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

[Tasks](#) | [Download](#) | [Baselines](#) | [Leaderboard](#) | [Website](#) | [Paper](#) | [Translations](#)

This repository contains information about XTREME, code for downloading data, and implementations of baseline systems for the benchmark.

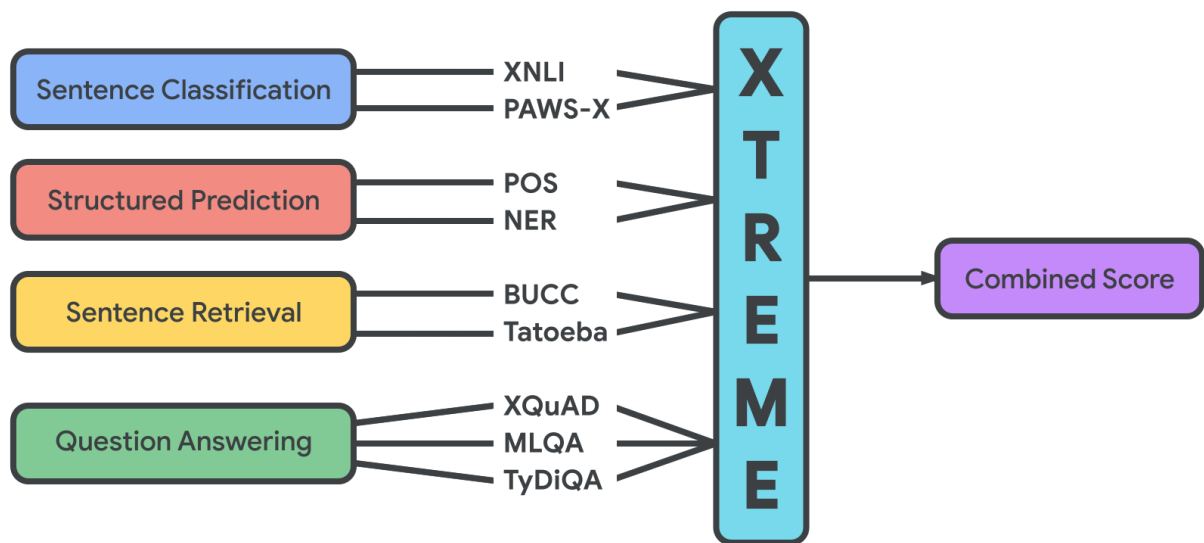
Introduction

The Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark is a benchmark for the evaluation of the cross-lingual generalization ability of pre-trained multilingual models. It covers 40 typologically diverse languages (spanning 12 language families) and includes nine tasks that collectively require reasoning about different levels of syntax and semantics. The languages in XTREME are selected to maximize language diversity, coverage in existing tasks, and availability of training data. Among these are many under-studied languages, such as the Dravidian languages Tamil (spoken in southern India, Sri Lanka, and Singapore), Telugu and Malayalam (spoken mainly in southern India), and the Niger-Congo languages Swahili and Yoruba, spoken in Africa.

For a full description of the benchmark, see the paper.

Tasks and Languages

The tasks included in XTREME cover a range of standard paradigms in natural language processing, including sentence classification, structured prediction, sentence retrieval and question answering. The full list of tasks can be seen in the image below.



In order for models to be successful on the XTREME benchmark, they must learn representations that generalize across many tasks and languages. Each of the tasks covers a subset of the 40 languages included in XTREME (shown here with their ISO 639-1 codes): 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. The languages were selected among the top 100 languages with the most Wikipedia articles to maximize language diversity, task coverage, and availability of training data. They include 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.

Download the data

In order to run experiments on XTREME, the first step is to download the dependencies. We assume you have installed [anaconda](#) and use Python 3.7+. The additional requirements including [transformers](#), [segeval](#) (for sequence labelling evaluation), [tensorboardx](#), [jieba](#), [kytea](#), and [pythainlp](#) (for text segmentation in Chinese, Japanese, and Thai), and [sacremoses](#) can be installed by running the following script:

```
1 bash install_tools.sh
```

The next step is to download the data. To this end, first create a `download` folder with `mkdir -p download` in the root of this project. You then need to manually download `panx_dataset` (for NER) from [here](#) (note that it will download as `AmazonPhotos.zip`) to the `download` directory. Finally, run the following command to download the remaining datasets:

```
1 bash scripts/download_data.sh
```

Note that in order to prevent accidental evaluation on the test sets while running experiments, we remove labels of the test data during pre-processing and change the order of the test sentences for cross-lingual sentence retrieval.

