XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization

Junjie Hu [* 1] **Sebastian Ruder** [* 2] **Aditya Siddhant** [3] **Graham Neubig** [1] **Orhan Firat** [3] **Melvin Johnson** [3]

Appendix

A. Languages

We show a detailed overview of languages in the cross-lingual benchmark including interesting typological differences in Table 1. Wikipedia information is taken from Wikipedia[1] and linguistic information from WALS Online[2]. XTREME includes members of the Afro-Asiatic, Austro-Asiatic, Austronesian, Dravidian, Indo-European, Japonic, Kartvelian, Kra-Dai, Niger-Congo, Sino-Tibetan, Turkic, and Uralic language families as well as of two isolates, Basque and Korean.

B. Hyper-parameters

Table 2 summarizes the hyper-parameters of baseline and state-of-the-art models. We refer to XLM-100 as XLM, and XLM-R-large as XLM-R in our paper to simplify the notation.

mBERT We use the cased version, which covers 104 languages, has 12 layers, 768 hidden units per layer, 12 attention heads, a 110k shared WordPiece vocabulary, and 110M parameters.[3] The model was trained using Wikipedia data in all 104 languages, oversampling low-resource languages with an exponential smoothing factor of 0.7. We generally fine-tune mBERT for two epochs, with a training batch size of 32 and a learning rate of 2e-5. For training BERT models on the QA tasks, we use the original BERT codebase. For all other tasks, we use the Transformers library (Wolf et al., 2019).

XLM and XLM-R We use the XLM and XLM-R Large versions that cover 100 languages, use a 200k shared BPE vocabulary, and that have been trained with masked language modelling.[4] We fine-tune both for two epochs with a learning rate of 3e-5 and an effective batch size of 16. In contrast to XLM, XLM-R does not use language embeddings. We use the Transformers library for training XLM and XLM-R models on all tasks.

C. Translations for QA datasets

We use an in-house translation tool to obtain translations for our datasets. For the question answering tasks (XQuAD and MLQA), the answer span is often not recoverable if the context is translated directly. We experimented with enclosing the answer span in the English context in quotes (Lee et al., 2018; Lewis et al., 2019) but found that quotes were often dropped in translations (at different rates depending on the language). We found that enclosing the answer span in HTML tags (e.g. and) worked more reliably. If this fails, as a back-off we fuzzy match the translated answer with the context similar to (Hsu et al., 2019). If the minimal edit distance between the closest match and the translated answer is larger than $\min(10, \texttt{answer\_len}/2)$, we drop the example. On the whole, using this combination, we recover more than 97% of all answer spans in training and test data.

D. Performance on translated test sets

We show results comparing the performance of mBERT and translate-train (mBERT) baselines on the XQuAD test sets with automatically translated test sets in Table 3. Performance on the automatically translated test sets underestimates the performance of mBERT by 2.9 F1 / 0.2 EM points but overestimates the performance of the translate-train baseline by 4.0 F1 / 6.7 EM points. The biggest part of this margin is explained by the difference in scores on the Thai test set. Overall, this indicates that automatically translated test sets are useful as a proxy for cross-lingual performance but may not be reliable for evaluating models that have been trained on translations as these have learnt to exploit the biases of *translationese*.

[*]Equal contribution [1]Carnegie Mellon University [2]DeepMind [3]Google Research. Correspondence to: Junjie Hu <junjieh@cs.cmu.edu>, Melvin Johnson <melvinp@google.com>.

[1]https://meta.wikimedia.org/wiki/List_of_Wikipedias

[2]https://wals.info/languoid

[3]https://github.com/google-research/bert/blob/master/multilingual.md

[4]https://github.com/facebookresearch/XLM

Table 1. Statistics about languages in the cross-lingual benchmark. Languages belong to 12 language families and two isolates, with Indo-European (IE) having the most members. Diacritics / special characters: Language adds diacritics (additional symbols to letters). Compounding: Language makes extensive use of word compounds. Bound words / clitics: Function words attach to other words. Inflection: Words are inflected to represent grammatical meaning (e.g. case marking). Derivation: A single token can represent entire phrases or sentences.

