mT6: Multilingual Pretrained Text-to-Text Transformer with Translation Pairs

Zewen Chi[†‡*], **Li Dong**[‡], **Shuming Ma**[‡], **Shaohan Huang**[‡]
Xian-Ling Mao[†], **Heyan Huang**[†], **Furu Wei**[‡]
[†]Beijing Institute of Technology
[‡]Microsoft Research
{czw,maoxl,hhy63}@bit.edu.cn
{lidong1,shumma,shaohanh,fuwei}@microsoft.com

Abstract

Multilingual T5 (MT5; Xue et al. 2020) pretrains a sequence-to-sequence model on massive monolingual texts, which has shown promising results on many cross-lingual tasks. In this paper, we improve **m**ultilingual **t**ext-**t**o-**t**ext **t**ransfer **T**ransformer with **t**ranslation pairs (MT6). Specifically, we explore three cross-lingual text-to-text pre-training tasks, namely, machine translation, translation pair span corruption, and translation span corruption. In addition, we propose a partially non-autoregressive objective for text-to-text pre-training. We evaluate the methods on eight multilingual benchmark datasets, including sentence classification, named entity recognition, question answering, and abstractive summarization. Experimental results show that the proposed MT6 improves cross-lingual transferability over MT5.

1 Introduction

Multilingual pretrained language models, such as mBERT (Devlin et al., 2019), have attracted increasing attention. They not only improve the performance on downstream multilingual NLP tasks (Conneau and Lample, 2019; Conneau et al., 2020; Liu et al., 2020; Chi et al., 2021c), but also show an impressive cross-lingual transferability (Wu and Dredze, 2019; K et al., 2020; Hu et al., 2020b; Chi et al., 2021a).

Multilingual pretrained models are typically trained on multilingual unlabeled text with unsupervised language modeling tasks, e.g., masked language modeling (Devlin et al., 2019), causal language modeling (Conneau and Lample, 2019), and span corruption (Raffel et al., 2020). These unsupervised tasks are built upon large-scale monolingual texts. In addition, several studies propose cross-lingual tasks that utilize translation data from multilingual parallel corpora, such as translation language modeling (Conneau and Lample, 2019), cross-lingual contrast (Chi et al., 2021a), and bidirectional word alignment (Hu et al., 2020a). Thanks to the translation data, the pretrained models produce better-aligned cross-lingual representations and obtain better cross-lingual transferability.

Recently, the multilingual text-to-text transfer Transformer (MT5; Xue et al. 2020) achieves state-of-the-art performance on several cross-lingual understanding benchmarks. MT5 inherits the benefits of T5 (Raffel et al., 2020) that treats every text processing problem as a text-to-text problem, i.e., the problem of generating some target text conditioned on the input text. Despite the effectiveness of MT5, how to improve MT5 with translation data is still an open problem.

In this paper, we present MT6, standing for improving multilingual text-to-text transfer Transformer with translation data. MT6 differs from MT5 in terms of both pre-training tasks and the training objective. We present three cross-lingual tasks for text-to-text Transformer pre-training, i.e., machine translation, translation pair span corruption, and translation span corruption. In the translation span corruption task, the model is trained to predict the text spans based on the input translation pair. The cross-lingual tasks encourage the model to align representations of different languages. We also propose a new objective for text-to-text pre-training, called partially non-autoregressive (PNAT) decoding. The PNAT objective divides the target sequence into several groups, and constrains that the predictions should be only conditioned on the source tokens and the target tokens from the same group.

We conduct experiments on both multilingual understanding and generation tasks. Our MT6 model yields substantially better performance than MT5 on eight benchmarks. We also provide an empirical comparison of the cross-lingual pre-training tasks, where we evaluate several variants of MT6 under the same pre-training and fine-tuning procedure.

[*]Contribution during internship at Microsoft Research.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1671–1683
November 7–11, 2021.

Moreover, our analysis indicates that the representations produced by MT6 are more cross-lingual transferable and better-aligned than MT5.

The contributions are summarized as follows:

- We introduce three cross-lingual tasks for text-to-text Transformer pre-training, which improves MT5 with translation data.
- We propose a partially non-autoregressive objective that pretrains the decoder to use more information from the source sequence.
- We provide extensive evaluation results of various pre-training tasks and training objectives.

2 Background on T5 and MT5

Multilingual text-to-text transfer Transformer (MT5; Xue et al. 2020) is the multilingual variant of T5 (Raffel et al., 2020) pretrained on the mC4 (Xue et al., 2020) dataset, which consists of natural text in 101 languages drawn from the public Common Crawl web scrape.

The backbone architecture of MT5 is the simple encoder-decoder Transformer (Vaswani et al., 2017), which is trained in a unified text-to-text manner. In specific, text-based NLP problems are formulated as text-to-text transfer, i.e., the model is trained to predict the target text conditioned on the input source text. For example, in text classification, the model predicts the label text rather than a class index. This feature enables the MT5 to be fine-tuned with the same training objective for every task. Formally, let x and y denote the input sequence and the output sequence, the loss function of training the $x \rightarrow y$ transfer is

$$\mathcal{L}(x \rightarrow y) = -\sum_{i=1}^{|y|} \log p(y_i|x, y_{<i}), \quad (1)$$

where $y_{<i} = y_1, \cdots, y_{i-1}$. With the unified text-to-text formulation, the pre-training task can be designed by constructing the input and output text sequences. Specifically, MT5 employs the span corruption task as the pre-training task, which is an unsupervised masked language modeling task. As shown in Figure 1, we provide an example of constructing the input and output sequences for span corruption. Given a natural sentence s, it first randomly selects several spans of s as the spans to be masked. Then, the input sequence is constructed by replacing the selected spans with unique mask tokens. The output sequence is the concatenation of the original tokens of the masked spans, each of which starts with a unique mask token to indicate the span to be decoded. We denote the above two operations as g_i and g_o, standing for converting the original sentence s into the input or the output formats of span corruption. Thus, the loss function of the span corruption task can be written as

