Document Flattening: Beyond Concatenating Context for Document-Level Neural Machine Translation

Minghao Wu$^\heartsuit$ **George Foster**$^\spadesuit$ **Lizhen Qu**$^\heartsuit$ **Gholamreza Haffari**$^\heartsuit$
$^\heartsuit$Monash University $^\spadesuit$Google Research
{firstname.lastname}@monash.edu   fosterg@google.com

Abstract

Existing work in document-level neural machine translation commonly concatenates several consecutive sentences as a *pseudo-document*, and then learns inter-sentential dependencies. This strategy limits the model's ability to leverage information from distant context. We overcome this limitation with a novel Document Flattening (DOCFLAT) technique that integrates FLAT-BATCH ATTENTION (FBA) and NEURAL CONTEXT GATE (NCG) into Transformer model to utilize information beyond the pseudo-document boundaries. FBA allows the model to attend to all the positions in the batch and learns the relationships between positions explicitly and NCG identifies the useful information from the distant context. We conduct comprehensive experiments and analyses on three benchmark datasets for English-German translation, and validate the effectiveness of two variants of DOCFLAT. Empirical results show that our approach outperforms strong baselines with statistical significance on BLEU, COMET and accuracy on the contrastive test set. The analyses highlight that DOCFLAT is highly effective in capturing the long-range information.

1 Introduction

Remarkable progress has been made in neural machine translation (NMT) (Sutskever et al., 2014; Vaswani et al., 2017; Chen et al., 2018), yet human translation still clearly outperforms NMT at the document level (Läubli et al., 2018; Freitag et al., 2021), because current sentence-level NMT systems ignore the inter-sentential relationships. To narrow this gap, numerous document-level NMT (DocNMT) approaches have been proposed in recent years to improve the context awareness by incorporating the contextual information during the translation (Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Wong et al., 2020).

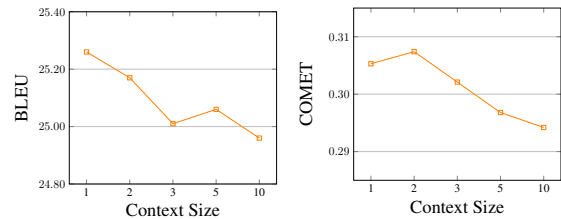Existing DocNMT systems commonly concatenate several consecutive sentences to form a

Figure 1: The change of BLEU (left) and COMET (right) given by DOC2DOC on TED with regard to the context size of *pseudo-document* (in sentences) based on the experimental setup described in Section 4.1.

pseudo-document, instead of processing the *entire document* (Zhang et al., 2018b; Voita et al., 2019; Junczys-Dowmunt, 2019; Fernandes et al., 2021). One typical pseudo-document contains the *current sentence* to be translated and the *surronding context*. Intuitively, larger context should result in better performance. In our preliminary study, the model performance does not always grow as the context size increases as shown in Figure 1. Liu et al. (2020) and Bao et al. (2021) also observe that Transformer's performance declines with longer inputs. We refer to this phenomenon as the *quality saturation problem* (Glaser and Strauss, 1967). Therefore, such formation of pseudo-document limits the DocNMT systems to leverage the information from a relatively small context. Consequently, once the entire original document is segmented into several pseudo-documents for reducing the sequence length, the information out of the pseudo-document's scope is no longer accessible to the current sentence. Therefore, a natural research question to ask is that, *is there a more effective way to model the parallel documents in DocNMT?*

In this work, we seek DocNMT approaches that could better expand the context scope and improve the corresponding translation performance. Instead of directly training DocNMT system on the entire document, we propose to store the document as multiple pseudo-documents in a single batch and

optimize the DocNMT models by leveraging the inter-pseudo-document relationships at the batch level. Inspired by Kossen et al. (2021), we propose a Document Flattening (DOCFLAT) technique that integrates FLAT-BATCH ATTENTION (FBA) and NEURAL CONTEXT GATE (NCG) into the Transformer model (Vaswani et al., 2017). FBA flattens all the current sentences in the batch with the original order into a sequence along the temporal dimension. It then applies the attention mechanism to the flattened sequence. The goal of this design is to preserve the linguistic structure of documents and expand the scope of context by explicitly learning the pseudo-document relationships. As there is both supportive and noisy information in the longer context, we introduce NCG, a simple feed-forward network, to identify the usefulness of contextual information and filter out the noise. With the combination of FBA and NCG, DOCFLAT effectively captures the information in the distant context. To the best of our knowledge, Morishita et al. (2021) propose mini-batch embedding (MBE), which is the only close work to ours. They compute the average representation for all the source tokens in the batch and prepend it to the source and target pseudo-documents. The compressed representation ignores the linguistic structure of documents, providing limited contextual information.

Our contributions are summarized as follows. Firstly, we propose a novel approach DOCFLAT that allows the model to attend the content beyond the pseudo-document boundaries using FBA and NCG. Secondly, we demonstrate that DOCFLAT outperforms strong baselines with statistical significance, in terms of BLEU, COMET and accuracy on the contrastive test set, on three Doc-NMT benchmark datasets, including `TED`, `News Commentary` and `Europarl`. Thirdly, we conduct comprehensive analyses to understand the effectiveness of DOCFLAT. The analyses highlight that DOCFLAT is highly effective in capturing the distant context.

2 Preliminaries

Sentence-level NMT (SentNMT) The sentence-level NMT model neglects the inter-sentential dependencies between the current sentence and its context. Its probability of translation is defined as:

$$P(\boldsymbol{y}_i|\boldsymbol{x}_i) = \prod_{t=1}^{d} P(y_{i,t}|\boldsymbol{y}_{i,<t},\boldsymbol{x}_i), \quad (1)$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the i-th source and target training sentence, $y_{i,t}$ denotes the t-th token in $\boldsymbol{y}_i$ and d is the sentence length of $\boldsymbol{y}_i$.

Document-level NMT (DocNMT) Given a document pair $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{M}$ where we denote the aligned sentence pair as $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ and M is the length of document in sentences, the i-th pseudo-document pair $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ can be defined as:

$$\begin{aligned} \boldsymbol{X}_i &= \text{Concat}([\boldsymbol{x}_{i-c^-},\dots,\boldsymbol{x}_i,\dots,\boldsymbol{x}_{i+c^+}]), \\ \boldsymbol{Y}_i &= \text{Concat}([\boldsymbol{y}_{i-c^-},\dots,\boldsymbol{y}_i,\dots,\boldsymbol{y}_{i+c^+}]), \end{aligned} \quad (2)$$

where c^- is the context size before the current sentence and c^+ is the context size after the current sentence. The translation probability of target *current* sentence $\boldsymbol{y}_i$ in the target pseudo-document $\boldsymbol{Y}_i$ given the source pseudo-document $\boldsymbol{X}_i$ in DocNMT can be written as:

$$P(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{C}_{-i}) = \prod_{t=1}^{d} P(y_{i,t}|\boldsymbol{y}_{i,<t},\boldsymbol{x}_i,\boldsymbol{C}_{-i}), \quad (3)$$

where $\boldsymbol{C}_{-i}$ is the collection of all the sentences in the pseudo-document pair except $(\boldsymbol{x}_i,\boldsymbol{y}_i)$, and $\boldsymbol{x}_i$ is the source *current* sentence. We do not consider the context after the current sentence in this work, so c^+ is 0.