Language	ISO 639-1 code	# Wikipedia articles (in millions)	Script	Language family	Diacritics / special characters	Extensive compound-ing	Bound words / clitics	Inflec-tion	Deriva-tion	# datasets with language
Afrikaans	af	0.09	Latin	IE: Germanic		X				3
Arabic	ar	1.02	Arabic	Afro-Asiatic	X		X	X		7
Basque	eu	0.34	Latin	Basque	X		X	X	X	3
Bengali	bn	0.08	Brahmic	IE: Indo-Aryan	X	X	X	X	X	3
Bulgarian	bg	0.26	Cyrillic	IE: Slavic	X		X	X		4
Burmese	my	0.05	Brahmic	Sino-Tibetan	X	X				1
Dutch	nl	1.99	Latin	IE: Germanic		X				3
English	en	5.98	Latin	IE: Germanic						9
Estonian	et	0.20	Latin	Uralic	X	X		X	X	3
Finnish	fi	0.47	Latin	Uralic				X	X	3
French	fr	2.16	Latin	IE: Romance	X		X			6
Georgian	ka	0.13	Georgian	Kartvelian				X	X	2
German	de	2.37	Latin	IE: Germanic		X		X		8
Greek	el	0.17	Greek	IE: Greek	X	X		X		5
Hebrew	he	0.25	Hebrew	Afro-Asiatic				X		3
Hindi	hi	0.13	Devanagari	IE: Indo-Aryan	X	X	X	X	X	6
Hungarian	hu	0.46	Latin	Uralic	X	X		X	X	4
Indonesian	id	0.51	Latin	Austronesian			X	X	X	4
Italian	it	1.57	Latin	IE: Romance	X		X			3
Japanese	ja	1.18	Ideograms	Japonic			X	X		4
Javanese	jv	0.06	Brahmic	Austronesian	X		X			1
Kazakh	kk	0.23	Arabic	Turkic	X			X	X	1
Korean	ko	0.47	Hangul	Koreanic		X		X	X	5
Malay	ms	0.33	Latin	Austronesian			X	X		2
Malayalam	ml	0.07	Brahmic	Dravidian	X	X	X	X		2
Mandarin	zh	1.09	Chinese ideograms	Sino-Tibetan		X				8
Marathi	mr	0.06	Devanagari	IE: Indo-Aryan			X	X		3
Persian	fa	0.70	Perso-Arabic	IE: Iranian		X				2
Portuguese	pt	1.02	Latin	IE: Romance	X		X			3
Russian	ru	1.58	Cyrillic	IE: Slavic				X		7
Spanish	es	1.56	Latin	IE: Romance	X		X			7
Swahili	sw	0.05	Latin	Niger-Congo			X	X	X	3
Tagalog	tl	0.08	Brahmic	Austronesian	X		X	X		1
Tamil	ta	0.12	Brahmic	Dravidian	X	X	X	X	X	3
Telugu	te	0.07	Brahmic	Dravidian	X	X	X	X	X	4
Thai	th	0.13	Brahmic	Kra-Dai	X					4
Turkish	tr	0.34	Latin	Turkic	X	X		X	X	5
Urdu	ur	0.15	Perso-Arabic	IE: Indo-Aryan	X	X	X	X	X	4
Vietnamese	vi	1.24	Latin	Austro-Asiatic	X					6
Yoruba	yo	0.03	Arabic	Niger-Congo	X					1

Table 2. Hyper-parameters of baseline and state-of-the-art models. We do not use XLM-15 and XLM-R-Base in our experiments.

Model	Parameters	Langs	Vocab size	Layers
BERT-large	364,353,862	1	28,996	24
mBERT	178,566,653	104	119,547	12
MMTE	191,733,123	103	64,000	6
XLM-15	346,351,384	15	95,000	12
XLM-100	827,696,960	100	200,000	12
XLM-R-Base	470,295,954	100	250,002	12
XLM-R-Large	816,143,506	100	250,002	24

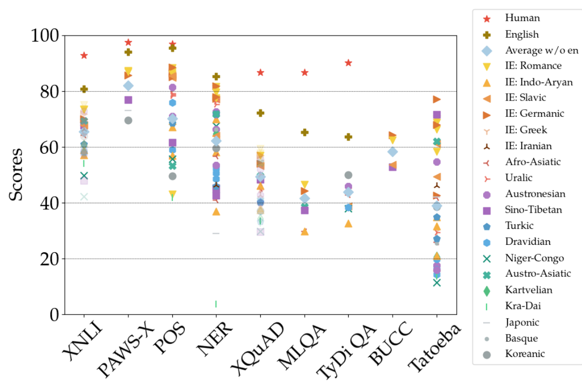

Figure 1. An overview of mBERT's performance on the XTREME tasks for the languages of each task. We highlight an estimate of human performance, performance on the English test set, the average of all languages excluding English, and the family of each language. Performance on pseudo test sets for XNLI and XQuAD is shown with slightly transparent markers.

E. mBERT performance across tasks and languages

We show the performance of mBERT across all tasks and languages of XTREME in Table 1.

