DOCmT5: Document-Level Pre-training of Multilingual Language Models

Anonymous ACL submission

Abstract

In this paper, we introduce **DOCmT5**, a multilingual sequence-to-sequence language model pre-trained with large scale parallel documents. While previous approaches have focused on leveraging sentence-level parallel data, we try to build a general-purpose pre-trained model that can understand and generate long documents. We propose a simple and effective pre-training objective - **D**ocument **r**eordering **M**achine **T**ranslation (DrMT), in which the input documents that are shuffled and masked need to be translated. DrMT brings consistent improvements over strong baselines on a variety of document-level generation tasks, including over 12 BLEU points for seen-language-pair document-level MT, over 7 BLEU points for unseen-language-pair document-level MT and over 3 ROUGE-1 points for seen-language-pair cross-lingual summarization. We achieve state-of-the-art (SOTA) on WMT20 De-En and IWSLT15 Zh-En document translation tasks. We also conduct extensive analysis on various factors for document pre-training, including (1) the effects of pre-training data quality and (2) The effects of combining mono-lingual and cross-lingual pre-training. We plan to make our model checkpoints publicly available.

1 Introduction

Multilingual pre-trained language models have been useful for a wide variety of NLP tasks. Pre-training on large-scale multilingual corpora facilitates transfer across languages and benefits low-resource languages.

Previously, sentence-level or word-level cross-lingual objectives have been considered for pre-training large language models (LLM), but not much effort has been put in document-level objectives for pre-training. In this work, we propose a multilingual sequence-to-sequence language model pre-trained with cross-lingual structure-aware document-level objectives. DOCmT5 is built on top of mT5 (Xue et al., 2021) and is further trained with parallel documents across multiple language pairs. To encourage the model to gain a deep understanding of the document structure and cross-lingual relationships, we consider a challenging translation scenario as a second-stage pre-training task: the input sentences are shuffled in a random order and random spans are masked. To effectively translate the input document, the model needs to reconstruct the document in the original order, making the model learn sentence relationships, and also recover the masked spans. This objective is effective on document-level generation tasks such as machine translation and cross-lingual summarization, outperforming previous best systems.

To enable cross-lingual pre-training at a large scale, we created a synthetic parallel document corpus. To avoid expensive human annotation, we use off-the-shelf neural machine translation (NMT) models to translate the documents in the mC4 corpus (Xue et al., 2021) into English. In our experimental results, this corpus is more effective for pre-training than existing large-scale automatically aligned corpora (e.g., CCAligned (El-Kishky et al., 2020)).

We also conduct extensive ablation studies and provide insights on document-level pre-training. We show that simple document-level pre-training is more useful than sentence-level pre-training for generative tasks. We also show that data quality matters when performing multilingual document pre-training. Finally, we don't observe improvements from combining mono-lingual and cross-lingual objectives when evaluating on two document-level translation tasks.

In summary, this paper makes the following contributions:

- We build a state-of-the-art multilingual document-level sequence-to-sequence language model pre-trained with a structure-aware cross-lingual objective.

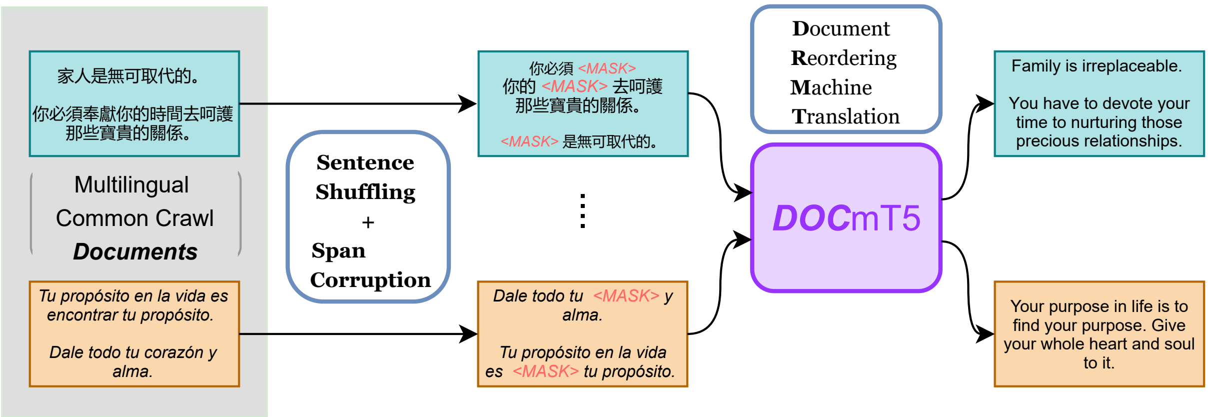

Figure 1: Overview of our proposed **D**ocument-**R**eordering **M**achine **T**ranslation (**DrMT**) pre-training. For each input document, the sentences are shuffled in random order and then randomly selected spans will be masked. The prediction target of DOCmT5 is to generate the translation of the input document.

