Zero-Shot and Translation Experiments on XQuAD, MLQA and TyDiQA

Julen Etxaniz
UPV/EHU
jetxaniz007@ikasle.ehu.eus

Oihane Cantero
UPV/EHU
ocantero003@ikasle.ehu.eus

Abstract

1 Introduction

Question answering (QA) is a popular area in NLP, with many datasets available to tackle the problem from various angles. Despite such popularity, QA datasets in languages other than English remain scarce, even for relatively high-resource languages. The main reason for this is that collecting such datasets at sufficient scale and quality is difficult and costly.

There are two reasons why this lack of data prevents internationalization of QA systems. First, we cannot measure progress on multilingual QA without relevant benchmark data. Second, we cannot easily train end-to-end QA models on the task, and most recent successes in QA have been in fully supervised settings.

There are a few datasets that try to address the first issue, by providing multilingual validation data. However, training data for multiple languages remains scarce and these datasets include few or no training data. That's why zero-shot and translation settings are popular options to test the performance of models on these datasets.

In this work we perform some zero-shot and translation experiments on 3 multilingual question answering datasets: XQuAD (Artetxe et al., 2019), MLQA (Lewis et al., 2019), and TyDi QA (Clark et al., 2020). The objective is to compare the results of zero shot, translate-train and translate-test settings on each datasets with different models. We use BERT and RoBERTa models of different sizes and their multilingual versions mBERT and XLM-R.

The QA datasets are described in the following section. Next sections explain the models we used and the experiments we performed. Then we present the baseline results and we discuss our results. Finally, we extract some conclusions.

2 Datasets

We perform experiments in 3 multilingual extractive question answering datasets: XQuAD (Artetxe et al., 2019), MLQA (Lewis et al., 2019), and TyDi QA (Clark et al., 2020) with 12, 7, and 9 languages respectively. Each dataset has unique features that justify experimenting with all of them to extarct conclusions.

2.1 XQuAD

XQuAD is a multilingual Question Answering dataset (Artetxe et al., 2019). It is composed of 240 paraghaps and 1190 question-answer pair from SQuAD v1.1[1]. SQuAD is based on a set of Wikipedia articles in English. Professional translations into 11 languages were added in XQuAD (Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi and Romanian). As the dataset is based on SQuAD v1.1, there are no unanswerable questions in the data.

We also added automatic translation from XTREME (Hu et al., 2020) for translate-train and translate-test experiments. The combined dataset can be found in HuggingFace [2].

2.2 MLQA

MLQA (Lewis et al., 2019) is another multilingual question answering evaluation benchmark. It has 5K extractive question-answering instances (12K in English) in seven languages (English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese). It is also based on Wikipedia articles, and the questions has been translated by professional translators, while he answers are directly taken form the different languages of the given Wikipedia article, to get parallel sentences. This allows testing settings where context and question languages are different.

[1] https://huggingface.co/datasets/squad
[2] https://huggingface.co/datasets/juletxara/xquad_xtreme

Automatic translation from XTREME (Hu et al., 2020) were also added for translate-train and translate-test experiments. The combined can be found in HuggingFace [3].

2.3 TyDiQA

TyDi QA is a question answering dataset covering 11 typologically diverse languages with 204K question-answer pairs (Clark et al., 2020). The languages of TyDi QA are diverse with regard to their typology – the set of linguistic features that each language expresses – such that we expect models performing well on this set to generalize across a large number of the languages in the world. It contains language phenomena that would not be found in English-only corpora. To provide a realistic information-seeking task and avoid priming effects, questions are written by people who want to know the answer, but don't know the answer yet, (unlike SQuAD and its descendents) and the data is collected directly in each language without the use of translation (unlike MLQA and XQuAD).

We also added automatic translation from XTREME (Hu et al., 2020) for translate-train and translate-test experiments. The combined can be found in HuggingFace [4].

3 Models

We use 7 different models in total for our experiments, 4 monolingual models and three multilingual models.

Monolingual models include the base and large versions of BERT and RoBERTa.

1. **BERT** (Devlin et al., 2018): It is a pretrained model on raw text in English. It has been pretrained with masked language modeling and next sentence prediction objectives to learn an inner representation of the English language. It has 12 layers, the size of the hidden layers is 768, it has 12 self-attention heads and 110M parameters.

