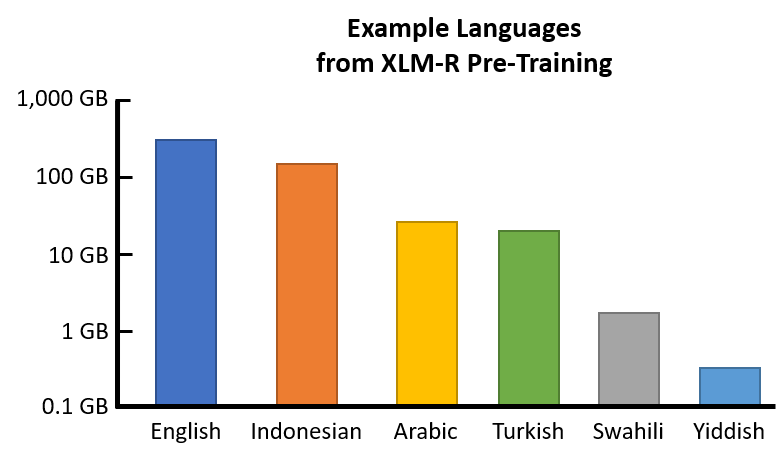
# **XLM-RoBERTa: The alternative for non-english NLP**

If you are doing NLP in a non-english language, you’ll often be agonising over the question “what language model should I use?” While there’s a growing number of monolingual models trained by the community, there’s also an alternative that seems to get less attention: multilingual models.A new model, called XLM-R, that uses self-supervised training techniques to achieve state-of-the-art performance in cross-lingual understanding, a task in which a model is trained in one language and then used with other languages without additional training data. Our model improves upon previous multilingual approaches by incorporating more training data and languages — including so-called low-resource languages, which lack extensive labeled and unlabeled data sets.

**Prerequisite:** Basic understanding ofTransformer model,XLM-100,Bert.

**Why multilingual models?**

XLM-Roberta comes at a time when there is a proliferation of non-English models such as Finnish BERT, French BERT(a.k.a. CamemBERT) and German BERT. There is a real need for cutting edge NLP technologies that work on languages other than English.We also see multilingual models as a great solution for companies who anticipate future expansion in different countries with non native english as their first language. XLM-R represents an important step toward our vision of providing the best possible experience on our platforms for everyone, regardless of what language they speak.

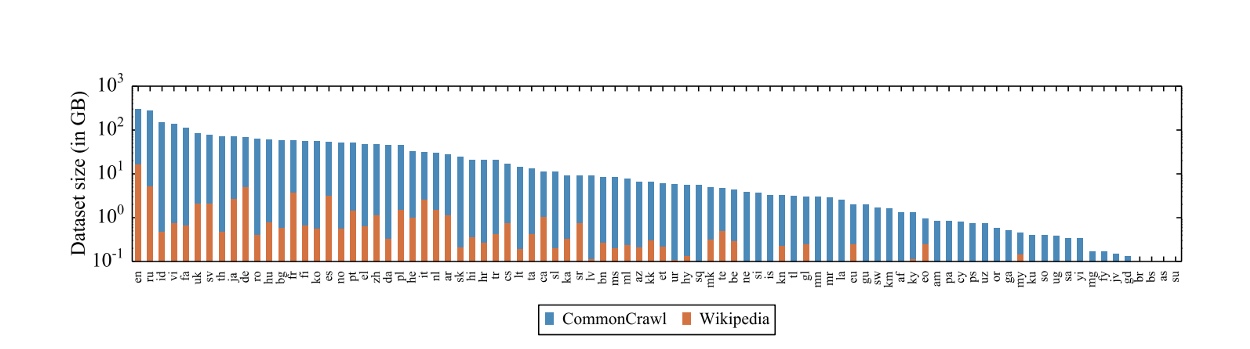
The following bar plot shows, for a small selection of languages, how much text data the authors of XLM-R were able to gather for pre-training.

