-R provides superior cross-lingual transfer abilities than MiniLM, a fact that we hypothesize due to distillation. Such loss of generalisation can be difficult to identify, and opens questions for future work.

QA, an unsolved task for lower resource languages Factoid QA tasks have been criticized for being a too easy task: the answer can often be identified given simple heuristics: e.g. a "When" question is answered by one of the "date" spans in the context (Kočiský et al., 2018; Kwiatkowski et al., 2019). SQuAD-v2 was for instance introduced to increase the difficulty of the task by adding unanswerable questions. The research community is now moving towards the construction of long context questions and non-factoid QA datasets (Dulceanu et al., 2018; Hashemi et al., 2019; Fan et al., 2019; Lewis et al., 2020b). In any case, the motivation of this work was to cope for the lack of training data for under-served languages in the QA domain which was severely impacting models performance. Therefore, potential criticisms regarding the simplicity of the task do not apply if seen from a lower-resource language scenario: our work deals with alleviating the lack of native training data, allowing us to focus our future work on further important issues such as domain adaptation, robustness and explainability in low-resource contexts.

6 Conclusion

In this work, we presented a method to generate synthetic QA dataset in a multilingual fashion, showing how QA models can benefit from it and reporting large improvements over the baselines. The proposed approach contributes to fill the gap between English and other languages, and is shown to generalize for languages not present in the synthetic corpus (e.g. French, Italian, Korean).

In future work, we plan to investigate whether the proposed data augmentation method could be applied to other multilingual tasks, such as classification. We will also experiment more in depth with different strategies to control the target language of a model, and extrapolate on unseen ones.

7 Acknowledgments

Djamé Seddah was partly funded by the French Research National Agency via the ANR project ParSiTi (ANR-16-CE33-0021), Arij Riabi was partly funded by Benoît Sagot's chair in the PRAIRIE institute as part of the French national agency ANR "Investissements d'avenir" programme (ANR-19-P3IA-0001) and by the Counter H2020 European project (grant 101021607).

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for arabic mt evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213.

Laurie Burchell, Jie Chi, Tom Hosking, Nina Markl, and Bonnie Webber. 2020. Querent intent in multi-sentence questions. *arXiv preprint arXiv:2010.08980*.

Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. 2020. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *AAAI*, pages 7570–7577.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Andrei Dulceanu, Thang Le Dinh, Walter Chang, Trung Bui, Doo Soon Kim, Manh Chien Vu, and Seokhwan Kim. 2018. PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Germain Forestier, François Petitjean, Hoang Anh Dau, Geoffrey I Webb, and Eamonn Keogh. 2017. Generating synthetic time series to augment sparse datasets. In *2017 IEEE international conference on data mining (ICDM)*, pages 865–870. IEEE.

David Golub, Po-Sen Huang, Xiaodong He, and Li Deng. 2017. Two-stage synthesis networks for transfer learning in machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 835–844, Copenhagen, Denmark. Association for Computational Linguistics.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2019. Antique: A non-factoid question answering benchmark.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: Building a native French question-answering dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5481–5490, Marseille, France. European Language Resources Association.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872, Florence, Italy. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Dongyub Lee, Myeongcheol Shin, Taesun Whang, Seungwoo Cho, Byeongil Ko, Daniel Lee, Eunggyun Kim, and Jaechoon Jo. 2020. Reference and document aware semantic evaluation methods for korean language summarization. In *Proceedings of COLING 2020, the 30th International Conference on Computational Linguistics: Technical Papers*. The COLING 2020 Organizing Committee.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Christian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6027–6032, Florence, Italy. Association for Computational Linguistics.

Thomas Scialom, Serra Sinem Tekiroglu, Jacopo Staiano, and Marco Guerini. 2020. Toward stance-based personas for opinionated dialogues. *arXiv preprint arXiv:2010.03369*.

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

Robert F Simmons. 1965. Answering english questions by computer: a survey. *Communications of the ACM*, 8(1):53–70.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 662–671. Springer.