2. **BERT-large**: It is the large version of BERT and has been pre-trained the same way. It has 24 layers, the size of the hidden layers is 1024, it has 12 self-attention heads and 340M parameters

3. **RoBERTa** (Liu et al., 2019): As the previous models, it has been pretrained on raw English data. It has been trained with dynamic masking, full-sentences without NSP loss, large mini-batches and a larger byte-level BPE. It has 125M parameters, 12 layer of hidden size 768 and 12 attention heads.

4. **RoBERTa-large**: It is the large version of RoBERTa. It has 355M parameters, 24 layers of hidden size 1024 and 16 attention heads.

Multilingual models include the multilingual versions of the previous models, mBERT and XLM-R.

1. **mBERT**: It has been trained the same way than BERT but using Wikipedia articles in 102 languages. It has 110M parameters, 12 layers, 768 hidden-states and 12 self-attention heads.

2. **XLM-R** (Conneau et al., 2019): It has been pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. It has 125M parameters with 12 layers, 768 hidden-states, 3072 feed-forward hidden-states, 8 self-attention heads.

3. **XLM-R-large**: It is the large version of XLM-R. It has 355M parameters with 24 layers, 1027 hidden-states, 4096 feed-forward hidden-states, and 16 self-attention heads,

4 Experiments

We perform 6 experiments in total with the models in the previous section. These experiments include zero-shot, translate-test, translate-train, fine-tuning and data-augmentation. The models we use are already fine-tuned and available on Huggingface. This way we save training time and we can do more experiments. All the code to replicate the experiments can be found on GitHub [5].

1. **Zero-shot**: We fine-tune the multilingual models on SQuAD, and evaluate them on the XQuAD, MLQA and TiDyQA test data for other languages. This is known as cross-lingual zero-shot transfer (XLT). For MLQA, we also evaluate models on generalised cross-lingual zero-shot transfer (G-XLT). In this setting, different languages are used for context and question.

[3] https://huggingface.co/datasets/mlqa

[4] https://huggingface.co/datasets/juletxara/tydiqa_xtreme

[5] https://github.com/juletx/XQuAD-MLQA

2. **Translate-test Monolingual**: We fine-tune the monolingual models on SQuAD, and evaluate them on translated XQuAD, MLQA and TiDyQA test data.

3. **Translate-test Multilingual**: We fine-tune the multilingual models on SQuAD, and evaluate them on translated XQuAD, MLQA and TiDyQA test data.

4. **Translate-train**: We fine-tune multilingual models on translated SQuAD, and evaluate them on XQuAD, MLQA and TiDyQA test data.

5. **Fine-tuning**: For XQuAD dataset, we fine-tune multilingual models on XQuAD and evaluate them on XQuAD test data. As we used already fine-tuned models, we know that the fine-tuning has been done with a part of the test data, but we don't know which one, so the results will be very high, as shown in section 6.1. For TiDyQA, the multilingual models are fine-tuned on TiDyQA, and evaluated on TiDyQA test set.

6. **Data augmentation**: In XQuAD, we fine tune multilingual models on augmented XQuAD, and evaluate them on XQuAD test data. In this case also, we have the same issue than in the previous one, because we don't know in which part of the test data the fine-tuning has been done. For TiDyQA, we use an model that has been trained on augmented XQuAD dataset and then fine-tuned on TyDiQA, and evaluate it on TyDiQA test set.

5 Baselines

Each dataset provides a few baseline results as reference. In this section, we comment those results so that we can compare our results with them in the next section.

5.1 XQuAD

In Table 1, we can see the baseline results, zero shot results are found by directly fine-tuning mBERT and XLM-R Large in the English SQuAD v1.1 training data and evaluating them on XQuAD test dataset. For translate-train, the models are fine-tuned on translated SQuAD, and for translate-test, BERT Large was fine-tuned on SQuAD training set and evaluated on translated XQuAD.

We can see that the best results are obtained by zero-shot with XLM-Large, and Translate train using BERT Large.

5.2 MLQA XLT

The baseline results for MLQA dataset are shown in Table 2. XLM performs best overall, transferring best in Spanish, German and Arabic, and competitively with translate-train with mBERT for Vietnamese and Chinese. However, XLM is weaker in English. There is a 39.8% drop in mean EM score (20.9% F1) over the English BERT-large baseline, showing significant room for improvement. All models generally struggle on Arabic and Hindi.

5.3 MLQA G-XLT

Table 3 shows results for XLM on G-XLT. For questions in a given language, the model performs best when the context language matches the question, except for Hindi and Arabic. For contexts in a given language, English questions tend to perform best, apart from Chinese and Vietnamese.

