Towards Zero-Shot Multilingual Synthetic Question and Answer Generation for Cross-Lingual Reading Comprehension

Siamak Shakeri [†]
Google Research
siamaks@google.com

Noah Constant
Google Research
nconstant@google.com

Mihir Sanjay Kale
Google Research
mihirkale@google.com

Linting Xue
Google Research
lintingx@google.com

Abstract

We propose a simple method to generate multilingual question and answer pairs on a large scale through the use of a single generative model. These synthetic samples can be used to improve the zero-shot performance of multilingual QA models on target languages. Our proposed multi-task training of the generative model only requires labeled training samples in English, thus removing the need for such samples in the target languages, making it applicable to far more languages than those with labeled data. Human evaluations indicate the majority of such samples are grammatically correct and sensible. Experimental results show our proposed approach can achieve large gains on the XQuAD dataset, reducing the gap between zero-shot and supervised performance of smaller QA models across various languages.

1 Introduction

Generating question and answers from raw text has always been a challenging problem in natural language generation. Recently, there have been numerous efforts around question generation (Du et al., 2017; Song et al., 2018; Klein and Nabi, 2019; Wang et al., 2020; Ma et al., 2020; Chen et al., 2020; Tuan et al., 2019).

Using such synthetic samples to improve the performance of question answering models has been explored by Puri et al. (2020), Alberti et al. (2019), and Shakeri et al. (2020), who show that reading comprehension (RC) models can be improved by generating large-scale synthetic training data. These promising results combined with the recent surge in the development of powerful generative models such as GPT-3 (Brown et al., 2020), BART (Lewis et al., 2020a), and T5 (Raffel et al., 2020) suggest that the need for large manually labeled datasets can be reduced.

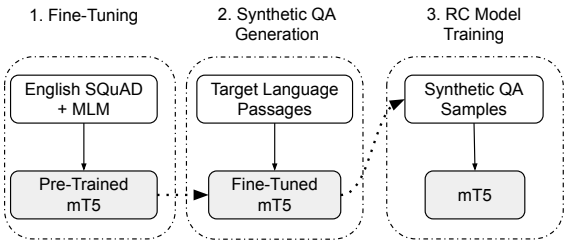

Figure 1: End-to-End pipeline: 1) Fine-tuning the generative model using SQuAD English samples and multilingual MLM. 2) Generating synthetic samples from Wikipedia passages of the target language using the fine-tuned generative model. 3) Training the downstream reading comprehension model using synthetic samples.

Although synthetic question-answer (QA) generation is well explored in English, the efficacy of such methods in the other languages remains an open question. Considering the lack of manually labeled QA datasets in many languages other than English, QA generation techniques can be applied to improve RC models in those languages. The emergence of multilingual generative models such as mBART (Liu et al., 2020a) and mT5 (Xue et al., 2021) facilitates such endeavors.

In this work, we propose generating multilingual question answer pairs to improve the performance of RC models in languages other than English. Besides unlabeled articles and questions, our proposed method only requires labeled training samples in English, thus completely removing the need to acquire new labeled datasets. Our approach can easily be extended to any language, as long as the multilingual generative model supports the language, and unlabeled questions and articles, such as Wikipedia, books, etc., exist in that language.

To enable zero-shot QA generation, the generative model should be able to produce non-English QA samples from non-English inputs when only

[†] Corresponding author.

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 35–45,
Aberdeen, Scotland, UK, 20-24 September 2021.

trained on English samples. Inspired by the work of Artetxe et al. (2020); Gururangan et al. (2020); Liu et al. (2020b), we propose a multi-task learning setting, where during the fine-tuning stage, we train on two tasks in parallel: the target question-answer generation task, and the multilingual masked language modeling (MLM) task that was used in pre-training the generative model. Our experimental results show that including the MLM task is crucial in enabling the zero-shot capability of the fine-tuned generative model.

We propose fine-tuning a pre-trained multilingual T5 model on the SQuAD 1.1 (Rajpurkar et al., 2016) training set. The fine-tuned model is then used to generate a large set of synthetic question-answer pairs from Wikipedia passages in the target language. Fig. 1 illustrates the end-to-end pipeline. We show that such synthetic samples can significantly boost RC models trained only on the English samples, with improvements up to 9 absolute points on F1. To summarize, our contributions are:

- Improving the zero-shot performance of multilingual RC models on multilingual QA tasks through generation of synthetic multilingual QA pairs.
- Proposing a multi-task fine-tuning of the multilingual generative model which is crucial for enabling zero-shot multilingual generation.
- Our approach is entirely zero-shot. No manually-labeled sample is used in fine-tuning the generative model on target languages, making our method applicable to both high and low resource languages.
- Demonstrating grammatical correctness and sensibility of generated questions through human evaluations.

