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]

Abstract

Much recent progress in applications of machine learning models to NLP has been driven by benchmarks that evaluate models across a wide variety of tasks. However, these broad-coverage benchmarks have been mostly limited to English, and despite an increasing interest in multilingual models, a benchmark that enables the comprehensive evaluation of such methods on a diverse range of languages and tasks is still missing. To this end, we introduce the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark, a multi-task benchmark for evaluating the cross-lingual generalization capabilities of multilingual representations across 40 languages and 9 tasks. We demonstrate that while models tested on English reach human performance on many tasks, there is still a sizable gap in the performance of cross-lingually transferred models, particularly on syntactic and sentence retrieval tasks. There is also a wide spread of results across languages. We release the benchmark[1] to encourage research on cross-lingual learning methods that transfer linguistic knowledge across a diverse and representative set of languages and tasks.

1. Introduction

In natural language processing (NLP), there is a pressing urgency to build systems that serve *all* of the world's approximately 6,900 languages to overcome language barriers and enable universal information access for the world's citizens (Ruder et al., 2019; Aharoni et al., 2019; Arivazhagan et al., 2019). At the same time, building NLP systems for most of these languages is challenging due to a stark lack of data. Luckily, many languages have similarities in syntax or vocabulary, and multilingual learning approaches that train on multiple languages while leveraging the shared structure of the input space have begun to show promise as ways to alleviate data sparsity. Early work in this direction focused on single tasks, such as grammar induction (Snyder et al., 2009), part-of-speech (POS) tagging (Täckström et al., 2013), parsing (McDonald et al., 2011), and text classification (Klementiev et al., 2012). Over the last few years, there has been a move towards *general-purpose multilingual representations* that are applicable to many tasks, both on the word level (Mikolov et al., 2013; Faruqui & Dyer, 2014; Artetxe et al., 2017) or the full-sentence level (Devlin et al., 2019; Lample & Conneau, 2019). Despite the fact that such representations are intended to be general-purpose, evaluation of them has often been performed on a very limited and often disparate set of tasks—typically focusing on translation (Glavaš et al., 2019; Lample & Conneau, 2019) and classification (Schwenk & Li, 2018; Conneau et al., 2018b)—and typologically similar languages (Conneau et al., 2018a).

To address this problem and incentivize research on truly general-purpose cross-lingual representation and transfer learning, we introduce the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark. XTREME covers 40 typologically diverse languages spanning 12 language families and includes 9 tasks that require reasoning about different levels of syntax or semantics.[2] In addition, we introduce *pseudo* test sets as diagnostics that cover all 40 languages by automatically translating the English test set of the natural language inference and question-answering dataset to the remaining languages.

XTREME focuses on the *zero-shot cross-lingual transfer* scenario, where annotated training data is provided in English but none is provided in the language to which systems must transfer.[3] We evaluate a range of state-of-the-art machine

*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] The benchmark is publicly available at https://sites.research.google/xtreme. The codes used for downloading data and training baseline models are available at https://github.com/google-research/xtreme.

[2] By typologically diverse, we mean languages that span a wide set of linguistic phenomena such as compounding, inflection, derivation, etc. which occur in many of the world's languages.

[3] This is done both for efficiency purposes (as it only requires testing, not training, on each language) and practical considerations (as annotated training data is not available for many languages).

translation (MT) and multilingual representation-based approaches to performing this transfer. We find that while state-of-the-art models come close to human performance in English on many of the tasks we consider, performance drops significantly when evaluated on other languages. Overall, performance differences are highest for syntactic and sentence retrieval tasks. Further, while models do reasonably well in most languages in the Indo-European family, we observe lower performance particularly for Sino-Tibetan, Japonic, Koreanic, and Niger-Congo languages.

In sum, our contributions are the following: (i) We release a suite of 9 cross-lingual benchmark tasks covering 40 typologically diverse languages. (ii) We provide an online platform and leaderboard for the evaluation of multilingual models. (iii) We provide a set of strong baselines, which we evaluate across all tasks, and release code to facilitate adoption. (iv) We provide an extensive analysis of limitations of state-of-the-art cross-lingual models.

2. Related Work

Cross-lingual representations Early work focused on learning cross-lingual representations using either parallel corpora (Gouws et al., 2015; Luong et al., 2015) or a bilingual dictionary to learn a linear transformation (Mikolov et al., 2013; Faruqui & Dyer, 2014). Later approaches reduced the amount of supervision required using self-training (Artetxe et al., 2017) and unsupervised strategies such as adversarial training (Conneau et al., 2018a), heuristic initialisation (Artetxe et al., 2018), and optimal transport (Zhang et al., 2017). Building on advances in monolingual transfer learning (McCann et al., 2017; Howard & Ruder, 2018; Peters et al., 2018; Devlin et al., 2019), multilingual extensions of pretrained encoders have recently been shown to be effective for learning deep cross-lingual representations (Eriguchi et al., 2018; Pires et al., 2019; Wu & Dredze, 2019; Lample & Conneau, 2019; Siddhant et al., 2020).

