mT5: A massively multilingual pre-trained text-to-text transformer

Linting Xue[*] Noah Constant[*] Adam Roberts[*]
Mihir Kale Rami Al-Rfou Aditya Siddhant Aditya Barua
Colin Raffel

Google Research

Abstract

The recent "Text-to-Text Transfer Transformer" (T5) leveraged a unified text-to-text format and scale to attain state-of-the-art results on a wide variety of English-language NLP tasks. In this paper, we introduce mT5, a multilingual variant of T5 that was pre-trained on a new Common Crawl-based dataset covering 101 languages. We describe the design and modified training of mT5 and demonstrate its state-of-the-art performance on many multilingual benchmarks. All of the code and model checkpoints used in this work are publicly available.[1]

1 Introduction

Current natural language processing (NLP) pipelines often make use of transfer learning, where a model is pre-trained on a data-rich task before being fine-tuned on a downstream task of interest (Ruder et al., 2019). The success of this paradigm is partially thanks to the release of parameter checkpoints for pre-trained models. These checkpoints allow members of the NLP community to quickly attain strong performance on many tasks without needing to perform expensive pre-training themselves. As one example, the pre-trained checkpoints for the "Text-to-Text Transfer Transformer" (T5) model released by Raffel et al. (2019) have been used to achieve state-of-the-art results on many benchmarks (Khashabi et al., 2020; Roberts et al., 2020; Kale, 2020; Izacard and Grave, 2020; Nogueira et al., 2020; Narang et al., 2020, etc.).

Unfortunately, many of these language models were pre-trained solely on English-language text. This significantly limits their use given that roughly 80% of the world population does not speak English (Crystal, 2008). One way the community has addressed this English-centricity has been to release dozens of models that have instead been pre-trained on a single non-English language (Carmo et al., 2020; de Vries et al., 2019; Le et al., 2019; Martin et al., 2019; Delobelle et al., 2020; Malmsten et al., 2020; Nguyen and Nguyen, 2020; Polignano et al., 2019, etc.). A more general solution is to produce multilingual models that have been pre-trained on a mixture of many languages. Popular models of this type are mBERT (Devlin, 2018), mBART (Liu et al., 2020), and XLM-R (Conneau et al., 2019), which are multilingual variants of BERT (Devlin et al., 2018), BART (Lewis et al., 2019a), and RoBERTa (Liu et al., 2019), respectively.

In this paper, we continue this tradition by releasing mT5, a multilingual variant of T5. Our goal with mT5 is to produce a massively multilingual model that deviates as little as possible from the recipe used to create T5. As such, mT5 inherits all of the benefits of T5 (described in section 2), such as its general-purpose text-to-text format, its design based on insights from a large-scale empirical study, and its scale. To train mT5, we introduce a multilingual variant of the C4 dataset called mC4. mC4 comprises natural text in 101 languages drawn from the public Common Crawl web scrape. To validate the performance of mT5, we include results on several benchmark datasets, showing state-of-the-art performance in many cases. We release our pre-trained models and code so that the community can leverage our work.

[*] Equal Contribution. Please direct correspondence to lintingx@google.com, nconstant@google.com, adarob@google.com, and craffel@google.com

[1] http://goo.gle/mt5-code

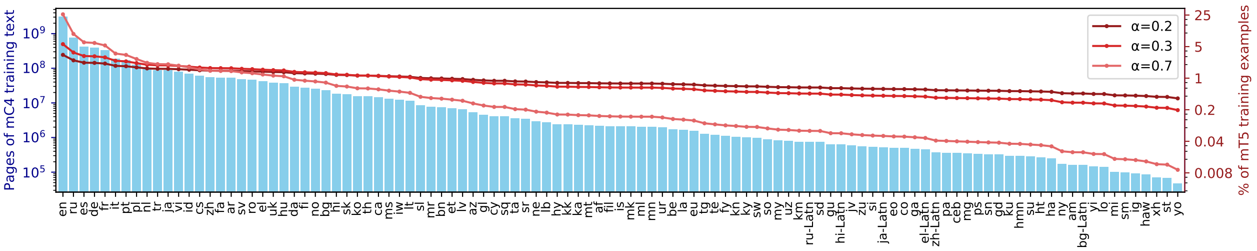

Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents α (right axis). Our final model uses α=0.3.

2 Background on T5 and C4

In this section, we provide a short overview of T5 and the C4 pre-training dataset. Further details are available in Raffel et al. (2019).

T5 is a pre-trained language model whose primary distinction is its use of a unified "text-to-text" format for all text-based NLP problems. This approach is natural for generative tasks (such as machine translation or abstractive summarization) where the task format requires the model to generate text conditioned on some input. It is more unusual for classification tasks, where T5 is trained to output the literal text of the label (e.g. "positive" or "negative" for sentiment analysis) instead of a class index. The primary advantage of this approach is that it allows the use of exactly the same training objective (teacher-forced maximum-likelihood) for every task, which in practice means that a single set of hyperparameters can be used for effective fine-tuning on any downstream task. Similar unifying frameworks were proposed by Keskar et al. (2019) and McCann et al. (2018). Given the sequence-to-sequence structure of this task format, T5 uses a basic encoder-decoder Transformer architecture as originally proposed by Vaswani et al. (2017). T5 is pre-trained on a masked language modeling "span-corruption" objective, where consecutive spans of input tokens are replaced with a mask token and the model is trained to reconstruct the masked-out tokens.

An additional distinguishing factor of T5 is its scale, with pre-trained model sizes available from 60 million to 11 billion parameters. These models were pre-trained on around 1 trillion tokens of data. Unlabeled data comes from the C4 dataset, which is a collection of about 750GB of English-language text sourced from the public Common Crawl web scrape. C4 includes heuristics to extract only natural language (as opposed to boilerplate and other gibberish) in addition to extensive deduplication. The pre-training objective, model architecture, scaling strategy, and many other design choices for T5 were chosen based on a large-scale empirical study described in detail in Raffel et al. (2019).

3 mC4 and mT5

Our goal in this paper is to create a massively multilingual model that follows T5's recipe as closely as possible. Towards this end, we develop an extended version of the C4 pre-training dataset that covers 101 languages and integrates changes into T5 to better suit this multilinguality.