The rest of the paper is organized as follows. In section 2, we discuss the process designed to train the generative model and produce synthetic samples. Section 3 discusses related work in the area of multilingual question-answer generation. In section 4, we present experiments to measure the quality of generated samples. Section 5 focuses on the application of synthetic question-answer samples to downstream reading comprehension models. Finally, we conclude in section 6.

2 End-to-End Question-Answer Generation and Filtering

2.1 Modeling

We use pre-trained "multilingual T5" (mT5) (Xue et al., 2021) as our generative model. The mT5 model is based on T5 (Raffel et al., 2020), which is an encoder-decoder sequence-to-sequence model.

2.2 QA Generation Task

We follow the probability distribution factorization suggested by Shakeri et al. (2020), where:

$$p(Q, A|P) = p(Q|P) \times p(A|Q, P)$$

Sampling from the above factorization is performed as follows:

$$q \sim p(Q|P), a \sim p(A|Q, P)$$

where Q, P, A refer to question, passage, and answer, respectively. During fine-tuning, passage tokens are fed as inputs, and the targets are a concatenation of the question and answer tokens. During sampling, candidate passages are passed as inputs to the fine-tuned generative model, and question-answer pairs are sampled from the decoder.

Fig. 2 depicts the fine-tuning and sampling processes. We prepend *"question"* to the question tokens and *"answer"* to the answer tokens, to help the model distinguish one from the other.

2.3 Masked Language Modeling Task

The mT5 model is pre-trained on the large multilingual "mC4" dataset (Xue et al., 2021) built from Common Crawl data, and trained using a Masked Language Modeling (MLM) task. This task involves replacing contiguous spans of input tokens with unique sentinel tokens (one per span). The decoder is then trained to reconstruct all the masked spans in the input, using a standard cross-entropy loss with teacher forcing. We use a variant of this MLM task, where we remove all "sentinel" tokens (corresponding to non-masked spans in the input text) from the target sequence, as we find this improves the quality of generated QAs.

2.4 Multi-Task Fine-Tuning

To perform zero-shot generation, the model needs not only to learn the QA Generation task, but also to retain its multilingual generation capabilities achieved during pre-training. To avoid *catastrophic*

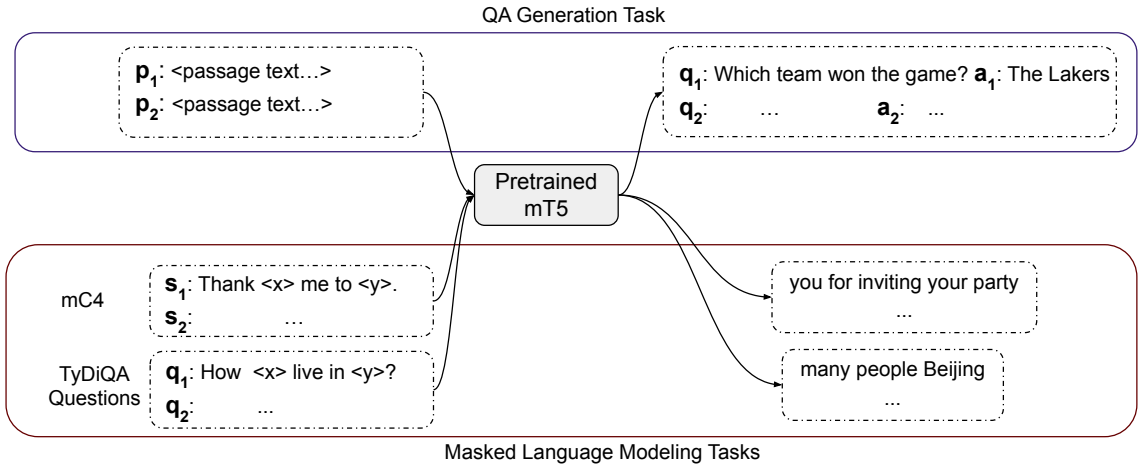

Figure 2: Multi-task fine-tuning of the multilingual pre-trained mT5 model. 1) QA generation task, which uses SQuAD English samples. 2) MLM task on a subset of mC4. 3) MLM on only the questions from the TyDiQA Gold Passage Task. The MLM variant used does not include sentinel tokens in the decoder output.

forgetting (French, 1999), which could lead to degraded generation capability, we propose a multi-task setting, where a predetermined percentage of fine-tuning examples come from the QA Generation task, while the remaining examples (trained in parallel) are from a mixture of two MLM tasks: 1) MLM on a subset of mC4 which is a continuation of mT5 pre-training, 2) MLM on only the questions from TyDiQA Gold Passage dev and training sets. The MLM task on mC4 helps the fine-tuned model retain its multilingual generation capabilities, while the MLM task on TyDiQA questions further improves the question generation capabilities of the generative model. Note that the only supervised QA training data is SQuAD 1.1. The MLM task on TyDiQA questions is not conditioned on the associated passages of the questions. Experimental results in section 4 demonstrate the efficacy of our proposed approach. Fig. 2 illustrates the multi-task fine-tuning process.