Appendix

A On target language control for text generation

When relying on the translated versions of SQuAD, the target language for generating synthetic questions can easily be controlled, and results in fluent and relevant questions in the different languages. However, one limitation of this approach is that synthetic questions can only be generated in the languages that were available during training: the <LANG> prompts are special tokens that are randomly initialised when fine-tuning QG on SQuAD$_{trans;qg}$: before fine-tuning, they bear no semantic relation with the corresponding language names ("English", "Español" etc.), thus the learned representations for the <LANG> tokens are limited to the languages present in the training set.

To the best of our knowledge, no method allows so far to generalize this target control to an unseen language. It would be valuable, for instance, to be able to generate synthetic data in Korean, French and Italian, without having to translate the entire SQuAD$-en$ dataset in these three languages to then fine-tune the QG model.

To this purpose, we report – alas, as a negative result – the following attempt: instead of controlling the target language with a special, randomly initialised, token, we used a token semantically related to the language-word: "English", "Español" for Spanish, or "中文" for Chinese. The intuition is that the model might adopt the correct language at inference, even for a target language unseen during training.[9] A similar intuition has been explored in GPT-2: the authors report an improvement for summarization when the input text is followed by "TL;DR" (i.e. Too Long Didn't Read).

At inference time, we evaluated this approach on French with the prompt `language=Français`. Unfortunately, the model did not succeed to generate text in French. Controlling the target language in the context of multilingual text generation remains under-explored, and progress in this direction could have direct applications to improve this work, and beyond.

B Question Generation Scores

We report the BLEU-4 scores for MLQA on QG$_{\text{synth+trans}}$ on Table 5 and QG$_{\text{synth}}$ on Table 6.

[9] With *unseen during training*, we mean *not present in the QG dataset*; obviously, the language should have been present in the first self-supervised stage.

q/c	en	es	de	ar	hi	vi	zh
en	14.5	8.9	7.2	5.9	6.5	8.4	6.0
es	9.0	10.	6.6	4.2	5.9	6.3	4.6
de	6.2	4.8	6.3	3.1	3.7	5.0	3.2
ar	2.8	2.2	2.4	3.3	2.0	2.3	2.1
hi	7.9	6.7	6.6	5.8	8.3	6.6	5.2
vi	9.1	7.3	7.2	6.0	6.5	12.3	6.1
zh	9.2	8.0	7.8	6.1	7.2	8.0	15.0

Table 5: BLEU4 scores on MLQA test for QG$_{\text{synth+trans}}$ model.

q/c	en	es	de	ar	hi	vi	zh
en	21.7	4.73	4.58	2.47	2.58	3.02	3.11
es	0.8	1.23	0.38	0.0	0.0	0.22	0.11
de	1.4	0.85	1.32	0.0	0.0	0.22	0.0
ar	0.0	0.0	0.0	0.0	0.0	0.0	0.0
hi	0.0	0.21	0.0	0.0	0.0	0.0	0.0
vi	0.55	0.5	0.25	0.0	0.0	0.34	0.0
zh	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 6: BLEU-4 scores on MLQA test for QG$_{\text{synth}}$ model.

q/c	en	es	de	ar	hi	vi	zh
en	83.9	74.8	70.1	67.8	72.3	74.1	69.3
es	78.0	74.3	68.1	63.6	67.4	67.6	69.1
de	75.6	72.9	70.1	65.0	67.5	68.5	70.4
ar	61.3	58.4	56.8	66.7	57.7	56.5	69.8
hi	70.9	62.1	58.9	56.8	70.9	61.3	69.2
vi	71.0	64.1	60.5	59.7	63.7	74.7	69.9
zh	67.1	62.5	59.0	56.7	60.7	63.2	69.3

Table 7: F1 score on MLQA for XLM-R model finetuned on SQuAD$_{\text{en}}$.

q/c	en	es	de	ar	hi	vi	zh
en	83.9	75.4	70.9	68.9	72.8	75.6	66.8
es	81.0	74.4	71.8	66.6	70.2	72.5	65.7
de	81.4	75.2	70.8	69.3	70.2	74.4	65.1
ar	76.3	68.9	67.3	66.6	65.4	70.4	61.7
hi	78.5	70.5	64.5	63.6	70.8	71.1	63.2
vi	78.0	72.0	66.4	65.2	68.5	74.7	64.5
zh	77.8	70.1	67.0	64.9	68.1	71.8	67.7

Table 8: F1 score on MLQA for XLM-R$_{\text{+synth-trans}}$ model.