3.1 mC4

The C4 dataset was explicitly designed to be English only: any page that was not given a probability of at least 99% of being English by `langdetect`[2] was discarded. In contrast, for mC4 we use `cld3`[3] to identify over 100 languages. Since some of these languages are relatively scarce on the internet, we make use of all of the 71 monthly web scrapes released so far by Common Crawl. This is dramatically more source data than was used for C4, for which the April 2019 web scrape alone was enough to provide plenty of English-language data.

An important heuristic filtering step in C4 was the removal of lines that did not end in an English terminal punctuation mark. As this is inappropriate for many languages, we instead apply a "line length filter" that requires pages to contain at least three lines of text with 200 or more characters. Otherwise, we follow C4's filtering by deduplicating lines across documents and filtering pages containing bad words.[4] Finally, we detect each page's primary language using `cld3` and remove pages where the confidence is below 70%.

After these filters are applied, we group the remaining pages by language and include in the corpus

[2] https://pypi.org/project/langdetect/
[3] https://github.com/google/cld3
[4] https://github.com/LDNOOBW/

Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	110M	104	Wikipedia
XLM (Lample and Conneau, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2019)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2019a)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

Table 1: Comparison of mT5 to existing massively multilingual pre-trained language models. Multiple versions of XLM and mBERT have been released; we refer here to the ones that cover the most languages. Note that XLM-R counts five Romanized variants as separate languages, while we ignore six Romanized variants in the mT5 language count.

all languages with 10,000 or more pages. This produces text in 107 "languages" as defined by `cld3`. However, we note that six of these are just script variants of the same spoken language (e.g. `ru` is Russian in Cyrillic script and `ru-Latn` is Russian in Latin script). A histogram of the page counts for each language is shown in fig. 1. Detailed dataset statistics including per-language token counts are shown in table 5 (appendix).

3.2 mT5

The model architecture and training procedure that we used for mT5 closely follows that of T5. Specifically, we based mT5 on the "T5.1.1" recipe,[5] which improves upon T5 by using GeGLU nonlinearities (Shazeer, 2020), scaling $\mathrm{d}_{\mathrm{model}}$ instead of d_{ff} in the larger models, and pre-training on unlabeled data only with no dropout. For brevity, we refer to Raffel et al. (2019) for further details on T5.

A major factor in pre-training multilingual models is how to sample data from each language. Ultimately, this choice is a zero-sum game: If low-resource languages are sampled too often, the model may overfit; if high-resource languages are not trained on enough, the model will underfit. We therefore take the approach used in (Devlin, 2018; Conneau et al., 2019; Arivazhagan et al., 2019) and boost lower-resource languages by sampling examples according to the probability $p(L) \propto |L|^\alpha$, where $p(L)$ is the probability of sampling text from a given language during pre-training and $|L|$ is the number of examples in the language. The hyperparameter α (typically with $\alpha < 1$) allows us to control how much to "boost" the probability of training on low-resource languages. Values used by prior work include $\alpha = 0.7$ for mBERT (Devlin, 2018), $\alpha = 0.3$ for XLM-R (Conneau et al., 2019), and $\alpha = 0.2$ for MMNMT (Arivazhagan et al., 2019). We tried all three of these values and found $\alpha = 0.3$ to give a reasonable compromise between performance on high- and low-resource languages.

The fact that our model covers over 100 languages necessitates a larger vocabulary. Following XLM-R (Conneau et al., 2018), we increase the vocabulary size to 250,000 wordpieces. As in T5, we use SentencePiece (Kudo and Richardson, 2018; Kudo, 2018) wordpiece models that are trained with the same language sampling rates used during training. To accommodate languages with large character sets like Chinese, we use a character coverage of 0.99999, but also enable SentencePiece's "byte-fallback" feature to ensure that any string can be uniquely encoded.

3.3 Comparison to related models

To contextualize our new model, we provide a brief comparison with existing massively multilingual pre-trained language models. For brevity, we focus on models that support more than a few dozen languages. Table 1 gives a high-level comparison of mT5 to the most similar models.

mBERT (Devlin, 2018) is a multilingual version of BERT (Devlin et al., 2018). Similar to our approach with mT5, mBERT follows the BERT recipe as closely as possible (same architecture, objective, etc.). The primary difference is the training set: Instead of training on English Wikipedia and the Toronto Books Corpus, mBERT is trained on up to 104 languages from Wikipedia. **XLM** (Lample and Conneau, 2019) is also based on BERT, but includes improved methods for pre-training multilingual language models. Many pre-trained versions of XLM have been released; the most massively-multilingual variant was trained on 100 languages from Wikipedia. **XLM-R** (Conneau et al., 2019) is an improved version of XLM that is based on the RoBERTa model

[5] `https://github.com/google-research/text-to-text-transfer-transformer/blob/master/released_checkpoints.md#t511`

Model	Sentence pair		Structured	Question answering		
	XNLI	PAWS-X	WikiAnn NER	XQuAD	MLQA	TyDiQA-GoldP
Metrics	Acc.	Acc.	F1	F1 / EM	F1 / EM	F1 / EM
Cross-lingual zero-shot transfer (models fine-tuned on English data only)						
mBERT	65.4	81.9	62.2	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9
XLM	69.1	80.9	61.2	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1
InfoXLM	81.4	-	-	- / -	73.6 / 55.2	- / -
X-STILTs	80.4	87.7	64.7	77.2 / 61.3	72.3 / 53.5	76.0 / 59.5
XLM-R	79.2	86.4	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0
mT5-Small	67.5	82.4	51.0	58.1 / 42.5	54.6 / 37.1	34.9 / 23.9
mT5-Base	75.4	87.4	56.6	67.0 / 49.0	64.6 / 45.0	58.1 / 42.8
mT5-Large	81.1	89.6	58.8	77.8 / 61.5	71.2 / 51.7	57.8 / 41.1
mT5-XL	82.9	**90.2**	65.7	79.5 / 63.6	73.5 / 54.5	77.3 / 61.5
mT5-XXL (75% trained)	**84.8**	89.2	**69.2**	**81.9 / 65.7**	**75.5 / 56.9**	**80.8 / 66.3**
Translate-train (models fine-tuned on English data plus translations in all target languages)						
XLM-R	82.6	90.4	-	80.2 / 65.9	72.8 / 54.3	66.5 / 47.7
Filter + Self-Teaching	83.9	91.4	-	82.4 / 68.0	76.2 / 57.7	68.3 / 50.9
mT5-Small	64.7	87.8	-	64.3 / 49.5	56.6 / 38.8	48.2 / 34.0
mT5-Base	75.9	90.2	-	75.3 / 59.7	67.6 / 48.5	64.0 / 47.7
mT5-Large	81.8	91.3	-	81.2 / 65.9	73.9 / 55.2	71.1 / 54.9
mT5-XL	84.8	91.3	-	82.7 / 68.1	75.1 / 56.6	79.9 / 65.3
mT5-XXL (75% trained)	**87.2**	**92.0**	-	**85.0 / 70.8**	**76.6 / 58.2**	**82.0 / 67.8**