3 Document Flattening

In this section, we firstly describe the overview of DOCFLAT (Section 3.1). We then introduce DOCFLAT's core components, FLAT-BATCH ATTENTION (FBA; Section 3.2) and NEURAL CONTEXT GATE (NCG; Section 3.3). Finally, we discuss the practical considerations (Section 3.4 and Section 3.5) of DOCFLAT along with a concrete example.

3.1 Overview of DOCFLAT

We present the overall architecture of DOCFLAT in Figure 2. Given a sequence-to-sequence Transformer with L encoder layers and L decoder layers, we apply the FBA and NCG to the input word embeddings with the residual connection and Layer Normalization, instead of directly feeding the embeddings into either encoder or decoder. DOCFLAT's translation probability of the i-th target current sentence $\boldsymbol{y}_i$ of the original document in the i-th target pseudo-document $\boldsymbol{Y}_i$ given the i-th source pseudo-document $\boldsymbol{X}_i$ in the batch

$\boldsymbol{B} = \{(\boldsymbol{X}_j, \boldsymbol{Y}_j)\}_{j=1}^n$, where n is the batch size, is defined as:

$$P(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{C}_{-i}, \boldsymbol{B}_{-i}) = \prod_{t=1}^d P(y_{i,t}|\boldsymbol{y}_{i,<t}, \boldsymbol{x}_i, \boldsymbol{C}_{-i}, \boldsymbol{B}_{-i}), \quad (4)$$

where $\boldsymbol{C}_{-i}$ is defined as in Equation 3 and $\boldsymbol{B}_{-i}$ is the collection of all the current sentences in the batch except $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. We categorize the context for the current sentence into two groups, the global context (GC) from other pseudo-documents $\boldsymbol{B}_{-i}$ and the local context (LC) from its own pseudo-document $\boldsymbol{C}_{-i}$ (See Figure 3).

3.2 FLAT-BATCH ATTENTION

Multi-Head Self-Attention (MHSA) Scaled dot-product attention is the core mechanism of Transformer model with the inputs of query $\boldsymbol{Q}$, key $\boldsymbol{K}$ and $\boldsymbol{V}$ (Vaswani et al., 2017). The attention mechanism computes the attention weights by comparing queries $\boldsymbol{Q}$ with keys $\boldsymbol{K}$ and then updates the representations of queries by computing the weighted sum of values $\boldsymbol{V}$ with the attention weights, which is described as follows:

$$\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{e}})\boldsymbol{V}, \quad (5)$$

where e is the hidden state dimension. Multi-head self-attention (MHSA) then allows the model to jointly attend to information from different hidden subspaces by concatenating a sequence of independent attention heads as follows:

$$\text{MHSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_k), \quad (6)$$

where head_j is the scaled dot-product attention in Equation 5 with independent parameters and $j \in \{1, \ldots, k\}$ for each head j.

FLAT-BATCH ATTENTION (FBA) To leverage the contextual information beyond the pseudo-document boundaries, we propose FLAT-BATCH ATTENTION (FBA). It explicitly transforms the stacked instances in the batch to a single flattened sequence of tokens as shown in Figure 3. Given a batch of hidden representations $\boldsymbol{H} \in \mathbb{R}^{n \times d \times e}$ consisting of n instances padded to the length of d with the hidden dimension of e, FBA operates as
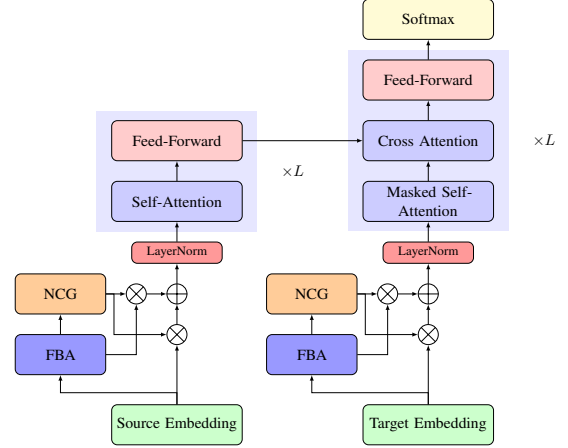

Figure 2: The model architecture of DOCFLAT. $\otimes$ denotes the element-wise multiplication. $\oplus$ denotes the element-wise addition. More details are in Section 3.

follows:

$$\begin{aligned}
\hat{\boldsymbol{H}}_{\text{flat}} &= \text{Flatten}(\boldsymbol{H}) \in \mathbb{R}^{(n \times d) \times e}, \\
\hat{\boldsymbol{H}}_{\text{mhsa}} &= \text{MHSA}(\hat{\boldsymbol{H}}_{\text{flat}}, \hat{\boldsymbol{H}}_{\text{flat}}, \hat{\boldsymbol{H}}_{\text{flat}}), \\
\hat{\boldsymbol{H}}_{\text{rsh}} &= \text{Reshape}(\hat{\boldsymbol{H}}_{\text{mhsa}}) \in \mathbb{R}^{n \times d \times e}, \\
\hat{\boldsymbol{H}} &= \text{LN}((1 - \boldsymbol{g}) \otimes \boldsymbol{H} + \boldsymbol{g} \otimes \hat{\boldsymbol{H}}_{\text{rsh}}).
\end{aligned} \quad (7)$$

As shown in Equation 7, we first flatten $\boldsymbol{H} \in \mathbb{R}^{n \times d \times e}$ to $\hat{\boldsymbol{H}}_{\text{flat}} \in \mathbb{R}^{(n \times d) \times e}$, where $(n \times d)$ indicates the flattened sequence length. The $\hat{\boldsymbol{H}}_{\text{flat}}$ is then fed into a MHSA layer and reshaped back to $\hat{\boldsymbol{H}}_{\text{rsh}} \in \mathbb{R}^{n \times d \times e}$. We then add a residual connection with $\boldsymbol{g}$ given by NCG $\boldsymbol{\psi}$ followed by a sigmoid function σ and apply the Layer Normalization (LN; Ba et al., 2016) following the reshape operation. $\otimes$ denotes the element-wise multiplication. We discuss the details of NCG in Section 3.3. Note that FBA at the decoder side is associated with a causal mask to preserve the auto-regressive property. By attending to all the other current sentences in the batch, FBA effectively allows the current sentences to access a much larger context than the self-attention on the pseudo-documents. In addition, this does not increase the input length of each instance, preventing the quality saturation problem as shown in Figure 1.