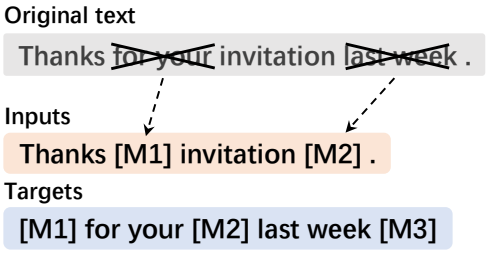

Figure 1: Example of the span corruption task (Raffel et al., 2020) used in T5 and MT5.

$$\mathcal{L}_{\text{SC}}(s) = \mathcal{L}(g_i(s) \rightarrow g_o(s)). \quad (2)$$

3 Methods

In this section, we first present three text-to-text pre-training tasks for improving MT5 with translation data. Then, we introduce the partially non-autoregressive decoding objective, and provide the detailed fine-tuning procedures for the classification, question answering, and named entity recognition tasks.

3.1 Cross-lingual Pre-training Tasks with Translation Pairs

As shown in Figure 2, we illustrate an overview of our cross-lingual text-to-text pre-training tasks. Given the same translation pair, the three tasks construct different input and output sequences.

3.1.1 Machine Translation

Machine translation (MT) is a typical text-to-text task with the goal of translating a sentence from the source language into a target language. It is a natural design to use MT as a text-to-text pre-training task for sequence-to-sequence learning (Chi et al., 2020). Let e and f denote a sentence and its corresponding translation. We directly use e and f as the input and output sequences, respectively. The loss function of MT is

$$\mathcal{L}_{\text{MT}}(e, f) = \mathcal{L}(e \rightarrow f). \quad (3)$$

Figure 2: Overview of three cross-lingual text-to-text pre-training tasks. For each task, we provide an example of the input and target text. The words marked with "×" are randomly replaced with unique mask tokens like $[\mathrm{M}_1]$. Notice that in the translation span corruption task, we mask tokens only in one language.

3.1.2 Translation Pair Span Corruption

Inspired by the translation masked language modeling (Conneau and Lample, 2019) task, we propose the translation pair span corruption (TPSC) task that aims to predict the masked spans from a translation pair instead of a monolingual sentence. Let e and f denote a sentence and its corresponding translation. We concatenate e and f as a single sentence, and perform the span corruption on the concatenated sentence. Formally, we construct the input and output sequences by $g_i([e; f])$ and $g_o([e; f])$, where $[e; f]$ stands for the concatenation of e and f. With the resulting input and output sequences, the loss function of TPSC can be written as

$$\mathcal{L}_{\text{TPSC}}(e, f) = \mathcal{L}(g_i([e; f]) \rightarrow g_o([e; f])). \quad (4)$$

3.1.3 Translation Span Corruption

A potential issue of translation pair span corruption is that the spans in the target sequence can be organized in unnatural word order. As shown in Figure 2, the output sequence of TPSC is organized as "$[M_1]$ *for your* $[M_2]$ *last week* $[M_3]$ *invitation* $[M_4]$". It can be found that the French word "invitation" is after the English word "week", which could harm the language model of the decoder. This motivates us to propose the translation span corruption (TSC) task where we only mask and predict the spans in one language. Given a translation pair (e, f), we randomly select the e or f to perform span corruption. Without loss of generality, we consider e as the sentence for span corruption. Then, the input and output sequences are constructed by $[g_i(e); f]$ and $g_o(e)$, respectively. With the resulting input and output sequences, the loss function of TSC can be written as

$$\mathcal{L}_{\text{TSC}}(e, f) = \mathcal{L}([g_i(e); f] \rightarrow g_o(e))). \quad (5)$$

3.2 Pre-training Objective: Partially Non-autoregressive Decoding

Recall that the predictions in MT5 are conditioned on both the source tokens and the target tokens to the left. When predicting the tokens closer to the end, the model can use more information from the target sequence, resulting in the insufficient training of the encoder.

To encourage the model to utilize more information from the encoding side while preserving the ability of autoregressive decoding, we propose a new training objective for text-to-text training, called partially non-autoregressive decoding (PNAT). In Figure 3, we provide an example for PNAT. Specifically, given a target sequence containing several spans, we divide the target sequence into groups, and train the model to decode each group separately. With the PNAT objective, a prediction is only conditioned on the source tokens and the target tokens from the same group. Consider the target sequence consisting of m spans. We divide the spans into n_g groups, each of which contains m/n_g consecutive spans. For the j-th group, we denote l_j and r_j as the start position and the end position, respectively. The PNAT objective is defined as

$$\mathcal{L}^{\text{PNAT}}(x \rightarrow y) = -\sum_{j=1}^{n_g} \sum_{i=l_j}^{r_j} \log p(y_i | x, y_{l_j} \dots y_{i-1}).$$

The text-to-text loss $\mathcal{L}(x \rightarrow y)$ is a specially case of $\mathcal{L}^{\text{PNAT}}(x \rightarrow y)$ with $n_g = 1$.