Table 2: Results on XTREME sentence-pair classification, structured prediction and question answering tasks. Apart from mT5 (ours), all metrics are from Fang et al. (2020), though Conneau et al. (2019) report better performance of XLM-R on XNLI (80.9). For the "translate-train" setting, we include English training data, so as to be comparable with Fang et al. (2020). This differs from XTREME "translate-train" setup of Hu et al. (2020). Full results for all languages in all tasks are provided in tables 6 to 11 (appendix).

(Liu et al., 2019). XLM-R is trained with a cross-lingual masked language modeling objective on data in 100 languages from Common Crawl. To improve the quality of the pre-training data, an n-gram language model is first trained on Wikipedia and a page from Common Crawl is only retained if it is assigned a high likelihood by the n-gram model (Wenzek et al., 2019). **mBART** (Liu et al., 2020) is a multilingual encoder-decoder model that is based on BART (Lewis et al., 2019a). mBART is trained with a combination of span masking and sentence shuffling objectives on a subset of 25 languages from the same data as XLM-R. **MARGE** (Lewis et al., 2020) is a multilingual encoder-decoder model that is trained to reconstruct a document in one language by retrieving documents in other languages. It uses data in 26 languages from Wikipedia and CC-News (Liu et al., 2019).

4 Experiments

To validate the performance of mT5, we evaluate our models on 6 tasks from the XTREME multilingual benchmark (Hu et al., 2020): the XNLI (Conneau et al., 2018) entailment task covering 14 languages; the XQuAD (Artetxe et al., 2019), MLQA (Lewis et al., 2019b), and TyDi QA (Clark et al., 2020) reading comprehension benchmarks with 10, 7, and 11 languages respectively; the Named Entity Recognition (NER) dataset of WikiAnn (Pan et al., 2017) with the 40 languages from XTREME (Hu et al., 2020); and the PAWS-X (Yang et al., 2019) paraphrase identification dataset with 7 languages. We cast all tasks into the text-to-text format, i.e. generating the label text (XNLI and PAWS-X), entity tags and labels (WikiAnn NER), or answer (XQuAD, MLQA, and TyDi QA) directly. For NER, if there are multiple entities then they are concatenated in the order they appear, and if there are no entities then the target text is 'None'. We consider variants of these tasks where the model is fine-tuned only on English data ("zero-shot") or on data that has been machine-translated from English into each target language ("translate-train"). For brevity, we refer to Hu et al. (2020) for further details on these benchmarks.

Following the original T5 recipe, we consider five model sizes: *Small* (≈ 300M parameters), *Base* (600M), *Large* (1B), *XL* (4B), and *XXL* (13B). The increase in parameter counts compared to the corresponding T5 model variants comes from the larger vocabulary used in mT5. We pre-train our models for 1 million steps on batches of 1024 length-1024 input sequences, corresponding to roughly 1 trillion input tokens total. This is the same amount of pre-training as T5 and about $\frac{1}{6}$ as much as XLM-R. Due to time constraints, we report results with mt5-XXL trained for only 750 thousand steps. Final results and further experiments will be updated on our public codebase.[1]

We use the same inverse square-root learning rate schedule used by T5 during pre-training, with the learning rate set to $1/\sqrt{\max(n,k)}$ where n is the current training iteration and $k = 10^4$ is the number of warm-up steps. Following the T5.1.1 recipe, we do not apply dropout during pre-training. We use the same self-supervised objective as T5, with 15% of tokens masked and an average noise span length of 3. We ablate some of these experimental details in section 4.2.

For fine-tuning, we use a constant learning rate of 0.001 and dropout rate of 0.1 for all tasks. In the zero-shot setting, we use a batch size of 2^{20} for XNLI and 2^{16} for PAWS-X, NER, XQuAD, MLQA, and TyDi QA. For early stopping, we save checkpoints every 200 steps and choose the checkpoint with the highest performance on the validation set.

4.1 Results

Table 2 presents our main results, with per-language breakdowns for each task given in tables 6 to 11 (appendix). Our largest model mT5-XXL reaches state-of-the-art on all of the tasks we consider. Note that unlike our model, InfoXLM (Chi et al., 2020) benefits from parallel training data, while X-STILTs (Phang et al., 2020) leverages labeled data from tasks similar to the target task. Overall, our results highlight the importance of model capacity in cross-lingual representation learning and suggest that scaling up a simple pre-training recipe can be a viable alternative to more complex techniques relying on LM filtering, parallel data, or intermediate tasks.

In the "translate-train" setting, we also match or exceed state-of-the-art on all XTREME classification and QA tasks. For these tasks, we fine-tune on the combination of the labeled English data and machine translations thereof.[6] This allows direct comparison with both FILTER (Fang et al., 2020) as well as the XLM-R baseline of Fang et al. (2020). Note however that this setup differs from the XTREME "translate-train" (Hu et al., 2020), which excludes the English data.

	T5	mT5
Small	87.24/ 79.10	84.74 / 76.39
Base	92.08 / 85.44	89.55 / 83.83
Large	93.79 / 86.66	93.04 / 87.00
XL	94.95 / 88.53	94.48 / 88.88
XXL	96.22 / 91.26	95.41 / 90.17

Table 3: Comparison of T5 vs. mT5 on SQuAD (F1/EM). Note, mT5-XXL is only 75% trained.