- Our proposed model achieves strong results on cross-lingual summarization and document-level machine translation for seen and unseen language paris, including SOTA on WMT20 De-En and IWSLT2015 Zh-En tasks.
- We also conduct extensive experiments to study what works and what doesn't work in document-level multilingual pre-training.

2 Related Work

2.1 Multilingual Pre-training

Multilingual pre-trained models provide a set of parameters that can be quickly fine-tuned for different downstream tasks (Ruder et al., 2021). Some popular models are: mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) which pre-train with masked language modeling objective using only monolingual data, mT5 (Xue et al., 2021) and mBART (Liu et al., 2020) which use a sequence-to-sequence language model and pre-train on large-scale mono-lingual corpora across many languages. Our proposed model uses mT5 as a backbone and further utilizes pseudo-parallel documents to learn better cross-lingual representations.

To capture cross-lingual information, translation language modeling (Conneau and Lample, 2019) and its variants (VECO (Luo et al., 2021), ERNIE-M (Ouyang et al., 2021)) was proposed to leverage sentence-level parallel data. AMBER (Hu et al., 2021) use two explicit alignment objectives that align representations at the word and sentence level. HICTL (Wei et al., 2020) pre-trains on parallel sentences with word and sentence-level contrastive losses. mBART50 (Tang et al., 2021), mT6 (Chi et al., 2021) and nmT5 (Kale et al., 2021) focus on second-stage of pre-training using large-scale sentence-level translation data. Our model goes beyond the sentence and focuses on document-level understanding.

While sentence-level pre-training has received a lot of attention, document-level pre-training has been under-studied. Unicoder (Huang et al., 2019) replaces alternating sentences in a document with translations and pre-trains with masked language modeling. MARGE (Lewis et al., 2020) adopts the retriever-generator paradigm and pre-trains with an unsupervised translation objective on automatically retrieved documents. Our model considers a challenging supervised translation objective on parallel documents.

2.2 Multilingual Parallel Data Sources

OPUS-100 (Aharoni et al., 2019; Zhang et al., 2020a) is collected from a variety of domains and is human labeled but it is at the sentence level. ML50 (Tang et al., 2021) is collected from different machine translation challenges and other publicly available corpora such as OPUS, but most of the data is at the sentence level. CCMatrix (Schwenk et al., 2021b) and Wikimatrix (Schwenk et al., 2021a) use multilingual sentence embedding to automatically mine parallel sentences. Perhaps the most closest to our proposed corpus is CCAligned (El-Kishky et al., 2020), which is also automatically mined but its quality is in question (Caswell et al.). Our MTmC4 corpus does not require human annotation and instead was produced by NMT models.

Language	Architecture	Parameters	# Languages	Monolingual Data	Cross-lingual Data	Parallel Docs
mBERT	Encoder-only	180M	104	Wikipedia	✗	✗
RemBERT	Encoder-only	980M	110	Wikipedia and Common Crawl	✗	✗
XLM	Encoder-only	570M	100	Wikipedia	Misc.	✗
XLM-R	Encoder-only	270M - 550M	100	Common Crawl (CCNet)	✗	✗
mBART	Encoder-decoder	680M	25	Common Crawl (CC25)	✗	✗
mBART50	Encoder-decoder	680M	50	Common Crawl (CC25)	ML50	✓
MARGE	Encoder-decoder	960M	26	Wikipedia or CC-News	✗	✗
mT5	Encoder-decoder	300M - 13B	101	Common Crawl (mC4)	✗	✗
nmT5	Encoder-decoder	800M - 3B	101	Common Crawl (mC4)	OPUS-100	✗
DOCmT5 (ours)	Encoder-decoder	580M - 800M	25	Common Crawl (mC4)	MTmC4	✓

Table 1: Comparisons of DOCmT5 to previous multilingual language models.

Language	Size/GB	Language	Size/GB
De⋆	44	Ar	58
Es⋆	52	Az	42
Tr⋆	45	Bn	66
Ru⋆	58	Bn	66
Vi⋆	50	Fa	54
Fi	47	Ko	87
Fr	43	Lt	48
Hi	20	Mr	125
It	40	Nl	38
Ja	120	Pl	45
Pt	40	Th	63
Ro	53	Uk	66
Zh	41		

Table 2: Statistics of the MTmC4 corpus. ⋆ indicates that the language is used in DOCmT5-5.

2.3 Document-level Machine Translation

There are different ways to incorporate document context into translation model. Just to name a few, previous works have explored concatenation-based methods (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019; Sun et al., 2020; Lopes et al., 2020), multi-source context encoder (Zhang et al., 2018; Jean et al., 2017), and hierarchical networks (Zheng et al., 2020; Zhang et al., 2020b; Chen et al., 2020). This line of research focuses on architectural modifications of neural translation models. We focus on how to design a generalized pre-training objective and furthermore, our model can be fine-tuned for various downstream tasks (e.g. summarization) without task-specific changes.