Table 4 shows results for mBERT on G-XLT. XLM outperforms mBERT for most language pairs, with a mean G-XLT performance of 53.4 F1 compared to 47.2 F1 (mean of off-diagonal elements). Multilingual BERT exhibits more of a preference for English than XLM for G-XLT. It also has a bigger performance drop going from XLT to G-XLT (10.5 mean drop in F1 compared to 8.2).

5.4 TyDiQA GoldP

Table 5 shows baseline results for TyDiQA from Clark et al. (2020). First, they fine tune mBERT jointly on all languages of the TyDiQA gold passage training data and evaluate on its dev set. F1 scores remain low, leaving headroom for future improvement.

Second, they fine tune on the 100k English-only SQuAD 1.1 training set and evaluate on the full TyDiQA gold passage dev set, following the XQuAD evaluation zero-shot setting. F1 scores are somewhat lower than the ones observed in the XQuAD zero-shot setting of Artetxe et al. (2019). Even the English performance is significantly lower, demonstrating that the style of question-answer pairs in SQuAD have very limited value in training a model for TyDiQA questions.

6 Results

In this section, we explain our results and compare them with the baseline results from the previous

Model	en	ar	de	el	es	hi	ru	th	tr	vi	zh	ro	avg
Zero-shot													
mBERT	83.5 / 72.2	61.5 / 45.1	70.6 / 54.0	62.6 / 44.9	75.5 / 56.9	59.2 / 46.0	71.3 / 53.3	42.7 / 33.5	55.4 / 40.1	69.5 / 49.6	58.0 / 48.3	72.7 / 59.9	65.2 / 50.3
XLM-R Large	86.5 / 75.7	68.6 / 49.0	**80.4** / 63.4	**79.8** / 61.7	82.0 / 63.9	**76.7** / 59.7	**80.1** / 64.3	**74.2 / 62.8**	**75.9 / 59.3**	**79.1** / 59.0	59.3 / 50.0	**83.6 / 69.7**	**77.2**/ 61.5
Translate-test													
BERT Large	**87.9 / 77.1**	**73.7 / 58.8**	79.8 / **66.7**	79.4 / **65.5**	**82.0 / 68.4**	74.9 / **60.1**	79.9 / **66.7**	64.6 / 50.0	67.4 / 49.6	76.3 / **61.5**	**73.7 / 59.1**		76.3 / **62.1**
Translate-train													
mBERT	83.5 / 72.2	68.0 / 51.1	75.6 / 60.7	70.0 / 53.0	80.2 / 63.1	69.6 / 55.4	75.0 / 59.7	36.9 / 33.5	68.9 / 54.8	75.6 / 56.2	66.2 / 56.6		70.0 / 56.0

Table 1: Baseline results (F1/EM) for each language in XQuAD dataset.

Model	en	es	de	ar	hi	vi	zh	avg
Zero-shot								
BERT-Large	**80.2 / 67.4**							80.2 / 67.4
mBERT	77.7 / 65.2	64.3 / 46.6	57.9 / 44.3	45.7 / 29.8	43.8 / 29.7	57.1 / 38.6	57.5 / 37.3	57.7 / 41.6
XLM	74.9 / 62.4	**68.0 / 49.8**	**62.2 / 47.6**	**54.8 / 36.3**	48.8 / 27.3	61.4 / 41.8	61.1 / **39.6**	61.1 / 43.5
Translate-test								
BERT-Large		65.4 / 44.0	57.9 / 41.8	33.6 / 20.4	23.8 / 18.9	58.2 / 33.2	44.2 / 20.3	58.4 / 39.9
Translate-train								
mBERT		53.9 / 37.4	62.0 / 47.5	51.8 / 33.2	**55.0 / 40.0**	**62.0 / 43.1**	**61.4** / 39.5	47.2 / 29.7
XLM		65.2 / 47.8	61.4 / 46.7	54.0 / 34.4	50.7 / 33.4	59.3 / 39.4	59.8 / 37.9	57.7 / 40.1

Table 2: Baseline results (F1/EM) for each language in MLQA dataset.

c/q	en	es	de	ar	hi	vi	zh
en	74.9	65.0	58.5	50.8	43.6	55.7	53.9
es	69.5	68.0	61.7	54.0	49.5	58.1	56.5
de	70.6	67.7	62.2	57.4	49.9	60.1	57.3
ar	60.0	57.8	54.9	54.8	42.4	50.5	43.5
hi	59.6	56.3	50.5	44.4	48.8	48.9	40.2
vi	60.2	59.6	53.2	48.7	40.5	61.4	48.5
zh	52.9	55.8	50.0	40.9	35.4	46.5	61.1