In addition, we report the F1 scores for XLM-R finetuned on SQuAD$_{\text{en}}$ and XLM-R$_{\text{+synth-trans}}$ on all the language pairs, on both MLQA and XQuAD in Tables 7, 8, 9 and 10.

C Qualitative Evaluation

We report in Tables 11, 12, and 13 different examples that we analysed in our manual qualitative analysis, discussed at the end of section 5.1 in the main paper.

q/c	en	es	de	ar	hi	vi	zh	ru	th	tr	el
en	87.4	82.0	80.7	75.6	79.0	78.1	73.0	79.0	73.2	74.8	79.9
es	80.6	84.2	76.2	71.3	72.7	72.8	67.5	76.4	69.6	70.6	75.8
de	79.8	76.4	83.5	71.0	72.9	72.4	68.1	75.6	68.3	71.4	75.5
ar	65.3	63.1	62.0	77.8	64.1	61.9	58.8	63.9	62.8	57.4	65.2
hi	72.9	64.8	65.7	63.7	75.8	64.0	61.8	66.9	63.5	59.6	67.2
vi	74.5	69.8	70.7	67.7	67.7	80.6	66.9	70.5	66.6	64.8	70.6
zh	70.7	65.8	64.1	64.1	65.6	67.4	81.8	67.4	64.8	60.8	66.9
ru	79.8	78.0	76.7	70.2	72.5	74.6	67.3	81.1	69.6	70.7	75.6
th	49.6	42.1	44.1	49.8	51.9	49.0	53.9	50.4	74.9	37.1	48.8
tr	72.6	64.9	68.5	65.5	68.1	64.1	62.4	70.3	64.3	76.7	71.3
el	70.8	69.1	69.3	63.3	65.6	65.0	63.0	70.8	64.2	62.3	81.3

Table 9: F1 score on XQuAD for XLM-R model finetuned on SQuAD_{en}.

q/c	en	es	de	ar	hi	vi	zh	ru	th	tr	el
en	87.0	82.9	81.6	77.8	81.5	81.0	77.5	80.9	77.7	74.8	81.4
es	83.5	84.1	79.2	74.9	78.0	78.7	76.5	78.9	76.3	72.7	78.8
de	82.8	80.3	83.1	75.1	77.3	78.3	75.1	78.4	75.7	72.9	78.3
ar	78.4	77.2	75.6	77.5	72.8	74.3	72.1	74.2	72.5	69.5	74.2
hi	80.6	77.5	76.4	71.6	77.5	75.2	73.9	76.3	73.6	70.2	75.9
vi	81.3	79.6	77.5	73.0	75.6	80.7	74.7	76.5	74.5	71.8	76.0
zh	80.2	78.3	76.9	72.1	74.3	76.1	85.1	77.0	74.1	70.1	75.7
ru	82.2	80.4	78.8	74.1	75.9	78.4	76.5	80.8	75.4	72.9	78.6
th	79.1	76.7	75.1	71.0	73.5	75.1	73.6	75.3	77.2	68.8	74.1
tr	80.2	77.7	76.4	71.9	74.6	74.2	74.7	77.8	73.3	75.2	75.6
el	82.0	80.1	78.5	73.8	77.3	77.3	75.4	78.6	75.3	72.2	80.1

Table 10: F1 score on XQuAD for $\text{XLM-R}_{\text{+synth-trans}}$ model.

Paragraph Kochi was the centre of Indian spice trade for many centuries, and was known to the Yavanas (Greeks and Romans) as well as Jews, Syrians, Arabs, and Chinese since ancient times. It rose to significance as a trading centre after the port Muziris around Kodungallur (Cranganore) was destroyed by **massive flooding of Periyar** in 1341. The earliest documented references to Kochi occur in books written by Chinese voyager **Ma Huan** during his visit to Kochi in the 15th century as part of Admiral Zheng He's treasure fleet. There are also references to Kochi in accounts written by Italian traveller Niccolò Da Conti, who visited Kochi in 1440.
Answer massive flooding of Periyar **$\text{QG}_{\text{synth+trans}}$ (target language = de) Welche Veranstaltung hat den Angriff auf Kochi im Jahr 1341 verursacht?** (tr: *Which event caused the attack on Kochi in 1341?*) **$\text{QG}_{\text{synth+trans}}$ (target language = ar)** ما هو الحدث الذي تسبب في تدمير كاتماندو؟ (tr: *What event caused the destruction of Kathmandu?*)
Answer Ma Huan **$\text{QG}_{\text{synth+trans}}$ (target language = de) Welcher chinesische traveller hat die frühesten Erinnerungen an Kochi geschrieben?** (tr: *Which Chinese traveler wrote the earliest memories of Kochi?*) **$\text{QG}_{\text{synth+trans}}$ (target language = ar)** من كتب اول اشارة الى كوتي في القرن ال 15؟ (tr: *Who wrote the first reference to Coty in the 15th century?*)