Cross-lingual evaluation One pillar of the evaluation of cross-lingual representations has been translation, either on the word level (*bilingual lexicon induction*) or on the sentence level (*machine translation*). In most cases, evaluation has been restricted to typologically related languages and similar domains; approaches have been shown to fail in less favorable conditions (Glavaš et al., 2019; Vulić et al., 2019; Guzmán et al., 2019). Past work has also reported issues with common datasets for bilingual lexicon induction (Czarnowska et al., 2019; Kementchedjhieva et al., 2019) and a weak correlation with certain downstream tasks (Glavaš et al., 2019). Translation, however, only covers one facet of a model's cross-lingual generalization ability. For instance, it does not capture differences in classification performance that are due to cultural differences (Mohammad et al., 2016; Smith et al., 2016).

On the other hand, cross-lingual approaches have been evaluated on a wide range of tasks, including dependency parsing (Schuster et al., 2019), named entity recognition (Rahimi et al., 2019), sentiment analysis (Barnes et al., 2018), natural language inference (Conneau et al., 2018b), document classification (Schwenk & Li, 2018), and question answering (Artetxe et al., 2020; Lewis et al., 2019). Evaluation on a single task is problematic as past work has noted potential issues with standard datasets: MLDoc (Schwenk & Li, 2018) can be solved by matching keywords (Artetxe et al., 2020), while MultiNLI, the dataset from which XNLI (Conneau et al., 2018b) was derived, contains superficial cues that can be exploited (Gururangan et al., 2018). Evaluation on multiple tasks is thus necessary to fairly compare cross-lingual models. Benchmarks covering multiple tasks like GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) have arguably spurred research in monolingual transfer learning. In the cross-lingual setting, such a benchmark not only needs to cover a diverse set of tasks but also languages. XTREME aims to fill this gap.

3. XTREME

3.1. Design principles

Given XTREME's goal of providing an accessible benchmark for the evaluation of cross-lingual transfer learning on a diverse and representative set of tasks and languages, we select the tasks and languages that make up the benchmark based on the following principles:

Task difficulty Tasks should be sufficiently challenging so that cross-language performance falls short of human performance.

Task diversity Tasks should require multilingual models to transfer their meaning representations at different levels, e.g. words, phrases and sentences. For example, while classification tasks require sentence-level transfer of meaning, sequence labeling tasks like part-of-speech (POS) tagging or named entity recognition (NER) test the model's transfer capabilities at the word level.

Training efficiency Tasks should be trainable on a single GPU for less than a day. This is to make the benchmark accessible, in particular to practitioners working with low-resource languages under resource constraints.

Multilinguality We prefer tasks that cover as many languages and language families as possible.

Sufficient monolingual data Languages should have sufficient monolingual data for learning useful pre-trained representations.

Accessibility Each task should be available under a permissive license that allows the use and redistribution of the

Table 1. Characteristics of the datasets in XTREME for the zero-shot transfer setting. For tasks that have training and dev sets in other languages, we only report the English numbers. We report the number of test examples per target language and the nature of the test sets (whether they are translations of English data or independently annotated). The number in brackets is the size of the intersection with our selected languages. For NER and POS, sizes are in sentences. Struct. pred.: structured prediction. Sent. retrieval: sentence retrieval.

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517–11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

data for research purposes.

3.2. Tasks

XTREME consists of nine tasks that fall into four different categories requiring reasoning on different levels of meaning. We give an overview of all tasks in Table 1, and describe the task details as follows.

XNLI The Cross-lingual Natural Language Inference corpus (Conneau et al., 2018b) asks whether a premise sentence entails, contradicts, or is neutral toward a hypothesis sentence. Crowd-sourced English data is translated to ten other languages by professional translators and used for evaluation, while the MultiNLI (Williams et al., 2018) training data is used for training.

PAWS-X The Cross-lingual Paraphrase Adversaries from Word Scrambling (Yang et al., 2019) dataset requires to determine whether two sentences are paraphrases. A subset of the PAWS dev and test sets (Zhang et al., 2019) was translated to six other languages by professional translators and is used for evaluation, while the PAWS training set is used for training.

POS We use POS tagging data from the Universal Dependencies v2.5 (Nivre et al., 2018) treebanks, which cover 90 languages. Each word is assigned one of 17 universal POS tags. We use the English training data for training and evaluate on the test sets of the target languages.

NER For NER, we use the `Wikiann` (Pan et al., 2017) dataset. Named entities in Wikipedia were automatically annotated with `LOC`, `PER`, and `ORG` tags in IOB2 format using a combination of knowledge base properties, cross-lingual and anchor links, self-training, and data selection. We use the balanced train, dev, and test splits from Rahimi et al. (2019).

XQuAD The Cross-lingual Question Answering Dataset (Artetxe et al., 2020) requires identifying the answer to a question as a span in the corresponding paragraph. A subset of the English SQuAD v1.1 (Rajpurkar et al., 2016) dev set was translated into ten other languages by professional translators and is used for evaluation.

MLQA The Multilingual Question Answering (Lewis et al., 2019) dataset is another cross-lingual question answering dataset similar to XQuAD. The evaluation data for English and six other languages was obtained by automatically mining target language sentences that are parallel to sentences in English from Wikipedia, crowd-sourcing annotations in English, and translating the question and aligning the answer spans in the target languages. For both XQuAD and MLQA, we use the SQuAD v1.1 training data for training and evaluate on the test data of the corresponding task.

