An Empirical Study of Document-to-document Neural Machine Translation

Anonymous ACL submission

Abstract

This paper does not aim at introducing a novel method for document NMT. Instead, we head back to the original Transformer model with document-level training and hope to answer the following question: Is the capacity of current models strong enough for document-level NMT? Interestingly, we observe that the original transformer with appropriate training techniques can achieve strong results for document translation, even with a length of 2000 words. We evaluate this model and several recent approaches on nine document-level datasets and two sentence-level datasets across six languages. Experiments show that the original Transformer model outperforms sentence-level models and many previous methods in a comprehensive set of metrics, including BLEU, four lexical indices, three newly proposed assistant linguistic indicators, and human evaluation.

1 Introduction

Neural machine translation (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017) has achieved great progress and reached near human-level performance. However, most current sequence-to-sequence NMT models translate sentences individually. In such cases, discourse phenomena, such as pronominal anaphora, lexical consistency, and document coherence that depend on long-range context going further than a few previous sentences, are neglected (Bawden et al., 2017). As a result, Läubli et al. (2018) find human raters still show a markedly stronger preference for human translations when evaluating at the level of documents.

Many methods have been proposed to improve document-level neural machine translation (DNMT). Among them, the mainstream works focus on the model architecture modification, including hierarchical attention (Wang et al., 2017; Miculicich et al., 2018; Tan et al., 2019), additional context extraction encoders or query layers (Jean et al., 2017; Bawden et al., 2017; Zhang et al., 2018; Voita et al., 2018; Kuang and Xiong, 2018; Maruf et al., 2019; Yang et al., 2019; Jiang et al., 2019; Zheng et al., 2020; Yun et al., 2020; Xu et al., 2020), and cache-like memory network (Maruf and Haffari, 2018; Kuang et al., 2018; Tu et al., 2018).

These studies come up with different structures in order to include discourse information, specifically speaking, introducing adjacent sentences into the encoder or decoder as document contexts. Experimental results show effective improvements on universal translation metrics like BLEU (Papineni et al., 2002) and document-level linguistic indices (Tiedemann and Scherrer, 2017; Bawden et al., 2017; Werlen and Popescu-Belis, 2017; Müller et al., 2018; Voita et al., 2018, 2019).

Unlike previous works, this paper does not aim at introducing a novel method. Instead, we hope to answer the following question: Is the basic sequence-to-sequence model strong enough to directly handle document-level translation? To this end, we head back to the original Transformer (Vaswani et al., 2017) and conduct literal document-to-document (Doc2Doc) training. We leverage the full document information, making the model capture the full context in both source and target sides.

Though many studies report less promising results of naive Doc2Doc translation (Zhang et al., 2018; Liu et al., 2020), we successfully activate it with **Multi-resolutional Training**, which involves multiple levels of sequences. It turns out that end-to-end document translation is not only feasible but also better functioning than sentence-level models and previous works. Furthermore, if assisted by extra sentence-level corpus, which can be much more easily obtained, the model can significantly improve the translation performance and achieve state-of-the-art results. It is worth noting that we do not change the model architecture and need no extra parameters.

Our experiments are conducted on nine

document-level datasets, including TED (ZH-EN, EN-DE), News (EN-DE, ES-EN, FR-EN, RU-EN), Europarl (EN-DE), Subtitles (EN-RU), and a newly constructed News dataset (ZH-EN). Additionally, two sentence-level datasets are adopted in further experiments, including Wikipedia (EN-DE) and WMT (ZH-EN). Experiment results show that our strategy outperforms previous methods in a comprehensive set of metrics, including BLEU, four lexical indices, three newly proposed assistant linguistic indicators, and human evaluation. In addition to serving as improvement evidence, our newly proposed document-level datasets and metrics can also be a boosting contribution to the community.

2 Doc2Doc: End-to-End DNMT

In this section, we attempt to analyze the different training patterns for DNMT. Firstly, let us formulate the problem. Let $D_x = \{x^{(1)}, x^{(2)}, \cdots, x^{(M)}\}$ be a source-language document containing M source sentences. The goal of the document-level NMT is to translate the document D_x in language x to a document D_y in language y. $D_y = \{y^{(1)}, y^{(2)}, \cdots, y^{(N)}\}$. We use $L_y^{(i)}$ to denote the sentence length of $y^{(i)}$.

Previous works translate a document sentence-by-sentence, regarding DNMT as a step-by-step sentence generating problem (Doc2Sent) as:

$$\mathcal{L}_{\text{Doc2Sent}} = -\sum_{i=1}^{N}\sum_{j=1}^{L_y^{(i)}} \log p_\theta(y_j^{(i)}|y_{(<j)}^{(i)}, x^{(i)}, S^{(i)}, T^{(i)}), \tag{1}$$

$S^{(i)}$ is the context in the source side, depending on the model architecture and is comprised of only two or three sentences in many works. Most current works focus on $S^{(i)}$, by utilizing hierarchical attention or extra encoders. And $T^{(i)}$ is the context in the target side, which is involved by only a couple of works. They usually make use of a topic model or word cache to form $T^{(i)}$.

