Zero-Shot and Translation Experiments on XQuAD

Julen Etxaniz and Oihane Cantero
UPV/EHU

XquAD Dataset:

- 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD.
- SQuAD is based on a set of English Wikipedia articles, for extractive questions answering.
- Professional translations into 11 languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi and Romanian.
- XTREME for automatically translated translate-train and translate-test data.

Results:

- Similar to XQuAD baseline results in first table.
- Zero-shot is better than translate-test for larger models and worse for smaller models.
- Monolingual models get better results than multilingual in translate-test.
- Larger versions of models get better results.
- Results from worst to better: translate-train, translate-test multi, translate-test monolingual, zero-shot, fine-tuning, data augmentation
- Best languages: English, Spanish, Romanian, Russian
- Worst languages: Chinese, Hindi, Thai, Turkish

Experiments:

- **Zero-Shot:** fine-tune multilingual models on SQuAD, evaluate on XQuAD test data.
- **Translate-test monolingual:** fine-tune monolingual models on SQuAD, evaluate on translated XQuAD test data.
- **Translate-test multilingual:** fine-tune multilingual models on SQuAD, evaluate on translated XQuAD test data.
- **Translate-train:** fine-tune multilingual models on translated SQuAD, evaluate on XQuAD test data.
- **Fine-tuning on XQuAD:** fine-tune multilingual models on XQuAD, evaluate on XQuAD test data.
- **Data augmentation on XQuAD:** fine-tune multilingual models on augmented XQuAD, evaluate on XQuAD test data.

Monolingual Models:

- BERT (110M)
- BERT-large (340M)
- RoBERTa (125M)
- RoBERTa-large (355M)

Multilingual Models:

- MBERT (110M)
- XLM-R (125M)
- XLM-R-large (355M)

Model Baseline F1 / EM	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
Zero-shot 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-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
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**

Model Ours F1 / EM	en	ar	de	el	es	hi	ru	th	tr	vi	zh	ro	avg
Zero-shot													
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
Zero-shot 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
Zero-shot 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													
Translate-test 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
Translate-test 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
Translate-test 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
Translate-test 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													
Translate-test 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
Translate-test 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
Translate-test 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													
Translate-train es XLM-R	**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
Translate-train de XLM-R	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													
Fine-tuning 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
Fine-tuning 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
Fine-tuning 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													
Data-augmentation 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