F. Correlation with pretraining data size

We show the Pearson correlation coefficient ρ of mBERT, XLM, and XLM-R with the number of Wikipedia articles in Table 5. XLM and mBERT were pretrained on Wikipedia, while XLM-R was pretrained on data from the web.

G. Generalization to unseen tag combinations

We show the performance of mBERT on POS tag trigrams and 4-grams that were seen and not seen in the English training data in Table 6.

H. Generalization to unseen entities

We show the performance of mBERT on entities in the target language NER dev data that were seen and not seen in the English NER training data in Table 7. For simplicity, we count an entity as occurring in the English training data if a subset of at least two tokens matches with an entity in the English training data. As most matching entities in the target language data only consist of up to two tokens, are somewhat frequent, and consist only of Latin characters, we provide the performance on all entities fitting each criterion respectively for comparison. For all target languages in the table except Spanish, entities that appeared in the English training data are more likely to be tagged correctly than ones that did not. The differences are largest for two languages that are typologically distant to English, Indonesian (id) and Swahili (sw). For most languages, entities that appear in the English training data are similarly likely to be correctly classified as entities that are either frequent, appear in Latin characters, or are short. However, for Swahili and Basque (eu), mBERT does much better on entities that appeared in the English training data compared to the comparison entities. Another interesting case is Georgian (ka), which uses a unique script. The NER model is very good at recognizing entities that are written in Latin script but performs less well on entities in Georgian script.

I. Sentence representations across all layers

For sentence retrieval tasks, we analyze whether the multilingual sentence representations obtained from all layers are well-aligned in the embedding spaces. Without fine-tuning on any parallel sentences at all, we explore three ways of extracting the sentence representations from all the models: (1) the embeddings of the first token in the last layer, also known as [CLS] token; (2) the average word embeddings in each layer; (3) the concatenation of the average word embeddings in the bottom, middle, and top 4 layers, i.e., Layer 1 to 4 (bottom), Layer 5 to 8 (middle), Layer 9 to 12 (top). Figure 2 shows the F1 scores of the average word embeddings in each layer of mBERT in the BUCC task. We observe that the average word embeddings in the middle layers, e.g., Layer 6 to 8, perform better than that in the bottom or the top layers. In Table 10, we show the performance of these three types of sentence embeddings in the BUCC task. The embeddings of the CLS token perform relatively bad in cross-lingual retrieval tasks. We conjecture that the CLS embeddings highly abstract the semantic meaning of a sentence, while they lose the token-level information which is important for matching two translated sentences in two languages. With respect to the concatenation of average word embeddings from four continuous layers, We also observe that embeddings from the middle layers perform better than that from the bottom and top layers. Average word embed-

Table 3. Comparison of F1 and EM scores of mBERT and translate-train (mBERT) baselines on XQuAD test sets (gold), which were translated by professional translators and automatically translated test sets (auto).

	Test set	es	de	el	ru	tr	ar	vi	th	zh	hi	avg
mBERT	gold	75.6 / 56.9	70.6 / 54.0	62.6 / 44.9	71.3 / 53.3	55.4 / 40.1	61.5 / 45.1	69.5 / 49.6	42.7 / 33.5	58.0 / 48.3	59.2 / 46.0	62.6 / 47.2
	auto	76.1 / 58.7	64.3 / 49.9	57.9 / 42.5	68.3 / 51.8	55.6 / 42.9	62.1 / 48.6	68.6 / 54.3	41.1 / 32.6	48.5 / 47.7	54.1 / 40.9	59.7 / 47.0
translate-train	gold	80.2 / 63.1	75.6 / 60.7	70.0 / 53.0	75.0 / 59.7	68.9 / 54.8	68.0 / 51.1	75.6 / 56.2	36.9 / 33.5	66.2 / 56.6	69.6 / 55.4	68.7 / 54.6
	auto	80.7 / 66.0	71.1 / 58.9	69.3 / 54.5	75.7 / 61.5	71.2 / 59.1	74.3 / 60.7	76.8 / 64.0	79.5 / 74.8	59.3 / 58.0	69.1 / 55.2	72.7 / 61.3

Table 4. Comparison of accuracy scores of mBERT baseline on XNLI test sets (gold), which were translated by professional translators and automatically translated test sets (auto) in 14 languages. BLEU and chrF scores are computed to measure the translation quality between gold and automatically translated test sets.