Massively multilingual models have been observed to underperform on a given language when compared to a similarly-sized "dedicated" model trained specifically for that language (Arivazhagan et al., 2019). To quantify this effect, we compare the performance of mT5 and T5 when fine-tuned on the SQuAD reading comprehension benchmark (Rajpurkar et al., 2016). The results are shown in table 3, with results for T5 reproduced from Raffel et al. (2019). While the *Small* and *Base* mT5 models fall short of their English T5 counterparts, we find that the larger models close the gap. This suggests there may be a turning point past which the model has enough capacity to effectively learn 101 languages without significant interference effects.

4.2 Ablation

We run six ablations, modifying various settings, using our *Large* model as a baseline: (i) increase dropout to 0.1 in hopes of mitigating overfitting to low-resource languages, (ii) decrease sequence length to 512 as was used in T5, (iii) increase the average noise span length in the pre-training objective to 10 since we observe fewer characters per token than T5, (iv) adjust language sampling exponent α to {0.2, 0.7} as used in MMNMT (Arivazhagan et al., 2019) and mBERT (Devlin, 2018), respectively, (v) turn off "line length filter" in the mC4 data pipeline, and (vi) supplement mC4 with Wikipedia data[7] from 103 languages.

The effect of these ablations on XNLI zero-shot accuracy is shown in table 4. In each case, the average XNLI score is lower than the mT5-Large baseline, justifying our chosen settings. The line length filter provides a +2 point boost, corroborating the findings of Conneau et al. (2019) and Raffel et al. (2019) that filtering low-quality pages from Common Crawl is valuable. Increasing the language sampling exponent α to 0.7 has the expected effect of improving performance in high-resource languages (e.g. Russian

[6] For PAWS-X, we use the translated data from the original dataset, for better comparability with FILTER. For other tasks, we use the translation data provided by Hu et al. (2020).

[7] We use the 2020 Wikipedia data from TensorFlow Datasets, selecting the same languages as mBERT. https://www.tensorflow.org/datasets/catalog/wikipedia

Model	Accuracy
Baseline (mT5-Large)	**81.1**
Dropout 0.1	77.6
Sequence length 512	80.5
Span length 10	78.6
$\alpha = 0.7$	80.7
$\alpha = 0.2$	80.7
No line length filter	79.1
Add Wikipedia data	80.3

Table 4: Average XNLI zero-shot accuracy of various ablations on our mT5-Large model. Per-language metrics are shown in table 12 (appendix).

81.5 → 82.8) while hurting low-resource languages (e.g. Swahili 75.4 → 70.6), with the average effect being negative. Conversely, lowering α to 0.2 boosts one tail language slightly (Urdu 73.5 → 73.9), but is harmful elsewhere. Detailed per-language metrics on XNLI, as well as the performance of our ablations on zero-shot XQuAD, are provided in table 12 and table 13 (appendix) respectively, showing largely the same trends.

5 Conclusion

In this paper, we introduced mT5 and mC4: massively multilingual variants of the T5 model and C4 dataset. We demonstrated that the T5 recipe is straightforwardly applicable to the multilingual setting, and achieve strong performance on a diverse set of benchmarks. We release all of the code and pretrained datasets used in this paper to facilitate future work on multilingual language understanding.[8]

Acknowledgements

We thank Melvin Johnson for tips on the translate-train procedure for XTREME and Itai Rolnick for help with infrastructure.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

David Crystal. 2008. Two thousand million? *English today*, 24(1):3–6.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Jacob Devlin. 2018. Multilingual BERT README. `https://github.com/google-research/bert/blob/master/multilingual.md`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2020. FILTER: An enhanced fusion method for cross-lingual language understanding. *arXiv preprint arXiv:2009.05166*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

[8] `https://goo.gle/mt5-code`

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Mihir Kale. 2020. Text-to-text pre-training for data-to-text tasks. *arXiv preprint arXiv:2005.10433*.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying question answering and text classification via span extraction. *arXiv preprint arXiv:1904.09286*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. *arXiv preprint arXiv:2006.15020*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019b. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden–making a swedish bert. *arXiv preprint arXiv:2007.01658*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Jason Phang, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, Iacer Calixto, and Samuel R Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. *arXiv preprint arXiv:2005.13013*.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CLiC-it*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Sebastian Ruder, Matthew Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. Tutorial at the North American Association for Computational Linguistics-Human Language Technologies Conference.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