Fig. 3 demonstrates examples of generated samples in five languages using an mT5-XL (3.7B parameter) model fine-tuned in the multi-task setting (§2.4). It can be observed that: 1) the generated questions are in the same language as the passage most of the time, 2) the answers are relevant to the generated questions, 3) the model is capable of generating long and non-trivial QA pairs.

Fig. 4 illustrates generated QA samples in Spanish and Arabic, when only the QA Generation task (§2.2) is included in the fine-tuning. We observe: 1) questions are primarily in English, not the target language, 2) outputs contain certain tokens and entities mentioned in the language of the passage, and 3) ignoring language issues, the outputs exhibit semantically well-formed QA correspondence.

2.5 Decoding and Filtering

Since the quality of the generated question answer pairs is vital in improving the performance of downstream models, the generated samples require a strong filtering technique. Using the F1 score of a trained RC model to perform filtering, a.k.a. *round-trip* filtering, has been previously explored by Puri et al. (2020) and Alberti et al. (2019). For a generated QA sample *(q, a, p)*, where q, a, and p indicate question, answer, and passage, the following steps are performed: 1) a trained RC model is applied to *(q, p)*, predicting a', and 2) the F1 score of a and a' is calculated, and if above a certain threshold, *(q, a, p)* is kept, otherwise dropped.

3 Related Work

Recent work has explored question-answer generation (Alberti et al., 2019; Puri et al., 2020; Lee et al., 2020; Shakeri et al., 2020), but limited in scope to English. We leverage the modeling and filtering approaches proposed by Shakeri et al. (2020) due to their simplicity and effectiveness.

Kumar et al. (2019) explores cross lingual question generation. In contrast to our work, this only generates questions, without the corresponding answers. Additionally, this approach requires a complicated pre-training process on the target languages, as well as gold samples to fine-tune the generative models, so it is not easily extensible to

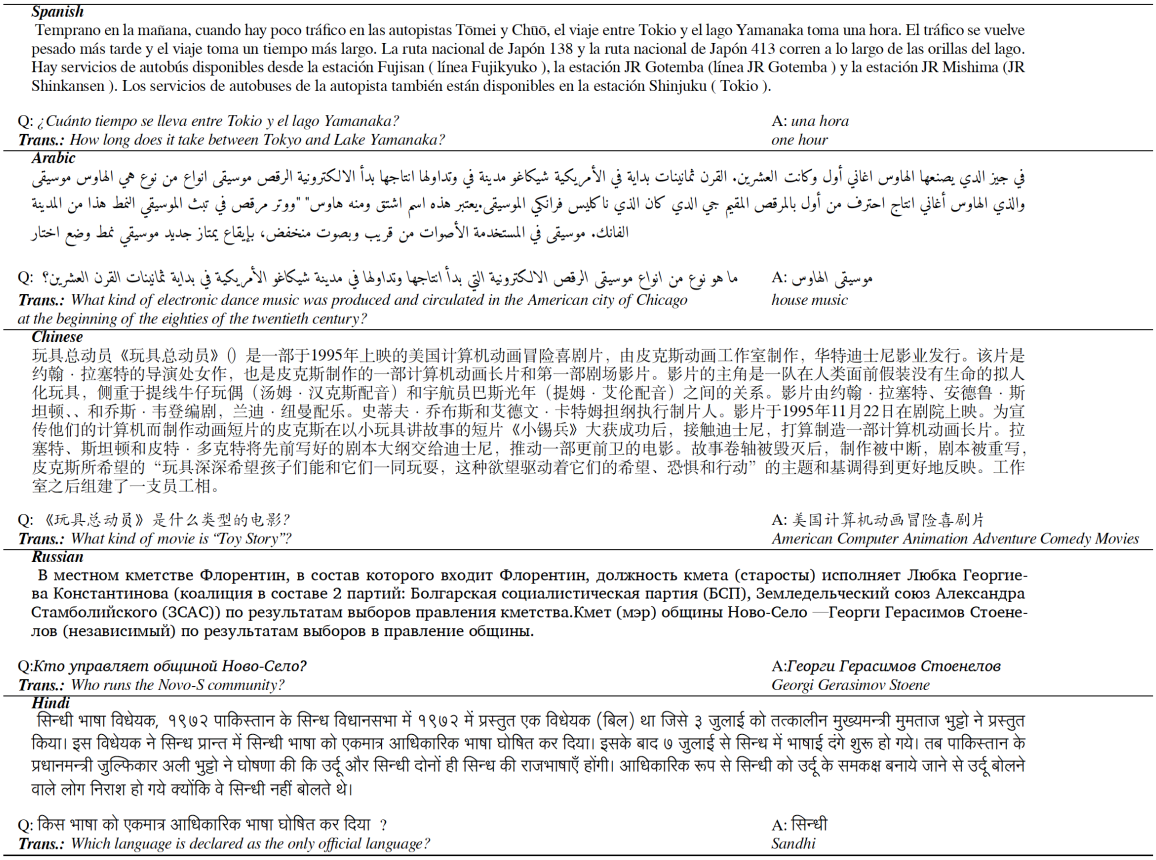

Figure 3: Samples of generated QAs in Spanish, Russian, Chinese, Arabic, and German. The generative model is mT5-XL fine-tuned on the mixture setting of section 2.4. ***Trans.*** refers to translations of the QA sample using Google Translate service.