TyDiQA-GoldP We use the gold passage version of the Typologically Diverse Question Answering (Clark et al., 2020) dataset, a benchmark for information-seeking question answering, which covers nine languages. The gold passage version is a simplified version of the primary task, which uses only the gold passage as context and excludes unanswerable questions. It is thus similar to XQuAD and MLQA, while being more challenging as questions have been written without seeing the answers, leading to 3× and 2× less lexical overlap compared to XQuAD and MLQA respectively. We use the English training data for training and evaluate on the test sets of the target languages.

BUCC The goal of the second and third shared task of the workshop on Building and Using Parallel Corpora (Zweigenbaum et al., 2017; 2018) is to extract parallel sentences from a comparable corpus between English and four other languages. The dataset provides train and test splits for each language. For simplicity, we evaluate representations on the test sets directly without fine-tuning and calculate similarity using cosine similarity.[4]

Tatoeba We use the Tatoeba dataset (Artetxe & Schwenk, 2019), which consists of up to 1,000 English-aligned sentence pairs covering 122 languages. We find the nearest neighbour using cosine similarity and calculate error rate.

3.3. Languages

As noted in Section 3.1, we choose our target languages based on availability of monolingual data, and typological diversity. We use the number of articles in Wikipedia as a proxy for the amount of monolingual data available online. In order to strike a balance between language diversity and availability of monolingual data, we select all languages out of the top 100 Wikipedias[5] with the most articles as of December 2019.[6] We first select all languages that appear in at least three of our benchmark datasets. This leaves us with 19 languages, most of which are Indo-European or major world languages. We now select 21 additional languages that appear in at least one dataset and come from less represented language families. Wherever possible, we choose at least two languages per family.[7]

In total, XTREME covers the following 40 languages (shown with their ISO 639-1 codes for brevity) belonging to 12 language families and two isolates: af, ar, bg, bn, de, el, en, es, et, eu, fa, fi, fr, he, hi, hu, id, it, ja, jv, ka, kk, ko, ml, mr, ms, my, nl, pt, ru, sw, ta, te, th, tl, tr, ur, vi, yo, and zh. We provide a detailed overview of these languages in terms of their number of Wikipedia articles, linguistic features, and coverage in XTREME in the appendix.

While XTREME covers these languages in the sense that there is gold standard data in at least one task in each language, this does not mean that it covers all aspects of each language that are necessary for transfer. Languages may reveal different characteristics based on the task, domain, and register in which they are used. XTREME thus only serves as a glimpse into a model's true cross-lingual generalization capability.

3.4. Pseudo test data for analyses

XTREME covers 40 languages overall. Evaluation across the majority of languages is only possible for a subset of tasks, i.e. POS, NER, and Tatoeba. As additional diagnositics and to enable a broader comparison across languages for a more diverse set of tasks, we automatically translate the English portions of a representative classification and QA task to the remaining languages using an in-house translation system.[8] We choose XNLI and XQuAD as both have test sets that are translations of the English data by professional translators.

We first verify that performance on the translated test sets is a good proxy for performance on the gold standard test sets. We report the detailed results in the appendix. For XQuAD, the automatically translated test sets underestimate mBERT's true performance by 3.0 F1 / 0.2 EM points, similar to the 2.6 F1 points reported by Agić & Schluter (2018) when translating the test data to other languages.[9] For XNLI, the automatically translated test sets overestimate the true prediction accuracy by 2.4 points. In order to measure the translation quality between the human-translated test data and our pseudo test data, we compute the BLEU score, and the chrF score (Popović, 2015), which is suitable for measuring the translation quality of some languages such as Chinese and Russian. For the 14 languages in XNLI, we obtain average scores of 34.2 BLEU and 58.9 chrF scores on our pseudo test data compared to the reference translations, which correlate with a Pearson's ρ of 0.57 and 0.28 respectively with mBERT performance.

Translating the English data to the remaining languages yields 40-way parallel pseudo test data that we employ for analyses in Section 5.

4. Experiments

4.1. Training and evaluation setup

XTREME focuses on the evaluation of multilingual representations. We do not place any restriction on the amount or nature of the monolingual data used for pretraining multilingual representations. However, we request authors to be explicit about the data they use for training, in particular any cross-lingual signal. In addition, we suggest authors should not use any additional labelled data in the target task beyond the one that is provided.

For evaluation, we focus on *zero-shot cross-lingual transfer* with English as the source language as this is the most common setting for the evaluation of multilingual representations and as many tasks only have training data available in English. Although English is not generally the best source language for cross-lingual transfer for all target languages (Lin et al., 2019), this is still the most practically useful setting. A single source language also facilitates evaluation as models only need to be trained once and can be evaluated

[4] Results can be improved using more sophisticated similarity metrics (Artetxe & Schwenk, 2019).

[5] `https://meta.wikimedia.org/wiki/List_of_Wikipedias`

[6] This also has the benefit that they are covered by state-of-the-art methods such as mBERT and XLM.