**What’s New in XLM-Roberta?**

The Facebook AI team released XLM-RoBERTa in November 2019 as an update to their original XLM-100 model. Both are transformer based language models, both rely on the Masked Language Model objective and both are capable of processing text from 100 separate languages. The biggest update that XLM-Roberta offers over the original is a significantly increased amount of training data. The cleaned CommonCrawl data that it is trained on takes up a whopping 2.5tb of storage! It is several orders of magnitude larger than the Wiki-100 corpus that was used to train its predecessor and the scale-up is particularly noticeable in the lower resourced languages. The “RoBERTa” part comes from the fact that its training routine is the same as the monolingual RoBERTa model, specifically, that the sole training objective is the Masked Language Model. There is no Next Sentence Prediction á la BERT or Sentence Order Prediction á la ALBERT.

During fine-tuning, we leveraged the ability of multilingual models to use labeled data in multiple languages in order to improve downstream task performance. This enabled our model to achieve state-of-the-art results on cross-lingual benchmarks while exceeding the per-language performance of monolingual BERT models.

XLM-Roberta now uses the one large shared Sentence Piece model to tokenize instead of having a slew of language specific tokenizers as was the case in XLM-100. Also validation perplexity is no longer used as the stopping criterion during training since the researchers found that downstream performance continues to improve even when perplexity does not.



The increase in size of the CommonCrawl dataset over Wikipedia per language (from XML-RoBERTa paper)

## 

## **XLM-R Vocabulary**

XLM-R has a very different vocabulary than the original BERT, in order to accomodate 100 different languages.

XLM-R has a vocabulary of 250,000 tokens, vs. BERT’s 30,000 tokens.

I’ve published a Notebook [here](https://colab.research.google.com/drive/1M7pDk5bbZh_wB4GMtVjDqVG2l9hCK1Wk) where I’ve poked around XLM-R’s vocabulary to get a sense for what it contains and to gather various statistics.

Here are some highlights:

* It contains an “alphabet” of 13,828 characters.
* It is 62% whole-words and 38% sub-words.
* To count English words, I tried looking up all whole-words in WordNet (a kind of comprehensive English dictionary), and found ~11,400 English words, which is only 5% of XLM-R’s vocabulary.

**APPLICATIONS**

XNLI (Cross-lingual Natural Language Inference)

The XNLI dataset is used. The model is evaluated on cross-lingual transfer from English to other languages. Moreover, it is also tuned on the following machine translation objectives:

translate-test: Dev and test sets are translated to English.

translate-train: The English training set is machine-translated to each language.

translate-train-all: The multilingual model is finetuned on a concatenation of all the training sets from translate-train.

**NER (Named Entity Recognition)**

The CoNLL-2002 and CoNLL-2003 datasets are used for English, Dutch, Spanish and German languages. The model is finetuned in the following ways:

Trained on the English set to evaluate cross-lingual transfer.

On each set to evaluate per-language performance.

On all sets to evaluate multi-lingual learning.

Cross-lingual Question Answering

The MLQ A benchmark is used, which is an extension to the standard SQuAD benchmark in Spanish, German, Arabic, Hindi, Vietnamese and Chinese languages.

**GLUE Benchmark**

And finally, since XLM-RoBERTa is a Language Model, it is evaluated on the standard GLUE Benchmark.