3 Multilingual Pre-training

3.1 Datasets

3.1.1 mC4

For pre-training, we use mC4 (Xue et al., 2021), a large scale corpus extracted from Common Crawl that covers over 100 languages.

3.1.2 MTmC4: Creating Parallel Documents with mC4

To create large-scale parallel documents, we take mC4 as a starting point and use in-house NMT models to translate documents from 25 languages into English. Each sentence in each document is translated independently. For each language, we sample 1 million documents, if there are more than that to start with, in mC4. Detailed data statistics for all the languages can be found in Table 2.

3.2 Document Reordering Machine Translation (DrMT)

We start by introducing two related pretraining objectives:

- *NMT Pre-training*: Tang et al. (2021) and Kale et al. (2021) proposed to perform a second-stage of pre-training using sentence-level MT data. The objective here is to perform sentence-level translation without any other changes to the input.

- *Monolingual Document Reordering (Dr) Pre-training*: This objective, proposed by mBART (Liu et al., 2020), changes the order of the sentences in each document. This is then followed by the original span corruption objective in T5. The decoder is required to generate the original document in order.

We combine these two objectives and propose **DrMT**. In DrMT, we introduce two types of noise on the input: **(i)** sentences in the document are randomly shuffled and **(ii)** randomly sampled spans are masked. In order to correctly translated the content, the model needs to decipher the corrupted document in order first. This enforces the models to gain deep understanding of the document structure. More formally, suppose we have N language pairs and each language has a set of parallel documents, the whole collection of document pairs are $D = \{D_1, D_2, ..., D_N\}$. And a pair of (x, y) is an instance in one of the language documents D_i. The overall learning objective is maximizing the likelihood of y given a corrupted $C(x)$, that is

$$\sum_{D_i \in D} \sum_{(x,y) \in D_i} \log P(y|C(x)). \quad (1)$$

3.3 DOCmT5

We use mT5 as the backbone model. mT5 is a sequence-to-sequence language model pre-trained with the span corruption objective in which random spans in the input are masked and the decoder is required to reconstruct the masked spans (see Raffel et al. (2020) and Xue et al. (2021) for further details). Our system, DOCmT5, incorporates a second-stage pre-training with a structure-aware cross-lingual objective(3.2) on pseudo parallel documents. Detailed comparisons with previous multilingual language models can be found in Table 1. We provide two variants of DOCmT5 with both Base and Large model settings:

- **DOCmT5-5** This model is pre-trained with 5 languages: {De, Ru, Tr, Vi and Es}. For all of the pre-training objective baselines in this paper, we pre-train with this set of languages, unless specified otherwise.
- **DOCmT5-25** This model is pre-trained with 25 languages. We show the full list of languages and their sizes in Table 2.

3.4 Implementation Details

We use mT5-Base[1] and mT5-Large[2] checkpoints at 1M steps as our pre-trained models. We perform a second-stage of pre-training for an additional 0.5M steps using batches of 256 examples each of max length 1024. The learning rate is determined by a inverse square root scheduler as defined in T5, with the learning rate set to $1/\sqrt{n}$ where n is the number of training step. We use the same span corruption objective as T5, with 15% of random tokens masked and an average noise span length of 3. For fine-tuning, we use a constant learning rate of 0.001 and dropout rate of 0.1 for all tasks until convergence. We adopt greedy decoding during inference.

[1]https://console.cloud.google.com/storage/browser/t5-data/pretrained_models/mt5/base/

[2]https://console.cloud.google.com/storage/browser/t5-data/pretrained_models/mt5/large/

4 Experiments

4.1 Baselines

- **Second-Stage Pre-training on 5 Languages** Language models pre-trained with huge numbers of languages suffer from curse of multilinguality. In order to make a fair comparison, we create a strong mT5 model by continuing to pre-train on the same 5 languages of mC4 as in DOCmT5-5 with the same number of steps using the original span corruption objective in mT5. Models pre-trained with this objective is denoted as **cont-5langs**.

- **Monolingual Document Reordering (Dr)** We briefly mention this objective in Section3.2. We use the mC4 corpus for this pre-training objective. Models pre-trained with this objective is denoted as **Dr** (**D**ocument **R**eordering).

- **Document TLM (DocTLM)** In Conneau and Lample (2019), the authors propose the translation language modeling(TLM) objective, which concatenates parallel sentences and applies masked language modeling to learn cross-lingual knowledge. Here we extend it to the document level by concatenating parallel documents. Instead of masking single tokens, we follow the span corruption objective in T5 and mask consecutive spans. The models are pretrained with this objective on MTmC4.

- **Document NMT (DocNMT)** We consider a standard document-level machine translation for pre-training. The source document is the input and the target translation is the output. We use MTmC4 for this pre-training objective.