Languages	zh	es	de	ar	ur	ru	bg	el	fr	hi	sw	th	tr	vi	avg
auto Acc.	69.1	74.7	72.8	66.5	64.5	71.6	70.2	67.7	74.3	65.1	50.2	54.5	60.0	72.7	66.7
gold Acc.	67.8	73.5	70.0	64.3	57.2	67.8	68.0	65.3	73.4	58.9	49.7	54.1	60.9	69.3	64.3
BLEU	40.92	43.46	30.94	32.35	20.13	22.62	45.04	60.29	47.91	29.55	31.25	10.65	15.39	56.93	34.82
chrF	35.96	67.92	60.28	59.64	48.21	50.38	67.52	75.34	69.58	53.85	59.84	54.89	51.46	69.37	58.87

Table 5. Pearson correlation coefficients (ρ) of zero-shot transfer performance and Wikipedia size across datasets and models.

	XNLI	PAWS-X	POS	NER	XQuAD	MLQA	TyDiQA-GoldP	BUCC	Tatoeba
mBERT	0.79	0.81	0.36	0.35	0.80	0.87	0.82	0.95	0.68
XLM	0.80	0.76	0.32	0.29	0.74	0.73	0.52	0.61	0.68
XLM-R	0.75	0.79	0.22	0.27	0.50	0.76	0.14	0.36	0.49

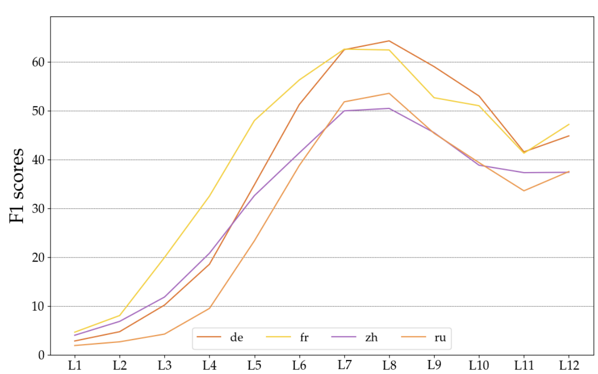

Figure 2. Comparison of mBERT's sentence representations by averaging word embeddings in each layer in the BUCC task.

dings in the middle individual layer perform comparative to the concatenated embeddings from the middle four layers.

I.1. Language Families and Scripts

We also report the performance of XLM-R in all the tasks across different language families and writing scripts in Figure 3.

J. Results for each task and language

We show the detailed results for all tasks and languages in Tables 8 (XNLI), 11 (PAWS-X), 16 (POS), 17 (NER), 13

Table 6. Accuracy of mBERT on the target language dev data on POS tag trigrams and 4-grams that appeared and did not appear in the English training data. We show the average performance across all non-English languages and the difference of said average compared to the English performance on the bottom.

	trigram, seen	trigram, unseen	4-gram, seen	4-gram, unseen
en	90.3	63.0	88.1	67.5
af	68.1	8.2	64.1	24.2
ar	22.0	0.7	14.9	4.6
bg	63.1	14.6	56.1	23.9
de	77.8	47.2	73.0	48.7
el	59.6	9.1	52.5	14.2
es	68.6	10.6	62.4	24.9
et	60.7	14.4	53.1	31.9
eu	32.8	7.1	28.7	8.1
he	52.7	35.7	44.0	27.4
hi	38.7	13.0	32.6	12.5
hu	55.5	28.8	46.9	23.7
id	60.8	16.6	54.7	21.6
it	75.5	12.8	71.8	23.5
ja	16.3	0.0	12.3	1.0
ko	22.0	2.9	14.7	3.8
mr	31.7	0.0	25.5	3.3
nl	75.5	24.1	71.0	37.8
pt	76.2	14.9	71.2	30.6
ru	69.1	4.8	63.8	20.6
ta	30.3	0.0	24.5	4.2
te	57.8	0.0	48.7	24.7
tr	41.2	6.2	33.9	10.1
ur	30.6	18.3	22.3	10.9
zh	29.0	0.0	21.7	3.9
avg	50.6	12.1	44.3	18.3
diff	39.7	50.9	43.7	49.2

Table 7. Comparison of accuracies for entities in the target language NER dev data that were seen in the English NER training data (a); were not seen in the English NER training data (b); only consist of up to two tokens (c); only consist of Latin characters (d); and occur at least twice in the dev data (e). We only show languages where the sets (a–e) contain at least 100 entities each. We show the difference between (a) and (b) and the minimum difference between (a) and (c-e).