Table 3: Baseline MLQA F1 results on G-XLT with XLM. Columns show question language, rows show context language.

c/q	en	es	de	ar	hi	vi	zh
en	77.7	64.4	62.7	45.7	40.1	52.2	54.2
es	67.4	64.3	58.5	44.1	38.1	48.2	51.1
de	62.8	57.4	57.9	38.8	35.5	44.7	46.3
ar	51.2	45.3	46.4	45.6	32.1	37.3	40.0
hi	51.8	43.2	46.2	36.9	43.8	38.4	40.5
vi	61.4	52.1	51.4	34.4	35.1	57.1	47.1
zh	58.0	49.1	49.6	40.5	36.0	44.6	57.5

Table 4: Baseline MLQA F1 results on G-XLT with mBERT. Columns show question language, rows show context language.

Model	en	ar	bn	fi	id	ko	ru	sw	te	avg
Zero-shot										
mBERT	73.4	60.3	57.3	56.2	60.8	52.9	50.0	64.4	49.3	56.4
Fine-tuning										
mBERT	76.8	81.7	75.4	79.4	84.8	81.9	69.2	76.2	83.3	79.0

Table 5: Baseline TyDiQA GoldP F1 results for each language.

section. We interpret the results of each dataset and extract some conclusions.

6.1 XQuAD

The results are obtained with XQuAD dataset are in Table 6. They are quite similar to those form the baseline and these are some conclusions we got.

We can see that zero-shot is better than translate-test for larger models and worse for smaller models. So we can deduce that larger models have more adaptability to unseen languages than smaller ones. Monolingual models get better results that multilingual ones translate-test, and as we might expect, larger models give better results than smaller ones.

Overall, the results from worst to better have been: Translate-train, Translate-test multilingual, monolingual, zero-shot, data augmentation and fine-tuning. The comparison of the results with fine tuning and data augmentation is not very pertinent because the fine tuning has been done with a part of the testing data. As we don't know which part has been used to fine tune the models, we couldn't remove them from the testing data.

The best languages have been English, Spanish,

Romanian and Russian and the worst ones have been Chinese, Hindi, Thai and Turkish, with some very bad results, as for example, 25.2 F1 score and 16.8 EM for fine tuned mBERT in Thai. This could be because these four languages are are not in Latin script, and because Thai is not supported by mBERT.

6.2 MLQA XLT

In the MLQA dataset also, we get higher results with the biggest models, as we can see in Table 7. The language that obtains the best score is English. It is not unexpected because the dataset has much more data in English than in the other languages. Comparing to the results we got with XQuAD dataset, the results are generally a little lower, but we get the best results with the same models: XLM-R Large for zero-shot and multilingual translate test, RoBERTa Large and BERT Large for monolingual translate test, and balanced between Spanish and German in translate train.

6.3 MLQA G-XLT

We made zero-shot experiments between all the seven languages of the MLQA dataset and here are the results we got with the three models we used.

Using mBERT, in Table 8 we see that no matter the language of the corpus, making the question in English always gives the best score. This is probably because the dataset is trained with more data in English than the other languages. In the cases of Spanish, German, Vietnamese and Chinese, the second best score corresponds to the case where the question is asked in the language itself, but for Arabic and Hindi, we get better results when the question language is Spanish or German, instead of Arabic or Hindi.

Using XLM-R, we see in Table 9 that the best scores are always those that have the same question and context language.

As we can see in Table 10, here also in most of the cases the best scores are when the question and the context are in the same language, even if the English scores are very close, and in some cases EM is better in English.

6.4 TyDiQA GoldP

In Table 11, we can see the results we get with TyDiQA dataset. As in the previous datasets, for zero-shot, the best results are obtained with the larger model, XLM-R large, for every language. For monolingual translate test, a large model also get the best scores, but it is BERT Large instead of RoBERTa Large. For multilingual translate test, XLM-R Large get almost every best results except for Korean en Finnish EM, where mBERT performs better. And for translate train, the results are here also balanced between Spanish and German. For fine-tuning and data augmentation, we get better results than in the rest of the experiments with this dataset, but a little lower than with XQuAD. Taking the case of Arabic, that appears both in MLQA and TyDiQA, we get better results with TyDiQA in every experiment.

7 Conclusions

References

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

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*.