4.2 Cross-Lingual Summarization

We evaluate *DOCmT5* on cross-lingual summarization as it is challenging for the model to summarize a long document and translate the salient information at the same time. We use Wikilingua, a cross-lingual summarization dataset, in which a document from a arbitrary language must be summarized in English. We adopt the GEM (Gehrmann

Pre-trained Model	Es-En	Ru-En	Tr-En	Vi-En	Average
			Previous Systems		
mBART	**38.30 / 15.40 / 32.40**	33.10 / 11.90 / 27.80	34.40 / 13.00 / 28.10	32.00 / 11.10 / 26.40	34.45 / 12.85 / 28.67
			Mono-Lingual		
mT5	29.97 / 10.65 / 25.70	27.91 / 8.90 / 22.60	29.98 / 11.96 / 24.56	24.38 / 7.39 / 19.59	28.06 / 9.72 / 23.11
w. cont-5langs	34.50 / 12.83 / 28.37	30.20 / 10.30 / 24.77	32.12 / 13.71 / 26.40	28.95 / 9.74 / 23.76	31.44 / 11.64 / 25.82
w. Dr	36.22 / 14.18 / 30.31	32.29 / 11.64 / 26.63	34.25 / 14.93 / 28.50	30.07 / 10.46 / 25.00	33.20 / 12.80 / 27.61
			Cross-Lingual		
w. DocNMT	33.45 / 12.56 / 29.04	30.93 / 11.01 / 25.82	33.32 / 14.10 / 27.54	27.60 / 9.26 / 22.52	31.40 / 11.59 / 26.12
w. DocTLM	35.40/ 13.76 / 29.71	30.26 / 10.33 / 24.78	34.85 / 15.35 / 28.88	30.35 / 10.86 / 25.03	32.71 / 12.57 / 27.10
DOCmT5-5	36.60 / 14.55 / 30.64	32.90 / 12.09 / 27.41	37.02 / 16.64 / 30.97	32.13 / 11.81 / 26.72	34.66 / 13.77 / 28.93
DOCmT5-5-Large	36.34 / 14.69 / 31.14	33.15 / 12.32 / 27.80	37.11 / 16.40 / 30.63	**33.29 / 12.35 / 27.50**	34.97 / 13.94 / 29.26
DOCmT5-25	36.42 / 14.47 / 30.51	30.99 / 10.94 / 25.78	35.99 / 16.13 / 29.67	31.71 / 11.53 / 26.40	33.77 / 13.26 / 28.09
DOCmT5-25-Large	36.79 / 15.04 / 31.48	**33.56 / 12.77 / 28.46**	**37.66 / 16.68 / 31.37**	32.43 / 11.87 / 27.04	**35.11 / 14.09 / 29.58**

Table 3: Results of four seen langauges paris {Es, Tr, Ru, Vi} on Wikilingua. Each cell demonstrates three metrics: ROUGE-1, ROUGE-2 and ROUGE-L in order. The mBART results are taken from the GEM(Gehrmann et al., 2021) paper for a strong baseline model.

Pre-trained Model	Fr-En	Id-En	Hi-En	Average
		Mono-Lingual		
mT5	29.66 / 9.96 / 24.37	29.08 / 9.87 / 23.83	26.18 / 8.51 / 20.91	28.30 / 9.44 / 23.03
w. cont-5langs	32.78 / 11.79 / 27.29	32.21 / 11.65 / 26.36	28.93 / 10.06 / 23.37	31.30 / 11.16 / 25.67
w. Dr	34.47 / 12.67 / 28.58	34.05 / 12.87 / 27.96	31.13 / 11.18 / 25.16	33.21 / 12.24 / 27.23
		Cross-Lingual		
w. DocNMT	33.22 / 12.33 / 27.97	31.97 / 11.80 / 27.11	29.33 / 10.12 / 23.86	31.50 / 11.41 / 26.31
w. DocTLM	32.79 / 11.75 / 27.12	33.35 / 12.24 / 27.37	30.48 / 11.24 / 24.92	32.20 / 11.74 / 26.47
DOCmT5-5	34.02 / 12.57 / 28.21	34.31 / 13.09 / 28.56	32.24 / 11.84 / 26.06	33.52 / 12.50 / 27.61
DOCmT5-5-Large	**36.28 / 14.27 / 30.78**	34.52 / 13.45 / 29.22	33.15 / 12.68 / 27.35	34.65 / 13.46 / 29.11
DOCmT5-25	34.56 / 13.10 / 29.03	34.16 / 13.04 / 28.23	32.33 / 11.99 / 26.25	33.68 / 12.71 / 27.83
DOCmT5-25-Large	35.66 / 13.99 / 30.26	**35.15 / 13.70 / 29.47**	**34.16 / 13.26 / 27.93**	**34.99 / 13.65 / 29.22**

Table 4: Results of three unseen langauges paris {Fr, Id, Hi} on Wikilingua.

et al., 2021) version where the data is re-split to avoid train-test overlap between languages. We use a special prefix for cross-lingual summarization: *"Summarize X to Y"*, where X and Y are the source and target language names respectively.