other languages. This is in contrast to our approach, which does not require any gold QA samples in any language other than English. Another distinguishing factor is that we demonstrate improved performance on downstream QA tasks, while Kumar et al. (2019) only measure the quality of the generated samples on automatic metrics such as BLEU, and human evaluations.

Similarly, Chi et al. (2020) explore cross-lingual question-only generation using SQuAD English samples. They propose cross-lingual pre-training on the source and target languages. Similar to Kumar et al. (2019), their focus is only on the quality of the generated questions, whereas we validate our approach directly through improvements on downstream QA tasks. Moreover, while Chi et al. (2020) depends on a complex pre-training recipe and parallel sentences in both source and target languages, our approach not only does not require such parallel corpus, but also the MLM task included in our fine-tuning setting is widely used and studied. This leads to our approach being more easily adaptable to other languages and pre-trained generative models.

Most closely related to our work is the multilingual synthetic question generation approach of Riabi et al. (2020). However, there are two important differences between the two approaches. Firstly, our work includes both question and answer generation using a single model, while theirs only focuses on question generation. We believe generating the question and answer jointly is a richer problem that better harnesses the capabilities of pre-trained language models. Their question generation is conditioned on the selected answers, which further limits the generation. Secondly, their proposed method depends on translating SQuAD to target languages to fine-tune the generative model, hence limiting the application of their approach to languages where such translation data exists. Furthermore, even when translated data is available, the quality of samples generated by a model trained on such data is highly affected by the quality of the translations. This could lead to low quality QA samples in low resource languages. This is in contrast to our zero-shot approach, which does not require any training data in the target language.

Spanish

Richard Anton Patrick Connel O'Ferrall (Paramaribo, 21 de julio 1855 - 28 de octubre de 1936) fue un maestro, escritor y miembro del Parlamento de Surinam. Durante la última década del siglo XIX y principios del siglo XX O'Ferrall fue una de las figuras centrales en la vida cultural de Paramaribo. Se graduó de maestro en holandés, francés e inglés y en 1881 escribió acerca de un método para la enseñanza de la lectura inicial. Fue director de una escuela especial para la Educación Primaria Avanzada (ULO) y también director de la puesta en marcha en 1888 de la escuela pública de formación de artesanos que publicó desde 1893 un catálogo de obras y técnicas constructivas. Con los estudiantes de la clase más alta que formó el grupo teatral Ons Genoegen. Él escribió artículos para la revista De Ambachtsman (El Artesano), pero en ocasiones también para los periódicos locales.

Q: *Who was Richard Anton Patrick Connel O'Ferrall?* A: *maestro, escritor y miembro del Parlamento de Surinam*
Q: *When did Richard O'Ferrall die?* A: *28 de octubre de 1936*

Arabic

آبلة (بالإسبانية: Ávila) هي مدينة تقع في وسط إسبانيا، هي عاصمة مقاطعة آبلة التابعة لمنطقة قشتالة وليون.الديموغرافيا بلغ عدد سكان مدينة آبلة 008.59 نسمة عام 2011 (وفقاً للمعهد الوطني للإحصاء الإسباني).توأمة لآبلة اتفاقيات توأمة مع كل من: فيلنوف سور لوت روي-ماليزون تيرامو رودسأعلام توماس دي توركيمادا خيسوس هرنانديز أوييدا سانتشث ألبورنوث فيليسيانو ريفيلا خيسوس هرنانديز جيل غونزاليس جيل غونزاليس دافيلا توماس لويس دي فيكتوريا كارلوس ساستري خوليو خيمينيزمراجعانظر أيضًا قائمة مواقع التراث العالمي في إسبانيا موقع التراث العالمي قائمة ...

Q: *Where is* آبلة *located?* A: فوسط إسبانيا
Q: *How many people lived in* آبلة *in 2011?* A: *59.008*

Figure 4: Samples of generated QAs in Spanish and Arabic. The mT5-XL model is unable to generate valid questions in the target language, as in this case it was fine-tuned exclusively on the English QA generation task from section 2.2.

4 Experimental Setup and Results

4.1 Datasets

SQuAD (Rajpurkar et al., 2016) is an English QA dataset consisting of 100k samples. The passages are extracted from Wikipedia. We use the train and dev splits of SQuAD 1.1 in this work.