Different from Doc2Sent, we propose to resolve document translation with the end-to-end, namely document-to-document (Doc2Doc) pattern as:

$$\mathcal{L}_{\text{Doc2Doc}} = -\sum_{i=1}^{\sum L_y} \log p_\theta(y_i|y_{<i}, D_x), \tag{2}$$

where D_x is the complete context in the source side, and $y_{<i}$ is the complete historical context in the target side.

2.1 Why We Dive into Doc2Doc?

Full Source Context: Firstly, though some Doc2Sent works utilize a full source-side context (Maruf and Haffari, 2018; Maruf et al., 2019; Tan et al., 2019), some studies show that more sentences beyond can harm the results (Miculicich et al., 2018; Zhang et al., 2018; Tu et al., 2018). Therefore, many works of Doc2Sent are more of "a couple of sentences to sentence" since they only involve two or three preceding sentences as context. However, broader contexts provide more information, which shall lead to more improvements. We attempt to re-visit involving the full context. Correspondingly, we pick Doc2Doc, as it is required to take account of all the source-side context.

Full Target Context: Secondly, though some Doc2Sent works utilize a full target-side context (Maruf and Haffari, 2018; Zheng et al., 2020), many previous works abandon the target-side historical context, and some even claim that it is harmful to translation quality (Wang et al., 2017; Zhang et al., 2018; Tu et al., 2018). Whether to utilize target-side contexts is controversial for Doc2Sent. However, once the cross-sentence language model is discarded, some problems, such as tense mismatch (especially when the source language is tenseless like Chinese), may occur. We attempt to re-visit involving the full context. Correspondingly, we pick Doc2Doc, as it treats the whole document as a sequence and can naturally take advantage of all the target-side historical context.

Loose Training: Thirdly, Doc2Sent restricts the training scene. The previous works focus on adjusting the model structure to feed preceding source sentences, so the training data has to be in the form of consecutive sentences so as to meet the model entrance. As a result, it is hard to use large numbers of piecemeal parallel sentences. Such a rigid form of training data also greatly hinders the model potential because the scale of parallel sentences can be tens of times of parallel documents. On the contrary, Doc2Doc can naturally absorb all kinds of sequences, including sentences and documents.

Simplicity: Lastly, Doc2Sent inevitably introduces extra model modules with extra parameters in order to capture contextual information. It complicates the model architecture, making it hard to renovate or generalize. On the contrary, Doc2Doc does not change the model structure and brings in no additional parameters.

Group	Datasets	Source	Language	N_Sent	N_Doc	Development Sets	Test Sets
Main Experiments	TED	IWSLT 2015	ZH-EN	205K	1.7K	dev2010	tst2010-2013
	TED	IWSLT 2017	EN-DE	206K	1.7K	dev2010+tst201[0-5]	tst2016-2017
	News	News Commentary v11	EN-DE	236K	6.1K	newstest2015	newstest2016
	Europarl	Europarl v7	EN-DE	1.67M	118K	(Maruf et al., 2019)	
Other Languages	News	News Commentary v14	ES-EN	355K	9.2K	newstest2012	newstest2013
	News	News Commentary v14	FR-EN	303K	7.8K	newstest2013	newstest2014
	News	News Commentary v14	RU-EN	226K	6.0K	newstest2018	newstest2019
Sentence-level Corpus	Wiki	Wikipedia	EN-DE	2.40M	-	-	-
	WMT	WMT 2019	ZH-EN	21M	-	-	-
Contrastive Experiments	Subtitles	OpenSubtitles	EN-RU	6M	1.5M	(Voita et al., 2019)	
Our New Datasets	PDC	FT/NYT	ZH-EN	1.39M	59K	newstest2019	released soon

Table 1: The detailed information of the used datasets in this paper with downloading links on their names.

2.2 Multi-resolutional Doc2Doc NMT

Although Doc2Doc seems more concise and promising in multiple terms, it is not widely recognized. Zhang et al. (2018); Liu et al. (2020) conduct experiments by directly feeding the whole documents into the model. We refer to it as **Single-resolutional Training** (denoted as SR Doc2Doc). Their experiments report extremely negative results unless pre-trained in advance. The model either has a large drop in performance or does not work at all. As pointed out by Koehn and Knowles (2017), one of the six challenges in neural machine translation is the dramatic drop of quality as the length of the sentences increases.

However, we find that Doc2Doc can be activated on any datasets and obtain better results than Doc2Sent models as long as we employ **Multi-resolutional Training**, mixing documents with shorter segments like sentences or paragraphs (denoted as MR Doc2Doc).

Specifically, we split each document averagely into k parts multiple times and collect all the sequences together, $k \in \{1, 2, 4, 8, ...\}$. For example, a document containing eight sentences will be split into two four-sentences segments, four two-sentences segments, and eight single-sentence segments. Finally, fifteen sequences are all gathered and fed into sequence-to-sequence training $(15 = 1 + 2 + 4 + 8)$.

In this way, the model can acquire the ability to translate long documents since it is assisted by easier and shorter segments. As a result, multi-resolutional Doc2Doc is able to translate all forms of sequences, including extremely long ones such as a document with more than 2000 tokens, as well as shorter ones like sentences. In the following sections, we conduct the same experiments as the aforementioned studies by translating the whole document directly and atomically.