[7] For the Austro-Asiatic, Kartvelian, and Kra-Dai families as well as for isolates, we only obtain one language.

[8] Details of our translation system are provided in the appendix.

[9] Note that even human translated test sets may underestimate a model's true cross-lingual generalization ability as such *translationese* has been shown to be less lexically diverse than naturally composed language (Koppel & Ordan).

on all other languages.[10]

Concretely, pretrained multilingual representations are fine-tuned on English labelled data of an XTREME task. The model is then evaluated on the test data of the task in the target languages.

4.2. Baselines

We evaluate a number of strong baselines and state-of-the-art models. The approaches we consider learn multilingual representations via self-supervision or leverage translations—either for representation learning or for training models in the source or target language. We focus on models that learn deep contextual representations as these have achieved state-of-the-art results on many tasks. For comparability among the representation learning approaches, we focus on models that learn a multilingual embedding space between all languages in XTREME. We encourage future work to focus on these languages to capture as much language diversity as possible. We report hyper-parameters in the appendix. All hyper-parameter tuning is done on English validation data. We encourage authors evaluating on XTREME to do the same.

mBERT Multilingual BERT (Devlin et al., 2019) is a transformer model (Vaswani et al., 2017) that has been pretrained on the Wikipedias of 104 languages using masked language modelling (MLM).

XLM XLM (Lample & Conneau, 2019) uses a similar pretraining objective as mBERT with a larger model, a larger shared vocabulary, and trained on the same Wikipedia data covering 100 languages.

XLM-R XLM-R Large (Conneau et al., 2020) is similar to XLM but was trained on more than a magnitude more data from the web covering 100 languages.

MMTE The massively multilingual translation encoder is part of an NMT model that has been trained on in-house parallel data of 103 languages extracted from the web (Arivazhagan et al., 2019). For transfer, we fine-tune the encoder of the model (Siddhant et al., 2020).

Translate-train For many language pairs, an MT model may be available, which can be used to obtain data in the target language. To evaluate the impact of using such data, we translate the English training data into the target language using our in-house MT system. We then fine-tune mBERT on the translated data. We provide details on how we align answer spans in the source and target language for the QA tasks in the appendix. We do not provide translation-based baselines for structured prediction tasks due to an abundance of in-language data and a requirement for annotation projection.

Translate-train multi-task We also experiment with a multi-task version of the translate-train setting where we fine-tune mBERT on the combined translated training data of all languages jointly.

Translate-test Alternatively, we train the English BERT-Large (Devlin et al., 2019) model on the English training data and evaluate it on test data that we translated from the target language to English using our in-house MT system.

In-language model For the POS, NER, and TyDiQA-GoldP tasks where target-language training data is available, we fine-tune mBERT on monolingual data in the target language to estimate how useful target language labelled data is compared to labelled data in a source language.

In-language few-shot In many cases, it may be possible to procure a small number of labelled examples in the target language (Eisenschlos et al., 2019). To evaluate the viability of such an approach, we additionally compare against an mBERT model fine-tuned on 1,000 target language examples for the tasks where monolingual training data is available in the target languages.

In-language multi-task For the tasks where monolingual training data is available, we additionally compare against an mBERT model that is jointly trained on the combined training data of all languages.

Human performance For XNLI, PAWS-X, and XQuAD, we obtain human performance estimates from the English datasets they are derived from, MNLI, PAWS-X, and SQuAD respectively (Nangia & Bowman, 2019; Zhang et al., 2019; Rajpurkar et al., 2016).[11] For TyDiQA-GoldP, we use the performance estimate of Clark et al. (2020). For MLQA, as answers are annotated using the same format as SQuAD, we employ the same human performance estimate. For POS tagging, we adopt 97% as a canonical estimate of human performance based on Manning (2011). We are not able to obtain human performance estimates for NER as annotations have been automatically generated and for sentence retrieval as identifying a translation among a large number of documents is too time-consuming.

4.3. Results

Overall results We show the main results in Table 2. XLM-R is the best-performing zero-shot transfer model and generally improves upon mBERT significantly. The improvement is smaller, however, for the structured prediction tasks. MMTE achieves performance competitive with mBERT on most tasks, with stronger results on XNLI, POS, and BUCC.

[10] Future work may also consider multi-source transfer, which is interesting particularly for low-resource languages, and transfer to unknown languages or unknown language-task combinations.

[11] Performance may differ across languages due to many factors but English performance still serves as a reasonable proxy.

Table 2. Overall results of baselines across all XTREME tasks. Translation-based baselines are not meaningful for sentence retrieval. We provide in-language baselines where target language training data is available. Note that for the QA tasks, translate-test performance is not directly comparable to the other scores as a small number of test questions were discarded and alignment is measured on the English data.