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

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

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

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Model	en	ar	de	el	es	hi	ru	th	tr	vi	zh	ro	avg
Zero-shot													
mBERT	85.0 / 73.5	57.8 / 42.2	72.6 / 55.9	62.2 / 45.2	76.4 / 58.1	55.3 / 40.6	71.3 / 54.7	35.1 / 26.3	51.1 / 34.9	68.1 / 47.9	58.2 / 47.3	72.4 / 59.5	63.8 / 48.8
XLM-R	84.4 / 73.8	67.9 / 52.1	75.3 / 59.8	74.3 / 57.0	77.0 / 59.2	69.0 / 52.5	75.1 / 58.6	68.0 / 56.4	68.0 / 51.8	73.6 / 54.5	65.0 / 55.0	80.0 / 66.3	73.1 / 58.1
XLM-R Large	**86.5 / 75.9**	**75.0 / 58.0**	**79.9 / 63.8**	**79.1 / 61.3**	**81.0 / 62.7**	**76.0 / 60.8**	**80.3 / 63.1**	**72.8 / 61.7**	**74.1 / 58.3**	**79.0 / 59.3**	**66.8 / 58.0**	**83.5 / 70.2**	**77.8 / 62.8**
Translate-test monolingual													
BERT		69.4 / 55.0	75.7 / 62.7	75.0 / 60.6	77.2 / 62.6	69.7 / 53.7	74.9 / 60.5	60.5 / 46.5	59.9 / 41.8	72.2 / 58.3	69.9 / 56.0		70.4 / 55.8
BERT Large		73.6 / 59.1	80.4 / 66.4	80.2 / 66.8	81.9 / 68.7	**75.3 / 61.7**	80.1 / 67.0	**67.5 / 53.9**	**66.3 / 47.3**	**76.4 / 62.1**	74.0 / 59.5		75.6 / 61.2
RoBERTa		71.6 / 57.0	77.0 / 62.4	76.8 / 63.9	80.0 / 64.6	72.0 / 55.6	77.2 / 62.4	62.2 / 46.6	63.4 / 44.1	72.4 / 56.6	72.4 / 57.9		72.5 / 57.1
RoBERTa Large		**74.8 / 61.1**	**80.4 / 67.1**	**80.8 / 68.0**	**83.1 / 69.4**	75.1 / 61.0	**81.2 / 68.0**	65.3 / 51.0	66.0 / 46.9	76.4 / 62.0	**74.0 / 59.9**		**75.7 / 61.4**
Translate-test multilingual													
mBERT		70.4 / 55.8	76.7 / 63.3	76.0 / 61.9	78.7 / 65.1	70.6 / 55.8	76.6 / 63.1	60.0 / 45.9	61.6 / 42.7	70.6 / 55.6	70.1 / 56.6		71.2 / 56.6
XLM-R		70.4 / 56.5	79.0 / 65.8	77.8 / 65.0	79.3 / 66.4	72.4 / 57.6	77.4 / 63.6	60.3 / 45.4	63.4 / 44.3	73.0 / 58.4	71.1 / 57.4		72.4 / 58.0
XLM-R Large		**72.9 / 59.1**	**80.1 / 66.6**	**79.6 / 66.2**	**81.5 / 67.1**	**74.2 / 60.1**	**79.7 / 65.7**	**61.7 / 46.0**	**66.2 / 48.2**	**75.1 / 61.5**	**73.6 / 58.8**		**74.5 / 59.9**
Translate-train													
XLM-R-es	**80.4** / 66.1	**67.0** / 47.9	74.2 / 56.4	**73.5**/ 52.4	**76.3** / 56.6	**66.9**/ 48.2	72.4 / 54.2	**68.7 / 58.5**	**66.2** / 46.5	73.2 / 52.0	63.4 / 50.3	**76.0** / 59.2	**71.5** / 54.0
XLM-R-de	79.8 / **67.1**	65.9 / **48.2**	**74.3 / 58.8**	72.3 / **54.4**	75.9 / **57.9**	66.4 / **50.6**	**73.1 / 56.4**	65.4 / 56.8	65.8 / **50.8**	72.7 / **53.2**	**64.7 / 55.0**	75.3 / **61.1**	71.0 / **55.9**
Fine-tuning XQuAD													
mBERT	97.3 / 95.3	90.0 / 84.3	94.2 / 90.0	92.2 / 87.0	96.2 / 92.4	88.2 / 77.5	94.4 / 90.1	25.2 / 16.8	89.9 / 84.4	93.4 / 87.6	87.5 / 84.4	95.5 / 91.3	87.0 / 81.8
XLM-R	98.5 / 97.5	92.5 / 88.2	95.1 / 91.8	96.0 / 91.8	97.8 / 93.6	92.6 / 88.6	95.2 / 90.8	94.0 / 92.4	92.0 / 87.3	95.5 / 91.3	94.0 / 92.9	97.7 / 94.8	95.1 / 91.8
XLM-R Large	**99.7 / 99.2**	**97.0 / 94.2**	**98.1 / 95.6**	**97.8 / 94.4**	**98.5 / 95.8**	**96.5 / 93.6**	**98.1 / 96.0**	**96.1 / 95.1**	**95.9 / 92.3**	**97.6 / 94.0**	**96.3 / 95.7**	**98.9 / 97.1**	**97.5 / 95.2**
Data-augmentation XQuAD													
mBERT	99.7 / 99.2	97.1 / 94.4	98.9 / 97.9	97.0 / 94.6	99.6 / 98.9	97.7 / 95.1	98.5 / 97.3	87.3 / 84.9	98.8 / 97.4	98.9 / 97.5	97.5 / 96.8	90.6 / 81.6	96.8 / 94.6