3 Experiment Settings

3.1 Datasets

For our main experiments, we follow the datasets provided by Maruf et al. (2019) and Zheng et al. (2020), including *TED* (ZH-EN/EN-DE), *News* (EN-DE), and *Europarl* (EN-DE). The Chinese-English and English-German TED datasets are from IWSLT 2015 and 2017 evaluation campaigns, respectively. For ZH-EN, we use dev2010 as the development set and tst2010-2013 as the test set. For TED (EN-DE), we use tst2016-2017 as the test set and the rest as the development set. For News (EN-DE). the training/develop/test sets are: News Commentary v11, WMT newstest2015, and WMT newstest2016. For Europarl (EN-DE). The corpus is extracted from the Europarl v7 according to the method proposed in Maruf et al. (2019). [1]

Experiments on Spanish, French, Russian to English are also conducted, whose training sets are News Commentary v14 , with the development sets and test sets are newstest2012 / newstest2013 (ES-EN), newstest2013 / newstest2014 (FR-EN), newstest2018 / newstest2019 (RU-EN), respectively.

Besides, two additional sentence-level datasets are also adopted. For EN-DE, we use *Wikipedia* , a corpus containing 2.4 million pairs of sentences. For ZH-EN, we extract one-tenth of WMT 2019 , around 2 million sentence pairs.

Additionally, a document-level dataset with contrastive test sets in EN-RU (Voita et al., 2019) is used to evaluate lexical coherence.

Lastly, we propose a new document-level dataset in this paper, whose source, scales, and benchmark will be illustrated in the subsequent sections.

For sentences without any ending symbol inside documents, periods are manually added. For our Doc2Doc experiments, the development and test sets are documents merged by sentences. We list

[1] EN-DE datasets are from https://github.com/sameenmaruf/selective-attn

Models	ZH-EN		EN-DE					
	TED		TED		News		Europarl	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU
Sent2Sent (Zheng et al., 2020)	17.0	-	23.10	-	22.40	-	29.40	-
Sent2Sent (Our implementation)	19.2	25.8	25.19	29.16	24.98	27.03	31.70	33.83
DocT (Zhang et al., 2018)	-	-	24.00	-	23.08	-	29.32	-
HAN (Miculicich et al., 2018)	17.9	-	24.58	-	25.03	-	28.60	-
SAN (Maruf et al., 2019)	-	-	24.42	-	24.84	-	29.75	-
QCN (Yang et al., 2019)	-	-	25.19	-	22.37	-	29.82	-
MCN (Zheng et al., 2020)	19.1	25.7	25.10	29.09	24.91	26.97	30.40	32.63
G-Transformer (Bao et al., 2021)	-	-	25.12	27.17	**25.52**	**27.11**	**32.39**	34.08
SR Doc2Doc	-	8.62	-	4.70	-	21.18	-	34.16
MR Doc2Sent	**19.4**	25.8	**25.24**	29.20	25.00	26.70	32.11	34.18
MR Doc2Doc	-	**25.9**	-	**29.27**	-	26.71	-	**34.48**
Sent2Sent ++	21.9	27.9	27.12	30.74	27.85	29.41	32.14	34.20
SR Doc2Doc ++	-	27.0	-	29.96	-	30.61	-	34.38
MR Doc2Sent ++	**22.0**	28.1	**27.34**	30.98	**29.50**	31.17	**32.44**	34.52
MR Doc2Doc ++	-	**28.4**	-	**31.37**	-	**32.59**	-	**34.91**

Table 2: Experiment results of document translation. "-" means not provided. We choose the best hyper-parameters (specifically, dropout) on the development sets for our models as well as baselines. "++" indicates using additional sentence corpus. From the upper part, though SR Doc2Doc yields disappointing translation and even fails on *TED*, MR Doc2Doc achieves much better results, proving the feasibility of Doc2Doc. From the lower part, extra sentence-level corpus can activate SR Doc2Doc and boost MR Doc2Doc, yielding the best results.

all the detailed information of used datasets in Table 1, including languages, scales, and downloading URLs for reproducibility.

3.2 Models

For the model setting, we follow the base version of Transformers (Vaswani et al., 2017), including 6 layers for both encoders and decoders, 512 dimensions for model, 2048 dimensions for ffn layers, 8 heads for attention. For all experiments, we use subword (Sennrich et al., 2016) with 32K merge operations on both sides and cut out tokens appearing less than five times. The models are trained with a batch size of 32000 tokens on 8 Tesla V100 GPUs. Parameters are optimized by using Adam optimizer (Kingma and Ba, 2015), with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. The learning rate is scheduled according to the method proposed in Vaswani et al. (2017), with $warmup_steps = 4000$. Label smoothing (Szegedy et al., 2016) of value=0.1 is also adopted. We set dropout $= 0.3$ for small datasets like *TED* and *News*, and $dropout = 0.1$ for larger datasets like *Europarl*, unless stated elsewise. The training ends when the maximum BLEU on development sets has not been surpassed for 30 times ($\times$ 1000 steps).