The MT6 model is jointly pretrained on both monolingual and parallel corpora, where we use the span corruption and one of the three cross-lingual text-to-text tasks. For both tasks, we use the partially non-autoregressive decoding as the training objective where we divide the target sequence into

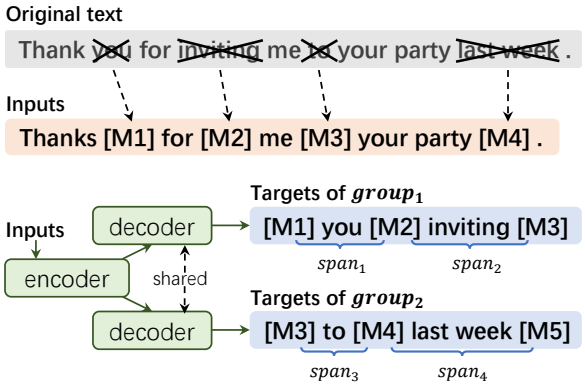

Figure 3: Partially non-autoregressive objective.

n_g groups. The overall pre-training objective is to minimize

$$\mathcal{L}_{\text{MT6}} = \mathcal{L}_{\text{SC}}^{\text{PNAT}}(s) + \mathcal{L}_{\text{X}}^{\text{PNAT}}(e, f), \quad X \in \{\text{MT}, \text{TPSC}, \text{TSC}\}, \tag{6}$$

where $\mathcal{L}_{\text{X}}^{\text{PNAT}}$ stands for the one of the loss functions of machine translation (MT; Section 3.1.1), translation pair span corruption (TPSC; Section 3.1.2) and translation span corruption (TSC; Section 3.1.3), with PNAT as the training objective.

3.3 Cross-lingual Fine-tuning

We fine-tune all parameters of the MT6 model with Equation (1) regardless of the end task. Unlike language generation tasks, language understanding tasks should be pre-processed as the text-to-text format. We introduce how to convert the following three types of the language understanding task into the text-to-text format, i.e., constructing the input and output sequences from the original examples.

Classification The goal of the text classification task is to predict the label of a given text. Following T5 (Raffel et al., 2020), we directly use the label text as the output text sequence. We provide an example for the MNLI natural language inference task (Williams et al., 2018). Given an input sentence pair of "*You have access to the facts .*" and "*The facts are accessible to you .*", the goal is to classify the input into the relationships of "*entailment*", "*contradiction*", or "*neutral*". The input and target sequences are constructed as

Input: $\langle bos \rangle$ *You have access to the facts.* $\langle eos \rangle$ *The facts are accessible to you.* $\langle eos \rangle$

Output: $\langle bos \rangle$ *entailment* $\langle eos \rangle$

Since multi-task fine-tuning is not the focus of this work, we do not prepend a task prefix in the input text. We also adopt a constrained decoding process, where the decoded text is constrained to be one of the labels.

Question Answering For the extractive question answering (QA) task, we concatenate the passage and the question as the input, and directly use the answer text as the target instead of predicting the answer span positions. We provide an example of converting a QA training example into the text-to-text format.

Input: $\langle bos \rangle$ *It has offices in Seoul, South Korea.* $\langle eos \rangle$ *Where is the office in South Korea?* $\langle eos \rangle$

Output: $\langle bos \rangle$ *Seoul* $\langle eos \rangle$

We use the constrained decoding for the QA tasks where we use the tokens shown in the input passage as the decoding vocabulary.

Named Entity Recognition In named entity recognition (NER), we do not directly use the original tag sentence as the output. We find that the model tends to repeat decoding the "*O*" tag if the model directly learns to decode the tag sequences. Alternately, we construct the target text by concatenating the entity spans, each of which starts with the entity tag and ends with the entity tokens. We show an example of converting a NER training example into the text-to-text format.

Input: $\langle bos \rangle$ *Italy recalled Marcello Cuttitta .* $\langle eos \rangle$

Output: $\langle bos \rangle$ $\langle loc \rangle$ *Italy* $\langle sep \rangle$ $\langle per \rangle$ *Marcello Cuttitta* $\langle sep \rangle$ $\langle eos \rangle$

$\langle loc \rangle$ and $\langle per \rangle$ are entity tags denoting location and person. The $\langle sep \rangle$ tag means the end of entity span. We use the following constrained decoding rules: (1) The model should decode entity tags or the end-of-sentence tag ($\langle eos \rangle$) after a $\langle bos \rangle$ token or a $\langle sep \rangle$ token; (2) Otherwise, the model should decode the tokens from the input sentence or the $\langle sep \rangle$ token for the other situations.

4 Experiments

4.1 Setup

Data Following previous work on cross-lingual pre-training (Conneau et al., 2020; Chi et al., 2021a), we use the natural sentences from CC-Net (Wenzek et al., 2019) in 94 languages for monolingual text-to-text tasks. For cross-lingual text-to-text tasks, we use parallel corpora of 14 English-centric language pairs, collected from MultiUN (Ziemski et al., 2016), IIT Bombay (Kunchukuttan et al., 2018), OPUS (Tiedemann, 2012), and WikiMatrix (Schwenk et al.,

2019). Details of the pre-training data are described in Appendix.

Training Details In the experiments, we consider the small-size Transformer model (Xue et al., 2020), with $d_{\text{model}} = 512, d_{\text{ff}} = 1,024$, 6 attention heads, and 8 layers for both the encoder and the decoder[1]. We use the vocabulary provided by XLM-R (Conneau et al., 2020), and extend it with 100 unique mask tokens for the span corruption tasks. We pretrain our MT6 for 0.5M steps with batches of 256 length-512 input sequences. The model is optimized by the Adam optimizer (Kingma and Ba, 2015) with a linear learning rate scheduler. The pre-training procedure takes about 2.5 days on an Nvidia DGX-2 Station. Details of the pre-training hyperparameters are described in Appendix.