Table 6: XQuAD results (F1/EM) for each language.

Model	en	es	de	ar	hi	vi	zh	avg
Zero-shot								
mBERT	80.3 / 67.0	64.9 / 43.6	59.4 / 43.8	44.9 / 28.0	46.2 / 30.0	58.8 / 39.6	37.4 / 36.8	56.0 / 41.3
XLM-R	80.8 / 68.0	66.5 / 46.1	62.2 / 46.7	54.6 / 36.0	61.4 / 44.2	67.2 / 46.3	40.0 / 39.3	61.8 / 46.7
XLM-R Large	**84.0 / 71.2**	**72.1 / 50.2**	**68.5 / 52.4**	**62.0 / 42.1**	**69.8 / 51.3**	**73.1 / 51.8**	**45.7 / 45.1**	**67.9 / 52.0**
Translate-test monolingual								
BERT		65.0 / 43.2	54.4 / 35.7	51.0 / 27.7	52.8 / 32.0	53.6 / 32.1	47.8 / 26.6	54.1 / 32.9
BERT Large		67.2 / 45.2	56.7 / 37.2	52.7 / 28.9	55.2 / 33.8	**56.7** / 34.7	50.1 / **27.8**	56.4 / 34.6
RoBERTa		66.0 / 43.4	54.1 / 34.1	51.4 / 27.6	52.3 / 31.0	54.0 / 32.4	47.6 / 25.2	54.3 / 32.3
RoBERTa Large		**68.0 /45.9**	**57.4 / 38.0**	**53.7 / 29.4**	**55.7 / 33.9**	56.3 / **34.9**	**50.6** / 27.7	**56.9 / 35.0**
Translate-test multilingual								
mBERT		64.3 / 43.0	53.6 / 34.8	49.5 / 27.0	51.9 / 31.2	53.4 / 32.0	45.9 / 24.5	53.1 / 32.1
XLM-R		64.8 / 43.0	53.6 / 34.9	50.4 / 27.7	52.8 / 32.0	54.2 / 33.4	47.7 / 26.1	53.9 / 32.9
XLM-R Large		**68.6 / 46.5**	**56.6 / 37.4**	**53.1 / 29.2**	**55.6 / 34.5**	**56.6/ 34.5**	**50.0 / 27.6**	**56.7 / 35.0**
Translate-train								
XLM-R-es	77.2 / 61.5	**68.0** / 44.8	61.4 / 44.9	**54.1** / 34.1	**60.2** / 40.7	**66.2** / 45.0	36.2 / 35.4	**60.5** / 43.8
XLM-R-de	**77.3 / 63.6**	65.6 / **45.0**	**62.4 / 46.7**	53.6 / **35.6**	60.1 / **43.8**	65.0 / **45.2**	**38.1 / 37.4**	60.3 / **45.3**

Table 7: MLQA results (F1/EM) for each language.