	af	de	el	en	es	et	eu	fi	fr	he	hu	id	it	ka	ms	nl	pt	ru	sw	tr	vi
(a) Seen	94.7	88.3	91.4	91.9	76.3	88.3	83.6	85.3	90.5	78.2	90.7	89.4	88.4	92.3	88.6	93.5	88.6	83.9	96.3	85.2	91.4
(b) Not seen	82.1	80.2	74.8	84.6	80.4	78.9	69.4	79.8	80.1	56.5	78.3	58.0	81.5	70.2	75.0	82.9	82.3	68.5	66.6	73.7	73.4
(a) − (b)	12.6	8.1	16.5	7.2	-4.1	9.4	14.1	5.5	10.4	21.7	12.3	31.5	6.9	22.1	13.6	10.6	6.4	15.4	29.7	11.6	18.0
(c) Short	86.5	82.9	80.3	88.2	86.6	81.7	72.5	83.9	88.6	66.3	83.7	85.8	87.2	72.5	89.1	87.6	87.8	78.0	65.7	83.1	84.6
(d) Latin	83.6	81.2	87.5	86.2	80.0	79.5	70.3	80.3	81.1	77.2	79.9	61.8	82.6	89.6	76.3	84.2	83.0	83.8	70.0	75.0	74.9
(e) Freq	87.3	80.6	81.9	91.6	83.4	79.4	68.8	85.7	77.3	66.8	86.0	56.5	88.8	74.3	81.3	87.1	84.4	76.5	49.1	81.9	78.6
min((a) − (c–e))	7.4	5.4	3.9	0.3	3.7	6.6	11.0	0.4	1.9	1.0	4.7	3.6	0.4	2.7	0.5	5.9	0.8	0.1	26.4	2.2	6.8

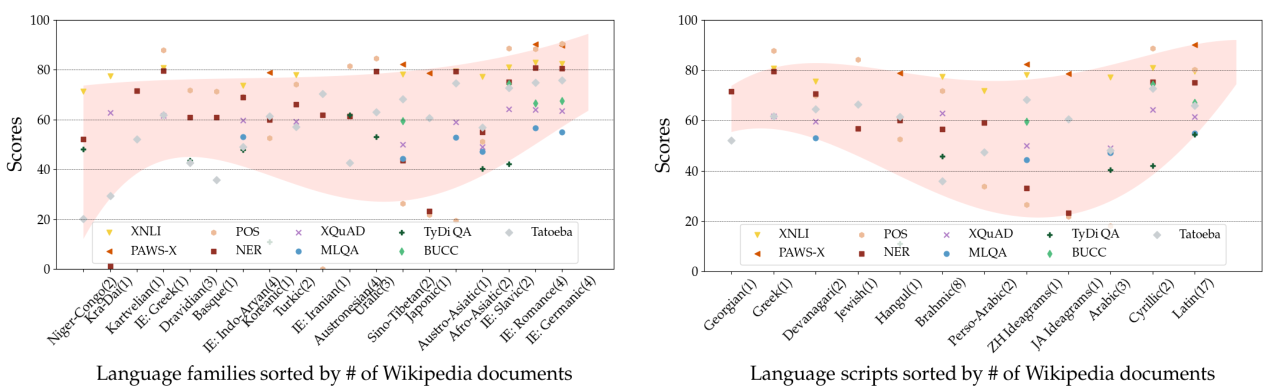

Figure 3. Performance of XLM-R across tasks grouped by language families (left) and scripts (right). The number of languages per group is in brackets and the groups are from low-resource to high-resource on the x-axis. We additionally plot the 3rd order polynomial fit for the minimum and maximum values for each group.

Table 8. XNLI accuracy scores for each language.

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	**avg**
mBERT	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
XLM	82.8	66.0	71.9	72.7	70.4	75.5	74.3	62.5	69.9	58.1	65.5	66.4	59.8	70.7	70.2	69.1
XLMR	**88.7**	**77.2**	**83.0**	**82.5**	**80.8**	**83.7**	**82.2**	**75.6**	**79.1**	**71.2**	**77.4**	**78.0**	**71.7**	**79.3**	**78.2**	**79.2**
MMTE	79.6	64.9	70.4	68.2	67.3	71.6	69.5	63.5	66.2	61.9	66.2	63.6	60.0	69.7	69.2	67.5
Translate-train (multi-task)	81.9	**73.8**	**77.6**	**77.6**	**75.9**	**79.1**	**77.8**	70.7	**75.4**	**70.5**	70.0	74.3	**67.4**	**77.0**	**77.6**	**75.1**
Translate-train	80.8	73.6	76.6	77.4	75.7	78.1	77.4	**71.9**	75.2	69.4	**70.9**	**75.3**	67.2	75.0	74.1	74.6
Translate-test	**85.9**	73.1	76.6	76.9	75.3	78.0	77.5	69.1	74.8	68.0	67.1	73.5	66.4	76.6	76.3	76.8