3.3 Evaluation

For inference, we generate the hypothesis with a beam size of 5. Following previous related works, we adopt tokenized case-insensitive BLEU (Papineni et al., 2002). Specifically, we follow the methods in Liu et al. (2020), which calculate sentence-level BLEU (denoted as s-BLEU) and document-level BLEU (denoted as d-BLEU), respectively. For d-BLEU, the computing object is either the concatenation of generated sentences or the directly generated documents. Since our documents are generated atomically and hard to split into sentences, we only report d-BLEU for Doc2Doc.

4 Results and Analysis

4.1 MR Doc2Doc Improves Performance

MR matters. It can be seen from the upper part of Table 2 that SR Doc2Doc indeed has a severe drop on *News* and even fails to generate normal results on *TED*, which accords with the findings of Zhang et al. (2018); Liu et al. (2020). It seems too hard to learn long-range document translation directly. However, once equipped with our training technique, MR Doc2Doc can yield the best results, outperforming our strong baseline and previous works on *TED* and *Europarl*. We suggest that NMT is able to acquire the capacity of translating long-range context, as long as it cooperates with some shorter segments as assistance. With the multi-resolutional help of easier patterns, the model can gradually master how to generate complicated sequences.

Doc2Doc matters. We also compare MR Doc2Doc to a intuitive baseline: MR Doc2Sent.

The latter one is trained in a typical Doc2Sent way: the source is the whole past context, the target is the current sentence. From the experimental results, we can see Doc2Doc outperforms it due to much broader contexts.

To show the universality of MR Doc2Doc, we also conduct the experiments on other language pairs: Spanish, French, Russian to English. As shown in Table 3, MR Doc2Doc can be achieved on all language pairs and obtains comparable or better results compared with Sent2Sent.

Models	ES-EN	FR-EN	RU-EN
Sent2Sent	**29.55**	28.69	23.22
SR Doc2Doc	26.79	23.86	16.47
MR Doc2Sent	29.23	28.75	23.48
MR Doc2Doc	29.37	**28.85**	**23.98**

Table 3: Document translation experiments on more languages, showing the comprehensive effectiveness.

It is worth noting that all our results are obtained without any adjustment of model architecture or any extra parameters.

4.2 Additional Sentence Corpus Helps

Furthermore, introducing extra sentence-level corpus is also an effective technique. This can be regarded as another form of multi-resolutional training, as it supplements more sentence-level information. This strategy makes an impact in two ways: activating SR Doc2Doc and boosting MR Doc2Doc.

We merge the datasets mentioned above and *Wikipedia* (EN-DE), WMT (ZH-EN), two out-of-domain sentence-level datasets to do experiments. [2]

As shown in the lower part of Table 2, on the one hand, SR Doc2Doc models are activated and can reach comparable levels with Sent2Sent models as long as assisted with additional sentences. On the other hand, MR Doc2Doc obtains the best results on all datasets and further widens the gap with the sentence corpus's boost. Even out-of-domain sentences can leverage the learning ability of document translation. It again proves the importance of multi-resolutional assistance.

In addition, as analyzed in the previous section, Doc2Sent models are not compatible with sentence-level corpus since the model entrance is specially designed for consecutive sentences. However, Doc2Doc models can naturally draw on the merits of any parallel pairs, including piecemeal sentences. Considering the amount of parallel sentence-level data is much larger than the document-level one, MR Doc2Doc has a powerful application potential compared with Doc2Sent.

4.3 Further Analysis on MR Doc2Doc

4.3.1 Improved Discourse Coherence

Except for BLEU, whether Doc2Doc truly learns to utilize the context to resolve discourse inconsistencies has to be verified. We use the contrastive test sets proposed by Voita et al. (2019), which include deixis, lexicon consistency, ellipsis (inflection), and ellipsis (verb phrase) on English-Russian. Each instance contains a positive translation and a few negative ones, whose difference is only one specific word. With force decoding, if the score of the positive one is the highest, then this instance is counted as correct.

As shown in Table 4, MR Doc2Doc achieves significant improvements and obtain the best results, which proves MR Doc2Doc indeed well captures the context information and maintain the cross-sentence coherence.

Models	deixis	lex.c	ell.infl	ell.VP
Sent2Sent	51.1	45.6	55.4	27.4
Zheng et al. (2020)	61.3	46.1	61.0	35.6
MR Doc2Doc	**64.7**	**46.3**	**65.9**	**53.0**

Table 4: Discourse phenomena evaluation on the contrastive test sets. Our Doc2Doc shows a much better capacity for building the document coherence.

4.3.2 Strong Context Sensibility

Li et al. (2020) find the performance of previous context-aware systems does not decrease with intentional incorrect context and suspect the context usage of context encoders. To verify whether Doc2Doc truly takes advantage of the contextual information in the document, we also conduct the inference with the wrong context deliberately. If the model neglects discourse dependency, then there should be no difference in the performance.