ISO **Code**	**Language**	**Tokens** (B)	**Pages** (M)	**mT5** (%)
en	English	2,733	3,067	5.67
ru	Russian	713	756	3.71
es	Spanish	433	416	3.09
de	German	347	397	3.05
fr	French	318	333	2.89
it	Italian	162	186	2.43
pt	Portuguese	146	169	2.36
pl	Polish	130	126	2.15
nl	Dutch	73	96	1.98
tr	Turkish	71	88	1.93
ja	Japanese	164	87	1.92
vi	Vietnamese	116	79	1.87
id	Indonesian	69	70	1.80
cs	Czech	63	60	1.72
zh	Chinese	39	55	1.67
fa	Persian	52	54	1.67
ar	Arabic	57	53	1.66
sv	Swedish	45	49	1.61
ro	Romanian	52	46	1.58
el	Greek	43	42	1.54
uk	Ukrainian	41	39	1.51
hu	Hungarian	39	37	1.48
da	Danish	29	29	1.38
fi	Finnish	25	27	1.35
no	Norwegian	27	25	1.33
bg	Bulgarian	22	23	1.29
hi	Hindi	24	19	1.21
sk	Slovak	18	18	1.19
ko	Korean	26	16	1.14
th	Thai	11	15	1.14
ca	Catalan	13	14	1.12
ms	Malay	13	13	1.09
iw	Hebrew	17	12	1.06
lt	Lithuanian	11	11	1.04
sl	Slovenian	8.8	8.5	0.95
mr	Marathi	14	7.8	0.93
bn	Bengali	7.3	7.4	0.91
et	Estonian	6.9	6.9	0.89
lv	Latvian	7.0	6.4	0.87
az	Azerbaijani	4.4	5.3	0.82
gl	Galician	2.4	4.6	0.79
cy	Welsh	4.9	4.1	0.76
sq	Albanian	4.0	4.1	0.76
ta	Tamil	3.4	3.5	0.73
sr	Serbian	4.3	3.4	0.72
ne	Nepali	3.2	2.9	0.69
lb	Luxembourgish	1.0	2.7	0.68
hy	Armenian	2.4	2.4	0.65
kk	Kazakh	3.1	2.4	0.65
ka	Georgian	2.5	2.3	0.64
mt	Maltese	5.2	2.3	0.64
af	Afrikaans	1.7	2.2	0.63
fil	Filipino	2.1	2.1	0.62
is	Icelandic	2.6	2.1	0.62
mk	Macedonian	1.8	2.1	0.62
ml	Malayalam	1.8	2.1	0.62
mn	Mongolian	2.7	2.1	0.62
ur	Urdu	2.4	1.9	0.61
be	Belarusian	2.0	1.7	0.59
la	Latin	1.3	1.7	0.58
eu	Basque	1.4	1.6	0.57
tg	Tajik	1.4	1.3	0.54
te	Telugu	1.3	1.2	0.52
fy	West Frisian	0.4	1.1	0.51
kn	Kannada	1.1	1.1	0.51
ky	Kyrgyz	1.0	1.0	0.50
sw	Swahili	1.0	1.0	0.50
so	Somali	1.4	0.9	0.48
my	Burmese	0.9	0.8	0.47
uz	Uzbek	0.9	0.8	0.46
km	Khmer	0.6	0.8	0.46
-	Russian (Latin)	0.9	0.7	0.46
sd	Sindhi	1.6	0.7	0.45
gu	Gujarati	0.8	0.6	0.43
-	Hindi (Latin)	0.6	0.6	0.43
jv	Javanese	0.3	0.6	0.42
zu	Zulu	0.2	0.6	0.42
si	Sinhala	0.8	0.5	0.41
-	Japanese (Latin)	0.3	0.5	0.41
eo	Esperanto	0.7	0.5	0.40
co	Corsican	0.2	0.5	0.40
ga	Irish	0.5	0.5	0.40
-	Greek (Latin)	0.4	0.4	0.39
-	Chinese (Latin)	0.2	0.4	0.37
pa	Punjabi	0.6	0.4	0.37
ceb	Cebuano	0.2	0.4	0.36
mg	Malagasy	0.2	0.3	0.36
ps	Pashto	0.4	0.3	0.36
sn	Shona	0.2	0.3	0.35
gd	Scottish Gaelic	0.4	0.3	0.35
ku	Kurdish	0.4	0.3	0.34
hmn	Hmong	0.2	0.3	0.34
su	Sundanese	0.1	0.3	0.34
ht	Haitian Creole	0.2	0.3	0.33
ha	Hausa	0.2	0.2	0.33
ny	Chichewa	0.1	0.2	0.29
am	Amharic	0.3	0.2	0.29
-	Bulgarian (Latin)	0.09	0.2	0.29
yi	Yiddish	0.3	0.1	0.28
lo	Lao	0.1	0.1	0.28
mi	Maori	0.1	0.1	0.25
sm	Samoan	0.09	0.1	0.25
ig	Igbo	0.09	0.09	0.24
haw	Hawaiian	0.09	0.08	0.24
xh	Xhosa	0.06	0.07	0.22
st	Sotho	0.08	0.07	0.22
yo	Yoruba	0.05	0.05	0.20

Table 5: Statistics of the mC4 corpus, totaling 6.6B pages and 6.3T tokens. The "mT5" column indicates the percentage of mT5 training data coming from a given language, using the default exponential smoothing value of α=0.3. We list 107 "languages" as detected by `cld3`, but note six of these (marked "Latin") are just Romanized variants of existing languages.

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
Cross-lingual zero-shot transfer (models fine-tune on English data only)																
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
XLM-R	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
mT5-Small	79.6	65.2	71.3	69.2	68.6	72.7	70.7	62.5	70.1	59.7	66.3	64.4	59.9	66.3	65.8	67.5
mT5-Base	84.7	73.3	78.6	77.4	77.1	80.3	79.1	70.8	77.1	69.4	73.2	72.8	68.3	74.2	74.1	75.4
mT5-Large	89.4	79.8	84.1	83.4	83.2	84.2	84.1	77.6	81.5	75.4	79.4	80.1	73.5	81.0	80.3	81.1
mT5-XL	90.6	82.2	85.4	85.8	85.4	81.3	85.3	80.4	83.7	78.6	80.9	82.0	77.0	81.8	**82.7**	82.9
mT5-XXL (75%)	**92.3**	**84.4**	**87.5**	**87.3**	**87.0**	**88.3**	**87.3**	**82.5**	**84.3**	**80.4**	**81.7**	**83.3**	**79.2**	**84.2**	**82.7**	**84.8**
Translate-train (models fine-tune on English training data plus translations in all target languages)																
Filter + Self-Teaching	89.5	83.6	86.4	85.6	85.4	86.6	85.7	81.1	83.7	78.7	81.7	83.2	79.1	83.9	83.8	83.9
mt5-Small	69.5	63.7	67.5	65.7	66.4	67.5	67.3	61.9	66.4	59.6	63.9	63.5	60.4	63.3	64.5	64.7
mt5-Base	82.0	74.4	78.5	77.7	78.1	79.1	77.9	72.2	76.5	71.5	75.0	74.8	70.4	74.5	76.0	75.9
mt5-Large	88.3	80.3	84.1	84.0	83.7	84.9	83.8	79.8	82.0	76.4	79.9	81.0	75.9	81.3	81.7	81.8
mt5-XL	90.9	84.2	86.8	86.8	86.4	87.4	86.8	83.1	84.9	81.3	82.3	84.4	79.4	83.9	84.0	84.8
mT5-XXL (75%)	**92.6**	**86.4**	**89.4**	**89.0**	**88.7**	**90.0**	**88.7**	**86.1**	**86.8**	**83.3**	**84.9**	**86.3**	**82.6**	**86.8**	**86.8**	**87.2**

Table 6: XNLI accuracy scores for each language.