## **An Overview of Attention-based Fusion**

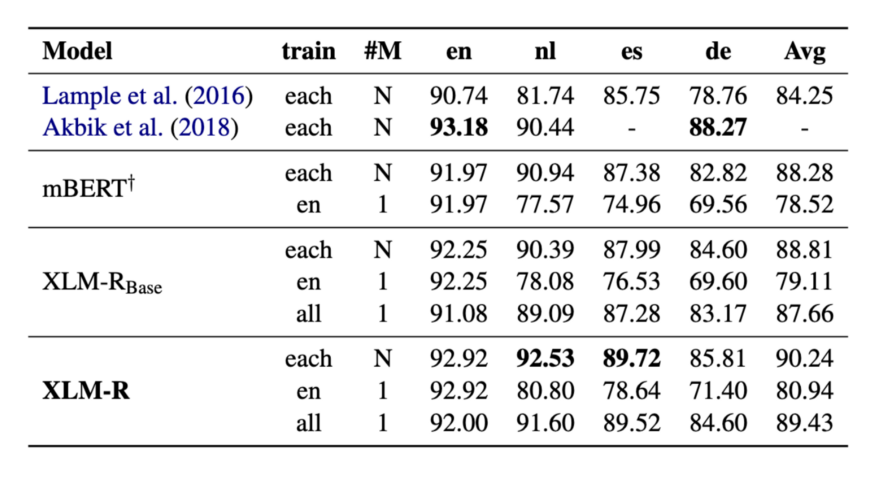
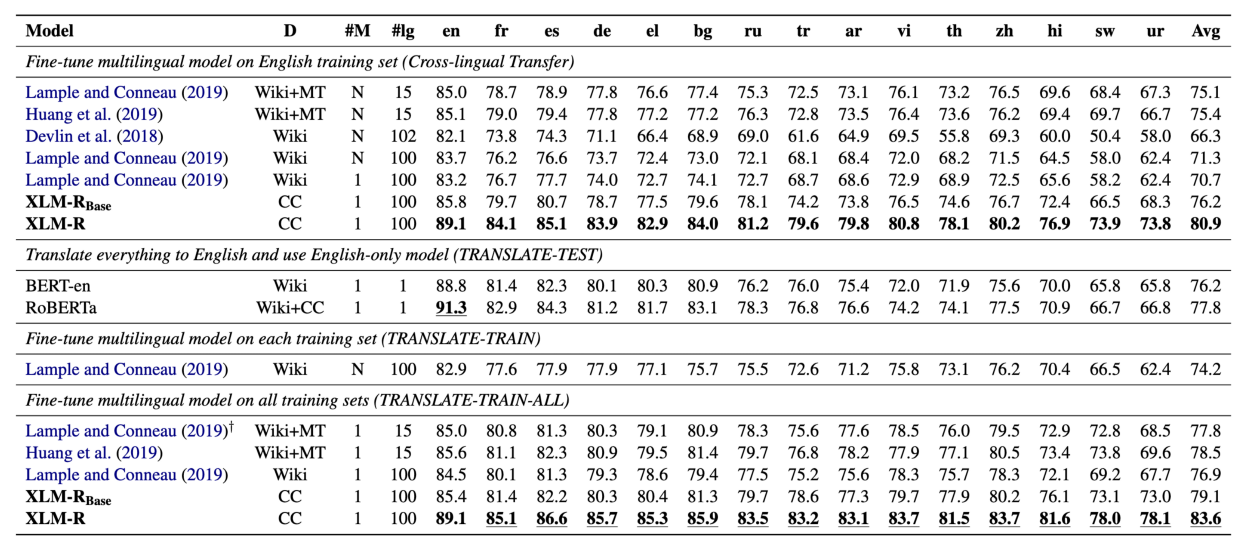
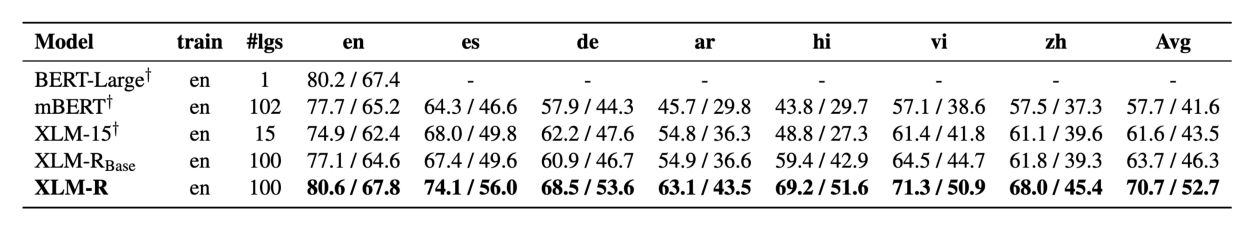
* We connected the XLM-R with the Transformer based NMT system using the attention-based fusion following the work
* We call it XLM-R-fused NMT system which consists of an additional XLM-R module other than the standard NMT-encoder and the NMT-decoder.
* An input sentence is passed to both the XLM-R and the NMT-encoder, which gives two different representations for the sentence. The contextualised word representation from the XLM-R is fused with the NMT-encoder representation using an attention-based fusion. Similarly, the XLM-R representation is fused with the decoder.