4.2 Results

4.2.1 XTREME Cross-lingual Understanding

To validate the performance of MT6, we evaluate the pretrained models on XTREME (Hu et al., 2020b), which is a widely used benchmark for cross-lingual understanding. Following MT5 (Xue et al., 2020), we consider six downstream tasks included by XTREME: the named entity recognition (NER) task on the WikiAnn (Pan et al., 2017; Rahimi et al., 2019) dataset in 40 languages, the question answering (QA) task on MLQA (Lewis et al., 2020b), XQuAD (Artetxe et al., 2020), and TyDiQA-GoldP (Clark et al., 2020), the cross-lingual natural language inference task on XNLI (Conneau et al., 2018), and cross-lingual paraphrase adversaries on PAWS-X (Yang et al., 2019). The models are evaluated under the cross-lingual transfer setting (Conneau et al., 2020; Hu et al., 2020b). Under this setting, the models should be fine-tuned only on English training data but evaluated on all target languages. Moreover, for each pretrained model, only one model is used for all languages rather than selecting fine-tuned models separately. Details of the fine-tuning hyperparameters are described in Appendix.

As shown in Table 1, we present the evaluation results of the pretrained models on the XTREME benchmark. We observe that MT6 achieves the best performance on XTREME, improving the average score from 45.0 to 50.4, as we go from MT5 to MT6. It is worth mentioning that pre-training the model only with the machine translation task performs even worse than MT5. We have noticed that several target languages in TyDiQA and WikiAnn are not covered by our parallel corpora. However, the NMT pretrained model still shows poor results on the other four tasks, where all target languages are covered by the training data. Detailed results can be found in Appendix.

4.2.2 Comparison of Pre-training Tasks

To provide a clear comparison among the pre-training tasks, we implement the text-to-text pre-training methods presented in Section 3, and pre-train variants of MT6 with the same training data and resources for fair comparisons.

Table 1 compares the evaluation results of the models pretrained with seven different combinations of span corruption (SC), machine translation (MT), translation pair span corruption (TPSC), translation span corruption (TSC), and partially non-autoregressive decoding (PNAT). It can be observed that jointly training SC+TSC with PNAT achieves the best overall performance on the XTREME benchmark, with substantial gains over the models trained on monolingual data only. The same trend can be observed for the other models pretrained on both monolingual data and parallel data. This demonstrates that introducing translation data to text-to-text pre-training can improve the performance on the end tasks of cross-lingual understanding. Moreover, PNAT provides consistent gains over SC and SC+TSC, showing that PNAT is effective on both monolingual and cross-lingual tasks. Surprisingly, SC+PNAT obtains comparable results to SC+MT without any parallel data. Comparing TSC with MT and TPSC, we observe that SC+TSC brings noticeable improvements on question answering tasks. Although SC+MT shows competitive results on XNLI, the results on the other tasks are relatively low, indicating that simply jointly training SC with MT is not the most effective way to pretrain MT6.

4.3 Abstractive Summarization

Multilingual Summarization In addition to language understanding tasks, we also evaluate our MT6 model on the abstractive summarization task. Abstractive summarization aims to generate a summary of the input document while preserving its original meaning. We use the Gigaword dataset provided by Chi et al. (2020). The dataset is constructed by extracting the first sentences and head-

[1]Notice that the "small-size" defined in T5 and MT5 are different. Here we follow the setting of MT5-small.

Model	Configuration					Structured (F1)	Question Answering (F1)			Classification (Acc.)	
	SC	PNAT	MT	TPSC	TSC	WikiAnn	XQuAD	MLQA	TyDiQA	XNLI	PAWS-X
NMT	✗	✗	✓	✗	✗	27.3	12.5	14.9	16.8	**64.8**	55.0
MT5	✓	✗	✗	✗	✗	43.1	42.1	37.6	30.7	57.2	78.0
MT6 (ours)	✓	✓	✗	✗	✓	**44.7**	**50.4**	**44.1**	**36.0**	64.7	**82.2**
Ablations	✓	✓	✗	✗	✗	43.7	45.1	38.5	32.3	57.9	77.5
	✓	✗	✓	✗	✗	43.9	38.5	33.3	29.4	65.9	79.3
	✓	✗	✗	✓	✗	42.3	46.2	40.8	35.3	64.0	78.9
	✓	✗	✗	✗	✓	43.8	47.6	40.5	36.7	65.4	80.3
Pre-training with larger batch size and more training steps											
MT5 (Xue et al., 2020)						50.5	58.1	54.6	35.2	67.5	82.4

Table 1: Evaluation results on XTREME under the cross-lingual transfer setting, where models are only fine-tuned on the English training data but evaluated on all target languages. We pretrain models with different combinations of span corruption (SC), machine translation (MT), translation pair span corruption (TPSC), translation span corruption (TSC), and partially non-autoregressive decoding (PNAT). All results are averaged over five runs.