Model	Avg	Pair sentence		Structured prediction		Question answering			Sentence retrieval	
		XNLI	PAWS-X	POS	NER	XQuAD	MLQA	TyDiQA-GoldP	BUCC	Tatoeba
Metrics		Acc.	Acc.	F1	F1	F1 / EM	F1 / EM	F1 / EM	F1	Acc.
Cross-lingual zero-shot transfer (models are trained on English data)										
mBERT	59.8	65.4	81.9	71.5	62.2	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9	56.7	38.7
XLM	55.7	69.1	80.9	71.3	61.2	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1	56.8	32.6
XLM-R Large	68.2	79.2	86.4	73.8	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0	66.0	57.3
MMTE	59.5	67.4	81.3	73.5	58.3	64.4 / 46.2	60.3 / 41.4	58.1 / 43.8	59.8	37.9
Translate-train (models are trained on English training data translated to the target language)										
mBERT	-	74.0	86.3	-	-	70.0 / 56.0	65.6 / 48.0	55.1 / 42.1	-	-
mBERT, multi-task	-	75.1	88.9	-	-	72.4 / 58.3	67.6 / 49.8	64.2 / 49.3	-	-
Translate-test (models are trained on English data and evaluated on target language data translated to English)										
BERT-large	-	76.8	84.4	-	-	76.3 / 62.1	72.9 / 55.3	72.1 / 56.0	-	-
In-language models (models are trained on the target language training data)										
mBERT, 1000 examples	-	-	-	87.6	77.9	-	-	58.7 / 46.5	-	-
mBERT	-	-	-	89.8	88.3	-	-	74.5 / 62.7	-	-
mBERT, multi-task	-	-	-	91.5	89.1	-	-	77.6 / 68.0	-	-
Human	-	92.8	97.5	97.0	-	91.2 / 82.3	91.2 / 82.3	90.1 / -	-	-

If a strong MT system is available, translating the training sets provides improvements over using the same model with zero-shot transfer. Translating the test data provides similar benefits compared to translating the training data and is particularly effective for the more complex QA tasks, while being more expensive during inference time. While using an MT system as a black box leads to strong baselines, the MT system could be further improved in the context of data augmentation.

For the tasks where in-language training data is available, multilingual models trained on in-language data outperform zero-shot transfer models. However, zero-shot transfer models nevertheless outperform multilingual models trained on only 1,000 in-language examples on the complex QA tasks as long as more samples in English are available. For the structured prediction tasks, 1,000 in-language examples enable the model to achieve performance that is similar to being trained on the full labelled dataset, similar to findings for classification (Eisenschlos et al., 2019). Finally, multi-task learning on the Translate-train and In-language setting generally improves upon single language training.

Cross-lingual transfer gap For a number of representative models, we show the cross-lingual transfer gap, i.e. the difference between the performance on the English test set and all other languages in Table 3.[12] While powerful models such as XLM-R reduce the gap significantly compared to mBERT for challenging tasks such as XQuAD and MLQA, they do not have the same impact on the syntactic structured prediction tasks. On the classification tasks, the transfer learning gap is lowest, indicating that there may be less headroom for progress on these tasks. The use of MT reduces the gap across all tasks. Overall, a large gap remains for all approaches, which indicates much potential for work on cross-lingual transfer.

Table 3. The cross-lingual transfer gap (lower is better) of different models on XTREME tasks. The transfer gap is the difference between performance on the English test set and the average performance on the other languages. A transfer gap of 0 indicates perfect cross-lingual transfer. For the QA datasets, we only show EM scores. The average gaps are computed over the sentence classification and QA tasks.

Model	XNLI	PAWS-X	XQuAD	MLQA	TyDiQA-GoldP	Avg	POS	NER
mBERT	16.5	14.1	25.0	27.5	22.2	21.1	25.5	23.6
XLM-R	10.2	12.4	16.3	19.1	13.3	14.3	24.3	19.8
Translate-train	7.3	9.0	17.6	22.2	24.2	16.1	-	-
Translate-test	6.7	12.0	16.3	18.3	11.2	12.9	-	-

5. Analyses

We conduct a series of analyses investigating the limitations of state-of-the-art cross-lingual models.

Best zero-shot model analysis We show the performance of the best zero-shot transfer model, XLM-R Large broken down by task and language in Figure 1. The figure illus-

[12] This comparison should be taken with a grain of salt, as scores across languages are not directly comparable for the tasks where test sets differ, i.e. POS, NER, MLQA, and TyDiQA-GoldP and differences in scores may not be linearly related.

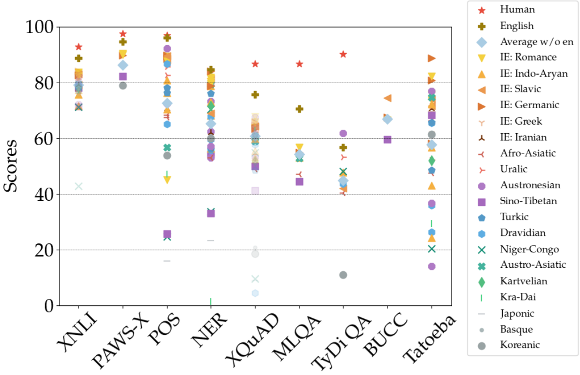

Figure 1. An overview of XLM-R's performance on the XTREME tasks across all languages in each task. We highlight an estimate of human performance, performance on the English test set, the average of all languages excluding English, and the family of each language. Performance on pseudo test sets for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

