Toward Making the Most of Context in Neural Machine Translation

Zaixiang Zheng[1*†] , **Xiang Yue**[1*] , **Shujian Huang**[1] , **Jiajun Chen**[1] and **Alexandra Birch**[2]

[1]National Key Laboratory for Novel Software Technology, Nanjing University
[2]ILCC, School of Informatics, University of Edinburgh
{zhengzx,xiangyue}@smail.nju.edu.cn, {huangsj,daixinyu,chenjj}@nju.edu.cn, a.birch@ed.ac.uk

Abstract

Document-level machine translation manages to outperform sentence level models by a small margin, but have failed to be widely adopted. We argue that previous research did not make a clear use of the global context, and propose a new document-level NMT framework that deliberately models the local context of each sentence with the awareness of the global context of the document in both source and target languages. We specifically design the model to be able to deal with documents containing any number of sentences, including single sentences. This unified approach allows our model to be trained elegantly on standard datasets without needing to train on sentence and document level data separately. Experimental results demonstrate that our model outperforms Transformer baselines and previous document-level NMT models with substantial margins of up to 2.1 BLEU on state-of-the-art baselines. We also provide analyses which show the benefit of context far beyond the neighboring two or three sentences, which previous studies have typically incorporated.

1 Introduction

Recent studies suggest that neural machine translation (NMT) [Sutskever *et al.*, 2014; Bahdanau *et al.*, 2015; Vaswani *et al.*, 2017] has achieved human parity, especially on resource-rich language pairs [Hassan *et al.*, 2018]. However, standard NMT systems are designed for sentence-level translation, which cannot consider the dependencies among sentences and translate entire documents. To address the above challenge, various document-level NMT models, viz., context-aware models, are proposed to leverage context beyond a single sentence [Wang *et al.*, 2017; Miculicich *et al.*, 2018; Zhang *et al.*, 2018; Yang *et al.*, 2019] and have achieved substantial improvements over their context-agnostic counterparts.

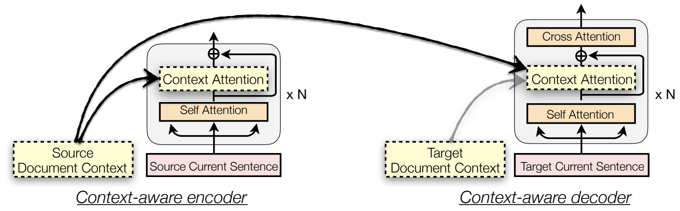

Figure 1: Illustration of typical Transformer-based context-aware approaches (some of them do not consider target context (grey line)).

Figure 1 briefly illustrates typical context-aware models, where the source and/or target document contexts are regarded as an additional input stream parallel to the current sentence, and incorporated into each layer of encoder and/or decoder [Zhang *et al.*, 2018; Tan *et al.*, 2019]. More specifically, the representation of each word in the current sentence is a deep hybrid of both *global* document context and *local* sentence context in every layer. We notice that these hybrid encoding approaches have two main weaknesses:

- *Models are context-aware, but do not fully exploit the context*. The deep hybrid makes the model more sensitive to noise in the context, especially when the context is enlarged. This could explain why previous studies show that enlarging context leads to performance degradation. Therefore, these approaches have not taken the best advantage of the entire document context.
- *Models translate documents, but cannot translate single sentences*. Because the deep hybrid requires global document context as additional input, these models are no longer compatible with sentence-level translation based on the solely local sentence context. As a result, these approaches usually translate poorly for single sentence documents without document-level context.

In this paper, we mitigate the aforementioned two weaknesses by designing a general-purpose NMT architecture which can fully exploit the context in documents of arbitrary number of sentences. To avoid the deep hybrid, our architecture balances *local context* and *global context* in a more deliberate way. To be more specific, our architecture independently encodes local context in the source sentence, instead of mixing it with global context from the beginning so it is robust to when the global context is large and noisy. Furthermore our architecture translates in a sentence-by-sentence

[*] Equal contribution.

[†] This work was done when Zaixiang was visiting at the University of Edinburgh.

manner with access to the partially generated document translation as the target global context which allows the local context to govern the translation process for single-sentence documents.

We highlight our contributions in three aspects:

- We propose a new NMT framework that is able to deal with documents containing any number of sentences, including single-sentence documents, making training and deployment simpler and more flexible.
- We conduct experiments on four document-level translation benchmark datasets, which show that the proposed unified approach outperforms Transformer baselines and previous state-of-the-art document-level NMT models both for sentence-level and document-level translation.
- Based on thorough analyses, we demonstrate that the document context really matters; and the more context provided, the better our model translates. This finding is in contrast to the prevailing consensus that a wider context deteriorates translation quality.

