# Zero-Shot and Translation Experiments on XQuAD and MLQA

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

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

## Abstract

## 1 Introduction

In this project we performed some zero-shot and translation experiments on Multilingual Question Answering. The objective is to compare the results of zero shot, translation test and translation test on different datasets, with different models. The datasets we used are XQuAD and MLQA, and the models are monolingual or multilingual:

- Monolingual models:
  1. BERT (110M)
  2. BERT-large (340M)
  3. RoBERTa
  4. RoBERTa-large
- Multilingual models
  1. mBERT (110M)
  2. XLM-R
  3. XLM-R-large

Most of the models we used are already pretrained and available on Huggingface.

## 2 Related Work

## 3 Data

### 3.1 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, and professional translations into 11 languages were added in XQuAD (Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi and Romanian). We also used XTREME (Hu et al., 2020) for automatically translated translate-train and translate-test data. As the dataset is based on SQuAD v1.1, there are no unanswerable questions in the data. We chose this setting so that models can focus on cross-lingual transfer.

The dataset can be found in HuggingFace. [2]

### 3.2 MLQA

MLQA (Lewis et al., 2019) has 5K extractive question-answering instances (12K in English) in seven languages (English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese)

The dataset can be found in HuggingFace: [3]

## 4 Methods

### 4.1 Zero-shot

### 4.2 Translate Train

### 4.3 Translate Test

All the code can be found on GitHub [4]

## 5 Results

### 5.1 XQuAD

### 5.2 MLQA

## 6 Conclusions

## References

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

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.

[1] https://huggingface.co/datasets/squad

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

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

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

| Model F1 / EM | 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 1: XQuAD results (F1/EM) for each language.

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