Complexity Given a Transformer model with L encoder layers and L decoder layers, suppose the average sentence length is n, the pseudo-document contains c consecutive sentences, and the batch size is b. The complexity of self-attention layer in the concatenation-based DOC2DOC is $\mathcal{O}(L(cn)^2)$. The extra complexity introduced by FBA is $\mathcal{O}((bn)^2)$. Lc^2 and b^2 have the same order

450
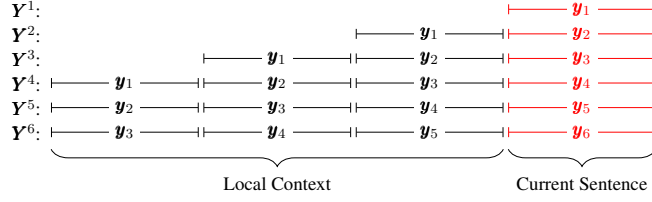
(a) An example batch of pseudo-documents $\boldsymbol{B}_{\text{tgt}} = \{\boldsymbol{Y}^1, \boldsymbol{Y}^2, \boldsymbol{Y}^3, \boldsymbol{Y}^4, \boldsymbol{Y}^5, \boldsymbol{Y}^6\}$ at the target side. Each $\boldsymbol{Y}^j$ contains four consecutive sentences and $\boldsymbol{y}_i$ indicates the i-th sentence of the same original document. The segments in red indicate the current sentence of each pseudo-document.
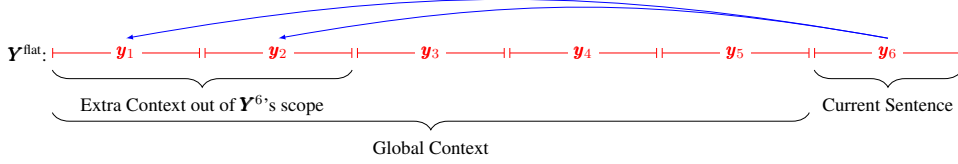
(b) An example of the flattened sequence $\boldsymbol{Y}^{\text{flat}}$ transformed from $\boldsymbol{B}_{\text{tgt}}$ in Figure 3a with FLAT-BATCH ATTENTION. For the current sentence $\boldsymbol{y}_6$, The blue arrows indicate the extra inter-sentential attention for $\boldsymbol{y}_6$ that our approach can model. $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the extra context introduced by our approach.

Figure 3: An example batch of pseudo-documents at the target side and its flattened sequence. Another example at the source side can be found at Appendix A.

of magnitude. The batch size b is set to be constant in practice and the self-attention operation in FBA is highly parallelizable, so integrating FBA into Transformer does not significantly increase the computational cost. Empirically, DOCFLAT is only 3% slower in training and 15% slower in inference, compared with DOC2DOC (See Section 4.2).

3.3 NEURAL CONTEXT GATE

The distant context can contain both supportive and noisy information. Supportive information can assist the translation of the current sentence, while the noise may damage the model predictions. To address this issue, we introduce a novel NEURAL CONTEXT GATE (NCG) to automatically identify the context usefulness and control the information flow from the distant context.

In this work, NCG $\boldsymbol{\psi}$ is a single-layer element-wise feed-forward neural network followed by a sigmoid function σ. Given a batch of hidden representations $\boldsymbol{H}$, the operations are defined as follows:

$$
\begin{aligned}
\boldsymbol{g} &= \sigma(\boldsymbol{\psi}(\text{FBA}(\boldsymbol{H}))), \\
\boldsymbol{H}^o &= (1 - \boldsymbol{g}) \otimes \boldsymbol{H} + \boldsymbol{g} \otimes \text{FBA}(\boldsymbol{H}),
\end{aligned}
\tag{8}
$$

where $\boldsymbol{g}$ is the information gate given by NCG $\boldsymbol{\psi}$ and the sigmoid function σ, $\boldsymbol{H}^o$ is output of the residual connection and $\otimes$ denotes the element-wise multiplication. The values of $\boldsymbol{g}$ are continuous, so we denote DOCFLAT with NCG described in Equation 8 as DOCFLAT$_{\text{C}}$.

However, the continuous gate may result in the noise leakage. In a long document, the noise at different positions may accumulate, even if they are only associated with very small gating values. The accumulated noise can make a substantial negative impact on the model predictions. Hence, we propose the discrete NCG as follows:

$$
\begin{aligned}
\boldsymbol{g}_{\text{D}} &= \mathbb{1}_\gamma(\sigma(\boldsymbol{\psi}(\text{FBA}(\boldsymbol{H})))), \\
\boldsymbol{H}^o &= (1 - \boldsymbol{g}_{\text{D}}) \otimes \boldsymbol{H} + \boldsymbol{g}_{\text{D}} \otimes \text{FBA}(\boldsymbol{H}),
\end{aligned}
\tag{9}
$$

where $\mathbb{1}_\gamma(\cdot)$ is indicator function defined as:

$$
\mathbb{1}_\gamma(g) = \begin{cases} 1 & \text{if } g \geq \gamma, \\ 0 & \text{otherwise,} \end{cases}
\tag{10}
$$

where γ is the threshold for binarizing the gating values. We denote DOCFLAT with the discrete NCG in Equation 9 as DOCFLAT$_{\text{D}}$ and set $\gamma = 0.5$ in this work. We expect DOCFLAT$_{\text{D}}$ is more robust against the noise in the context.

3.4 Data Shuffling

DOCFLAT aims to leverage the distant context with FBA by flattening a batch of sequences into a single sequence. As the ordering information among sentences is critical in DocNMT, we do not shuffle the pseudo-documents during the training and inference to preserve the linguistic structure of the original document. For each sentence, we replace the <BOS> symbol with its global index i in the document to preserve the ordering information. Figure 3a is an example at the target side to demonstrate how the pseudo-documents in the batch is

451

	Train	Valid	Test
TED	$204.4K/1.7K$	$8.9K/93$	$2.2K/23$
News	$242.4K/6.1K$	$2.3K/81$	$3.2K/155$
Europarl	$1.8M/117.9K$	$3.8K/240$	$5.5K/360$

Table 1: The number of sentences/documents of each split of the parallel corpora.

organized in this work. Figure 3b demonstrates how FBA flattens a batch of pseudo-documents. Since the pseudo-documents are not shuffled, $\boldsymbol{y}_6$ can attend to $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ which are not in the pseudo-document $\boldsymbol{Y}^6$. We apply the causal mask to FBA at the decoder side for preserving the auto-regressive property. Note that the pseudo-documents in a batch are mostly from the same original document. The batches crossing the document boundaries are relatively rare and have little effect on performance in our preliminary study.