Specifically, we firstly shuffle the sentence order inside each document randomly, marking it as *Local Shuffle*. Furthermore, we randomly swap sentences among all the documents to make the context more disordered, marking it as *Global Shuffle*. As shown in Table 5, the misleading context results

[2] Sentences and documents in non-MR settings are oversampled for six times to keep the same data ratio with the MR settings, which is proved helpful to the performance in Appendix A. Due to the larger scale, we find the settings of dropout=0.2 for *TED*, *News* and dropout=0.1 for *Europarl* yield the best results for both Sent2Sent and Doc2Doc.

in a significant drop for the Doc2Doc model in BLEU. Besides, Global Shuffle brings more harm than Local Shuffle, showing that more chaotic contexts lead to more harm. After all, Local Shuffle still reserves some general information, like topic or tense. These experiments prove the usage of the context.

Models	ZH-EN	EN-DE		
	TED	TED	News	Europarl
MR Doc2Doc	25.84	29.27	26.71	34.48
Local Shuffle	24.10	27.48	25.22	33.52
Global Shuffle	23.69	27.17	24.96	32.47

Table 5: Misleading contexts can bring negative effects to Doc2Doc, proving the dependent usage of the context information. And more chaotic contexts harm more (Global vs. Local).

4.3.3 Compatible with Sentences

The performance with sequence length is also analyzed in this study. Taking *Europarl* as an example, we randomly split documents into shorter paragraphs in different lengths and evaluate them with our models, as shown in Figure 1. Obviously, the model trained only on sentence-level corpus has a severe drop when translating long sequences, while the model trained only on document-level corpus shows the opposite result, which reveals the importance of data distribution. However, the model trained with our multi-resolutional strategy can sufficiently cope with all situations, breaking the limitation of sequence length in translation. By conducting MR Doc2Doc, we obtain an all-in-one model that is capable of translating sequences of any length, avoiding deploying two systems for sentences and documents, respectively.

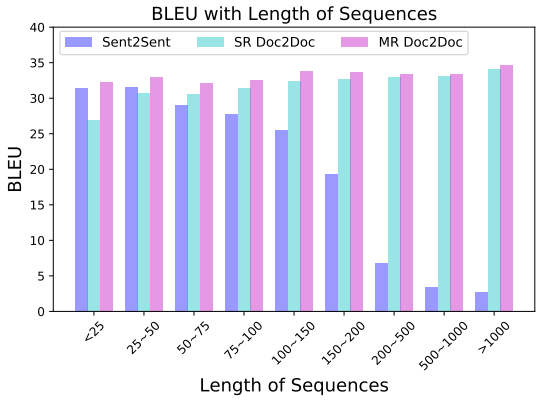

Figure 1: The model trained only on sentence-level or document-level corpus fails to translate sequences in unseen lengths while the MR model yields the best results in all scenarios.

5 Further Evidence with Newly Proposed Datasets and Metrics

To further verify our conclusions and push the development of this field, we also contribute a new dataset along with new metrics. Specifically, we propose a package of a large and diverse parallel document corpus, three deliberately designed metrics, and correspondingly constructed test sets, which will be released soon. On the one hand, they make our conclusions more solid. On the other hand, they may benefit future researches to expand the comparison scenes.

5.1 Parallel Document Corpus

We crawl bilingual news corpus from two websites[3] [4] with both English and Chinese content provided. The detailed cleaning procedure is in Appendix B. Finally, 1.39 million parallel sentences within almost 60 thousand parallel documents are collected. The corpus contains large-scale data with internal dependency in different lengths and diverse domains, including politics, finance, health, culture, etc. We name it **PDC** (Parallel Document Corpus).

5.2 Metrics

To inspect the coherence improvement, we sum up three common linguistic features in document corpus that the Sent2Sent model can not handle:

Tense Consistency (TC): If the source language is tenseless (e.g. Chinese), it is hard for Sent2Sent models to maintain the consistency of tense.

Conjunction Presence (CP): Traditional models ignore cross-sentence dependencies, and the sentence-level translation may cause the missing of conjunctions like "And" (Xiong et al., 2018).

Pronoun Translation (PT): In pro-drop languages such as Chinese and Japanese, pronouns are frequently omitted. When translating from a pro-drop language into a non-pro-drop language (e.g., Chinese-to-English), invisible dropped pronouns may be missing (Wang et al., 2016b,a, 2018a,b).

Afterward, we collect documents that contain abundant verbs in the past tense, conjunctions, and pronouns, as test sets. These words, as well as their positions, are labeled. Some cases are in Appendix C.

For each word-position pair $< w, p >$, we check whether w appears in the generated documents

[3] https://cn.nytimes.com
[4] https://cn.ft.com

within a rough span. And we calculate the appearance percentage as the evaluation score, Specifically:

$$\text{TC / CP / PT} = \frac{\sum_i^n \sum_j^{|W_i|} \mathbb{I}(w_{ij} \in y_i^{\text{span}})}{\sum_i^n |W_i|} \quad (3)$$

$$\text{span} = [\alpha_i p_{ij} - d, \alpha_i p_{ij} + d] \quad (4)$$

n indicates the number of sequences in the test set, W_i indicates the labeled word set of sequence$_i$, w indicates labeled words, y_i indicates output$_i$, p_{ij} indicates the labeled position of w_{ij} in the reference$_i$, α_i indicates the length ratio of translation and reference, d indicates the span radius. We set $d = 20$ in this paper, and calculate the geometric mean as the overall score denoted as **TCP**.