Model	en	de	es	fr	ja	ko	zh	avg
Cross-lingual zero-shot transfer (models fine-tune on English data only)								
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
XLM-R	94.7	89.7	90.1	90.4	78.7	79.0	82.3	86.4
mT5-Small	92.2	86.2	86.1	86.6	74.7	73.5	77.9	82.4
mT5-Base	95.4	89.4	89.6	91.2	79.8	78.5	81.1	87.4
mT5-Large	**96.1**	91.3	92	**92.7**	82.5	82.7	84.7	89.6
mT5-XL	96.0	**92.8**	**92.7**	92.4	**83.6**	**83.1**	86.5	**90.2**
mT5-XXL (75%)	95.7	91.6	91.6	92.3	83.4	82.9	**86.8**	89.2
Translate-train (models fine-tune on English training data plus translations in all target languages)								
Filter + Self-Teaching	95.9	92.8	93.0	93.7	**87.4**	87.6	89.6	91.4
mT5-Small	94.9	89.2	90.3	90.3	83.7	81.1	85.1	87.8
mT5-Base	95.6	91.7	92.4	93.2	85.3	86.2	86.9	90.2
mT5-Large	96.6	92.7	93.2	93.8	86.9	87.4	88.3	91.3
mT5-XL	96.2	92.8	93.2	94.2	86.4	87	88.8	91.3
mT5-XXL (75%)	**96.4**	**92.9**	**93.8**	**94.7**	87.2	**89**	**89.8**	**92**

Table 7: PAWS-X accuracy scores for each language.

Model	af	ar	bg	bn	de	el	en	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	
Cross-lingual zero-shot transfer (models fine-tune on English data only)																					
mBERT	77.4	41.1	77.0	70.0	78.0	72.5	85.2	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	74.9	44.8	76.7	70.0	78.1	73.5	82.6	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	
XLM-R	78.9	53.0	81.4	78.8	78.8	**79.5**	84.7	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	
mT5-Small	67.4	36.6	64.6	60.4	66.1	59.1	80.7	63.6	58.4	42.3	25.3	64.5	74.6	39.6	57.9	61.5	46.7	73.4	28.8	50.6	
mT5-Base	73.8	48.4	68.2	67.1	72.5	63.5	83.2	71.7	67.3	49.2	31.9	68.6	78.6	47.4	67.6	64.7	49.7	78.9	35.3	56.9	
mT5-Large	74.7	55.0	60.6	64.5	75.2	68.2	84.2	74.2	67.0	48.7	51.4	66.4	82.4	55.8	69.0	67.3	51.1	80.7	43.0	57.1	
mT5-XL	79.8	60.2	81.0	78.1	80.6	78.3	86.3	74.7	71.8	52.2	61.5	70.1	86.2	**65.5**	76.5	71.9	56.8	83.3	48.0	64.5	
mT5-XXL	**80.4**	**66.2**	**85.1**	**79.3**	**81.7**	79.0	**86.7**	**86.0**	73.5	57.6	58.8	70.4	**86.8**	65.1	**77.8**	74.2	**73.5**	**85.8**	**50.7**	**66.4**	

	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	avg
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	62.2
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	61.2
XLM-R	**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	65.4
mT5-Small	53.2	23.4	26.6	39.4	39.4	70.0	30.1	75.4	70.8	46.5	54.8	37.5	32.6	7.2	69.4	56.0	26.4	63.8	58.8	37.9	51.0
mT5-Base	50.1	23.4	33.9	48.2	43.8	72.6	37.0	80.1	76.0	55.4	62.4	41.2	42.7	9.5	74.6	58.4	38.4	73.0	59.3	41.5	56.6
mT5-Large	58.2	23.3	36.2	46.3	46.5	69.4	32.2	82.7	79.6	50.2	72.4	46.4	44.5	**10.5**	79.0	65.1	44.2	77.1	48.4	44.0	58.8
mT5-XL	66.0	31.6	38.1	54.1	57.6	74.8	42.6	85.7	85.2	66.9	72.8	49.0	54.7	9.6	84.1	67.7	64.7	79.6	59.9	54.4.	65.7
mT5-XXL	66.0	38.7	43.5	54.5	63.1	**77.6**	44.7	**87.7**	**86.9**	**72.0**	**72.9**	56.5	**59.5**	10.4	**85.2**	71.4	**80.7**	**84.6**	**70.0**	**56.8**	**69.2**

Table 8: NER F1 scores for each language.

Model	en	ar	de	el	es	hi	ru	th	tr	vi	zh	avg
Cross-lingual zero-shot transfer (models fine-tune on English data only)												
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
XLM-R	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
mT5-Small	78.5 / 66.1	51.4 / 34.0	63.8 / 45.9	53.8 / 33.4	67.0 / 50.3	47.8 / 34.5	50.5 / 30.1	54.0 / 44.5	55.7 / 38.9	58.1 / 41.3	58.9 / 48.7	58.1 / 42.5
mT5-Base	84.6 / 71.7	63.8 / 44.3	73.8 / 54.5	59.6 / 35.6	74.8 / 56.1	60.3 / 43.4	57.8 / 34.7	57.6 / 45.7	67.9 / 48.2	70.7 / 50.3	66.1 / 54.1	67.0 / 49.0
mT5-Large	88.4 / 77.3	75.2 / 56.7	80.0 / 62.9	77.5 / 57.6	81.8 / 64.2	73.4 / 56.6	74.7 / 56.9	73.4 / 62.0	76.5 / 56.3	79.4 / 60.3	75.9 / 65.5	77.8 / 61.5
mT5-XL	88.8 / 78.1	77.4 / 60.8	80.4 / 63.5	80.4 / 61.2	82.7 / 64.5	76.1 / 60.3	76.2 / 58.8	**74.2** / 62.5	77.7 / 58.4	80.5 / 60.8	80.5 / 71.0	79.5 / 63.6
mT5-XXL (75%)	**90.8** / **79.7**	**79.7** / **62.7**	**83.6** / **66.8**	**82.0** / **64.0**	**84.7** / **67.9**	**81.0** / **64.8**	**80.2** / 64.0	72.9 / 58.1	**80.0** / **59.9**	**83.0** / **62.9**	**82.5** / **71.6**	**81.9** / **65.7**
Translate-train (models fine-tune on English training data plus translations in all target languages)												
Filter + Self-Teaching	86.4 / 74.6	79.5 / 60.7	83.2 / 67.0	83.0 / 64.6	85.0 / 67.9	83.1 / 66.6	82.8 / 67.4	**79.6** / **73.2**	80.4 / 64.4	83.8 / 64.7	79.9 / 77.0	82.4 / 68.0
mT5-Small	74.0 / 61.2	61.0 / 45.0	66.0 / 50.2	64.1 / 47.2	67.5 / 50.8	60.2 / 43.7	64.4 / 46.7	58.9 / 52.9	59.0 / 39.4	63.5 / 46.0	68.2 / 61.2	64.3 / 49.5
mT5-Base	83.1 / 70.3	72.4 / 55.2	76.9 / 59.7	76.8 / 58.8	79.0 / 61.2	71.4 / 53.4	76.1 / 58.5	67.9 / 62.0	72.5 / 51.4	75.9 / 56.3	76.9 / 69.7	75.3 / 59.7
mT5-Large	87.3 / 75.5	79.4 / 62.7	82.7 / 66.0	81.8 / 63.5	83.8 / 66.1	78.0 / 59.8	81.9 / 66.3	74.7 / 68.2	80.2 / 59.2	80.4 / 60.8	83.2 / 76.9	81.2 / 65.9
mT5-XL	88.5 / 77.1	80.9 / 65.4	83.4 / 66.7	83.6 / 64.9	84.9 / 68.2	79.6 / 63.1	82.7 / 67.1	78.5 / 72.9	82.4 / 63.8	82.4 / 64.1	83.2 / 75.9	82.7 / 68.1
mT5-XXL (75%)	**90.8** / **79.8**	**83.6** / **68.7**	**85.5** / **68.7**	**85.7** / **68.5**	**86.9** / **70.3**	**83.3** / **67.0**	**84.9** / **69.4**	78.8 / 72.4	**84.4** / **66.8**	**85.0** / **66.2**	**86.2** / **80.5**	**85.0** / **70.8**