Table 11: Paragraph 1 and different questions analysed during out human evaluation.

Paragraph Portuguese navigator, **Pedro Álvares Cabral** founded the first European settlement in India at Kochi in 1500. From 1503 to 1663, Fort Kochi (Fort Emmanuel) was ruled by Portugal. This Portuguese period was a harrowing time for the Saint Thomas Christians and the Jews, as the Inquisition was active in Portuguese India. Kochi hosted the grave of Vasco da Gama, the first European explorer to set sail for India, who was buried at **St. Francis Church** until his remains were returned to Portugal in 1539. The Portuguese rule was followed by that of the Dutch who renamed Fort Immanuel as Fort Stormsburg. In meantime, the Royal Family of Kochi relocated the capital of Kochi Kingdom to Thrissur, leaving nominal authority over Islands of Kochi. In 1664, Fort Kochi Municipality was established by Dutch, making it the first municipality in Indian subcontinent, which got dissolved when Dutch authority got weaker in the 18th century. The remaining part of Kochi were governed by governors of Kochi Kingdom. By 1773, the Mysore ruler **Hyder Ali** extended his conquest in the Malabar region to Kochi forcing it to become a tributary of Mysore. The hereditary Prime Ministership of Kochi held by the Paliath Achans ended during this period.

Answer Pedro Álvares Cabral
$QG_{synth+trans}$ (target language = de) Wer hat die erste europäische Siedlung in Kochi gegründet? (tr: *Who founded the first European settlement in Kochi?*)
$QG_{synth+trans}$ (target language = ar) من الذي اسس اول تسوية الاوروبية في الهند في عام 1500؟ (tr: *Who founded the first European settlement in India in the year 1500?*)

Answer St. Francis Church
$QG_{synth+trans}$ (target language = de) Wo wurde Vasco da Gama begraben? (tr: *Where was Vasco da Gama buried?*)
$QG_{synth+trans}$ (target language = ar) اين تم دفن فازاو da Gama ؟ (tr: *Where was Vazaw da Gama buried?*)

Answer Hyder Ali
$QG_{synth+trans}$ (target language = de) Wer war der Herrscher von Mysore im Jahr 1773? (tr: *Who was the ruler of Mysore in 1773?*)
$QG_{synth+trans}$ (target language = ar) من كان الحاكم في ميسور في عام 1773؟ (tr: *Who was the ruler of Mysore in 1773?*)

Table 12: Paragraph 2 and different questions analysed during out human evaluation.

Paragraph Taiwan is an island country in East Asia. The main island, known historically as Formosa, makes up **99%** of the area controlled by the ROC, measuring **35,808 square kilometres** (13,826 sq mi) and lying some **180 kilometres** (112 mi) across the Taiwan Strait from the southeastern coast of mainland China. The East China Sea lies to its north, the Philippine Sea to its east, the **Luzon Strait** directly to its south and the **South China Sea** to its southwest. Smaller islands include a number in the Taiwan Strait including the Penghu archipelago, the Kinmen and Matsu Islands near the Chinese coast, and some of the South China Sea Islands.