4.2.1 Results on Seen Language Pairs

We show the fine-tuning results of language pairs that are in the second stage of pre-training in Table 3. We use the same four languages that are in Wikilingua's original release {Es, Ru, Tr, Vi}. The *Dr* objective brings significant improvements over *cont-5langs* in all four languages, justifying the importance of structure-aware objectives. As for cross-lingual objectives, *DocTLM* is better than *DocNMT* in almost all languages except for Russian. DOCmT5-5 significantly outperforms *DocNMT* and *DocTLM*, showing that our proposed pre-training objective leads to improved cross-lingual learning. The results of *DOCmT5-25* are inferior to *DOCmT5-5* and this is possibly due to capacity dilution (Arivazhagan et al., 2019). As we increase the capacity, we see that *DOCmT5-25-Large* outperforms *DOCmT5-5-Large*. *DOCmT5-25-Large* is the best overall model outperforming the strong prior system: mBART.

4.2.2 Results on Unseen Language Pairs

We show the fine-tuning results of language pairs that are not in the second-stage of pre-training stage in Table 4. We use three languages {Fr, Id, Hi}[3]. Once again, we see that the *Dr* objective brings significant improvements over *cont-5langs*. Surprisingly, without directly pre-training on the same

[3] We choose French to study the transfer ability of the cross-lingual models on high-resource and same-script (latin) languages. Indonesian is for studying high-resource and different-script language. Hindi is for studying low-resource and different-script language.

language pairs, *DOCmT5-5* leads to significant improvements over strong baselines. This shows that our pre-training objectives are able to generalize to other languages. *DOCmT5-25* pre-trains on French and Hindi but not Indonesian and hence we observe improvements of average results over *DOCmT5-5*. The improvements of *DOCmT5* are not so significant and sometimes even hurt performance in high-resource languages: French and Indonesian, which have 44556 and 33237 training examples respectively and there are only 6942 examples in Hindi. *DOCmT5-25-Large* obtains the best results in almost all 3 languages except for French.

Pre-trained Model	**d-BLEU**
Previous Systems	
NTT (Kiyono et al., 2020)	43.80
PROMT (Molchanov, 2020)	39.60
OPPO (Shi et al., 2020)	42.20
Mono-Lingual	
mT5	29.08
w. cont-5langs	32.24
w. Dr	36.71
Cross-Lingual	
w. DocNMT	41.23
w. DocTLM	37.74
DOCmT5-5	42.19
DOCmT5-5-Large	**44.73**
DOCmT5-25	40.99
DOCmT5-25-Large	43.49

Table 5: Fine-tuning results on WMT20 De-En.

Pre-trained Model	**d-BLEU**
Previous Systems	
HAN	24.00
mBART	29.60
MARGE	28.40
Mono-Lingual	
mT5	24.24
w. cont-5langs	24.22
w. Dr	23.75
Cross-Lingual	
w. DocNMT	26.17
w. DocTLM	25.87
DOCmT5-5	28.97
DOCmT5-5-Large	30.52
DOCmT5-25	30.99
DOCmT5-25-Large	**31.40**

Table 6: Unseen language pair results on IWSLT 2015 Zh-En. Chinese is in the second-stage pre-training language set of DOCmT5-25 but not in those of DOCmT5-5. DOCmT5-25-Large achieves SOTA.

4.3 Document-Level Machine Translation

We evaluate DOCmT5 on document translation. We split each document into chunks with a max length of 512 tokens. During inference, the decoded chunks are concatenated together to form the final document. We use prefix *"Translate X to Y"* for translation, where X and Y are the source and target language names respectively.

4.3.1 Seen Language Pair: WMT20 De-En

WMT20 De-En is a document-level machine translation task. We use parallel training data from WMT20 without using additional monolingual data. From the results in Table 5[4], we see that *Dr* provides large gains. *DocNMT* outperforms *DocTLM*. This is probably due to the fact that *DocNMT* is more close to the document-level translation task. *DOCmT5-5* once again outperforms Dr and other strong cross-lingual baselines. *DOCmT5-5* is better than *DOCmT5-25* again because of capacity dilution as noted in Aharoni et al. (2019). As expected, *DOCmT5-5-Large* outperforms *DOCmT5-5* and to the best of our knowledge, achieves the SOTA. Note that previous systems use one or more of the following techniques: additional monolingual data, back-translation, ensembling or re-ranking tailored to a single translation pair.