Table 4. Accuracy of mBERT on POS tag trigrams and 4-grams in the target language dev data that appeared and did not appear in the English training data. We show the performance on English, the average across all other languages, and their difference.

	trigram, seen	trigram, unseen	4-gram, seen	4-gram, unseen
en	90.3	63.0	88.1	67.5
avg w/o en	50.6	12.1	44.3	18.3
difference	39.7	50.9	43.7	49.2

trates why it is important to evaluate general-purpose multilingual representations across a diverse range of tasks and languages: On XNLI, probably the most common standard cross-lingual evaluation task, and PAWS-X, scores cluster in a relatively small range—even considering pseudo test sets for XNLI. However, scores for the remaining tasks have significantly wider spread, particularly as we include pseudo test sets. For TyDiQA-GoldP, English performance is lowest in comparison; the high performance on members of the Austronesian and Uralic language families (Indonesian and Finnish) may be due to less complex Wikipedia context passages for these languages. Across tasks, we generally observe higher performance on Indo-European languages and lower performance for other language families, particularly for Sino-Tibetan, Japonic, Koreanic, and Niger-Congo languages. Some of these difficulties may be due to tokenisation and an under-representation of ideograms in the joint sentencepiece vocabulary, which has been shown to be important in a cross-lingual model's performance (Artetxe et al., 2020; Conneau et al., 2020). We observe similar trends for mBERT, for which we show the same graph in the appendix.

Correlation with pretraining data size We calculate the Pearson correlation coefficient ρ of the model performance and the number of Wikipedia articles (see appendix) in each language and show results in Figure 2.[13] For mBERT, which was pretrained on Wikipedia, we observe a high correlation for most tasks ($\rho \approx 0.8$) except for the structured prediction tasks where $\rho \approx 0.35$. We observe similar trends for XLM and XLM-R, with lower numbers for XLM-R due to the different pretraining domain (see the appendix). This indicates that current models are not able to fully leverage the information extracted from the pretraining data to transfer to syntactic tasks.

Analysis of language characteristics We analyze results based on different language families and writing scripts in Figure 3. For mBERT, we observe the best transfer performance on branches of the Indo-European language family such as Germanic, Romance and Slavic languages. In contrast, cross-lingual transfer performance on low-resource language families such as Niger-Congo and Kra-Dai is still low. Looking at scripts, we find that the performance on syntactic tasks differs among popular scripts such as Latin and ideogram scripts. For example in the NER task, mBERT performs better on data in Latin script than that in Chinese or Japanese ideograms. This indicates that the current models still have difficulty transferring word-level syntactic information across languages written in different scripts.

Errors across languages For XNLI and XQuAD where the other test sets are translations from English, we analyze whether approaches make the same type of errors in the source and target languages. To this end, we explore whether examples that are correctly and incorrectly predicted in English are correctly predicted in other languages. On the XNLI dev set, mBERT correctly predicts on average 71.8% of examples that were correctly predicted in English. For examples that were misclassified, the model's performance is about random. On average, predictions on XNLI are consistent between English and another language for 68.3% of examples. On the XQuAD test set, mBERT correctly predicts around 60% of examples that were correclty predicted in English and 20% of examples that were incorrectly predicted. While some of these are plausible spans, more work needs to focus on achieving consistent predictions across languages.

Generalization to unseen tag combinations and entities We analyze possible reasons for the less successful transfer on structured prediction tasks. The Universal Dependencies dataset used for POS tagging uses a common set of 17 POS tags for all languages, so a model is not required to

[13] We observe similar correlations when using the number of tokens in Wikipedia instead.

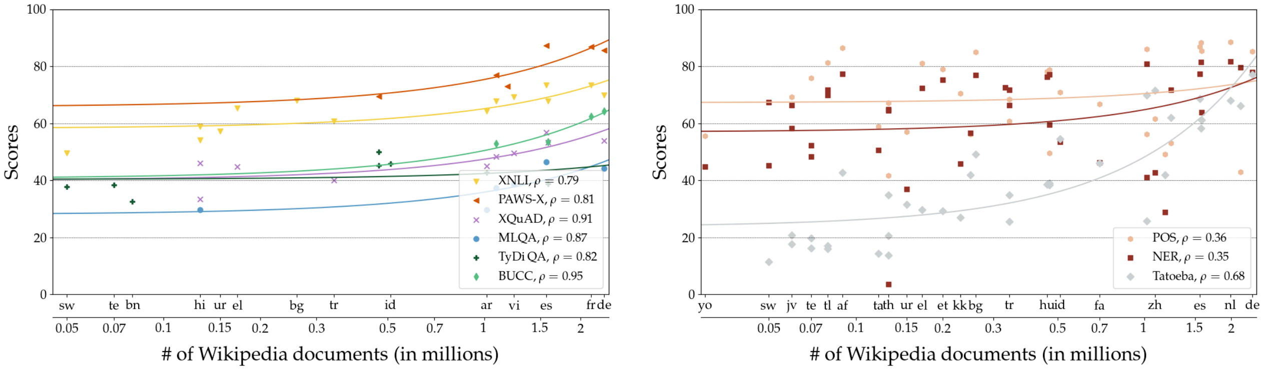

Figure 2. Performance of mBERT across tasks and languages in comparison to the number of Wikipedia articles for each language. We show tasks with a Pearson correlation coefficient $\rho > 0.7$ on the left and others on the right. Numbers across tasks are not directly comparable. We remove the x axis labels of overlapping languages for clarity. We additionally plot the linear fit for each task (curved due to the logarithmic scale of the x axis).