Answer 99%
$QG_{synth+trans}$ (target language = de) Welcher Prozentsatz der Gebiete von Taiwan wird von der ROK kontrolliert? (tr: *What percentage of the areas of Taiwan is controlled by the ROK?*)
$QG_{synth+trans}$ (target language = ar) ما هي نسبة الاراضي التي تسيطر عليها ال (ديوان)؟ (tr: *What is the percentage of lands controlled by the (Divan)?*)

Answer 35,808 square kilometres
$QG_{synth+trans}$ (target language = de) Wie groß ist die RAF? (tr: *How big is the RAF?*)
$QG_{synth+trans}$ (target language = ar) كم ميل مربع هي الجزيرة؟ (tr: *How many square miles is the island?*)

Answer Luzon Strait
$QG_{synth+trans}$ (target language = de) Welcher Fluss ist direkt zum Süden von Taiwan? (tr: *Which river is directly to the south of Taiwan?*)
$QG_{synth+trans}$ (target language = ar) ما هو اسم السكك الحديدية التي تقع بالقرب منها؟ (tr: *What is the name of the nearby railway?*)

Answer South China Sea
$QG_{synth+trans}$ (target language = de) Welches Meer ist im Südwesten von Taiwan? (tr: *Which sea is in southwest of Taiwan?*)
$QG_{synth+trans}$ (target language = ar) ما هو المحيط الذي ينظر اليه على الجانب الغربي من تايوان؟ (tr: *What is the ocean seen on the western side of Taiwan?*)

Answer 180 kilometres
$QG_{synth+trans}$ (target language = de) Wie weit ist die RAF von Taiwan aus der südlichen Küste von China? (tr: *How far is the RAF from Taiwan from the southern coast of China?*)
$QG_{synth+trans}$ (target language = ar) كم من الوقت تبعد جزر تايوان عن ساحل الصين؟ (tr: *How long are the Taiwan Islands from the coast of China?*)

Table 13: Paragraph 3 and different questions analysed during out human evaluation.

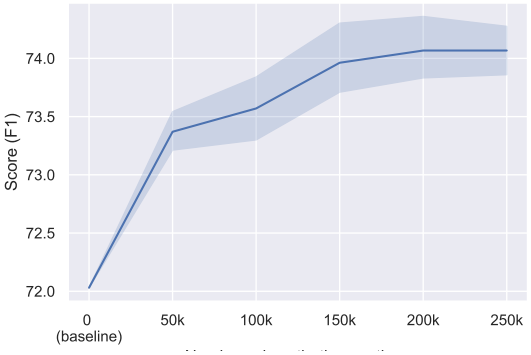

Figure 3: SQuAD-it

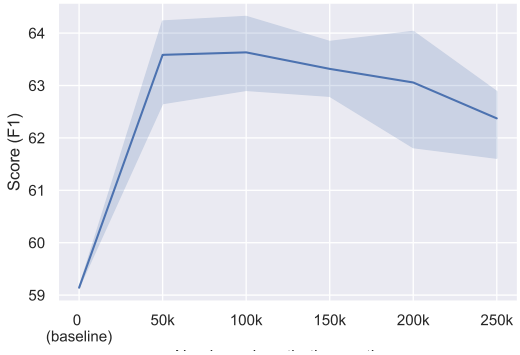

Figure 4: PIAF (fr)

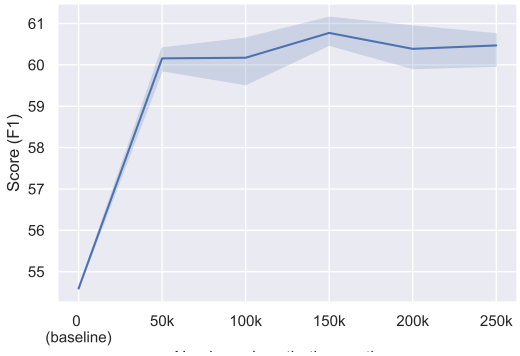

Figure 5: KorQuAD

D Learning Curves for Unseen Languages

We show on Figures 3,4,5 the results of three learning curves for respectively SQuAD-it, PIAF (fr) and KorQuad where the models are trained on different amount of synthetic questions in our *All Languages* setting.

The synthetic questions are sampled among all the five languages in MLQA . The standard deviation over four different seeds for the sampling are displayed through the confidence interval (light blue) around the averaged main curves. We observe that for SQuAD-it and KorQuAD the performances increase significantly at the beginning, then remain mostly stable, while for PIAF (fr) the best performances are obtained with 100k of additional synthetic data, a slight improvement from 50k additional questions before starting to decrease.