Table 9. Tatoeba results (Accuracy) for each language

Lang.	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja
BERT	42.7	25.8	49.3	17	77.2	29.8	68.7	29.3	25.5	46.1	39	66.3	41.9	34.8	38.7	54.6	58.4	42
XLM	43.2	18.2	40	13.5	66.2	25.6	58.4	24.8	17.1	32.2	32.2	54.5	32.1	26.5	30.1	45.9	56.5	40
XLMR	**58.2**	**47.5**	**71.6**	**43**	**88.8**	**61.8**	**75.7**	**52.2**	**35.8**	**70.5**	**71.6**	**73.7**	**66.4**	**72.2**	**65.4**	**77**	**68.3**	**60.6**

	jv	ka	kk	ko	ml	mr	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
BERT	17.6	20.5	27.1	38.5	19.8	20.9	68	69.9	61.2	11.5	14.3	16.2	13.7	16	34.8	**31.6**	62	**71.6**
XLM	**22.4**	22.9	17.9	25.5	20.1	13.9	59.6	63.9	44.8	12.6	20.2	12.4	**31.8**	14.8	26.2	18.1	47.1	42.2
XLMR	14.1	**52.1**	**48.5**	**61.4**	**65.4**	**56.8**	**80.8**	**82.2**	**74.1**	**20.3**	**26.4**	**35.9**	29.4	**36.7**	**65.7**	24.3	**74.7**	68.3

Table 10. Three types of sentence embeddings from mBERT in BUCC tasks: (1) CLS token embeddings in the last layer; (2) Average word embeddings in the middle layers, i.e., Layer 6, 7, 8; (3) the concatenation of average word embeddings in the continuous four layers, i.e., Layer 1-4 (bottom layers), Layer 5-8 (middle layers), Layer 9-12 (top layers).

Type	de	fr	zh	ru
CLS	3.88	4.73	0.89	2.15
Layer 6	51.29	56.32	41.38	38.81
Layer 7	62.51	62.62	49.99	51.84
Layer 8	64.32	62.46	50.49	53.58
Layer 1-4	6.98	12.3	12.05	4.33
Layer 5-8	63.12	63.42	52.84	51.67
Layer 9-12	53.97	52.68	44.18	43.13

Table 11. PAWS-X accuracy scores for each language.

Model	en	de	es	fr	ja	ko	zh	**avg**
mBERT	94.0	85.7	87.4	87.0	73.0	69.6	77.0	81.9
XLM	94.0	85.9	88.3	87.4	69.3	64.8	76.5	80.9
XLMR	**94.7**	**89.7**	**90.1**	**90.4**	**78.7**	**79.0**	**82.3**	**86.4**
MMTE	93.1	85.1	87.2	86.9	72.0	69.2	75.9	81.3
Translate-train	94.0	87.5	89.4	89.6	78.6	81.6	83.5	86.3
Translate-train (multi-task)	**94.5**	**90.5**	**91.6**	**91.7**	**84.4**	**83.9**	**85.8**	**88.9**
Translate-test	93.5	88.2	89.3	87.4	78.4	76.6	77.6	84.4

(XQuAD), 15 (MLQA), 14 (TyDiQA-GoldP), 12 (BUCC), and 9 (Tatoeba).

Table 12. BUCC results (F1 scores) for each languages.

Model	de	fr	ru	zh	avg
BERT	62.5	62.6	51.8	50.0	56.7
XLM	56.3	63.9	60.6	46.6	56.8
XLMR	67.5	**66.5**	**73.5**	**56.7**	**66.0**
MMTE	**67.9**	63.9	54.3	53.3	59.8

References

Hsu, T.-y., Liu, C.-l., and Lee, H.-y. Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model. In *Proceedings of EMNLP 2019*, pp. 5935–5942, 2019.

Lee, K., Yoon, K., Park, S., and Hwang, S. W. Semi-supervised training data generation for multilingual question answering. In *Proceedings of LREC 2018*, pp. 2758–2762, 2018.

Lewis, P., Ouz, B., Rinott, R., Riedel, S., and Schwenk, H. MLQA: Evaluating Cross-lingual Extractive Question Answering. *arXiv preprint arXiv:1910.07475*, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Table 13. XQuAD results (F1 / EM) for each language.