2 Related Work

Context beyond the current sentence is crucial for machine translation. Bawden *et al.* [2018], Läubli *et al.* [2018], Müller *et al.* [2018], and Voita *et al.* [2018] show that without access to the document-level context, NMT is likely to fail to maintain lexical, tense, deixis and ellipsis consistencies, resolve anaphoric pronouns and other discourse characteristics, and propose corresponding testsets for evaluating discourse phenomena in NMT.

Most of the current document-level NMT models can be classified into two main categories, context-aware model, and post-processing model. The post-processing models introduce an additional module that learns to refine the translations produced by context-agnostic NMT systems to be more discourse coherence [Xiong *et al.*, 2019; Voita *et al.*, 2019]. While this kind of approach is easy to deploy, the two-stage generation process may result in error accumulation.

In this paper, we pay our attention mainly on context-aware models, while post-processing approaches can be incorporated with and facilitate any NMT architectures. Tiedemann and Scherrer [2017] and Junczys-Dowmunt [2019] use the concatenation of multiple sentences (usually a small number of preceding sentences) as NMT's input/output. Going beyond simple concatenation, Jean *et al.* [2017] introduce a separate context encoder for a few previous source sentences. Wang *et al.* [2017] includes a hierarchical RNN to summarize source-side context. There are other approaches using a dynamic cache memory to store representations of previously translated contents [Tu *et al.*, 2018; Kuang *et al.*, 2018; Kuang and Xiong, 2018; Maruf and Haffari, 2018]. Miculicich *et al.* [2018], Zhang *et al.* [2018], Yang *et al.* [2019], Maruf *et al.* [2019] and Tan *et al.* [2019] extend context-aware model to Transformer architecture with additional context related modules.

While claiming that modeling the whole document is not necessary, these models only take into account a few surrounding sentences [Maruf and Haffari, 2018; Miculicich *et al.*, 2018; Zhang *et al.*, 2018; Yang *et al.*, 2019], or even only monolingual context [Zhang *et al.*, 2018; Yang *et al.*, 2019; Tan *et al.*, 2019], which is not necessarily sufficient to translate a document. On the contrary, our model can consider the entire arbitrary long document and simultaneously exploit contexts in both source and target languages. Furthermore, most of these document-level models cannot be applied to sentence-level translation, lacking both simplicity and flexibility in practice. They rely on variants of components specifically designed for document context (e.g., encoder/decoder-to-context attention embedded in all layers [Zhang *et al.*, 2018; Miculicich *et al.*, 2018; Tan *et al.*, 2019]), being limited to the scenario where the document context must be the additional input stream. Thanks to our general-purpose modeling, the proposed model manages to do general translation regardless of the number of sentences of the input text.

3 Background

Sentence-level NMT Standard NMT models usually model sentence-level translation (SENTNMT), adopting an *encoder-decoder* framework [Bahdanau *et al.*, 2015]. Here SENTNMT models aim to approximate the conditional distribution $\log p(y|x;\theta)$ over a target sentence $y = \langle y_1, \ldots, y_T \rangle$ given a source sentence $x = \langle x_1, \ldots, x_I \rangle$. Training criterion for a sentence-level NMT model is to maximize the conditional log-likelihood $\log p(y|x;\theta)$ on abundant parallel bilingual data $\mathcal{D}_\text{s} = \{x^{(m)}, y^{(m)}\}_{m=1}^M$ of i.i.d observations.

$$\mathcal{L}(\mathcal{D}_\text{s};\theta) = \sum_{m=1}^{M} \log p(y^{(m)}|x^{(m)};\theta)$$

Document-level NMT Given a document-level parallel dataset $D_\text{d} = \{X^{(m)}, Y^{(m)}\}_{m=1}^M$, where $X^{(m)} = \langle x_k^{(m)} \rangle_{k=1}^n$ is a source document containing n sentences while $Y^{(m)} = \langle y_k^{(m)} \rangle_{k=1}^n$ is a target document with n sentences, the training criterion for document-level NMT model (DOCNMT) is to maximize the conditional log-likelihood over the pairs of document translation sentence by sentence by:

$$\begin{aligned}\mathcal{L}(\mathcal{D}_\text{d};\theta\}) &= \sum_{m=1}^{M} \log p(Y^{(m)}|X^{(m)};\theta) \\ &= \sum_{m=1}^{M}\sum_{k=1}^{n} \log p(y_k^{(m)}|y_{<k}^{(m)}, x_k^{(m)}, x_{-k}^{(m)}, ;\theta)\end{aligned}$$

where $y_{<k}^{(m)}$ denotes the history translated sentences prior to $y_k^{(m)}$, while $x_{-k}^{(m)}$ means the rest of the source sentences other than the current k-th source sentence $x_k^{(m)}$.