XQuAD (Artetxe et al., 2020) is a multilingual QA dataset consisting of 240 paragraphs and 1190 question-answers in Arabic, Chinese, German, Greek, Hindi, Russian, Spanish, Thai, Turkish and Vietnamese. These samples have been professionally translated from the SQuAD 1.1 dev set.

MLQA (Lewis et al., 2020b) is a benchmark dataset for evaluating cross-lingual question answering performance. This dataset contains over 5k QA instances (12k in English) following the SQuAD format in each of Arabic, Chinese, English, German, Hindi, Spanish and Vietnamese. We use the test split in our evaluations.

TyDiQA (Clark et al., 2020) is another multilingual QA dataset consisting of 200k QA pairs from 11 typologically diverse languages. There is less lexical overlap between questions and answers compared to XQuAD and MLQA. We use the Gold Passage task, which includes ∼50k samples in the train split and between 130 and 1,100 samples for each language in the development set.

XTREME (Hu et al., 2020) is a multilingual benchmark consisting of nine tasks spanning 40 typologically diverse languages. This dataset includes machine translated SQuAD 1.1 train and dev samples, which we employed in our experiments. We refer to such samples as `translate-train`.

4.2 Generative Model Fine-Tuning

We used the official mT5-XL model (Xue et al., 2021) with 3.7 billion parameters as our generative model. The official pre-trained checkpoint is fine-tuned using the mixture of tasks described in section 2.1. We chose the task mixing ratio to be 10:1, meaning for every 10 instances of the QA Generation task (§2.2), we mix one instance of the MLM task (§2.3). We experimented with mixing ratios of 100:1 and 1000:1 as well, both of which under-performed 10:1. The unsupervised MLM task covers text from two domains: 1) the subset of the mC4 corpus (Xue et al., 2021) covering Arabic, Bengali, English, Finnish, Indonesian, Korean, Russian, Swahili, and Telugu, and 2) questions from TyDiQA (Clark et al., 2020) train and dev sets, covering the same set of languages.

It is worth highlighting that we only fine-tune a single model to generate across all target languages. We do not apply language code prompts during fine-tuning or inference. We observe that by properly designing the fine-tuning mixture, the model is capable of generating samples that match the language of the input passage. Human evaluations in section 4.4 further verify this.

All of our models are fine-tuned for 5,000 steps with a batch size of 131,072 tokens, distributed over 64 TPU-v3 chips. We use the Adafactor optimizer (Shazeer and Stern, 2018) with constant learning rate of 1e-3. The final checkpoint is used to perform synthetic data generation.

4.3 Automatic Evaluation Results

To compute automatic metrics such as BLEU against QA samples of the development set, we modify the generation task to generate a question

Training Task	ar	de	en	es	hi	vi	zh
SQuAD en	1.7	3.0	23.4	3.6	3.2	4.4	1.2
Mixture 1	12.2	14.9	**25.0**	18.4	10.6	13.8	**10.0**
Mixture 2	13.1	**15.2**	24.9	18.4	**11.1**	**13.9**	9.7
Mixture 3	**14.5**	14.8	**25.0**	**18.6**	10.8	13.5	9.6

Table 1: Comparison of question generation quality (BLEU score) on the MLQA test set with mT5-XL: The Mixtures are as follows: *SQuAD en*: SQuAD en as the training data, *Mixture 1*: SQuAD en + MLM on mC4 subset, *Mixture 2*: SQuAD en + TyDiQA questions, *Mixture 3*: SQuAD en + MLM on mC4 subset + MLM on TyDiQA questions.

Model Size	ar	de	en	es	hi	vi	zh
Base (580M)	3.9	5.1	19.0	8.2	3.5	7.4	3.1
Large (1.2B)	10.3	5.7	23.9	5.9	4.3	6.2	3.9
XL (3.7B)	14.5	14.8	**25.0**	18.6	10.8	13.5	9.6
XXL (13B)	**15.8**	**16.2**	**24.9**	**19.3**	**12.2**	**15.6**	**10.2**

Table 2: Performance of question generation (mixture setting) on the MLQA test set for different mT5 model sizes.

given the passage and answer. Conditioning on the answer is needed, as without it, the generative model might generate samples that are of high quality but not related to the answers provided in the development set for a given passage. This would lead to difficulty in interpreting metrics such as BLEU.

Tab. 1 compares BLEU[1] performance of two fine-tuning settings on the MLQA test set. We report results using the mT5-XL model. As can be seen, including the MLM tasks has a large impact on performance, conveying large gains up to +15 absolute BLEU points. This is in line with our observations from section 2.5, where adding MLM fine-tuning task enabled the generative model to produce QA samples in the language of the target passage.