5.3 Test Sets

Along with the filtration of the aforementioned coherence indices, the test sets are built based on websites that are totally different from the training corpus to avoid overfitting. Meanwhile, to alleviate the bias of human translation, the English documents are selected as the reference and manually translated to the Chinese documents as the source. Finally, a total of nearly five thousand sentences within 148 documents is obtained.

5.3.1 Benchmark

Basic experiments with Sent2Sent and Doc2Doc are conducted based on our new datasets, along with full WMT ZH-EN corpus, a sentence-level dataset containing around 20 million pairs. [5] We use WMT newstest2019 as the development set and evaluate the models with our new test sets as well as metrics. The results are shown in Table 6.

Systems	d-BLEU	TC	CP	PT	TCP	Man
Sent2Sent	27.05	54.0	25.5	62.5	44.1	2.89
SR Doc2Doc	24.33	46.7	24.8	61.5	41.5	2.87
MR Doc2Doc	**27.80**	**56.9**	**25.7**	**63.9**	**45.4**	**3.02**
Sent2Sent ++	30.28	58.3	34.1	64.5	50.4	3.58
SR Doc2Doc ++	31.20	59.3	36.3	64.9	51.9	3.61
MR Doc2Doc ++	**31.62**	**59.7**	**36.3**	**65.9**	**52.3**	**3.69**

Table 6: Benckmark of our new datasets. "++" indicates using additional WMT corpus. "Man" refers to human evaluation. Doc2Doc shows much better results in all terms.

[5] We set dropout=0.2 for Sent2Sent and MR Doc2Doc without WMT, and dropout=0.1 for the rest settings according to the performance on the development set. Oversampling is done again, as aforementioned, to enhance the performance for non-MR settings.

BLEU: In terms of BLEU, MR Doc2Doc outperforms Sent2Sent, illustrating the positive effect of long-range context. Moreover, with extra sentence-level corpus, Doc2Doc shows significant improvements again.

Fine-grained Metrics: Our metrics show much clearer improvements. Considering the usage of contextual information, tense consistency is better guaranteed with Doc2Doc. Meanwhile, Doc2Doc is much more capable of translating the invisible pronouns by capturing original referent beyond the current sentence. Finally, the conjunction presence shows the same tendency.

Human Evaluation: Human evaluation is also conducted to illustrate the reliability of our metrics. One-fifth of translated documents are sampled and scored by linguistics experts from 1 to 5 according to not only translation quality but also translation consistency. As shown in Table 6, human evaluation shows a strong correlation with TCP. More specifically, the Pearson Correlation Coefficient (PCCs) between human scores and TCP is higher than that of BLEU (97.9 vs. 94.1).

5.4 Case Study

Table 7 shows an example of document translation. Sent2Sent model neglects the cross-sentence context and mistakenly translate the ambiguous word, which leads to a confusing reading experience. However, the Doc2Doc model can grasp a full picture of the historical context and make accurate decisions.

Source	与大多数欧洲人一样, 德国总理对美国总统的"美国优先"民族主义难以掩饰不屑。 ... 但她已进入第四个、也必定是最后一个总理任期。
Sent2Sent	Like most Europeans , the German chancellor has struggled to hide his disdain for the US president's "America First" nationalism. ... But she has entered a fourth and surely last term as prime minister.
Doc2Doc	Like most Europeans, the German chancellor's disdain for the US president's "America First" nationalism is hard to hide. ... But she has entered her fourth and certainly final term as chancellor.

Table 7: Coherence problem in document translation. Without discourse contexts, the Chinese word "总理" is usually translated to "prime minister", while in the context of "German", it should be translated into "chancellor".

Also, we manually switch the context information in the source side to test the model sensibility, as shown in Table 8. It turns out that Doc2Doc is able to adapt to different contexts.

Country	Sent2Sent	Doc2Doc	Oracle
Germany	prime minister	**chancellor**	chancellor
Italy	**prime minister**	**prime minister**	prime minister
Austria	prime minister	**chancellor**	chancellor
France	**prime minister**	**prime minister**	prime minister

Table 8: Further study of Table 7. We switch the country information in the source side like *German* → *Italian/Austrian/French*, *Berlin* → *Rome/Vienna/Paris*. Doc2Doc model shows strong sensibility to the discourse context.

6 Limitation

Though multi-resolutional Doc2Doc achieves direct document translation and obtains better results, there still exists a big challenge: efficiency. The computation cost of self-attention in Transformer rises with the square of the sequence length. As we feed the entire document into the model, the memory usage will be a bottleneck for larger model deployment. And the inference speed may be affected if no parallel operation is conducted. Recently, many studies focus on the efficiency enhancement on long-range sequence processing (Correia et al., 2019; Child et al., 2019; Kitaev et al., 2020; Wu et al., 2020; Beltagy et al., 2019; Rae et al., 2020). We leave reducing the computation cost to the future work.

7 Related Work

Document-level neural machine translation is an important task and has been abundantly studied with multiple datasets as well as methods.