Model	#Param	en			fr			zh		
		RG-1	RG-2	RG-L	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
Larger model size										
XLM (Chi et al., 2020)	800M	48.15	26.35	45.04	56.27	39.20	52.84	55.30	42.57	52.95
XNLG (Chi et al., 2020)	800M	48.76	26.82	45.57	57.84	40.81	54.24	57.65	44.93	54.95
Our re-implementation (Fine-tuning with full training data)										
MT5 (reimpl)	300M	46.58	24.45	43.32	54.12	36.78	50.61	57.30	44.08	54.65
MT6	300M	**46.82**	**24.65**	**43.50**	**54.82**	**37.61**	**51.30**	**57.38**	**44.20**	**54.66**
Our re-implementation (Fine-tuning with 1K training data)										
MT5	300M	28.00	10.89	26.13	32.56	17.25	29.75	44.16	31.20	41.86
MT6	300M	**28.80**	**11.44**	**26.45**	**35.07**	**18.70**	**31.39**	**46.48**	**33.17**	**44.02**

Table 2: Evaluation results on Gigaword multilingual abstractive summarization. RG is short for ROUGE. Results of XLM and XNLG are taken from (Chi et al., 2020). Results of MT5 and MT6 are averaged over three runs.

lines as the input documents and summaries, respectively. The dataset consists of examples in the languages of English, French, and Chinese. For each language, it contains 500K, 5K, and 5K examples for the training, validation, and test, respectively. We fine-tune the models for 20 epochs with a batch size of 32 and a learning rate of 0.00001. During decoding, we use the greedy decoding for all evaluated models.

As shown in Table 2, we report the ROUGE (Lin, 2004) scores of the models on Gigaword multilingual abstractive summarization. We observe that MT6 consistently outperforms MT5 on all the three target languages. Comparing with the XLM (Conneau and Lample, 2019) and XNLG (Chi et al., 2020) models with 800M parameters, our MT6 model achieves a similar performance with only 300M parameters. Besides, under the setting with fewer training data, MT6 shows more improvements over MT5.

Cross-Lingual Summarization The cross-lingual summarization task aims to generate summaries in a different language. We use the Wikilingua (Ladhak et al., 2020) dataset containing passage-summary pairs in four language pairs. We fine-tune the models for 100K steps with a batch size of 32 and a learning rate of 0.0001. We use the greedy decoding for all evaluated models. The evaluation results are shown in Table 3, where MT6 outperforms MT5 on the test sets of four language pairs.

Model	es-en	ru-en	vi-en	tr-en
MT5	11.36	8.77	8.98	10.57
MT6	**11.83**	**9.49**	**9.52**	**10.80**

Table 3: ROUGE-2 scores on Wikilingua cross-lingual summarization. Results are averaged over three runs.

Model	XQuAD	MLQA	TyDiQA	XNLI	PAWS-X
MT5	30.4	27.5	27.5	19.5	16.0
MT6	**28.6**	**27.2**	**25.9**	**14.6**	**13.2**

Table 4: The cross-lingual transfer gap scores on the XTREME tasks. A lower transfer gap score indicates better cross-lingual transferability. We use the EM scores to compute the gap scores for the QA tasks.

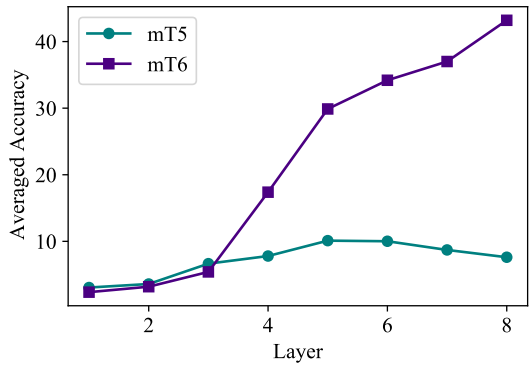

Figure 4: Evaluation results of different layers on Tatoeba cross-lingual sentence retrieval. We illustrate the average accuracy@1 scores on the Tatoeba test sets of the 14 language pairs covered by the parallel data.

4.4 Cross-lingual Transfer Gap

To explore whether our MT6 model achieves better cross-lingual transferability, we compare the cross-lingual transfer gap scores of our MT6 with MT5. Cross-lingual transfer gap (Hu et al., 2020b) is defined as the difference between the performance on the English test set and the average performance on the non-English test sets. The transfer gap indicates how much the end-task knowledge preserves when transferring from English to the other target languages. Empirically, a lower transfer gap score indicates better cross-lingual transferability. Following Hu et al. (2020b), we compute the transfer gap scores over the sentence classification and question answering tasks. As shown in Table 4, MT6 consistently reduces the transfer gap across all the five tasks, demonstrating that our model is more effective for cross-lingual transfer than MT5.

4.5 Cross-lingual Representations

We analyze the cross-lingual representations produced by our MT6 model. Following Chi et al. (2021a), we evaluate the representations on the Tatoeba (Artetxe and Schwenk, 2019) cross-lingual sentence retrieval task. The test sets consist of 14 English-centric language pairs covered by the parallel data in our experiments. Figure 4 illustrates the average accuracy@1 scores of cross-lingual sentence retrieval. The scores are averaged over 14 language pairs and both the directions of xx $\rightarrow$ en and en $\rightarrow$ xx. From the figure, we observe that MT5 shows a parabolic trend across different layers, which also appears in other cross-lingual encoder models (Jalili Sabet et al., 2020; Chi et al., 2021a). Differently, we obtain better performance as we use higher layers of our MT6 model. At layer-8, our MT6 model achieves an average accuracy@1 of 43.2, outperforming the MT5 model by 35.6, which means our MT6 model produces better-aligned text representations. We believe the better-aligned representations potentially improve the cross-lingual transferability. Furthermore, the results also indicate that our pre-training objective is more effective for training the encoder than MT5.