Build a baseline system

The evaluation setting in XTREME is zero-shot cross-lingual transfer from English. We fine-tune models that were pre-trained on multilingual data on the labelled data of each XTREME task in English. Each fine-tuned model is then applied to the test data of the same task in other languages to obtain predictions.

For every task, we provide a single script `scripts/train.sh` that fine-tunes pre-trained models implemented in the Transformers repo. To fine-tune a different model, simply pass a different `MODEL` argument to the script with the corresponding model. The current supported models are `bert-base-multilingual-cased`, `xlm-mlm-100-1280` and `xlm-roberta-large`.

Universal dependencies part-of-speech tagging

For part-of-speech tagging, we use data from the Universal Dependencies v2.5. You can fine-tune a pre-trained multilingual model on the English POS tagging data with the following command:

```
1 bash scripts/train.sh [MODEL] udpos
```

Wikiann named entity recognition

For named entity recognition (NER), we use data from the Wikiann (panx) dataset. You can fine-tune a pre-trained multilingual model on the English NER data with the following command:

```
1 bash scripts/train.sh [MODEL] panx
```

PAXS-X sentence classification

For sentence classification, we use the Cross-lingual Paraphrase Adversaries from Word Scrambling (PAWS-X) dataset. You can fine-tune a pre-trained multilingual model on the English PAWS data with the following command:

```
1 bash scripts/train.sh [MODEL] pawsx
```

XNLI sentence classification

The second sentence classification dataset is the Cross-lingual Natural Language Inference (XNLI) dataset. You can fine-tune a pre-trained multilingual model on the English MNLI data with the following command:

```
1 bash scripts/train.sh [MODEL] xnli
```

XQuAD, MLQA, TyDiQA-GoldP question answering

For question answering, we use the data from the XQuAD, MLQA, and TyDiQA-Gold Passage datasets. For XQuAD and MLQA, the model should be trained on the English SQuAD training set. For TyDiQA-Gold Passage, the model is trained on the English TyDiQA-GoldP training set. Using the following command, you can first fine-tune a pre-trained multilingual model on the corresponding English training data, and then you can obtain predictions on the test data of all tasks.

```
1 bash scripts/train.sh [MODEL] [xquad,mlqa,tydiqa]
```

BUCC sentence retrieval

For cross-lingual sentence retrieval, we use the data from the Building and Using Parallel Corpora (BUCC) shared task. As the models are not trained for this task but the representations of the pre-trained models are directly used to obtain similarity judgements, you can directly apply the model to obtain predictions on the test data of the task:

```
1 bash scripts/train.sh [MODEL] bucc2018
```

Tatoeba sentence retrieval

The second cross-lingual sentence retrieval dataset we use is the Tatoeba dataset. Similarly to BUCC, you can directly apply the model to obtain predictions on the test data of the task:

```
1 bash scripts/train.sh [MODEL] tatoeba
```

Leaderboard Submission

Submissions

To submit your predictions to **XTREME**, please create one single folder that contains 9 sub-folders named after all the tasks, i.e., `udpos`, `panx`, `xnli`, `pawsx`, `xquad`, `mlqa`, `tydiqa`, `bucc2018`, `tatoeba`. Inside each sub-folder, create a file containing the predicted labels of the test set for all languages. Name the file using the format `test-{language}.{extension}` where `language` indicates the 2-character language code, and `extension` is `json` for QA tasks and `tsv` for other tasks. You can see an example of the folder structure in `mock_test_data/predictions`.

Evaluation

We will compare your submissions with our label files using the following command:

```
1 python evaluate.py --prediction_folder [path] --label_folder [path]
```

Translations

As part of training translate-train and translate-test baselines we have automatically translated English training sets to other languages and tests sets to English. Translations are available for the following datasets: SQuAD v1.1 (only train and dev), MLQA, PAWS-X, TyDiQA-GoldP, XNLI, and XQuAD.

For PAWS-X and XNLI, the translations are in the following format: Column 1 and Column 2: original sentence pairs Column 3 and Column 4: translated sentence pairs Column 5: label

This will help make the association between the original data and their translations.

For XNLI and XQuAD, we have furthermore created pseudo test sets by automatically translating the English test set to the remaining languages in XTREME so that test data for all 40 languages is available. Note that these translations are noisy and should not be treated as ground truth.

All translations are available [here](#).

Paper

If you use our benchmark or the code in this repo, please cite our paper \cite{hu2020xtreme}.

```

1 @article{hu2020xtreme,
2   author    = {Junjie Hu and Sebastian Ruder and Aditya Siddhant
3               and Graham Neubig and Orhan Firat and Melvin Johnson},
4   title     = {XTREME: A Massively Multilingual Multi-task
5               Benchmark for Evaluating Cross-lingual Generalization},
6   journal   = {CoRR},
7   volume    = {abs/2003.11080},
8   year      = {2020},
9   archivePrefix = {arXiv},
10  eprint     = {2003.11080}
11 }

```

Please consider including a note similar to the one below to make sure to cite all the individual datasets in your paper.