3.5 Inference

We discuss the batch inference of DOCFLAT in this section. At the encoder side, each source current sentence can attend to its own local context (LC) and all other source current sentence as the global context (GC) in the batch during the inference, as it is at the training stage. At the decoder side, all the target current sentences are translated simultaneously, so each target current sentence can attend to its own LC and partially translated target GC. For example, all the target current sentences in Figure 3a are translated simultaneously during the inference. $\boldsymbol{y}_6$ is conditioned on its own LC, $\boldsymbol{y}_3$, $\boldsymbol{y}_4$ and $\boldsymbol{y}_5$, and partially translated target GC. Additionally, we use the batched inference as usual and there is no overlap between batches. For example, the first batch of sentences to be translated is $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_b\}$, and the second batch of sentences to be translated is $\{\boldsymbol{y}_{b+1}, \cdots, \boldsymbol{y}_{2b}\}$, where b is the inference batch size. For decoding, we used the iterative decoding method for decoding (Maruf and Haffari, 2018; Maruf et al., 2019). The initial translations of each sentence were generated by a Sent-NMT model, and then, we translate each sentence using the DOCNMT model with the translations in the first pass as the context.

4 Experiments

4.1 Setup

Datasets We conduct experiments on three benchmark datasets for English-German trans-

lation, including the small-scale datasets TED (Cettolo et al., 2012) and News Commentary (Tiedemann, 2012), and the large-scale dataset Europarl (Koehn, 2005). We tokenize the datasets with the Moses (Koehn et al., 2007) and apply BPE (Sennrich et al., 2016b) with $32K$ merges. Data statistics can be found in Table 1. We choose up to 3 previous sentences as the local context for each source and target sentence to form the pseudo-document unless otherwise specified.

Evaluation We report the detokenized BLEU (Papineni et al., 2002) using SacreBLEU (Post, 2018) and the neural-based COMET (Rei et al., 2020) to measure the translation quality.[1] We report the results with inference batch size of 16 and beam size of 5 for all the approaches, unless otherwise specified.

Contrastive Evaluation This evaluation paradigm is proposed to evaluate the contextual awareness of DOCNMT models with an independent test set, where each test example includes one correct translation and several incorrect translations. The model is required to identify the correct translation and its overall performance is measured by micro-average *Accuracy*. In this work, we use the large-scale English-German anaphoric pronoun test set from Müller et al. (2018), containing $12K$ contrastive examples. Given the provided context, the model of interest is required to identify the translation with the correct use of pronoun from *er*, *es* and *sie* in German.

Models All the models in this work are based on the standard Transformer base (Vaswani et al., 2017). Besides the direct comparisons with prior works, we also compare DOCFLAT with several re-implemented baselines, including SENT2SENT (Vaswani et al., 2017), DOC2DOC (Tiedemann and Scherrer, 2017), FLATTRANS (Ma et al., 2020), MBE (Morishita et al., 2021) and ABD (Kossen et al., 2021). We only apply ABD at the encoder side in this work unless otherwise specified, which is its best-performing setup as shown in Appendix C. We re-produce the results of GTRANS (Bao et al., 2021) with its official code and recommended hyperparameters. The optimization details are in Appendix B.

---

[1]SacreBLEU Signature: `nrefs:1|case:mixed| eff:no|tok:13a|smooth:exp|version:2.2.0` and COMET Signature: `wmt20-comet-da`.

	TED			News			Europarl			UPS
	BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.	
Reported										
DocTransformer (Zhang et al., 2018b)	24.00	—	—	23.08	—	—	29.32	—	—	—
HAN (Miculicich et al., 2018)	24.58	—	—	25.03	—	—	28.60	—	—	—
Selective (Maruf et al., 2019)	24.42	—	—	24.84	—	—	29.75	—	—	—
Hybrid (Zheng et al., 2020)	25.10	—	—	24.91	—	—	30.40	—	—	—
Re-produced (standard)										
SENT2SENT	24.78	0.2860	46.48	25.00	0.1993	47.71	31.24	0.5933	53.02	2.57
DOC2DOC	25.01	0.3021	66.99	24.95	0.1990	64.21	31.65	0.5929	78.18	0.86
FLATTRANS	24.71	0.2963	45.45	25.05	0.2020	48.54	31.58	0.5954	51.14	0.90
GTRANS	25.29	0.3058	—	25.59	0.2097	—	32.33	0.5904	—	—
Re-produced (batch-level)										
MBE	24.75	0.3032	68.12	24.86	0.1969	62.82	31.63	0.5954	77.08	0.64
ABD	24.97	0.3046	68.25	24.33	0.1772	62.52	31.98	0.5955	78.16	0.84
Ours										
DOCFLAT$_C$	25.31†	**0.3173**†	70.92†	**25.96**†	**0.2199**†	65.45†	**32.38**†	**0.6020**†	77.65	0.84
DOCFLAT$_D$	**25.41**†	0.3101†	**72.04**†	25.38	0.2119†	**66.70**†	32.16†	0.5990†	**79.68**†	0.84

Table 2: BLEU, COMET and accuracy on three benchmark datasets for English-German translation. UPS ($\uparrow$) indicates updates per second. The best results are highlighted in **bold**. — indicate the result is not available. $\dagger$ indicates the statistical significance at $p = 0.05$ against re-implemented DOC2DOC based on Koehn (2004).

4.2 Main Results

We present the main results in Table 2.

Comparisons with Baselines Compared with all the baselines regardless whether they utilize the batch information or not, both DOCFLAT$_C$ and DOCFLAT$_D$ substantially outperform these strong baseline approaches, especially in terms of the context awareness (accuracy) which is the main emphasis of this work. For the approaches that utilize the batch information, we observe that MBE and ABD only marginally improves the performance compared with DOC2DOC, suggesting the importance of preserving the linguistic structure in utilizing the batch-level information for DocNMT. We also observe the larger performance gain from DOCFLAT on small TED and News, implying DOCFLAT performs better in the low-resource settings.