The mainstream research in this field is the model architecture improvement. Specifically, several recent attempts extend the Sent2Sent approach to the Doc2Sent-like one. Wang et al. (2017); Miculicich et al. (2018); Tan et al. (2019) make use of hierarchical RNNs or Transformer to summarize previous sentences. Jean et al. (2017); Bawden et al. (2017); Zhang et al. (2018); Voita et al. (2018); Kuang and Xiong (2018); Maruf et al. (2019); Yang et al. (2019); Jiang et al. (2019); Zheng et al. (2020); Yun et al. (2020); Xu et al. (2020) introduce additional encoders or query layers with attention model and feed the history contexts into decoders. Maruf and Haffari (2018); Kuang et al. (2018); Tu et al. (2018) propose to augment NMT models with a cache-like memory network, which generates the translation depending on the decoder history retrieved from the memory.

Besides, some works intend to resolve this problem in other ways. Jean and Cho (2019) propose a regularization term for encouraging to focus more on the additional context using a multi-level pairwise ranking loss. Yu et al. (2020) utilize a noisy channel reranker with Bayes' rule. Garcia et al. (2019) extends the beam search decoding process with fusing an attentional RNN with an SSLM by modifying the computation of the final score. Saunders et al. (2020) present an approach for structured loss training with document-level objective functions. Liu et al. (2020); Ma et al. (2020) combine large-scale pre-train model with DNMT. Unanue et al. (2020); Kang et al. (2020) adopt reinforcement learning methods.

There are also some works sharing similar ideas with us. Tiedemann and Scherrer (2017); Bawden et al. (2017) explore concatenating two consecutive sentences and generate two sentences directly. Obviously, we leverage greatly longer information and capture the full context. Junczys-Dowmunt (2019) cut documents into long segments and feed them into training like BERT (Devlin et al., 2019). There are at least three main differences. Firstly, they need to add specific boundary tokens between sentences while we directly translate the original documents without any additional processing. Secondly, we propose a novel multi-resolutional training paradigm that shows consistent improvements compared with regular training. Thirdly, for extremely long documents, they restrict the segment length to 1000 tokens or make a truncation while we preserve entire documents and achieve literal document-to-document training and inference.

Finally, our work is also related to a series of studies in long sequence generation like GPT (Radford, 2018), GPT-2 (Radford et al., 2019), and Transformer-XL (Dai et al., 2019). We all suggest that the deep neural generation models have the potential to well process long-range sequences.

8 Conclusion

In this paper, we try to answer the question of whether Document-to-document translation works. It seems naive Doc2Doc can fail in multiple scenes. However, with the multi-resolutional training proposed in this paper, it can be successfully activated. Different from traditional methods of modifying the model architectures, our approach introduces no extra parameters. A comprehensive set of experiments on various metrics show the advantage of MR Doc2Doc. In addition, we contribute a new document-level dataset as well as three new metrics to the community.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *ACL*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. In *NAACL-HLT*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2019. Longformer: The long-document transformer. *arXiv*, abs/2004.05150.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv*, abs/1904.10509.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. 2019. Adaptively sparse transformers. In *EMNLP-IJCNLP*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Eva Martínez Garcia, C. Creus, and C. España-Bonet. 2019. Context-aware neural machine translation decoding. In *DiscoMT@EMNLP*.

Sébastien Jean and Kyunghyun Cho. 2019. Context-aware learning for neural machine translation. *arXiv*, abs/1903.04715.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Neural machine translation for cross-lingual pronoun prediction. In *DiscoMT@EMNLP*.

Shu Jiang, Rui Wang, Zuchao Li, Masao Utiyama, Kehai Chen, Eiichiro Sumita, Hai Zhao, and Bao-Liang Lu. 2019. Document-level neural machine translation with inter-sentence attention. *arXiv*, abs/1910.14528.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *WMT*.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *EMNLP*.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *ICLR*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *NMT@ACL*.

Shaohui Kuang and Deyi Xiong. 2018. Fusing recency into neural machine translation with an inter-sentence gate model. In *COLING*.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *COLING*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *EMNLP*.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *ACL*.

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

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *ACL*.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *ACL*.

Sameen Maruf, André F. T. Martins, and Gholamreza" Haffari. 2019. Selective attention for context-aware neural machine translation. In *NAACL-HLT*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *EMNLP*.

Mathias Müller, Annette Rios Gonzales, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *WMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *ICLR*.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2020. Using context in neural machine translation training objectives. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *EMNLP-IJCNLP*.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *DiscoMT@EMNLP*.

Zhaopeng Tu, Yang P. Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *TACL*.

Inigo Jauregi Unanue, Nazanin Esmaili, Gholamreza Haffari, and Massimo Piccardi. 2020. Leveraging discourse rewards for document-level neural machine translation. In *COLING*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *ACL*.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *ACL*.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018a. Translating pro-drop languages with reconstruction models. In *AAAI*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2018b. Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016a. A novel approach to dropped pronoun translation. In *NAACL-HLT*.

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Hang Li, and Qun Liu. 2016b. Dropped pronoun generation for dialogue machine translation. In *ICASSP*.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *DiscoMT@EMNLP*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, abs/1609.08144.

Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. 2020. Lite transformer with long-short range attention. In *ICLR*.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2018. Modeling coherence for discourse neural machine translation. In *AAAI*.

Hongfei Xu, Deyi Xiong, Josef van Genabith, and Qiuhui Liu. 2020. Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *DiscoMT@IJCAI*.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level nmt. In *EMNLP-IJCNLP*.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes' rule. *TACL*.

Hyeongu Yun, Yongkeun Hwang, and Kyomin Jung. 2020. Improving context-aware neural machine translation using self-attentive sentence embedding. In *AAAI*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang P. Liu. 2018. Improving the transformer translation model with document-level context. In *EMNLP*.

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

A Oversampling Illustration

When combining document-level datasets with sentence-level datasets (especially out-of-domain corpus), we employ oversampling for non-MR settings. This can keep them the same data ratio with the MR setting and is helpful for their performance. Since the data size of MR is around 6 times of non-MR ($\approx \log_2 64$), as shown in Table 9, we mainly oversample for 6 times. The contrastive experiments are in Table 10. We attribute the improvements to the reduction of the proportion of out-of-domain data.

Datasets	Ratio
TED (ZH-EN)	6.7
TED (EN-DE)	7.6
News (EN-DE)	5.9
Europal	4.6
News (ES-EN)	5.9
News (FR-EN)	5.9
News (RU-EN)	5.9
PDC	5.3
Mean	6.0

Table 9: Ratio of MR/non-MR in data size

Dataset	Sent2Sent		SR Doc2Doc	
	non-OS	OS	non-OS	OS
TED(ZH-EN)+WMT	27.52	**27.90**	26.05	**26.67**
TED(EN-DE)+Wiki	29.19	**30.74**	29.81	**29.96**
News+Wiki	27.77	**29.41**	30.15	**30.61**
Europarl+Wiki	33.93	**34.20**	34.25	**34.38**
PDC+WMT	29.52	**30.28**	29.60	**31.20**

Table 10: The contrastive results of oversampling when combining sentence-level corpus.

B Clean Procedure on PDC

We mainly crawl bilingual news corpus from two websites (https://cn.nytimes.com, https://cn.ft.com) with both English and Chinese content provided. Then three steps are followed to clean the corpus.

1. **Deduplication:** We deduplicate the documents that include almost the same content.
2. **Sentence Segmentation:** We use *Pragmatic Segmenter* [6] to segment paragraphs into sentences.
3. **Filtration:** We use *fast_align* [7] to align sentence pairs and label the pairs as misaligned ones if the alignment scores are less than 40%. Documents are finally removed if they contain misaligned sentence pairs.

[6] https://github.com/diasks2/pragmatic_segmenter

[7] https://github.com/clab/fast_align

Finally, we obtain 1.39 million parallel sentences within almost 60 thousand cleaned parallel documents. The dataset contains diverse domains including politics, finance, health, culture, etc.

C Cases of Our Test Sets

Apart from the statistic number in the main paper, we also provide some cases in our test sets to illustrate the value of our test sets and metrics, as shown in Table 11,12,13.

Src	1.双方在2017 年都向法庭提交了申请。 2.邓普顿奈特想要 报销他的租金。 3.伯德特想要 赶走邓普顿奈特。
Ref	1.Both parties had lodged applications with the tribunal in 2017. 2.Templeton-Knight wanted his rent reimbursed. 3.Burdett wanted to evict Templeton-Knight.
NMT	1.Both parties filed applications with the court in 2017. 2.Templeton Knight wants to reimburse his rent. 3.Burdett wants to get rid of Templeton Knight.

Table 11: Tense inconsistency problem in translating tenseless languages (e.g. Chinese) to tense-sensitive languages (e.g. English). Individual sentences are translated into present tense with sentence-level models while the history context has provided the signal of past tense.

Src	1.我女儿使用的胰岛素类型— — 世界上只有两家类似类型的制造商。 2.他们继续保持一致同时提高价格。
Ref	1.The type of insulin that my daughter uses — there are only two manufacturers worldwide of a similar type. 2.And they continue to increase their prices lockstep together.
NMT	1.The type of insulin my daughter uses - there are only two manufacturers of similar types in the world. 2.[conj miss] They continue to be consistent while raising prices.

Table 12: Conjunction missing problem in sentence-level translation. The sentences has strong semantic connection but are translated without any conjunction.

Src	1.根据市政府的说法，奥特里工厂的其他拟议功能似乎极不可能实施。 2.即使顾问和调查人推荐[pro drop]。
Ref	1.Other proposed features for Autrey Mill seem highly unlikely to be implemented according to the City Manager. 2.Even though consultants and surveys recommended them.
NMT_A	1.According to the city government, other proposed functions at the Autry plant appear highly unlikely to be implemented. 2.Even if consultants and surveys recommend [pro miss].
NMT_B	1.According to the municipal government , other proposed functions of the Autry plant seem highly impossible to implement . 2.Even if consultants and surveys recommended it.

Table 13: Pronoun drop problem in translating pro-drop languages (e.g. Chinese) to non-pro-drop languages (e.g. English). The pronoun is omitted or translated wrongly with sentence-level models..