We experiment on the XTREME benchmark \cite{hu2020xtreme}, a composite benchmark for multi-lingual learning consisting of data from the XNLI \cite{Conneau2018xnli}, PAWS-X \cite{Yang2019paws-x}, UD-POS \cite{nivre2018universal}, Wikiann NER \cite{Pan2017}, XQuAD \cite{artetxe2020cross}, MLQA \cite{Lewis2020mlqa}, TyDiQA-GoldP \cite{Clark2020tydiqa}, BUCC 2018 \cite{zweigenbaum2018overview}, Tatoeba \cite{Artetxe2019massively} tasks. We provide their BibTex information as follows.

```

1 @inproceedings{Conneau2018xnli,
2   title = "{XNLI}: Evaluating Cross-lingual Sentence Representations"
3   ,
4   author = "Conneau, Alexis and
5           Rinott, Ruty and
6           Lample, Guillaume and
7           Williams, Adina and
8           Bowman, Samuel and
9           Schwenk, Holger and
10          Stoyanov, Veselin",
11   booktitle = "Proceedings of EMNLP 2018",
12   year = "2018",
13   pages = "2475--2485",
14 }
15 @inproceedings{Yang2019paws-x,
16   title = "{PAWS-X}: A Cross-lingual Adversarial Dataset for
17           Paraphrase Identification",
18   author = "Yang, Yinfei and
19           Zhang, Yuan and
20           Tar, Chris and
21           Baldridge, Jason",
22   booktitle = "Proceedings of EMNLP 2019",
23   year = "2019",
24   pages = "3685--3690",

```

```

24 }
25
26 @article{nivre2018universal,
27   title={Universal Dependencies 2.2},
28   author={Nivre, Joakim and Abrams, Mitchell and Agi{\c}, {\v{Z}}eljko
           and Ahrenberg, Lars and Antonsen, Lene and Aranzabe, Maria Jesus
           and Arutie, Gashaw and Asahara, Masayuki and Ateyah, Luma and
           Attia, Mohammed and others},
29   year={2018}
30 }
31
32 @inproceedings{Pan2017,
33   author = {Pan, Xiaoman and Zhang, Boliang and May, Jonathan and Nothman
           , Joel and Knight, Kevin and Ji, Heng},
34   booktitle = {Proceedings of ACL 2017},
35   pages = {1946--1958},
36   title = {{Cross-lingual name tagging and linking for 282 languages}},
37   year = {2017}
38 }
39
40 @inproceedings{artetxe2020cross,
41   author = {Artetxe, Mikel and Ruder, Sebastian and Yogatama, Dani},
42   booktitle = {Proceedings of ACL 2020},
43   title = {{On the Cross-lingual Transferability of Monolingual
           Representations}},
44   year = {2020}
45 }
46
47 @inproceedings{Lewis2020mlqa,
48   author = {Lewis, Patrick and ĞOuz, Barlas and Rinott, Ruty and Riedel,
           Sebastian and Schwenk, Holger},
49   booktitle = {Proceedings of ACL 2020},
50   title = {{MLQA: Evaluating Cross-lingual Extractive Question Answering
           }},
51   year = {2020}
52 }
53
54 @inproceedings{Clark2020tydiqa,
55   author = {Jonathan H. Clark and Eunsol Choi and Michael Collins and Dan
           Garrette and Tom Kwiatkowski and Vitaly Nikolaev and Jennimaria
           Palomaki},
56   booktitle = {Transactions of the Association of Computational
           Linguistics},
57   title = {{TyDi QA: A Benchmark for Information-Seeking Question
           Answering in Typologically Diverse Languages}},
58   year = {2020}
59 }
60
61 @inproceedings{zweigenbaum2018overview,
62   title={Overview of the third BUCC shared task: Spotting parallel
           sentences in comparable corpora},

```

```
63   author={Zweigenbaum, Pierre and Sharoff, Serge and Rapp, Reinhard},
64   booktitle={Proceedings of 11th Workshop on Building and Using
        Comparable Corpora},
65   pages={39--42},
66   year={2018}
67 }
68
69 @article{Artetxe2019massively,
70   author = {Artetxe, Mikel and Schwenk, Holger},
71   journal = {Transactions of the ACL 2019},
72   title = {{Massively Multilingual Sentence Embeddings for Zero-Shot
        Cross-Lingual Transfer and Beyond}},
73   year = {2019}
74 }
```