DOCFLAT$_C$ vs. DOCFLAT$_D$ As shown in Table 2, DOCFLAT$_C$ and DOCFLAT$_D$ demonstrate different strengths: DOCFLAT$_C$ mainly improves the translation quality (BLEU and COMET), while DOCFLAT$_D$ improves the context awareness (accuracy). In the contrastive evaluation, we have no access to the entire document, so the model predictions are always conditioned on the golden local context (LC) and irrelevant global context (GC). DOCFLAT$_D$ outperforms DOCFLAT$_C$ in terms of accuracy, suggesting the discrete NCG is more robust against the noise in the context as we ex-

pected in Section 3.3. However, DOCFLAT$_D$ also aggressively filters out the supportive information in the context as demonstrated on its lower results in BLEU and COMET on News and Europarl. We believe tuning γ in Equation 9 can fix this issue.

Computational Efficiency As described in Section 3.2, FBA introduces additional computational overhead. We thus evaluate computational efficiency of DOCFLAT along with the baselines in terms of update per second (UPS) and report the results in Table 2. When the context size of the pseudo-document is the same, our approach DOCFLAT is almost as fast as the standard DOC2DOC on the identical computational infrastructure (one Tesla A40 GPU) with significant performance gain. Note that GTRANS (Bao et al., 2021) does not support FP16 mode, so its UPS is not reported. During the inference, DOCFLAT is only 15% slower than DOC2DOC.

4.3 Ablation Study

Ablation Study for FBA We conduct the ablation study for FBA and present the results in Table 3. Compared with DOC2DOC, FBA at either side can effectively improve the model performance of DOCFLAT in terms of BLEU, COMET and accuracy, although both FBAs does not demonstrate orthogonal effectiveness. We also observe that, when the FBA at the decoder side is removed, the contextual awareness (accuracy) is slightly im-

	Enc.	Dec.	BLEU	COMET	Acc.
Doc2Doc	∅	∅	24.86	0.2821	66.98
DocFlat$_\text{C}$			25.31	0.3173	70.92
		∅	25.58	0.3114	**71.83**
	∅		**25.70**	**0.3176**	71.06
DocFlat$_\text{D}$			25.41	0.3101	72.04
		∅	25.22	0.3112	**73.60**
	∅		**25.71**	0.3113	70.48

Table 3: Ablation study for FBA on TED. ∅ indicates FBA is removed. The best results for DocFlat$_\text{C}$ and DocFlat$_\text{D}$ are highlighted in **bold** respectively.

	BLEU	COMET	Acc.
DocFlat$_\text{C}$	25.31	**0.3173**	70.92
DocFlat$_\text{D}$	**25.41**	0.3101	**72.04**
DocFlat$_\text{I}$	25.07	0.3049	69.45

Table 4: Ablation study for NCG on TED. DocFlat$_\text{I}$ indicates DocFlat with identity mapping in NCG. The best results are highlighted in **bold**.

	GC	LC	BLEU	COMET	Acc.
Doc2Doc	∅	✓	25.01	0.3021	66.99
	✗	✗	—	—	—
DocFlat$_\text{C}$	✓	✓	**25.31**	**0.3173**	**70.92**
	✗	✓	24.86	0.2738	66.83
	✗	✗	—	—	—
DocFlat$_\text{D}$	✓	✓	**25.41**	**0.3101**	**72.04**
	✗	✓	24.81	0.2037	69.30
	✗	✗	—	—	—

Table 5: Ablation study for data shuffling on TED. ✓ indicates the golden context. ✗ indicates the irrelevant context. — indicates the model fails to converge. ∅ indicates Doc2Doc is not associated with GC. The best results for DocFlat$_\text{C}$ and DocFlat$_\text{D}$ are highlighted in **bold** respectively.

	avg	*er*	*es*	*sie*
Doc2Doc	66.99	56.82	89.20	54.95
MBE	68.12	52.57	89.72	62.07
ABD	68.25	55.30	**90.65**	58.82
DocFlat$_\text{C}$	70.92	56.65	89.52	**66.60**
DocFlat$_\text{D}$	**72.04**	**60.02**	89.67	66.42

Table 6: Accuracy (in %) on the contrastive test set for TED with regard to the anaphoric pronoun types. The best results are highlighted in **bold**.

proved. All these results demonstrate that FBA can effectively leverage the distant context beyond the pseudo-document boundaries.

Ablation Study for NCG We present the ablation study for NCG in Table 4. To probe the utility of NCG, we replace NCG in Equation 7 with identity mapping (He et al., 2016) and denote this variant of DocFlat as DocFlat$_\text{I}$. The results from DocFlat$_\text{I}$ support our argument that not all the information in the context is useful. Both DocFlat$_\text{C}$ and DocFlat$_\text{D}$ outperform DocFlat$_\text{I}$ on BLEU, COMET and accuracy, which confirms NCG can effectively filter out the noise in the context.

Ablation Study for Data Shuffling To preserve the linguistic structure of the original document, we do not shuffle examples during training. If the examples are shuffled, the predictions of the current sentence are conditioned on the gold local context (LC) and the irrelevant global context (GC). In this section, we investigate how the data shuffling affects DocFlat. We present the results in Table 5. We observe the performance reduction for DocFlat$_\text{C}$ and DocFlat$_\text{D}$ when the current sentence is conditioned on the gold LC but irrelevant GC. When conditioned on the irrelevant GC, DocFlat even performs worse than Doc2Doc which is free from the irrelevant GC. We also train DocFlat and Doc2Doc with the completely irrelevant context and find out that both models fail

to converge. Hence, we confirm that the relatedness between the context and current sentence is of vital importance in DocNMT and DocFlat can effectively leverage the information from the context beyond the scope of the pseudo-documents.

5 Analysis

In this section, we investigate the effectiveness of DocFlat on the contextual awareness and the quality saturation problem. We also demonstrate how the inference batch size affects the model predictions. A visualization of FBA attention map is presented in Appendix D.

Contextual Awareness In English-German translation, the choice of anaphoric pronoun types, including feminine *sie*, neutral *er* and masculine *es*, commonly depends on its context. We present the accuracy with regard to the anaphoric pronoun types given by the selected models trained on TED in Table 6. DocFlat$_\text{D}$ is the only approach that demonstrates substantial improvements on the neutral *er*. For the feminine *sie*, DocFlat$_\text{C}$ and DocFlat$_\text{D}$ both outperform Doc2Doc by approximately 12% accuracy. MBE and ABD only improves the accuracy on the feminine *sie* by 7%
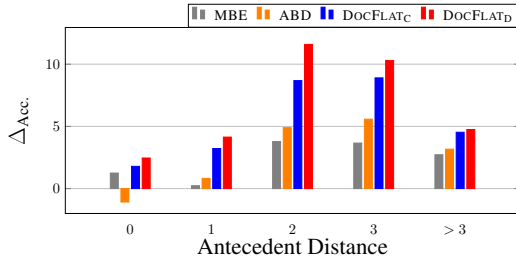
Figure 4: The change of accuracy (in %; Δ_{acc}) given by the selected models against DOC2DOC with regard to the antecedent distance (in sentences) on TED.
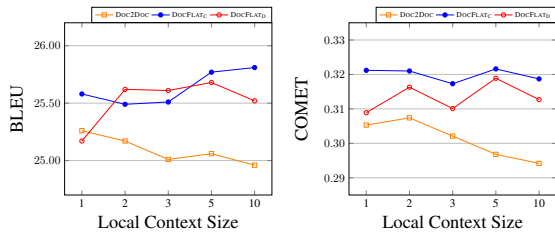