**Advantages:**

* Improve cross-lingual language understanding (XLU), by carefully studying the effects of training unsupervised cross-lingual representations at a very large scale.
* We present *XLM-R*, a transformer-based multilingual masked language model pre-trained on text in 100 languages, which obtains state-of-the-art performance on cross-lingual classification, sequence labeling and question answering.
* The benefits of scaling language model pretraining by increasing the size of the model as well as the training data has been extensively studied in the literature.

**Comparison:**

In addition to the comparison of XLM-R and RoBERTa, we provide the first comprehensive study to assess this claim on the XNLI benchmark. We extend our comparison between multilingual XLM models and monolingual BERT models on 7 languages and compare performance in Table 5. We train 14 monolingual BERT models on Wikipedia and CommonCrawl (capped at 60 GiB), and two XLM-7 models. We increase the vocabulary size of the multilingual model for a better comparison. We found that multilingual models can outperform their monolingual BERT counterparts. However, by making use of multilingual training (translate-trainall) and leveraging training sets coming from multiple languages, XLM-7 can outperform the BERT models: our XLM-7 trained on CC obtains 80.0% average accuracy on the 7 languages, while the average performance of BERT models trained on CC is 77.5%.



**Results**

In the end we evaluated XLM-RoBERTa on one classification and two NER tasks where it showed very impressive performance. XLM-RoBERTa Large is on par with the best submission of GermEval18 (Classification). On GermEval14 (NER) the model outperforms Flair by 2.35% F1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | XLM-RoBERTA Large | German BERT | Multilingual BERT Cased | Previous Best | From |
| Gern Eval 18(Coarse) | 77.3 | 74.7 | 71.0 | 76.8 | TU Wein |
| GermEval14 | 87.0 | 84.0 | 83.4 | 84.7 | Flair |

Results from evaluation. Here is the leaderboard of GermEval18 and here are the reported scores from Flair

These results were produced without extensive hyperparameter tuning and we expect that they could improve with more tweaking of learning rates and batch sizes. Also, for the NER tasks, we believe there are gains to be made by adding a CRF layer on top of XLM-RoBERTa.

XLM-R sets a new state of the art on XNLI. On cross-lingual transfer, XLM-R obtains 80.9% accuracy, outperforming the XLM-100 and mBERT open-source models by 10.2% and 14.6% average accuracy. On the Swahili and Urdu lowresource languages, XLM-R outperforms XLM-100 by 15.7% and 11.4%, and mBERT by 23.5% and 15.8%. While XLM-R handles 100 languages, we also show that it outperforms the former state of the art Unicoder (Huang et al., 2019) and XLM (MLM+TLM), which handle only 15 languages, by 5.5% and 5.8% average accuracy respectively. Using the multilingual training of translate-train-all, XLM-R further improves performance and reaches 83.6% accuracy, a new overall state of the art for XNLI, outperforming Unicoder by 5.1%.

**Conclusion**

The strength of these results show that multilingual models exhibit great performance even when evaluated on a single language and suggested that German NLP practitioners at least consider one of the XLM-Roberta variants when choosing a language model for their NLP systems. The importance of breaking the English-centric focus of NLP research is something that has already been [extensively covered](https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/) and we believe that research in non-English languages will only increase.

**REFERENCES**

[**XLM-R: State-of-the-art cross-lingual understanding through self-supervision (facebook.com)**](https://ai.facebook.com/blog/-xlm-r-state-of-the-art-cross-lingual-understanding-through-self-supervision/)

[**XLM-RoBERTa: The alternative for non-english NLP | by Branden Chan | deepset-ai | Medium**](https://medium.com/deepset-ai/xlm-roberta-the-multilingual-alternative-for-non-english-nlp-cf0b889ccbbf)

**https://github.com/ashwanitanwar/nmt-transfer-learning-xlm-r**