Model	en-de	en-fr	en-ro	Avg
MT5	35.84	19.05	45.24	33.38
MT6	**23.69**	**12.11**	**42.56**	**26.12**

Table 5: Evaluation results on word alignment. We report the alignment error rate scores (lower is better). We use the hidden vectors from the last encoder layer, and apply the SimAlign (Jalili Sabet et al., 2020) tool to obtain the resulting word alignments.

Noise Density	NER	QA	Classification	Avg
15%	41.7	33.5	71.9	47.4
30%	41.3	**35.9**	72.2	48.9
50%	43.8	35.5	**72.9**	**49.4**
100% (MT)	**43.9**	29.1	72.6	46.1

Table 6: Effects of noise density. We report the average results over different task types and the average results over all the six tasks on the XTREME benchmark. We vary the noise density of the translation span corruption task from 15% to 100%. All results are averaged over five runs.

4.6 Word Alignment

In addition to cross-lingual sentence retrieval that evaluates sentence-level representations, we also explore whether the representations produced by MT6 are better-aligned at token-level. Thus, we compare our MT6 with MT5 on the word alignment task, where the goal is to find corresponding word pairs in a translation pair. We use the hidden vectors from the last encoder layer, and apply the SimAlign (Jalili Sabet et al., 2020) tool to obtain the resulting word alignments. Table 5 shows the alignment error rate (AER) scores on the test sets provided by Jalili Sabet et al. (2020). Among the three language pairs, MT6 achieves lower AER scores than MT5, indicating that the cross-lingual representations produced by MT6 are also better-aligned at token-level.

4.7 Effects of Noise Density

In the translation span corruption (TSC) task, the input parallel sentences provide redundant information in two languages, which is different from the standard monolingual span corruption task. Thus, we explore the effects of noise density by varying the noise density in the translation span corruption task, with the other hyperparameters fixed. To reduce the computational load, we do not apply the partially non-autoregressive decoding, i.e., we pretrain the models with the original text-to-text objective. We pretrain MT6 models with the noise density of 0.15, 0.3, 0.5, and 1.0 respectively. It means 15%, 30%, 50%, or all of the source or target tokens are replaced with the masked tokens. Notice that setting the noise density as 1.0 is identical to machine translation, where the decoder is required to decode the whole target sentence.

In Table 6, we report the average scores on the XTREME benchmark. From the results, we observe that MT6 achieves the best results with the noise density of 0.5, rather than a higher noise density such as 1.0. The results indicate that the TSC task prefers a higher noise density, so that the model can learn to use more cross-lingual information. This finding is different from that reported by T5 (Raffel et al., 2020), where the span corruption task works better with the noise density of 0.15 under the monolingual setting.

5 Related Work

Cross-lingual LM Pre-training Cross-lingual language models are typically built with the Transformer (Vaswani et al., 2017) architecture, and pretrained with various pre-training tasks on large-scale text data. Multilingual BERT (mBERT; Devlin et al. 2019) and XLM-R (Conneau et al., 2020) are pretrained with masked language modeling (MLM; Devlin et al. 2019) on large-scale unlabeled text in about 100 languages. MASS (Song et al., 2019) and mBART (Liu et al., 2020) are pretrained in an auto-encoding manner, which provides improvements on the neural machine translation tasks. MT5 (Xue et al., 2020) is pretrained with the span corruption (Raffel et al., 2020) task under the text-to-text formulation (Raffel et al., 2020). Cross-lingual pretrained models also benefit from translation data. XLM (Conneau and Lample, 2019) jointly learns MLM and the translation language modeling (TLM) task. Unicoder (Huang et al., 2019) presents three cross-lingual tasks to learn mappings among languages. ALM (Yang et al., 2020) converts the translation pairs into code-switched sequences as the training examples. Word-aligned BERT models (Cao et al., 2020; Zhao et al., 2020) improves the cross-lingual representations by fine-tuning the mBERT with the objective of minimizing the distance between aligned tokens. AMBER (Hu et al., 2020a) propose to maximize the agreement between the forward and backward attention matrices of the input translation pair. InfoXLM (Chi et al., 2021a) proposes the cross-lingual contrastive learning task that maximizes the InfoNCE (Oord et al., 2018) lower bound of the mutual information between the input translation pair. XLM-Align (Chi et al., 2021b) leverages token-level alignments implied in translation pairs to improve cross-lingual transfer. XNLG (Chi et al., 2020) introduces the cross-lingual transfer for NLG tasks, and achieves zero-shot cross-lingual transfer for question generation and abstractive summarization. VECO (Luo et al., 2020) pretrains a variable cross-lingual pre-training model that learns unified language representations for both NLU and NLG. ERNIE-M (Ouyang et al., 2020) utilizes the back-translation masked language modeling task that generates pseudo parallel sentence pairs for learning TLM.

Encoder-Decoder Pre-training Raffel et al. (2020) use span corruption to pretrain text-to-text Transformer, where both language understanding and generation tasks are formulated as sequence-to-sequence fine-tuning. Song et al. (2019) propose masked sequence-to-sequence pre-training where the model predicts a randomly masked span. BART (Lewis et al., 2020a) design various denoised autoencoding tasks to recover the whole original sentence. PEGASUS (Zhang et al., 2020) introduces the gap sentence generation task for abstractive summarization pre-training. Chi et al. (2020) use both denoised autoencoding and machine translation for cross-lingual language generation. Another strand of research follows unified language model pre-training (Dong et al., 2019; Bao et al., 2020; Luo et al., 2020), where the encoder and the decoder share parameters. Ma et al. (2020, 2021) reuse pretrained multilingual encoder for sequence-to-sequence pre-training.