c/q	en	es	de	ar	hi	vi	zh	avg
en	**80.3 / 67.0**	67.4 / 52.8	66.4 / 52.5	44.1 / 31.1	39.3 / 26.3	53.7 / 39.1	55.8 / 41.4	58.1 / 44.3
es	**66.9 /46.4**	64.9 / 43.6	60.6 / 40.2	43.1 / 26.0	36.2 / 20.1	48.5 / 31.4	49.9 / 30.6	52.9 / 34.0
de	**62.4 / 46.7**	56.4 / 41.0	59.4 / 43.8	36.8 / 23.6	34.0 / 21.5	43.6 / 29.6	46.5 / 30.7	48.4 / 33.8
ar	**51.1 / 33.7**	45.4 / 28.7	46.3 / 30.5	44.9 / 28.0	30.8 / 17.3	35.9 / 20.1	36.8 / 21.3	41.6 / 25.7
hi	**52.9 / 37.1**	43.7 / 29.1	47.6 / 33.8	34.5 / 21.4	46.2 / 30.0	38.0 / 25.0	39.2 / 25.2	43.2 / 28.8
vi	**64.5 / 44.8**	53.9 / 37.5	53.7 / 36.6	32.5 / 19.3	35.1 / 19.7	25.8 / 39.6	50.3 / 32.3	49.8 / 32.8
zh	**38.3 / 37.7**	29.0 / 28.3	30.0 / 28.9	21.0 / 20.6	16.6 / 16.2	25.1 / 24.4	37.4 / 36.8	28.2 / 27.6
avg	59.5 / 44.8	51.5 / 37.3	52.0 / 38.0	36.7 / 24.3	34.0 / 21.6	43.4 / 29.9	45.1 / 31.2	46.0 / 32.4

Table 8: MLQA results (F1/EM) for each language in zero-shot with mBERT. Columns show question language, rows show context language.

c/q	en	es	de	ar	hi	vi	zh	avg
en	**80.8 / 68.0**	57.8 / 43.9	60.8 / 47.1	33.5 / 21.3	45.0 / 32.0	39.8 / 27.5	37.9 / 25.3	50.8 / 37.9
es	66.0 / 45.1	**66.5 / 46.1**	50.5 / 32.6	25.2 / 12.3	31.8 / 17.1	29.1 / 14.9	28.2 / 14.3	42.5 / 26.1
de	60.0 / 44.3	44.0 / 29.7	**62.2 / 46.7**	22.2 / 12.1	29.4 / 17.6	28.7 / 16.2	29.1 / 17.2	39.4 / 26.3
ar	51.5 / 33.8	27.0 / 13.5	34.2 / 19.8	**54.6 / 36.0**	15.6 / 5.8	15.0 / 5.7	14.1 / 5.1	30.3 / 17.1
hi	60.6 / 43.4	37.4 / 23.0	42.8 / 27.8	19.5 / 8.0	**61.4 / 44.2**	24.3 / 11.9	26.1 / 13.6	38.9 / 24.6
vi	63.6 / 44.6	32.6 / 19.1	41.9 / 25.7	17.8 / 6.6	29.2 / 15.0	**67.2 / 46.3**	27.4 / 13.8	40.0 / 24.4
zh	34.9 / 34.3	11.3 / 10.7	14.0 / 13.3	3.9 / 3.7	10.8 / 10.4	8.1 / 7.7	**40.0 / 39.3**	17.6 / 17.1
avg	59.6 / 44.8	39.5 / 26.6	43.8 / 30.4	25.2 / 14.3	31.9 / 20.3	30.3 / 18.6	29.0 / 18.4	37.0 / 24.8

Table 9: MLQA results (F1/EM) for each language in zero-shot with XLM-R. Columns show question language, rows show context language.

c/q	en	es	de	ar	hi	vi	zh	avg
en	**84.0 / 71.2**	77.2 / 64.2	77.7 / 65.1	32.4 / 22.1	43.6 / 30.7	61.6 / 48.5	33.8 / 21.1	58.6 / 46.1
es	**72.1 / 50.3**	**72.1** / 50.2	70.0 / 48.6	33.0 / 17.8	42.1 / 26.1	54.8 / 35.6	36.5 / 20.5	54.4 / 35.6
de	67.7 / 51.7	65.3 / 49.6	**68.5 / 52.4**	31.2 / 19.5	36.2 / 21.9	50.7 / 34.3	32.3 / 18.9	50.3 / 35.5
ar	61.7 / **42.2**	56.7 / 38.4	59.7 / 41.8	**62.0 / 42.1**	43.9 / 27.4	48.6 / 30.7	38.7 / 21.6	53.0 / 34.9
hi	70.5 / **52.6**	63.2 / 45.9	65.1 / 49.9	45.5 / 29.1	**69.8 / 51.3**	54.4 / 37.6	44.6 / 28.5	59.0 / 42.1
vi	72.1 / 50.9	64.7 / 45.5	67.7 / 48.2	35.8 / 20.9	42.0 / 25.3	**73.1 / 51.8**	39.2 / 21.7	56.4 / 37.8
zh	44.2 / 43.6	36.7 / 36.1	41.1 / 40.2	25.9 / 25.4	30.0 / 29.5	35.0 / 34.6	**45.7 / 45.1**	36.9 / 36.4
avg	67.5 / 51.8	62.3 / 47.1	64.3 / 49.5	38.0 / 25.3	43.9 / 30.3	54.0 / 39.0	38.7 / 25.3	52.7 / 38.3

Table 10: MLQA results (F1/EM) for each language in zero-shot with XLM-R-Large. Columns show question language, rows show context language.