Table 9: XQuAD results (F1/EM) for each language.

Model	en	ar	de	es	hi	vi	zh	avg
Cross-lingual zero-shot transfer (models fine-tune on English data only)								
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	74.0 / 52.9	62.1 / 37.0	71.6 / 53.2
mT5-Small	77.2 / 63.0	44.7 / 27.3	53.3 / 35.7	60.1 / 41.5	43.0 / 29.2	52.9 / 33.2	51.3 / 29.7	54.6 / 37.1
mT5-Base	81.7 / 66.9	57.1 / 36.9	62.1 / 43.2	67.1 / 47.2	55.4 / 37.9	65.9 / 44.1	61.6 / 38.6	64.4 / 45.0
mT5-Large	84.9 / 70.7	65.3 / 44.6	68.9 / 51.8	73.5 / 54.1	66.9 / 47.7	72.5 / 50.7	66.2 / 42.0	71.2 / 51.7
mT5-XL	85.5 / 71.9	68.0 / 47.4	70.5 / 54.4	75.2 / 56.3	70.5 / 51.0	74.2 / 52.8	70.5 / 47.2	73.5 / 54.4
mT5-XXL (75%)	**86.4** / **73.0**	**70.0** / **49.6**	**72.8** / **56.5**	**76.3** / **58.2**	**74.8** / **56.6**	**76.4** / **56.0**	**72.0** / **48.8**	**75.5** / **56.9**
Translate-train (models fine-tune on English training data plus translations in all target languages)								
Filter + Self-Teaching	84.0 / 70.8	**72.1** / **51.1**	**74.8** /**60.0**	78.1 / 60.1	**76.0** / **57.6**	**78.1** /**57.5**	70.5 / 47.0	76.2 / 57.7
mT5-small	70.5 / 56.2	49.3 / 31.0	55.6 / 40.6	60.5 / 43.0	50.4 / 32.9	55.2 / 36.3	54.4 / 31.6	56.6 / 38.8
mT5-base	80.7 / 66.3	61.1 / 40.7	65.5 / 49.2	70.7 / 52.1	63.6 / 44.3	68.0 / 47.6	63.5 / 39.4	67.6 / 48.5
mT5-large	85.3 / 72.0	68.5 / 47.7	71.6 / 55.8	75.7 / 57.1	71.8 / 52.6	74.3 / 54.0	70.1 / 47.1	73.9 / 55.2
mT5-xl	86.0 / 73.0	70.0 / 49.8	72.7 / 56.8	76.9 / 58.3	73.4 / 55.0	75.4 / 55.0	71.4 / 48.4	75.1 / 56.6
mT5-xxl	**86.9** / **74.0**	71.3 / 50.8	74.4 / 58.2	**78.5** / **60.2**	75.8 / 57.5	77.3 / 57.1	**72.3** / **49.4**	**76.6** / **58.2**

Table 10: MLQA results (F1/EM) for each language.