Interestingly, MLM on either mC4 or TyDiQA questions results in similar BLEU scores. Furthermore, using a mixture of the two does not yield additional gains. However, eyeballing the generated samples indicated that the model fine-tuned on the mixture of both MLM tasks and the supervised English task generates more well-structured and sensible questions and answers. Human evaluations in section 4.4 verify the high quality of generated samples from a model trained with this mixture.

To investigate the effect of the generative model size on the quality of data generation, we perform experiments using mT5 variants with different number of parameters: Base (580M), Large (1.2B) and XL (3.7B). We report results of the fine-tuned models with the mixture setting (§2.4) on the MLQA dataset in Tab. 2. Model performance improves dramatically with the size of the pre-trained model. Based on these results, for the remainder of the paper, we use the mT5-XL model fine-tuned using the mixture approach.

[1]All BLEU scores in this work are calculated using SacreBLEU v1.3.0 (Post, 2018), with "exp" smoothing and "intl" tokenization.

4.4 Human Evaluations

To perform manual quality evaluation of the generated questions, raters were presented with generated questions, and tasked with rating them according to the following criteria:

- *Is the question in the target language?* Raters could select *yes* or *no*.
- *grammatical correctness:* Raters could select a whole number from 1 (lowest) to 4 (highest).
- *sensibility:* Raters could select a whole number from 1 (lowest) to 4 (highest).

In total, 400 generated samples from 5 languages were randomly selected and rated by native speakers of each language. Each rater was assigned 40 samples. Two native speakers of each of the five languages were asked to perform the task. Tab. 3 shows the evaluation results.

The results show that the multilingual generative model is nearly perfect at generating samples that match the language of the input passage. Considering no language codes are used during fine-tuning, and only English supervised training data are used, the results show that our proposed mixture has enabled the model to perform zero-shot cross-lingual generation coherently.

Interestingly, Spanish samples achieve high scores in all of the categories. Considering the model is not trained on any Spanish samples, either in the MLM tasks or SQuAD 1.1, the model shows strong transfer learning capabilities. This implies that including the MLM task as proposed in our mixture setting not only prevents the generative model from catastrophic forgetting of its multilingual capability on the languages included in the MLM fine-tuning task, but also on those not included. The same argument partially applies to Hindi. While there are no Hindi samples in the fine-tuning mixture, Bengali (a related Indo-Aryan language) was seen in the MLM task.

	Target Language	Grammatical Correctness	Sensibility
Arabic	0.98	3.55	3.35
Chinese	1.00	3.60	3.60
Hindi	1.00	2.93	3.35
Russian	1.00	3.50	3.75
Spanish	1.00	3.10	3.05
Average	1.00	3.34	3.38

Table 3: Human evaluation metrics on the generated samples. Samples are randomly drawn, and rated by native speakers. "Target Language" scores are in the range 0–1, while the other columns range from 1–4.

5 Application of Synthetic Data to Multilingual Reading Comprehension

In this section, we describe experimental results that demonstrate the efficacy of using synthetic samples for improving multilingual reading comprehension (RC) models. This refers to the setting where given a passage and a question, the model is tasked with finding a span of the passage that answers the question.

5.1 Synthetic Data Generation

We randomly selected 10k paragraphs from Wikipedia, for each of Arabic (ar), German (de), Hindi (hi), Russian (ru) and Spanish (es). The selected paragraphs were restricted to have between 30 and 450 tokens, thereby removing passages that are too long or too short.

We fine-tune the mT5-XL model according to the mixture setting discussed in section 2.4 and the hyper-parameters from section 4.2, and then use this model to generate 20 questions per passage. We apply *top-k* sampling (Holtzman et al., 2020) with *k=10* and temperature of 0.5. The generated samples are processed to ensure: 1) each consists of a question followed by an answer, 2) the answer does exist in the passage. This was done to ensure answers are extractive. Non-extractive or no-answer QA are outside the scope of this work.

Finally, as discussed in section 2.5, round-trip filtering is applied to the generated QA samples. We use an mT5 XL model trained on SQuAD 1.1 (Rajpurkar et al., 2016) as the filtering model. The overall process results in approximately 10-20k synthetically generated samples in each target language. These generated samples are then used for training the RC models.

5.2 RC Model Fine-tuning

All of our reading comprehension models are initialized from the official mT5 (Xue et al., 2021) and later fine-tuned on the generated samples. We experimented with Base (580M), Large (1.2B), and XL (3.7B) parameter variants of mT5. We fine-tune using the TensorFlow framework. Each model was trained for 10,000 steps with a learning rate of 1e-3 and a batch size of 131,072 tokens. The models were trained on 16 TPU-v3 chips. In experiments where both the SQuAD 1.1 samples and synthetically generated samples are used to fine-tune the RC models, the model is trained on a mixture of the two, with a 1:1 mixing ratio. Adafactor optimizer (Shazeer and Stern, 2018) with constant learning rate of 1e-3 is used in all cases.