Model	en	ar	bn	fi	id	ko	ru	sw	te	avg
Zero-shot										
mBERT	77.8 / 69.8	60.6 / 45.1	59.5 / 47.8	60.6 / 50.1	63.0 / 50.6	47.9 / 39.5	65.3 / 47.3	61.0 / 49.5	48.9 / 41.4	60.5 / 49.0
XLM-R	75.2 / 65.9	66.7 / 52.8	67.5 / 51.3	72.6 / 62.3	75.8 / 61.6	62.6 / 53.6	67.6 / 48.9	68.8 / 59.1	74.6 / 58.6	70.2 / 57.1
XLM-R Large	**81.4 / 70.9**	**78.2 / 64.1**	**75.3 / 60.2**	**79.8 / 68.5**	**81.7 / 68.7**	**72.9 / 62.0**	**73.1 / 52.6**	**81.6 / 71.9**	**80.8 / 67.1**	**78.3 / 65.1**
Translate-test monolingual										
BERT		67.2 / 49.8	76.6 / **64.1**	71.2 / 57.5	74.2 / 60.3	70.1 / 58.9	71.9 / 55.5	76.0 / 63.8	63.5 / 51.1	71.3 / 57.6
BERT Large		**69.8 / 52.2**	**79.0 / 64.1**	**76.8 / 64.2**	**76.0 / 62.4**	**73.8 / 63.5**	**75.3 / 60.1**	**80.9 / 69.8**	**79.7 / 66.2**	**76.4 / 62.8**
RoBERTa		66.9 / 47.6	72.4 / 57.5	74.3 / 60.4	74.4 / 60.9	70.3 / 57.4	71.6 / 55.1	78.6 / 67.0	69.2 / 55.5	72.2 / 57.7
RoBERTa Large		66.9 / 48.1	78.1 / 63.0	75.4 / 60.6	73.0 / 56.8	73.2 / 61.6	74.3 / 58.1	80.2 / 69.6	78.0 / 63.2	74.9 / 60.1
Translate-test multilingual										
mBERT		66.7 / 48.9	70.4 / 56.4	73.2 / **61.6**	72.9 / 59.1	**71.9 / 60.3**	72.0 / 55.5	79.8 / 68.4	66.9 / 55.0	71.7 / 58.2
XLM-R		63.6 / 46.4	72.4 / 60.8	69.6 / 56.9	71.2 / 58.3	69.6 / 56.2	70.8 / 55.1	78.8 / 68.9	59.4 / 46.8	69.4 / 56.2
XLM-R Large		**68.6 / 52.5**	**73.3 / 58.0**	**75.2** / 61.3	**75.5 / 62.9**	68.5 / 56.7	**73.8 / 58.7**	**80.2 / 69.5**	**76.6 / 61.8**	**74.0 / 60.2**
Translate-train										
XLM-R-es	71.8 / 59.1	**68.2 / 52.1**	63.6 / 44.2	71.2 / 56.3	**73.1** / 57.4	53.8 / 40.6	67.2 / 43.6	67.2 / 55.9	71.7 / **54.4**	67.5 / 51.5
XLM-R-de	**73.6 / 63.2**	66.0 / 50.2	**64.7 / 49.6**	**72.4 / 60.1**	72.4 / **60.2**	**58.2 / 44.9**	**68.2 / 51.6**	**72.1 / 63.5**	**72.8** / 53.8	**68.9 / 55.2**
Fine-tuning										
mBERT	74.7 / 63.4	81.3 / 68.1	65.3 / 54.0	79.6 / 69.2	81.9 / 70.4	63.0 / 52.9	71.2 / 60.8	81.5 / 75.2	80.4 / 66.8	75.4 / 64.5
Data-augmentation										
mBERT	84.2 / 74.1	86.4 / 74.7	77.6 / 65.5	84.2 / 74.2	88.5 / 80.2	75.2 / 67.4	81.0 / 70.1	85.5 / 79.8	84.8 / 71.6	83.0 / 73.0

Table 11: TyDiQA GoldP results (F1/EM) for each language.