6 Conclusion

In this paper, we propose MT6 that improves the multilingual text-to-text transfer Transformer with

translation data. We introduce three text-to-text pre-training tasks that are built on parallel corpora, and a training objective for improving text-to-text pre-training. Nonetheless, we present a comprehensive comparison of the text-to-text tasks, and show that our MT6 model outperforms MT5 on both cross-lingual understanding and generation benchmarks. For future work, we would like to pretrain MT6 models at a larger scale, and explore more applications, such as machine translation.

Acknowledgements

We would like to acknowledge Bo Zheng for the helpful discussions. The work is supported by National Key R&D Plan (No. 2018YFB1005100), National Natural Science Foundation of China (No. 61751201, 61602197, 61772076, and 61732005), Natural Science Fund of Beijing (No. Z181100008918002), and the funds of Beijing Advanced Innovation Center for Language Resources (No. TYZ19005). Heyan Huang is the corresponding author.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7(0):597–610.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-masked language models for unified language model pre-training. *arXiv preprint arXiv:2002.12804*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7570–7577. AAAI Press.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021c. XLM-E: Cross-lingual language model pre-training via ELECTRA. *ArXiv*, abs/2106.16138.

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*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075. Curran Associates, Inc.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2020a. Explicit alignment objectives for multilingual bidirectional encoders. *arXiv preprint arXiv:2010.07972*.

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

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, San Diego, CA.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020b. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain.

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*.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2020. Veco: Variable encoder-decoder pre-training for cross-lingual understanding and generation. *arXiv preprint arXiv:2010.16046*.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *ArXiv*, abs/2106.13736.

Shuming Ma, Jian Yang, H. Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, Xia Song, Arul Menezes, and Furu Wei. 2020. XLM-T: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *ArXiv*, abs/2012.15547.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *arXiv preprint arXiv:2012.15674*.

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, Vancouver, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association.

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*, pages 5998–6008. Curran Associates, Inc.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, pages 1112–1122, New Orleans, Louisiana.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020. Inducing language-agnostic multilingual representations. *arXiv preprint arXiv:2008.09112*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *LREC*, pages 3530–3534.

A Pre-Training Data

We reconstruct CCNet[2] and follow (Conneau et al., 2020) to reproduce the CC-100 corpus for monolingual data. The resulting corpus contains 94 languages. We present the language codes and data size in Table 7 and Table 8 for the monolingual corpus and parallel corpus, respectively. Table 7 reports the language codes and data size in our work. We apply the multilingual sampling strategy (Conneau and Lample, 2019) with $\alpha = 0.7$ for both monolingual and parallel data.

B Hyperparameters for Pre-Training

As shown in Table 9, we present the hyperparameters for pre-training MT6. We extend the vocabulary of the XLM-R (Conneau et al., 2020) with external 100 unique mask tokens as the vocabulary of MT6 and our MT5 re-implementation.

C Hyperparameters for Fine-Tuning

In Table 10, we present the hyperparameters for fine-tuning MT6 on the end tasks.

[2] github.com/facebookresearch/cc_net

Code	Size (GB)	Code	Size (GB)	Code	Size (GB)
af	0.2	hr	1.4	pa	0.8
am	0.4	hu	9.5	pl	28.6
ar	16.1	hy	0.7	ps	0.4
as	0.1	id	17.2	pt	39.4
az	0.8	is	0.5	ro	11.0
ba	0.2	it	47.2	ru	253.3
be	0.5	ja	86.8	sa	0.2
bg	7.0	ka	1.0	sd	0.2
bn	5.5	kk	0.6	si	1.3
ca	3.0	km	0.2	sk	13.6
ckb	0.6	kn	0.3	sl	6.2
cs	14.9	ko	40.0	sq	3.0
cy	0.4	ky	0.5	sr	7.2
da	6.9	la	0.3	sv	60.4
de	99.0	lo	0.2	sw	0.3
el	13.1	lt	2.3	ta	7.9
en	731.6	lv	1.3	te	2.3
eo	0.5	mk	0.6	tg	0.7
es	85.6	ml	1.3	th	33.0
et	1.4	mn	0.4	tl	1.2
eu	1.0	mr	0.5	tr	56.4
fa	19.0	ms	0.7	tt	0.6
fi	5.9	mt	0.2	ug	0.2
fr	89.9	my	0.4	uk	13.4
ga	0.2	ne	0.6	ur	3.0
gl	1.5	nl	25.9	uz	0.1
gu	0.3	nn	0.4	vi	74.5
he	4.4	no	5.5	yi	0.3
hi	5.0	or	0.3	zh	96.8

Table 7: Statistics of CCNet used for pre-training.

ISO Code	Size (GB)	ISO Code	Size (GB)
en-ar	5.88	en-ru	7.72
en-bg	0.49	en-sw	0.06
en-de	4.21	en-th	0.47
en-el	2.28	en-tr	0.34
en-es	7.09	en-ur	0.39
en-fr	7.63	en-vi	0.86
en-hi	0.62	en-zh	4.02

Table 8: Parallel data used for pre-training.

D Results on XTREME Cross-Lingual Understanding

We present the detailed results of the MT6 and our re-implemented MT5 models on XTREME in Table 11-16.

E Results on Wikilingua Cross-Lingual Summarization

As shown in Table 17, we present the detailed results of the MT6 and our re-implemented MT5 models on Wikilingua cross-lingual summarization.