4.3.2 Unseen Language Pair: IWSLT 2015 Zh-En

We use IWSLT 2015 Zh-En, another document-level machine translation task, to examine the multilingual transferability of *DOCmT5* when the target transfer language (Chinese in this case) is of a very different script. Chinese is only in the first-stage pre-training of mT5 but not in our second-stage pre-training. We use parallel training data from IWSLT15 without using additional monolingual data. Following HAN (Werlen et al., 2018), we use 2010-2013 TED as the test set. The results are in Table 6. *DOCmT5-5* outperforms the strong cross-lingual and mono-lingual baselines, demonstrating impressive transfer capability . *DOCmT5-25* includes Chinese as one of the second-stage pre-training languages therefore obtains better numbers than *DOCmT5-5*. Unsurprisingly, large models are better than their corresponding base models. To the best of our knowledge, *DOCmT5-25-Large* achieves the SOTA on this task. We qualitatively

[4]For all the document translation experiments in this paper, the numbers are calculated using sacreBLEU https://github.com/mjpost/sacrebleu in document level.

analyze the translations of different systems in Appendix A.

Pre-trained Model	De-En	Ru-En	Pl-En	Ja-En
mT5				
w. DocNMT	44.09	40.48	3.13	0.92
w. DocTLM	0.31	0.11	0.23	0.22
DOCmT5-5	21.74	15.84	2.81	0.47
DOCmT5-5-Large	35.63	29.50	14.15	1.16
DOCmT5-25	22.00	14.62	17.40	16.93
DOCmT5-25-Large	28.24	24.34	23.18	19.17

Table 7: Document translation without finetuning on WMT20 De-En, Ru-En, Pl-En and Ja-En.

4.3.3 Document Translation Without Fine-tuning

We further show that *DOCmT5* is able to perform document translation without fine-tuning, i.e., evaluate the model right after second-stage pre-training without any fine-tuning on task-specific data. We show the results in Table 7. While the monolingual pre-trained models completely fail to produce meaningful translations, *DOCmT5-5* is able to achieve over 20 BLEU points in De-En and 15 in Ru-En. Not surprisingly, *DOCmT5-5-Large* further improves to over 35 and 29 respectively. *DOCmT5-25* includes Pl-En and Ja-En in the second-stage pre-training and therefore obtains competitive results on these two language pairs with either base or large model. Although *DOCmT5-5* is not pre-trained on Pl-En, the large model gets over 14 BLEU on this task. One hypothesis is that Polish uses the Latin script and shares common subwords with German and Spanish, allowing our model to transfer knowledge across languages. On the other hand, the *DOCmT5-5-Base* model fails to produce meaningful translations for Pl-En. This shows the importance of size when performing multilingual pre-training. The best model is *DocNMT* which obtains over 40 BLUE points in both De-En and Ru-En, outperforming *DOCmT5-5* and *DOCmT5-25*. This is reasonable because *DOCmT5* shuffles documents in pre-training and this is misaligned with the document translation task inputs. The impressive performance of both *DocNMT* and *DOCmT5* shows that our MTmC4 corpus is of very high-quality and is likely better than the parallel data provided by the specific tasks in question. Further analysis of the quality of this data will be an interesting avenue for future work.

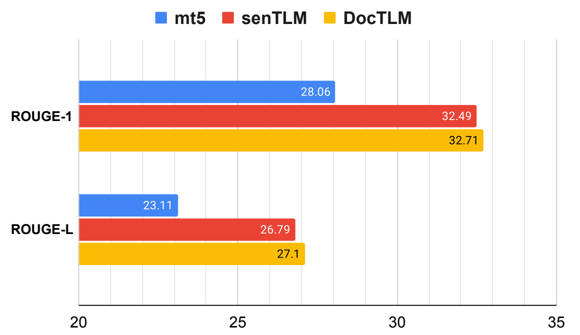

Figure 2: SenTLM and DocTLM fine-tuning results on Wikilingua. The numbers are average of four languages: {Es, Tr, Ru, Vi}.

5 Analysis

5.1 Are Document-level Models Better Than Sentence-level Models?

To demonstrate the benefits of pre-training with longer context, we pre-train mT5 using translation language modeling (TLM) on five languages: {De, Es, Tr, Vi, Ru} with two different inputs. In *DocTLM*, we concatenate the parallel documents into a single training sequence. As for *SenTLM*, we break down the document into individual sentences and find the alignments in the parallel document pair. Then we concatenate the single aligned sentence pair as a training sequence. We fine-tune these second-stage pre-trained models on Wikilingua and WMT20 De-En. The results are shown in Figure 2 and Table 8. We see that document-level models offer small improvements on summarization and very significant improvements on document-level translation, showing that the longer context is indeed useful.

Pretrained-Model	BLEU
mT5	29.08
w. SenTLM	34.68
w. **DocTLM**	**37.74**

Table 8: SenTLM and DocTLM fine-tuning results on WMT20 De-En.