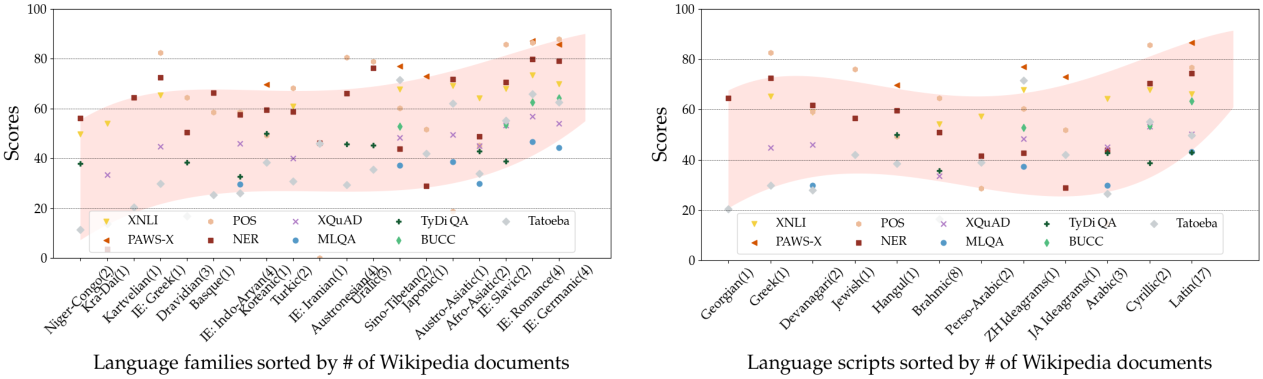

Figure 3. Performance of mBERT across tasks grouped by language families (left) and scripts (right). The number of languages per group is in brackets and the groups are from low-resource to high-resource on the x-axis. We additionally plot the 3rd order polynomial fit for the minimum and maximum values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

generalize to unseen tags at test time. However, a model may be required to generalize to unseen tag *combinations* at test time, for instance due to differences in word order between languages. We gauge how challenging such generalization is by computing a model's accuracy for POS tag n-grams in the target language dev data that were not seen in the English training data. We calculate values for tag trigrams and 4-grams and show accuracy scores for mBERT in Table 4. We observe the largest differences in performance for unseen trigrams and 4-grams, which highlights that existing cross-lingual models struggle to transfer to the syntactic characteristics of other languages. For NER, we estimate how well models generalize to unseen entities at test time. We compute mBERT's accuracy on entities in the target language dev data that were not seen in the English training data. We observe the largest difference between performance on seen and unseen entities for Indonesian and Swahili. Isolating for confounding factors such as entity length, frequency, and Latin script, we find the largest differences in performance for Swahili and Basque. Together, this indicates that the model may struggle to generalize to entities that are more characteristic of the target language. We show the detailed results for both analyses in the appendix.

6. Conclusions

As we have highlighted in our analysis, a model's cross-lingual transfer performance varies significantly both between tasks and languages. XTREME is a first step towards obtaining a more accurate estimate of a model's cross-lingual generalization ability. While XTREME is still inherently limited by the data coverage of its constituent tasks for many low-resource languages, XTREME nevertheless provides significantly broader coverage and more fine-grained analysis tools to encourage research on cross-lingual generalization ability of models. We have released the code for XTREME and scripts for fine-tuning models on tasks in XTREME, which should be to catalyze future research.

Acknowledgements

We'd like to thank Jon Clark for sharing with us the TyDiQA Gold Passage data and for valuable feedback. We would also like to thank Sam Bowman, Sebastian Goodman, and Tal Linzen for their feedback. JH and GN are sponsored by the Air Force Research Laboratory under agreement number FA8750-19-2-0200.

References

Agić, Ž. and Schluter, N. Baselines and test data for cross-lingual inference. In *Proceedings of LREC 2018*, 2018.

Aharoni, R., Johnson, M., and Firat, O. Massively Multilingual Neural Machine Translation. In *Proceedings of NAACL 2019*, 2019.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. *arXiv preprint arXiv:1907.05019*, 2019.

Artetxe, M. and Schwenk, H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the ACL 2019*, 2019.

Artetxe, M., Labaka, G., and Agirre, E. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL 2017*, pp. 451–462, 2017.

Artetxe, M., Labaka, G., and Agirre, E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL 2018*, pp. 789–798, 2018.

Artetxe, M., Ruder, S., and Yogatama, D. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of ACL 2020*, 2020.

Barnes, J., Klinger, R., and Schulte im Walde, S. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of ACL 2018*, pp. 2483–2493, Melbourne, Australia, 2018. Association for Computational Linguistics.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. In *Transactions of the Association of Computational Linguistics*, 2020.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. In *Proceedings of ICLR 2018*, 2018a.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pp. 2475–2485, 2018b.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL 2020*, 2020.