Figure 5: BLEU (left) and COMET (right) against the LC size of pseudo-document (in sentences) by DOC2DOC, DOCFLAT$_C$ and DOCFLAT$_D$ on TED.

and 4% respectively. We also present the change of accuracy given by the selected models against DOC2DOC with regard to the antecedent distance on TED in Figure 4. Compared with DOC2DOC, the approaches that leverage the batch-level information all effectively improves the accuracy on the distant context (antecedent distance ≥ 2). DOCFLAT$_C$ significantly outperforms DOC2DOC with regard to the accuracy on the distant context by more than 8%, while DOCFLAT$_D$ outperforms DOC2DOC by more than 10% on the distant context. All these results demonstrate that DOCFLAT can effectively improve the contextual awareness on the discourse phenomena.

Effect of DOCFLAT on Quality Saturation DOC2DOC suffers from the *quality saturation problem* as shown in Figure 1. We investigate if DOCFLAT also suffers from the same problem. We display the results in Figure 5 and observe that DOCFLAT$_C$ and DOCFLAT$_D$ perform consistently with regard to the LC size. We conjecture the reasons for this observation from two perspectives. When the LC size is small, the information from GC introduced by FBA complements the missing information in LC. When the LC size is large enough, most information from GC is already covered by LC and FBA functions as a regularizer.
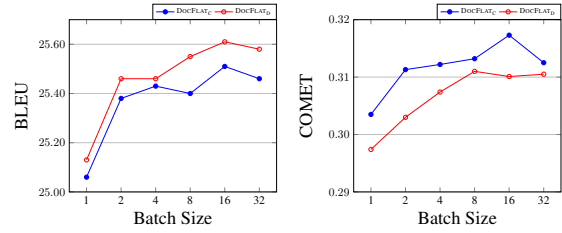

Figure 6: BLEU (left) and COMET (right) against the inference batch size (in sentences) given by DOCFLAT$_C$ and DOCFLAT$_D$ on TED.

Inference Batch Size At the inference stage, DOCFLAT is also able to leverage the batch-level information. We visualize how the inference batch size impacts the model performance in Figure 6. Overall, the model performance of DOCFLAT is positively correlated to the inference batch size. When the batch size is 1, DOCFLAT$_C$ and DOCFLAT$_D$ still outperform DOC2DOC, suggesting the FBA can help the model utilize the distant context during training. The performance gain on BLEU and COMET for both DOCFLAT$_C$ and DOCFLAT$_D$ diminishes as the inference batch size increases, and we do not observe further improvement when the inference batch size is larger than 16, suggesting the over-distant context is less influential to the predictions of the current sentence.

6 Related Work

Document-Level NMT Numerous document-level NMT approaches have been proposed in recent years. Tiedemann and Scherrer (2017) firstly proposed the simple concatenation-based DocNMT model. Existing works in the document-level NMT widely spread on a variety of research topics, including the model architecture (Miculicich et al., 2018; Maruf et al., 2019; Zhang et al., 2021), training methods (Sun et al., 2022; Lei et al., 2022), evaluation (Bawden et al., 2018; Jiang et al., 2022), etc. Zhang et al. (2018b) incorporate the contextual information using an independent context encoder. Bao et al. (2021) propose group attention that introduce a locality bias to force the model to focus on the recent context. Morishita et al. (2021) compute the average representation of all the source tokens, which is the only close work to ours. Maruf et al. (2021) present a detailed review on DocNMT.

Batch-Level Information Modeling instance relationships in the batch is relatively less explored. Prior works leveraging the instance relationships are mostly from the computer vision area. Ioffe

and Szegedy (2015) keep the running mean and variance in the batch to normalize the training and testing instances. Zhang et al. (2018a) linearly combine a random pair of instances to improve the model generalization. Mondal et al. (2021) use graph neural networks to aggregate information from similar images. Hou et al. (2022) propose BATCHFORMER to improve the long-tail recognition by combining different instances. Our work is directly inspired by Kossen et al. (2021) that computes the pairwise similarity among all the batched instances, with distinct motivation. We aim to utilize distant context beyond the pseudo-document boundaries, instead of finding the similar patterns.

7 Conclusion

In this work, we address the limitation of the pseudo-document formation in the DocNMT by utilizing the batch-level information. We propose a novel Document Flattening (DOCFLAT) technique that integrates FLAT-BATCH ATTENTION (FBA) and NEURAL CONTEXT GATE (NCG) into the Transformer model. FBA enables the current sentence to access the information beyond the pseudo-document boundaries and NCG identifies the usefulness of context and controls the information flow. We conduct comprehensive experiments and analyses on three benchmark datasets for English-German translation. We demonstrate that DOCFLAT outperforms several strong baselines with statistical significance. The analyses highlight that DOCFLAT can effectively alleviate the quality saturation problem in DocNMT and capture the long-range information.

8 Limitations

As suggested in Figure 6, the performance of DOCFLAT is positively correlated to the inference batch size. This is because large inference batch size could help DOCFLAT to better utilize distant context within the same inference batch. However, this property of DOCFLAT could become an issue when there are only limited inference computational resources available.

Acknowledgment

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. 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 3442–3455, Online. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. 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 6467–6478, Online. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of

human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1019–1027.

BG Glaser and AL Strauss. 1967. The discovery of grounded theory: strategies for qualitative research.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Zhi Hou, Baosheng Yu, and Dacheng Tao. 2022. Batchformer: Learning to explore sample relationships for robust representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7256–7266.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Jannik Kossen, Neil Band, Clare Lyle, Aidan N. Gomez, Thomas Rainforth, and Yarin Gal. 2021. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 28742–28756.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Yikun Lei, Yuqi Ren, and Deyi Xiong. 2022. CoDoNMT: Modeling cohesion devices for document-level neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5205–5216, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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 F. 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, Lisboa, Portugal. European Association for Machine Translation.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware 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 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

Lesly Miculicich, 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, Brussels, Belgium. Association for Computational Linguistics.

Arnab Kumar Mondal, Vineet Jain, and Kaleem Siddiqi. 2021. Mini-batch similarity graphs for robust image classification. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 194. BMVA Press.