5.2 Effect of Data Quality in Second-stage Pre-training

In our experiments, we observe big differences between different parallel corpora. We compare against the CCAligned corpus – a large automatically mined corpus from Common Crawl which is found to be very noisy (Caswell et al.). In contrast, MTmC4 is produced by using high-quality

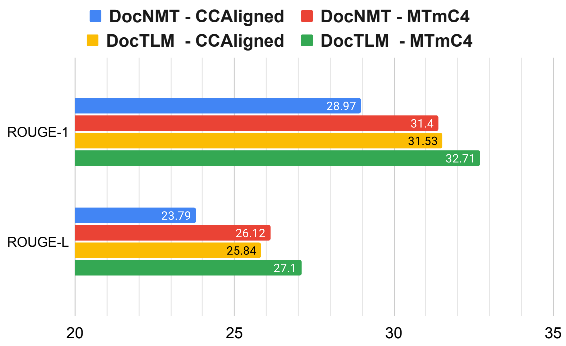

Figure 3: MTmC4 and CCAlgined fine-tuning results on Wikilingua. The numbers are average of four languages: {Es, Tr, Ru, Vi}.

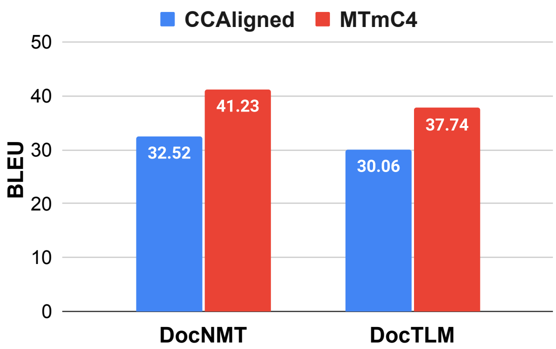

Figure 4: MTmC4 and CCAlgined fine-tuning results on WMT20 De-En.

translation systems. We pre-train mT5-Base on five languages: {De, Es, Tr, Vi, Ru} with these two corpora using *DocNMT* and *DocTLM*. We demonstrate the Wikilingua results in Figure 3 and WMT20 De-En results in Figure 4. Using our curated MTmC4 is consistently better regardless of pre-training objectives or tasks.

5.3 Does Combining Mono-Lingual and Cross-Lingual Pre-training Help?

Here we try to see if combining both monolingual and cross-lingual objectives helps. We try two different continual pre-training strategies for combining Dr and DrMT. We use five languages: {De, Ru, Tr, Vi, Es}. **(i)** Dr → DrMT: We first pre-train mT5 with Dr on mC4 for 0.5M steps and then pre-train with DrMT on MTmC4 for 0.5M steps. **(ii)** Dr + DrMT: We mix these two objectives with a 50-to-50% ratio and pre-train for 0.5M steps. In Table 9, we show that **(i)** slightly improves over only DrMT in both tasks and **(ii)** slightly improves on WMT20 De-En but seems to hurt performance on ISWLT15 Zh-En.

Pretrained-Model	WMT20 De-En	IWSLT15 Zh-En
mT5		
w. Dr	36.63	23.75
w. DrMT	42.05	28.00
***w.* Dr → DrMT**	**42.75**	**28.18**
w. Dr + DrMT	42.37	27.35

Table 9: Methods of combining mono-lingual and cross-lingual and their fine-tuning results on WMT20 De-En and IWSLT15 Zh-En.

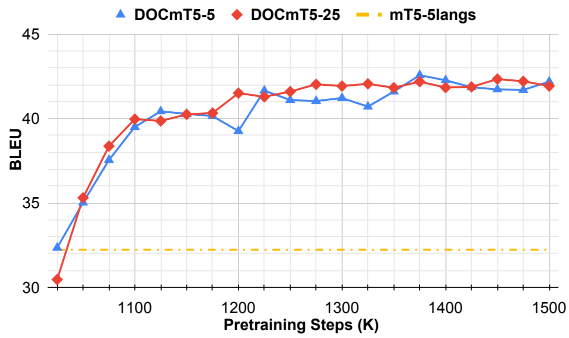

Figure 5: Fine-tuning results of WMT20 De-En along with pretraining steps. We use DOCmT5-5-base.

5.4 How Many Pre-training Steps is Required for DrMT?

To answer this question, we take different pre-training checkpoints of *DOCmT5-5* and *DOCmT5-25* and fine-tune with WMT20 De-En. The results are shown in Figure 5. After 50k steps of pre-training with *DrMT*, both systems outperform the *cont-5langs*. After 300k steps, both systems roughly converge and perform similarly.

6 Conclusion

In this paper, we present DOCmT5, a novel document-level multilingual pre-trained model. Our proposed objective, DrMT, is simple and effective and leads to large gains over strong baselines (e.g. mBART and MARGE) on cross-lingual summarization and document-level translation. DOCmT5 achieved SOTA on two competitive document-level translation tasks: WMT20 De-En and IWSLT15 Zh-En. We further analyze various factors that contribute to successful document-level pre-training. We plan to release the pre-trained model to facilitate future work on document-level language understanding.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884.