Model	en	ar	de	el	es	hi	ru	th	tr	vi	zh	**avg**
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	64.5 / 49.4
XLM	74.2 / 62.1	61.4 / 44.7	66.0 / 49.7	57.5 / 39.1	68.2 / 49.8	56.6 / 40.3	65.3 / 48.2	35.4 / 24.5	57.9 / 41.2	65.8 / 47.6	49.7 / 39.7	59.8 / 44.3
XLMR	**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**	**76.6 / 60.8**
MMTE	80.1 / 68.1	63.2 / 46.2	68.8 / 50.3	61.3 / 35.9	72.4 / 52.5	61.3 / 47.2	68.4 / 45.2	48.4 / 35.9	58.1 / 40.9	70.9 / 50.1	55.8 / 36.4	64.4 / 46.2
Translate-train	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-train (multi-task)	86.0 / 74.5	71.0 / 54.1	78.8 / 63.9	74.2 / 56.1	82.4 / 66.2	71.3 / 56.2	78.1 / 63.0	38.1 / 34.5	70.6 / 55.7	78.5 / 58.8	67.7 / 58.7	72.4 / 58.3
Translate-test	**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**

Table 14. TyDiQA-GoldP results (F1 / EM) for each language.

Model	en	ar	bn	fi	id	ko	ru	sw	te	avg
mBERT	**75.3 / 63.6**	62.2 / 42.8	49.3 / 32.7	59.7 / 45.3	64.8 / 45.8	**58.8 / 50.0**	60.0 / 38.8	57.5 / 37.9	49.6 / 38.4	59.7 / 43.9
XLM	66.9 / 53.9	59.4 / 41.2	27.2 / 15.0	58.2 / 41.4	62.5 / 45.8	14.2 / 5.1	49.2 / 30.7	39.4 / 21.6	15.5 / 6.9	43.6 / 29.1
XLM-R	71.5 / 56.8	**67.6 / 40.4**	**64.0 / 47.8**	**70.5 / 53.2**	**77.4 / 61.9**	31.9 / 10.9	**67.0 / 42.1**	**66.1 / 48.1**	**70.1 / 43.6**	**65.1 / 45.0**
MMTE	62.9 / 49.8	63.1 / 39.2	55.8 / 41.9	53.9 / 42.1	60.9 / 47.6	49.9 / 42.6	58.9 / 37.9	63.1 / 47.2	54.2 / 45.8	58.1 / 43.8
Translate-train	75.3 / 63.6	61.5 / 44.1	31.9 / 31.9	62.6 / 49.0	68.6 / 52.0	53.2 / 41.3	53.1 / 33.9	61.9 / 45.5	27.4 / 17.5	55.1 / 42.1
Translate-train (multi-task)	73.2 / 62.5	**71.8 / 54.2**	49.7 / 36.3	68.1 / 53.6	72.3 / 55.2	58.6 / 47.8	64.3 / 45.3	66.8 / 48.9	53.3 / 40.2	64.2 / 49.3
Translate-test	**75.9 / 65.9**	68.8 / 49.6	**66.7 / 48.1**	**72.0 / 56.6**	**76.8 / 60.9**	**69.2 / 55.7**	**71.4 / 54.3**	**73.3 / 53.8**	**75.1 / 59.2**	**72.1 / 56.0**
Monolingual	75.3 / 63.6	80.5 / 67.0	71.1 / 60.2	75.6 / 64.1	**81.3 / 70.4**	59.0 / 49.6	72.1 / 56.2	75.0 / 66.7	80.2 / 66.4	74.5 / 62.7
Monolingual few-shot	63.1 / 50.9	61.3 / 44.8	58.7 / 49.6	51.4 / 38.1	70.4 / 58.1	45.4 / 38.4	56.9 / 42.6	55.4 / 46.3	65.2 / 49.6	58.7 / 46.5
Joint monolingual	**77.6 / 69.3**	**82.7 / 69.4**	**79.6 / 69.9**	**79.2 / 67.8**	68.9 / 72.7	**68.9 / 59.4**	**75.8 / 59.2**	**81.9 / 74.3**	**83.4 / 70.3**	**77.6 / 68.0**

Table 15. MLQA results (F1 / EM) for each language.

Model	en	ar	de	es	hi	vi	zh	avg
mBERT	80.2 / 67.0	52.3 / 34.6	59.0 / 43.8	67.4 / 49.2	50.2 / 35.3	61.2 / 40.7	59.6 / 38.6	61.4 / 44.2
XLM	68.6 / 55.2	42.5 / 25.2	50.8 / 37.2	54.7 / 37.9	34.4 / 21.1	48.3 / 30.2	40.5 / 21.9	48.5 / 32.6
XLM-R	**83.5 / 70.6**	**66.6 / 47.1**	**70.1 / 54.9**	**74.1 / 56.6**	**70.6 / 53.1**	**74 / 52.9**	**62.1 / 37.0**	**71.6 / 53.2**
MMTE	78.5 / –	56.1 / –	58.4 / –	64.9 / –	46.2 / –	59.4 / –	58.3 / –	60.3 / 41.4
Translate-train	80.2 / 67.0	55.0 / 35.6	64.4 / 49.4	70.0 / 52.0	60.1 / 43.4	65.7 / 45.5	63.9 / 42.7	65.6 / 47.9
Translate-train (multi-task)	80.7 / 67.7	58.9 / 39.0	66.0 / 51.6	71.3 / 53.7	62.4 / 45.0	67.9 / 47.6	66.0 / 43.9	67.6 / 49.8
Translate-test	**83.8 / 71.0**	**65.3 / 46.4**	**71.2 / 54.0**	**73.9 / 55.9**	**71.0 / 55.1**	**70.6 / 54.0**	**67.2 / 50.6**	**71.9 / 55.3**