Makoto Morishita, Jun Suzuki, Tomoharu Iwata, and Masaaki Nagata. 2021. Context-aware neural machine translation with mini-batch embedding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2513–2521, Online. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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, Copenhagen, Denmark. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018a. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018b. 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, Brussels, Belgium. Association for Computational Linguistics.

Long Zhang, Tong Zhang, Haibo Zhang, Baosong Yang, Wei Ye, and Shikun Zhang. 2021. Multi-hop transformer for document-level machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3953–3963, Online. Association for Computational Linguistics.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3983–3989. ijcai.org.

A A Concrete Example at the Source Side

We present a concrete example at the source side in Figure 7.

B Optimization and Hyperparameters

We use a two-stage training routine following the previous works (Zhang et al., 2018b; Voita et al., 2019; Lopes et al., 2020; Bao et al., 2021).

Stage I We first train a SENT2SENT NMT model. The model is randomly initialized and optimized with Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and the learning rate $\alpha = 5 \times 10^{-4}$. The model is trained with the batch size of $32K$ tokens for both datasets and the dropout rate $p = 0.3$. The batch size of $32K$ tokens is achieved by using the batch size of 4096 tokens and updating the model for every 8 batches. The learning rate schedule is the same as described in Vaswani et al. (2017) with $4K$ warmup steps. We use early stopping on validation loss.

Stage II The document-level models are all fine-tuned from the best SENT2SENT model in the Stage I. With the same learning rate schedule as the Stage I, we set the learning rate $\alpha = 2 \times 10^{-4}$. All the other hyperparameters are identical. Training is early stopped on validation loss, and we average the last 5 checkpoints to report the model performance, following Vaswani et al. (2017). Following Bao et al. (2021), we apply word dropout (Gal and Ghahramani, 2016; Sennrich et al., 2016a) to the inputs with $p = 0.1$.

C ATTENTION BETWEEN DATAPOINTS

We adapt ATTENTION BETWEEN DATAPOINTS (ABD) proposed by Kossen et al. (2021) to the DocNMT. The model architecture is identical to DOCFLAT as shown in Figure 2 with FBA replaced with ABD. Given a batch of hidden representations $\boldsymbol{H} \in \mathbb{R}^{n \times d \times e}$, ABD is defined as follows:

$$
\begin{aligned}
\tilde{\boldsymbol{H}}_{\text{avg}} &= \text{AvgPool}(\boldsymbol{H}) \in \mathbb{R}^{n \times 1 \times e}, \\
\tilde{\boldsymbol{H}}_{\text{flat}} &= \text{Flatten}(\tilde{\boldsymbol{H}}_{\text{avg}}) \in \mathbb{R}^{1 \times n \times e}, \\
\tilde{\boldsymbol{H}}_{\text{mhsa}} &= \text{MHSA}(\tilde{\boldsymbol{H}}_{\text{flat}}, \tilde{\boldsymbol{H}}_{\text{flat}}, \tilde{\boldsymbol{H}}_{\text{flat}}) \in \mathbb{R}^{1 \times n \times e}, \\
\tilde{\boldsymbol{H}}_{\text{rsh}} &= \text{Reshape}(\text{Repeat}(\tilde{\boldsymbol{H}}_{\text{mhsa}})) \in \mathbb{R}^{n \times d \times e}, \\
\tilde{\boldsymbol{H}} &= \text{LN}(\boldsymbol{H} + \tilde{\boldsymbol{H}}_{\text{rsh}}) \in \mathbb{R}^{n \times d \times e}.
\end{aligned}
\tag{11}
$$

There is a noticeable difference in Equation 11 from Equation 7 that we apply the average pooling

	Enc.	Dec.	BLEU	COMET	Acc.
DOC2DOC	✗	✗	25.01	0.3021	66.99
ABD			18.57	-0.1202	66.55
	✗		18.46	-0.1123	66.47
		✗	24.97	0.3046	68.25

Table 7: Preliminary study on the usage of ABD on TED. ✗ indicates ABD is removed.

to the sequence to obtain the instance representation, instead of directly flattening the token representations into a single vector. ABD is originally designed for fixed-length data, and it is non-trivial to apply ABD to the variable-length inputs, and hence, we use the average pooling for simplicity.

We present the preliminary study of ABD on TED in Table 7. When ABD is applied at the decoder side, the model performance is significantly reduced. This observation suggests that the linguistic structure at the target side is of vital importance to DocNMT.
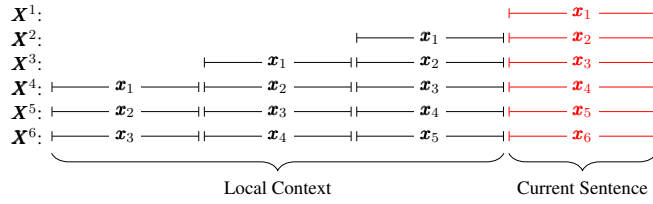
D Visualization of FBA

To better understand the behavior of FBA, we visualize the sentence-wise attention map learned by the FBA of DOCFLAT$_C$ and DOCFLAT$_D$ at the encoder side in Figure 8. The sample document for producing Figure 8 can be found in Table 8.

It is infeasible to visualize the token-wise attention map for a very long sequence, so we aggregate the token-wise attention scores into the sentence-level. We denote the token-wise attention map for the flattened sequence as $\mathcal{A}$. For each pair of sentences $\boldsymbol{s}_i$ attending to $\boldsymbol{s}_j$, their token-wise attention map is a patch of $\mathcal{A}$, denoted as $\mathcal{A}_{ij}^p$. We aggregate the token-level attention scores in the attention patch $\mathcal{A}_{ij}^p$ into a sentence-level score, as follows:

$$
\mathcal{A}_{\mathcal{S}}(i, j) = \frac{1}{\left| \mathcal{A}_{ij}^p \right|} \sum \mathcal{A}_{ij}^p
\tag{12}
$$

where $\left| \mathcal{A}_{ij}^p \right|$ is the size of $\mathcal{A}_{ij}^p$ and $\mathcal{A}_{\mathcal{S}}(i, j)$ is the sentence-level attention score for $\boldsymbol{s}_i$ attending to $\boldsymbol{s}_j$.

The FBA of DOCFLAT$_C$ at the encoder side considers all the sentences in the document to be equally important, while the one of DOCFLAT$_D$ approximately splits the whole documents into two parts. As shown by DOCFLAT$_D$ in Figure 8, sentences in the first half focus more on its neighbors in the same split but those in the second half

$\boldsymbol{X}^1$:
$\boldsymbol{X}^2$:
$\boldsymbol{X}^3$:
$\boldsymbol{X}^4$:
$\boldsymbol{X}^5$:
$\boldsymbol{X}^6$:

Local Context   Current Sentence

(a) An example batch of pseudo-documents $\boldsymbol{B}_{\text{src}} = \{\boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{X}^3, \boldsymbol{X}^4, \boldsymbol{X}^5, \boldsymbol{X}^6\}$ at the source side. Each $\boldsymbol{X}^j$ contains four consecutive sentences and $\boldsymbol{x}_i$ indicates the i-th sentence of the same original document. The segments in red indicate the current sentence of each pseudo-document.