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*.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. 2021. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

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.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, et al. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643.

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 (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233.

Mihir Kale, Aditya Siddhant, Rami Al-Rfou, Linting Xue, Noah Constant, and Melvin Johnson. 2021. nmT5 - is parallel data still relevant for pre-training massively multilingual language models? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 683–691, Online. Association for Computational Linguistics.

Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. 2020. Tohoku-aip-ntt at wmt 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 145–155.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

António Lopes, M Amin Farajian, Rachel Bawden, Michael Zhang, and André T Martins. 2020. Document-level neural mt: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. Veco: Variable and flexible cross-lingual pre-training for language understanding and generation. 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 3980–3994.

Alexander Molchanov. 2020. Promt systems for wmt 2020 shared news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 248–253.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. 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*, 21:1–67.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and Jie Hao. 2020. Oppo's machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 282–292.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2020. Capturing longer context for document-level neural machine translation: A multi-resolutional approach. *arXiv preprint arXiv:2010.08961*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On learning universal representations across languages. In *International Conference on Learning Representations*.

Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020b. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *IJCAI*.

Appendices

A Analysis of Document Translation

We take a deeper look at the translations produced by various systems to understand what makes *DOCmT5* better. We demonstrate an example in Table 6. We take the best system (*DOCmT5-25-Large*) and the *cont-5langs* baseline. We observe that *DOCmT5* uses time tenses better than the baseline, producing more coherent sentences (red-colored texts). Additionally, *DOCmT5* handles a compositional sentence more elegantly, instead of just using "and" (blue-colored texts). Finally, we observe that *cont-5langs* often makes minor translation mistakes while our *DOCmT5* makes much fewer of them.

Source Document	我11岁那年， 记得有一天早晨醒来，听见家里有愉悦的声音。 我的父亲在用他的灰色小收音机 听BBC新闻。 他面带笑容，这很少见， 因为大部分的新闻都只会使他沮丧。 "塔利班走了！" 父亲大声叫着。 我不知道那意味着什么， 但是我能看出父亲非常非常高兴。 "你现在可以去个真正的学校念书了。" 他说。 我永远不会忘记那个早晨。 一个真正的学校。 我6岁那年，塔利班占领阿富汗 并规定女孩上学是违法的。 所以在那之后5年，我女扮男装 陪着我姐姐去一个秘密学校 姐姐那时已经不被允许独自外出了。 这是我们俩人唯一的受教育方式。 我们每天要走不同的路线 这样才没有人会怀疑我们要去哪里。 ...
DOCmT5-25 translation	And when I was 11 years old, I remember waking up one morning to the sound of a happy voice in the house. My father was listening to the BBC on his little gray radio. He had a smile on his face, which is rare, because most of the news was depressing. "The Taliban are gone!" My father shouted. I didn't know what that meant, but I could see that my father was very, very happy. "You can go to a real school now," he said. And I will never forget that morning. A real school. When I was six years old, the Taliban occupied Afghanistan and made it illegal for girls to go to school. So for the next five years, I was a woman in a man's suit and went to a secret school with my sister, who was not allowed to go out alone. This was the only way we were educated. We had to go in different directions every day so no one would suspect where we were going. ...
mT5 translation	And I was 11 years old, and I remember awakefully waking up in the morning and hearing the familiar sound. My father was listening to the BBC news on his little radio. He was smiling, and it was rare, because most of the news was going to frustrate him. "Taliban go." The father went out. I don't know what that meant, but I can see that the father was very, very happy. "You can go to a real school now." He said. I'll never forget that morning. A real school. And I was six years old, and Taliban took Afghanistan and banned girls' schooling. So five years after that, my chick went to a secret school with my sister. And she wasn't allowed to go on a trip. It was the only way that we were educated. We walked on different roads every day so that nobody could suspect where we were. ...
Target Translation	When I was 11, I remember waking up one morning to the sound of joy in my house. My father was listening to BBC News on his small, gray radio. There was a big smile on his face which was unusual then, because the news mostly depressed him. "The Taliban are gone!" my father shouted. I didn't know what it meant, but I could see that my father was very, very happy. "You can go to a real school now," he said. A morning that I will never forget. A real school. You see, I was six when the Taliban took over Afghanistan and made it illegal for girls to go to school. So for the next five years, I dressed as a boy to escort my older sister, who was no longer allowed to be outside alone, to a secret school. It was the only way we both could be educated. Each day, we took a different route so that no one would suspect where we were going. ...

Figure 6: A comparison example of Zh-En document translation. DOCmT5 is able to produce consistent time tenses while mT5 baseline fails. DOCmT5 also produces longer and conherent sentences. Best viewed in color.