5.3 Results

Tabs. 4–6 demonstrate the F1 performance of the RC models trained on SQuAD 1.1 samples as well as synthetic data generated as described in 5.1 on mT5 Base, Large, and XL models. "*SQuAD en* " refers to the original SQuAD 1.1 (Rajpurkar et al., 2016) dataset in English. Our zero-shot baselines (denoted "ours") were slightly higher than those reported in Xue et al. (2021) (denoted "paper").

We observe that augmenting *SQuAD en* with synthetic samples leads to large gains with the Base model. An improvement of **+9** absolute points is observed for Russian. Furthermore, with the Base model, all average F1 scores are improved with the addition of synthetic data, regardless of which language the synthetic samples come from. The largest gain is seen when German samples are added **(+2.9)**.

As the size of the mT5 model increases, the gains from synthetic augmentation decrease, as shown in Tabs. 5 and 6. With the Large model, the maximum improvement in average F1 is **+1.2** absolute points. With the XL model, the average F1 scores are either the same as the zero-shot baseline or slightly lower. This is expected as when the model size increases, the gap between zero-shot and supervised also becomes smaller, hence less headroom exists when adding the synthetic samples. Fig. 5 demonstrates this scaling effect. Nonetheless, improvements of **+5.1**, **+2.2**, and **+3.4** are observed on Russian, Arabic, and Greek, respectively with the mT5 Large model. Similarly, smaller per-language gains can be seen with augmentation with the XL model, as shown in Tab. 6.

A surprising observation is that best per-language results are not necessarily achieved when augmenting with the synthetic samples from the

Dataset	en	ar	de	el	es	hi	ru	th	tr	vi	zh	avg
SQuAD en (paper)	84.6	63.8	73.8	59.6	74.8	60.3	57.8	57.6	67.9	70.7	66.1	67.0
SQuAD en (ours)	**85.5**	65.7	73.6	62.5	75.0	62.4	61.9	57.6	68.9	**71.9**	**71.1**	68.1
SQuAD en + ru	83.7	67.5	73.6	69.3	73.8	**66.2**	70.3	62.7	67.5	68.8	68.9	70.0
SQuAD en + hi	84.2	68.3	75.0	68.4	75.0	63.7	68.2	**64.5**	67.2	69.5	68.9	69.9
SQuAD en + de	84.6	**69.0**	71.8	**70.2**	**75.7**	**66.2**	**71.0**	63.5	**70.0**	70.9	**71.2**	**71.0**
SQuAD en + ar	84.5	64.0	74.4	69.4	74.4	65.1	65.1	62.5	67.9	70.0	70.2	70.2
SQuAD en + es	84.8	**69.1**	**76.1**	68.2	72.8	65.4	68.9	62.7	70.0	71.0	71.0	70.6
Supervised	83.1	72.4	76.9	76.8	79.0	71.4	76.1	67.9	72.5	75.9	76.9	75.3

Table 4: Performance of fine-tuned mT5 **Base** models on XQuAD. *Supervised* refers to training on SQuAD en + `translate-train` dataset of the target language.

Dataset	en	ar	de	el	es	hi	ru	th	tr	vi	zh	avg
SQuAD en (paper)	88.4	75.2	80.0	77.5	81.8	73.4	74.7	73.4	76.5	79.4	75.9	77.8
SQuAD en (ours)	88.6	75.0	80.4	76.5	81.6	73.9	74.1	73.8	**76.2**	**80.1**	76.4	77.4
SQuAD en + ru	88.2	76.6	81.2	79.1	82.6	76.1	77.6	72.1	75.1	78.4	77.4	**78.6**
SQuAD en + hi	**88.9**	76.7	81.5	79.4	**82.9**	73.4	78.7	**74.1**	75.0	79.1	78.0	**78.6**
SQuAD en + de	88.0	72.7	79.7	73.0	82.0	73.6	76.4	71.6	74.8	78.7	76.2	76.6
SQuAD en + ar	88.0	73.3	81.2	78.8	82.4	75.1	78.5	71.4	75.6	77.3	**78.2**	77.8
SQuAD en + es	88.2	**77.2**	**81.8**	**79.9**	81.3	**76.4**	**79.2**	72.3	75.8	79.5	77.7	**78.7**
Supervised	87.3	79.4	82.7	81.8	83.8	78.0	81.9	74.7	80.2	80.4	83.2	81.2

Table 5: Performance of fine-tuned mT5 **Large** models on XQuAD. *Supervised* refers to training on SQuAD en + `translate-train` dataset of the target language.