Table 16. POS results (Accuracy) for each language

Lang.	af	ar	bg	de	el	en	es	et	eu	fa	fi	fr	he	hi	hu	id	it
mBERT	86.6	56.2	85.0	85.2	81.1	95.5	86.9	79.1	60.7	66.7	78.9	84.2	56.2	67.2	78.3	71.0	88.4
XLM	88.5	63.1	85.0	85.8	84.3	95.4	85.8	78.3	62.8	64.7	78.4	82.8	65.9	66.2	77.3	70.2	87.4
XLMR	**89.8**	**67.5**	**88.1**	**88.5**	**86.3**	96.1	**88.3**	**86.5**	**72.5**	**70.6**	**85.8**	**87.2**	**68.3**	**76.4**	**82.6**	72.4	**89.4**
MMTE	86.2	65.9	87.2	85.8	77.7	**96.6**	85.8	81.6	61.9	67.3	81.1	84.3	57.3	76.4	78.1	**73.5**	89.2
	ja	kk	ko	mr	nl	pt	ru	ta	te	th	tl	tr	ur	vi	yo	zh	avg
mBERT	**49.2**	70.5	49.6	69.4	88.6	86.2	85.5	59.0	75.9	41.7	81.4	68.5	57.0	53.2	**55.7**	61.6	71.5
XLM	49.0	70.2	50.1	68.7	88.1	84.9	86.5	59.8	76.8	55.2	76.3	66.4	61.2	52.4	20.5	65.4	71.3
XLMR	15.9	**78.1**	53.9	**80.8**	**89.5**	**87.6**	**89.5**	**65.2**	**86.6**	**47.2**	**92.2**	**76.3**	**70.3**	**56.8**	24.6	25.7	**73.8**
MMTE	48.6	70.5	**59.3**	74.4	83.2	86.1	88.1	63.7	81.9	43.1	80.3	71.8	61.1	56.2	51.9	**68.1**	73.5

Table 17. NER results (F1 Score) for each language

Lang.	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv
mBERT	**85.2**	77.4	41.1	77.0	70.0	78.0	72.5	77.4	75.4	**66.3**	46.2	77.2	79.6	56.6	65.0	76.4	**53.5**	**81.5**	29.0	**66.4**
XLM	82.6	74.9	44.8	76.7	70.0	78.1	73.5	74.8	74.8	62.3	49.2	79.6	78.5	57.7	66.1	76.5	53.1	80.7	23.6	63.0
XLMR	84.7	**78.9**	**53.0**	**81.4**	**78.8**	**78.8**	**79.5**	**79.6**	**79.1**	60.9	**61.9**	**79.2**	**80.5**	**56.8**	**73.0**	**79.8**	53.0	81.3	23.2	62.5
MMTE	77.9	74.9	41.8	75.1	64.9	71.9	68.3	71.8	74.9	62.6	45.6	75.2	73.9	54.2	66.2	73.8	47.9	74.1	**31.2**	63.9

	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh
mBERT	64.6	45.8	59.6	52.3	58.2	**72.7**	45.2	81.8	80.8	64.0	67.5	50.7	48.5	3.6	71.7	71.8	36.9	71.8	**44.9**	**42.7**
XLM	67.7	**57.2**	26.3	59.4	62.4	69.6	47.6	81.2	77.9	63.5	68.4	53.6	49.6	0.3	**78.6**	71.0	43.0	70.1	26.5	32.4
XLMR	**71.6**	56.2	**60.0**	**67.8**	**68.1**	57.1	**54.3**	**84.0**	**81.9**	**69.1**	**70.5**	**59.5**	**55.8**	1.3	73.2	**76.1**	**56.4**	**79.4**	33.6	33.1
MMTE	60.9	43.9	58.2	44.8	58.5	68.3	42.9	74.8	72.9	58.2	66.3	48.1	46.9	**3.9**	64.1	61.9	37.2	68.1	32.1	28.9