Czarnowska, P., Ruder, S., Grave, E., Cotterell, R., and Copestake, A. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of EMNLP 2019*, pp. 973–982, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*, 2019.

Eisenschlos, J., Ruder, S., Czapla, P., Kadras, M., Gugger, S., and Howard, J. MultiFiT: Efficient Multi-lingual Language Model Fine-tuning. In *Proceedings of EMNLP 2019*, 2019.

Eriguchi, A., Johnson, M., Firat, O., Kazawa, H., and Macherey, W. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*, 2018.

Faruqui, M. and Dyer, C. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL 2014*, pp. 462–471, 2014.

Glavaš, G., Litschko, R., Ruder, S., and Vulić, I. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *Proceedings of ACL 2019*, 2019.

Gouws, S., Bengio, Y., and Corrado, G. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML 2015*, pp. 748–756, 2015.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of NAACL-HLT 2018*, 2018.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. The FLoRes Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English. In *Proceedings of EMNLP 2019*, pp. 6100–6113, 2019.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In *Proceedings of ACL 2018*, pp. 328–339, 2018.

Kementchedjhieva, Y., Hartmann, M., and Søgaard, A. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of EMNLP 2019*, pp. 3327–3332, 2019.

Klementiev, A., Titov, I., and Bhattarai, B. Inducing Crosslingual Distributed Representations of Words. In *Proceedings of COLING 2012*, 2012.

Koppel, M. and Ordan, N. Translationese and its dialects. In *Proceedings of ACL 2011, pages=1318–1326, year=2011, organization=Association for Computational Linguistics*.

Lample, G. and Conneau, A. Cross-lingual Language Model Pretraining. In *Proceedings of NeurIPS 2019*, 2019.

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

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., and Neubig, G. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of ACL 2019*, 2019.

Luong, T., Pham, H., and Manning, C. D. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 151–159, 2015.

Manning, C. D. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pp. 171–189. Springer, 2011.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. Learned in translation: Contextualized word vectors. In *Proceedings of NIPS 2017*, pp. 6294–6305, 2017.

McDonald, R., Petrov, S., and Hall, K. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP 2011*, pp. 62–72, 2011.

Mikolov, T., Le, Q. V., and Sutskever, I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130, 2016.

Nangia, N. and Bowman, S. R. Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark. In *Proceedings of ACL 2019*, pp. 4566–4575, 2019.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., et al. Universal dependencies 2.2. 2018.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL 2017*, pp. 1946–1958, 2017.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of NAACL 2018*, pp. 2227–2237, 2018.

Pires, T., Schlinger, E., and Garrette, D. How multilingual is Multilingual BERT? In *Proceedings of ACL 2019*, 2019.

Popović, M. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, 2015.

Rahimi, A., Li, Y., and Cohn, T. Massively Multilingual Transfer for NER. In *Proceedings of ACL 2019*, 2019.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of EMNLP 2016*, 2016.

Ruder, S., Vulić, I., and Søgaard, A. A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019.

Schuster, T., Ram, O., Barzilay, R., and Globerson, A. Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. In *Proceedings of NAACL 2019*, 2019.

Schwenk, H. and Li, X. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of LREC 2018*, 2018.

Siddhant, A., Johnson, M., Tsai, H., Arivazhagan, N., Riesa, J., Bapna, A., Firat, O., and Raman, K. Evaluating the Cross-Lingual Effectiveness of Massively Multilingual Neural Machine Translation. In *Proceedings of AAAI 2020*, 2020.

Smith, L., Giorgi, S., Solanki, R., Eichstaedt, J., Schwartz, H. A., Abdul-Mageed, M., Buffone, A., and Ungar, L. Does well-being translate on twitter? In *Proceedings of EMNLP 2016*, pp. 2042–2047, 2016.

Snyder, B., Naseem, T., and Barzilay, R. Unsupervised multilingual grammar induction. In *Proceedings of ACL 2009*, pp. 73–81, 2009.

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. In *Transactions of the Association for Computational Linguistics*, 2013.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention Is All You Need. In *Proceedings of NIPS 2017*, 2017.

Vulić, I., Glavaš, G., Reichart, R., and Korhonen, A. Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? In *Proceedings of EMNLP 2019*, 2019.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS 2019*, 2019a.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of ICLR 2019*, 2019b.

Williams, A., Nangia, N., and Bowman, S. R. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of NAACL-HLT 2018*, 2018.

Wu, S. and Dredze, M. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of EMNLP 2019*, 2019.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP 2019*, pp. 3685–3690, 2019.

Zhang, M., Liu, Y., Luan, H., and Sun, M. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of EMNLP 2017*, pp. 1934–1945, 2017.

Zhang, Y., Baldridge, J., and He, L. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of NAACL 2019*, pp. 1298–1308, 2019.

Zweigenbaum, P., Sharoff, S., and Rapp, R. Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pp. 60–67, 2017.

Zweigenbaum, P., Sharoff, S., and Rapp, R. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pp. 39–42, 2018.