Dataset	en	ar	de	el	es	hi	ru	th	tr	vi	zh	avg
SQuAD en (paper)	88.8	77.4	80.4	80.4	82.7	76.1	76.2	74.2	77.7	80.5	80.5	79.5
SQuAD en (ours)	**89.7**	**79.2**	80.9	80.9	83.2	**78.7**	78.4	74.3	78.4	79.5	**80.7**	**80.2**
SQuAD en + ru	89.1	78.6	**82.1**	**81.7**	82.7	78.6	79.4	74.3	**78.7**	80.6	79.2	**80.2**
SQuAD en + hi	89.1	**79.1**	81.7	80.9	**83.4**	76.1	79.0	**74.6**	77.6	81.0	80.4	79.9
SQuAD en + de	88.8	78.2	81.2	**81.7**	82.8	78.1	**79.6**	74.0	77.7	**81.2**	79.7	80.1
SQuAD en + ar	89.0	75.0	81.3	81.5	82.8	78.4	79.3	73.5	78.4	80.2	80.4	79.6
SQuAD en + es	88.8	79.0	**82.2**	81.3	82.6	**78.7**	78.8	73.8	78.3	**81.1**	80.5	**80.2**
Supervised	88.5	80.9	83.4	83.6	84.9	79.6	82.7	78.5	82.4	82.4	83.2	82.7

Table 6: Performance of fine-tuned mT5 **XL** models on XQuAD. *Supervised* refers to training on SQuAD en + `translate-train` dataset of the target language.

same target language. Our hypothesis is that strong multilingual models such as mT5 have already developed rich per-language representations. Adding non-English synthetic data enables the model to generalize well to non-English RC tasks by not overfitting to English RC samples.

Comparing the *Supervised* metrics vs. *SQuAD en* + ⟨*lang*⟩ indicates that with Base and Large, using synthetic samples reduces the gap between the zero-shot and supervised performance of the trained RC models. This gap is reduced from **7.2** to **4.2** absolute points with the Base model. However, there still exists a sizeable gap, which could likely be further reduced through the use of higher quality synthetic samples.

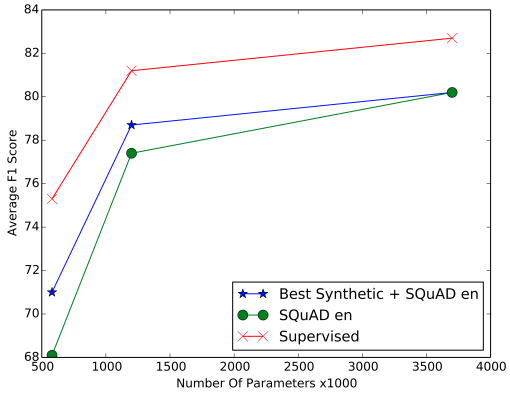

Figure 5: Scaling effect on augmentation using synthetic samples.

6 Conclusion

In this work, we presented a simple yet effective approach to generate large-scale synthetic multilingual question-answer pair data, which can be used to improve the zero-shot performance of multilingual reading comprehension (RC) models. Our experimental results showed large improvements in the performance of RC models trained on our synthetic multilingual datasets as compared to standard zero-shot baselines. Moreover, our zero-shot generation approach proved to be easily applied to any language, as long as the language is supported

by the pre-trained multilingual generative model.

While our results showed that using synthetic samples alongside English training data can significantly narrow the gap between zero-shot and supervised performance of RC models, the gap still remains. We are optimistic that future work can reduce this gap further through improved generation quality.

7 Ethical Considerations

Since the synthetic QA samples are generated by a generative model, it is possible that generated questions could include hallucinations and counterfactual information. We have employed the following safeguards: 1) The generative model is trained on the SQuAD dataset to learn question-answer generation. SQuAD is a well-studied and meticulously curated dataset. 2) The passages from which question-answer pairs are generated are selected from Wikipedia. 3) We apply round-trip filter on the generated question-answer pairs using the RC model. This approach ensures the questions are relevant to the passages. We believe these steps drastically reduce the chances of hallucinated and counterfactual samples. Nevertheless, there still exists the possibility that such bad samples could be generated. Future research efforts can explore such potential issues.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2020. Reinforcement learning based graph-to-sequence model for natural question generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7570–7577.

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.

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.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds. *CoRR*, abs/1911.02365.

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.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020b. 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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and

Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2020b. Exploring fine-tuning techniques for pre-trained cross-lingual models via continual learning. *CoRR*, abs/2004.14218.

Xiyao Ma, Qile Zhu, Yanlin Zhou, and Xiaolin Li. 2020. Improving question generation with sentence-level semantic matching and answer position inferring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8464–8471.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

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.

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

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 569–574, New Orleans, Louisiana. Association for Computational Linguistics.

Luu Anh Tuan, Darsh J. Shah, and Regina Barzilay. 2019. Capturing greater context for question generation. *CoRR*, abs/1910.10274.

Bingning Wang, Xiaochuan Wang, Ting Tao, Qi Zhang, and Jingfang Xu. 2020. Neural question generation with answer pivot. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9138–9145.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.