Hyperparameters	Value
Layers	8
Hidden size	512
FFN inner hidden size	1,024
Attention heads	6
Training steps	500K
Batch size	256
Input length	512
Adam ϵ	1e-6
Adam β	(0.9, 0.9999)
Learning rate	1e-4
Learning rate schedule	Linear
Warmup steps	10,000
Gradient clipping	1.0
Noise density	0.5
PNAT group number	3

Table 9: Hyperparameters used for pre-training MT6.

Hyperparameters	WikiAnn	XQuAD	MLQA	TyDiQA	XNLI	PAWS-X	Gigaword	Wikilingua
Batch size	32	32	32	32	32	32	32	32
Learning rate	7e-5	3e-5	3e-5	5e-5	2e-5	3e-5	1e-5	1e-4
LR schedule	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
Warmup	10%	10%	10%	10%	10%	10%	10K steps	2.5K steps
Epochs/Steps	5 epochs	3 epochs	3 epochs	40 epochs	10 epochs	10 epochs	20 epochs	100K steps

Table 10: Hyperparameters used for fine-tuning MT6 on the end tasks.

Model	ar	he	vi	id	jv	ms	tl	eu	ml	ta	te	af	nl	en	de	el	bn	hi	mr	ur
MT5	26.5	24.0	60.7	43.5	43.7	49.2	65.2	52.4	13.1	26.4	20.2	58.2	69.4	77.5	63.6	51.7	28.3	37.9	27.2	19.6
MT6	39.6	22.2	63.8	43.7	40.4	54.7	62.9	42.9	14.2	26.4	15.7	58.9	66.0	78.5	67.1	59.6	39.2	47.5	31.8	25.5

Model	fa	fr	it	pt	es	bg	ru	ja	ka	ko	th	sw	yo	my	zh	kk	tr	et	fi	hu	Avg
MT5	15.5	69.8	69.1	67.7	57.6	61.1	49.5	24.1	26.2	23.8	3.0	54.2	56.3	2.8	29.0	23.4	52.8	57.0	62.6	60.9	43.1
MT6	21.7	70.7	65.9	67.8	64.9	65.8	51.6	23.4	25.3	21.9	4.9	65.2	53.6	8.5	26.3	28.6	55.9	49.3	58.2	57.1	44.7

Table 11: Results on WikiAnn named entity recognition.

Model	en	es	de	el	ru	tr	ar	vi	th	zh	hi	Avg
MT5	68.6 / 56.7	50.2 / 35.6	47.2 / 34.1	30.3 / 18.5	41.4 / 28.5	35.9 / 21.9	25.1 / 14.7	48.6 / 31.6	31.7 / 24.6	54.7 / 34.9	29.7 / 18.6	42.1 / 29.1
MT6	74.2 / 62.4	57.8 / 43.1	53.1 / 38.7	41.6 / 28.2	51.1 / 35.6	39.2 / 26.0	40.4 / 25.2	53.6 / 35.2	41.9 / 33.9	61.7 / 45.8	39.8 / 26.0	50.4 / 36.4

Table 12: Results on XQuAD question answering.

Model	en	es	de	ar	hi	vi	zh	Avg
MT5	61.2 / 47.8	41.7 / 27.1	37.8 / 25.4	21.1 / 10.8	22.6 / 13.7	40.5 / 24.2	38.4 / 20.6	37.6 / 24.2
MT6	65.5 / 52.7	47.8 / 32.0	43.2 / 29.8	32.4 / 18.7	31.8 / 20.2	45.0 / 28.3	42.4 / 23.6	44.1 / 29.3

Table 13: Results on MLQA question answering.

Model	en	ar	bn	fi	id	ko	ru	sw	te	Avg
MT5	55.4 / 44.7	35.3 / 18.3	18.4 / 9.2	33.3 / 22.2	37.3 / 24.8	22.6 / 16.9	37.3 / 27.7	25.5 / 13.6	11.2 / 4.5	30.7 / 20.2
MT6	58.1 / 48.0	40.8 / 23.6	24.1 / 14.2	39.7 / 27.3	39.9 / 26.1	26.9 / 18.4	41.9 / 31.4	35.9 / 24.5	16.3 / 10.9	36.0 / 24.9

Table 14: Results on TyDiQA question answering.

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
MT5	75.4	62.0	62.1	58.9	58.9	57.7	59.0	55.7	52.7	58.4	55.0	55.2	53.6	42.4	50.7	57.2
MT6	78.4	70.6	69.8	64.8	65.7	66.6	65.8	61.6	63.3	66.6	63.1	66.2	60.3	51.5	56.9	64.7

Table 15: Results on XNLI natural language inference.

Model	en	fr	de	es	ja	ko	zh	Avg
MT5	91.6	81.2	79.9	80.7	70.7	68.2	73.5	78.0
MT6	93.5	87.0	85.4	87.3	72.4	70.1	79.8	82.2

Table 16: Results on PAWS-X cross-lingual paraphrase adversaries.

Model	**es-en**			**ru-en**			**vi-en**			**tr-en**		
	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
MT5	33.12	11.36	27.32	29.14	8.77	23.29	28.96	8.98	22.77	29.31	10.57	23.44
MT6	33.79	11.83	27.90	30.40	9.49	24.32	29.96	9.52	23.72	29.55	10.80	23.82

Table 17: Evaluation results on Wikilingua cross-lingual abstractive summarization. RG is short for ROUGE. Results of MT5 and MT6 are averaged over three runs.