$\boldsymbol{X}^{\text{flat}}$:

Extra Context   Current Sentence   Extra Context

Global Context   Global Context
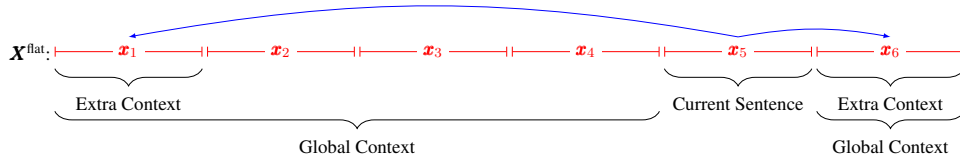
(b) An example of the flattened sequence $\boldsymbol{X}^{\text{flat}}$ transformed from $\boldsymbol{B}_{\text{tgt}}$ with FLAT-BATCH ATTENTION. For the current sentence $\boldsymbol{x}_5$, The blue arrows indicate the extra inter-sentential attention for $\boldsymbol{x}_5$ that our approach can model. $\boldsymbol{x}_1$ and $\boldsymbol{x}_6$ are the extra context introduced by our approach.

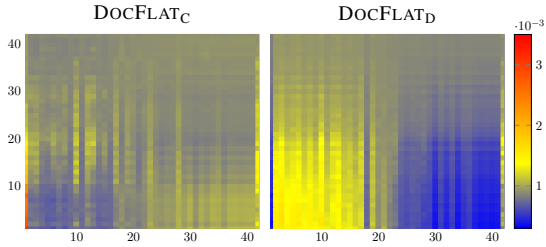Figure 7: An example batch of pseudo-documents at the source side and its flattened sequence.

Figure 8: Sentence-wise attention map produced by the FBA of DOCFLAT$_C$ and DOCFLAT$_D$ at the encoder side. x-axis indicates sentences as the keys of FBA. y-axis indicates sentences as the queries of FBA.

roughly attend to all the sentences in the documents. This observation implies that the latter context is more dependent on the former context.

461

idx	Context
1	<d>
2	We're at a tipping point in human history, a species poised between gaining the stars and losing the planet we call home.
3	Even in just the past few years, we've greatly expanded our knowledge of how Earth fits within the context of our universe.
4	NASA's Kepler mission has discovered thousands of potential planets around other stars, indicating that Earth is but one of billions of planets in our galaxy.
5	Kepler is a space telescope that measures the subtle dimming of stars as planets pass in front of them, blocking just a little bit of that light from reaching us.
6	Kepler's data reveals planets' sizes as well as their distance from their parent star.
7	Together, this helps us understand whether these planets are small and rocky, like the terrestrial planets in our own Solar System, and also how much light they receive from their parent sun.
8	In turn, this provides clues as to whether these planets that we discover might be habitable or not.
9	Unfortunately, at the same time as we're discovering this treasure trove of potentially habitable worlds, our own planet is sagging under the weight of humanity.
10	2014 was the hottest year on record.
11	Glaciers and sea ice that have been with us for millennia are now disappearing in a matter of decades.
12	These planetary-scale environmental changes that we have set in motion are rapidly outpacing our ability to alter their course.
13	But I'm not a climate scientist, I'm an astronomer.
14	I study planetary habitability as influenced by stars with the hopes of finding the places in the universe where we might discover life beyond our own planet.
15	You could say that I look for choice alien real estate.
16	Now, as somebody who is deeply embedded in the search for life in the universe, I can tell you that the more you look for planets like Earth, the more you appreciate our own planet itself.
17	Each one of these new worlds invites a comparison between the newly discovered planet and the planets we know best: those of our own Solar System.
18	Consider our neighbor, Mars.
19	Mars is small and rocky, and though it's a bit far from the Sun, it might be considered a potentially habitable world if found by a mission like Kepler.
20	Indeed, it's possible that Mars was habitable in the past, and in part, this is why we study Mars so much.
21	Our rovers, like Curiosity, crawl across its surface, scratching for clues as to the origins of life as we know it.
22	Orbiters like the MAVEN mission sample the Martian atmosphere, trying to understand how Mars might have lost its past habitability.
23	Private spaceflight companies now offer not just a short trip to near space but the tantalizing possibility of living our lives on Mars.
24	But though these Martian vistas resemble the deserts of our own home world, places that are tied in our imagination to ideas about pioneering and frontiers, compared to Earth Mars is a pretty terrible place to live.
25	Consider the extent to which we have not colonized the deserts of our own planet, places that are lush by comparison with Mars.
26	Even in the driest, highest places on Earth, the air is sweet and thick with oxygen exhaled from thousands of miles away by our rainforests.
27	I worry – I worry that this excitement about colonizing Mars and other planets carries with it a long, dark shadow: the implication and belief by some that Mars will be there to save us from the self-inflicted destruction of the only truly habitable planet we know of, the Earth.
28	As much as I love interplanetary exploration, I deeply disagree with this idea.
29	There are many excellent reasons to go to Mars, but for anyone to tell you that Mars will be there to back up humanity is like the captain of the Titanic telling you that the real party is happening later on the lifeboats.
30	Thank you.
31	But the goals of interplanetary exploration and planetary preservation are not opposed to one another.
32	No, they're in fact two sides of the same goal: to understand, preserve and improve life into the future.
33	The extreme environments of our own world are alien vistas.
34	They're just closer to home.
35	If we can understand how to create and maintain habitable spaces out of hostile, inhospitable spaces here on Earth, perhaps we can meet the needs of both preserving our own environment and moving beyond it.
36	I leave you with a final thought experiment: Fermi's paradox.
37	Many years ago, the physicist Enrico Fermi asked that, given the fact that our universe has been around for a very long time and we expect that there are many planets within it, we should have found evidence for alien life by now.
38	So where are they?
39	Well, one possible solution to Fermi's paradox is that, as civilizations become technologically advanced enough to consider living amongst the stars, they lose sight of how important it is to safeguard the home worlds that fostered that advancement to begin with.
40	It is hubris to believe that interplanetary colonization alone will save us from ourselves, but planetary preservation and interplanetary exploration can work together.
41	If we truly believe in our ability to bend the hostile environments of Mars for human habitation, then we should be able to surmount the far easier task of preserving the habitability of the Earth.
42	Thank you.

Table 8: The sample document used for producing Figure 8.