Model	en	ar	bn	fi	id	ko	ru	sw	te	avg
Cross-lingual zero-shot transfer (models fine-tune on English data only)										
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
mt5-Small	53.9 / 43.6	41.1 / 26.0	18.9 / 13.3	39.2 / 22.6	44.4 / 31.7	24.9 / 16.3	40.5 / 24.3	34.8 / 21.2	16.9 / 11.5	34.9 / 23.4
mt5-Base	71.8 / 60.9	67.1 / 50.4	40.7 / 22.1	67.0 / 52.2	71.3 / 54.5	49.5 / 37.7	54.9 / 32.6	60.4 / 43.9	40.6 / 31.1	58.1 / 42.8
mt5-Large	71.6 / 58.9	60.5 / 40.4	42.0 / 23.9	64.6 / 48.8	67.0 / 49.2	47.6 / 37.3	58.9 / 36.8	65.7 / 45.3	41.9 / 29.7	57.8 / 41.2
mt5-XL	80.3 / 70.9	81.7 / 65.5	74.5 / 57.5	79.4 / 65.3	83.5 / 70.4	70.0 / 60.5	71.6 / 47.8	77.3 / 59.7	77.9 / 55.8	77.4 / 61.5
mt5-XXL (75%)	**83.1** / **72.3**	**83.3** / **67.2**	**80.9** / **68.1**	**82.9** / **70.2**	**86.3** / **75.6**	**73.7** / **63.0**	**75.2** / **53.8**	**82.4** / **66.5**	**79.2** / **59.6**	**80.8** / **66.3**
Translate-train (models fine-tune on English training data plus translations in all target languages)										
Filter + Self-Teaching	72.4 / 59.1	72.8 / 50.8	70.5 / 56.6	73.3 / 57.2	76.8 / 59.8	33.1 / 12.3	68.9 / 46.6	77.4 / 65.7	69.9 / 50.4	68.3 / 50.9
mt5-Small	57.1 / 46.6	56.8 / 39.7	37.2 / 21.2	50.9 / 37.2	60.1 / 45.1	40.4 / 29.3	50.7 / 33.6	51.5 / 35.3	29.3 / 18.1	48.2 / 34.0
mt5-Base	71.1 / 58.9	68.0 / 50.2	57.4 / 35.4	68.8 / 55.2	73.5 / 57.2	56.5 / 43.8	64.0 / 45.8	65.8 / 48.3	51.2 / 34.1	64.0 / 47.7
mt5-Large	75.6 / 62.7	74.8 / 57.9	65.0 / 46.0	72.3 / 57.5	78.7 / 63.5	66.4 / 53.6	70.9 / 50.5	74.0 / 56.7	62.0 / 45.1	71.1 / 54.9
mt5-XL	**82.0** / 65.7	79.3 / 65.5	**80.4** / **68.9**	79.1 / 64.7	84.7 / 71.0	70.5 / 56.2	78.3 / 61.1	83.9 / 70.9	80.9 / 64.0	79.9 / 65.3
mt5-XXL (75%)	81.6 / **70.0**	**82.7** / **66.7**	80.2 / 66.4	**83.0** / **68.8**	**86.2** / **72.2**	**75.1** / **62.0**	**80.1** / **62.4**	**86.8** / **75.6**	**82.4** / **65.8**	**82.0** / **67.8**

Table 11: TyDi QA-GoldP results (F1/EM) for each language.

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
Baseline (mT5-large)	**79.8**	**84.1**	83.4	**83.2**	**89.4**	84.2	84.1	**77.6**	81.5	**75.4**	**79.4**	**80.1**	73.5	**81.0**	**80.3**	**81.1**
Dropout 0.1	76.4	82.1	81.7	81.0	88.0	70.8	80.3	74.4	79.0	72.3	75.8	75.9	70.6	78.6	76.5	77.6
Sequence length 512	78.1	83.4	83.1	82.1	88.8	84.5	82.8	77.3	81.2	**75.4**	78.2	79.6	73.8	80.0	78.9	80.5
Span length 10	77.6	81.5	80.5	81.2	87.2	83.0	81.2	74.7	79.8	73.6	76.7	75.9	71.3	78.6	76.5	78.6
$\alpha = 0.7$	79.3	**84.1**	**84.5**	83.1	**89.4**	**85.3**	**84.4**	76.4	**82.8**	70.6	78.7	79.8	71.7	80.3	79.9	80.7
$\alpha = 0.2$	78.7	83.8	83.3	82.5	89.3	83.4	83.6	77.3	81.2	**75.4**	78.6	79.4	**73.9**	79.9	79.7	80.7
No line length filter	78.4	83.3	81.5	81.4	88.9	83.8	82.5	74.4	80.5	69.4	77.6	76.9	71.3	78.8	78.3	79.1
Add Wikipedia data	79.3	83.1	83.1	82.7	88.6	80.1	83.2	77.3	81.4	75.0	78.9	79.3	73.5	80.2	79.2	80.3

Table 12: XNLI zero-shot accuracy of various ablations on our mT5-Large model.

Model	ar	de	el	en	es	hi	ru	th	tr	vi	zh	avg
Baseline(mT5-large)	75.2 / 56.7	80.0 / 62.9	77.5 / 57.6	88.4 / 77.3	81.8 / **64.2**	73.4 / 56.6	74.7 / 56.9	**73.4** / **62.0**	**76.5** / 56.3	79.4 / **60.3**	75.9 / 65.5	77.8 / 61.5
Dropout 0.1	55.0 / 33.1	77.1 / 59.0	64.1 / 38.3	87.2 / 75.7	78.2 / 58.6	59.2 / 41.0	59.2 / 38.3	65.4 / 50.4	73.8 / 52.4	75.8 / 55.0	76.3 / 63.7	70.1/ 51.4
Sequence length 512	**77.9** / **60.9**	**80.9** / **63.3**	**81.1** / **62.1**	88.3 / 77.1	81.8 / **64.2**	**75.4** / **58.6**	**78.1** / **59.8**	72.7 / 59.7	76.0 / 54.8	**79.6** / 59.5	77.7 / 66.2	**79.0** / **62.4**
Span length 10	68.9 / 50.1	77.0 / 58.5	68.6 / 44.5	88.0 / 76.6	80.0 / 61.6	65.8 / 46.3	66.8 / 45.4	67.8 / 55.5	74.0 / 53.3	78.3 / 57.7	77.3 / **66.8**	73.9 / 56.0
$\alpha = 0.7$	75.9 / 59.2	78.8 / 60.1	75.3 / 54.2	**89.3** / **78.8**	77.0 / 57.7	74.1 / 56.9	70.6 / 50.9	71.1 / 60.3	75.4 / 55.3	79.5 / 58.5	77.0 / 65.9	76.7 / 59.8
$\alpha = 0.2$	75.0 / 56.5	80.1 / 61.3	75.0 / 52.4	87.9 / 76.5	**81.9** / 63.5	73.7 / 55.9	69.9 / 48.7	71.6 / 59.3	75.6 / 55.2	79.1 / 58.4	77.5 / 65.4	77.0 / 59.4
No line length filter	73.2 / 53.6	79.4 / 61.3	70.2 / 46.1	88.3 / 77.5	81.3 / 64.1	71.0 / 53.4	71.8 / 52.7	68.7 / 58.0	76.0 / 55.3	78.2 / 59.2	77.7 / 66.4	76.0 / 58.9
Add Wikipedia data	71.4 / 50.7	78.6 / 60.4	60.4 / 37.3	88.1 / 77.1	79.9 / 60.6	72.1 / 54.2	67.8 / 48.1	71.2 / 59.0	**76.5** / **56.7**	78.8 / 59.0	**77.9** / **66.8**	74.8 / 57.3

Table 13: XQuAD zero-shot F1/EM of various ablations on our mT5-Large model.