4 Approach

By the definition of local and global contexts, general translation can be seen as a hierarchical natural language understanding and generation problem based on local and global contexts. Accordingly, we propose a general-purpose architecture to exploit context machine translation to a better extent. Figure 2 illustrates the idea of our proposed architecture:

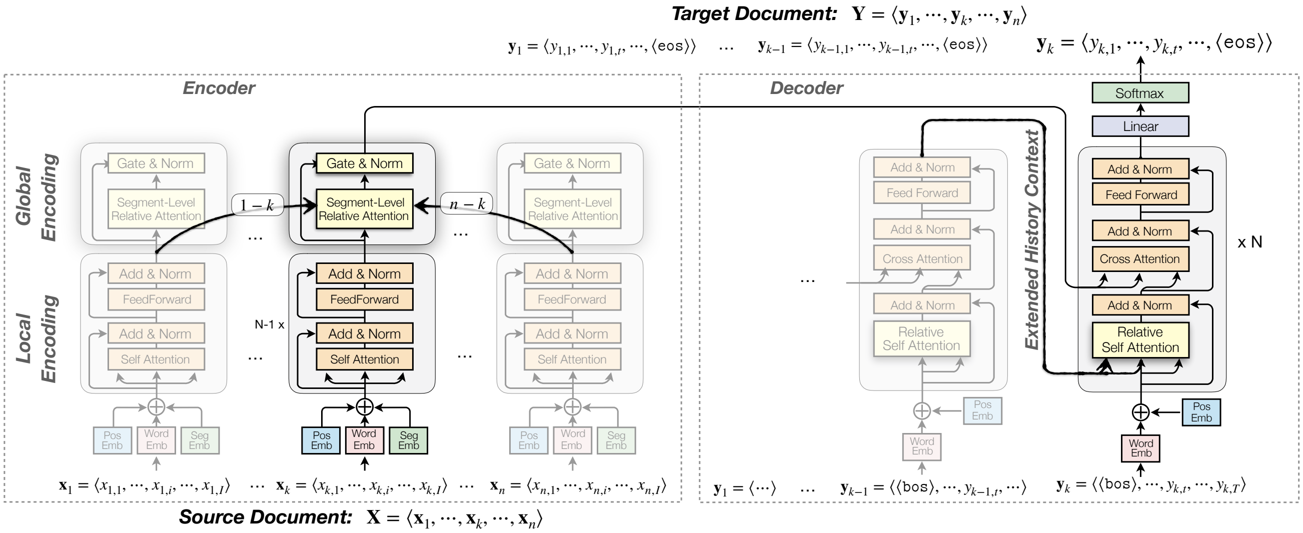
Target Document: $\mathbf{Y} = \langle \mathbf{y}_1, \cdots, \mathbf{y}_k, \cdots, \mathbf{y}_n \rangle$
$\mathbf{y}_1 = \langle y_{1,1}, \cdots, y_{1,t}, \cdots, \langle \text{eos} \rangle \rangle$ … $\mathbf{y}_{k-1} = \langle y_{k-1,1}, \cdots, y_{k-1,t}, \cdots, \langle \text{eos} \rangle \rangle$
$\mathbf{y}_k = \langle y_{k,1}, \cdots, y_{k,t}, \cdots, \langle \text{eos} \rangle \rangle$
Encoder
Decoder
Global Encoding
Local Encoding
Gate & Norm
Segment-Level Relative Attention
$1-k$
$n-k$
Add & Norm
FeedForward
Self Attention
N-1 x
Pos Emb
Word Emb
Seg Emb
$\mathbf{x}_1 = \langle x_{1,1}, \cdots, x_{1,i}, \cdots, x_{1,I} \rangle$ … $\mathbf{x}_k = \langle x_{k,1}, \cdots, x_{k,i}, \cdots, x_{k,I} \rangle$ … $\mathbf{x}_n = \langle x_{n,1}, \cdots, x_{n,i}, \cdots, x_{n,I} \rangle$
Source Document: $\mathbf{X} = \langle \mathbf{x}_1, \cdots, \mathbf{x}_k, \cdots, \mathbf{x}_n \rangle$
Softmax
Linear
Feed Forward
Cross Attention
Relative Self Attention
x N
Extended History Context
$\mathbf{y}_1 = \langle \cdots \rangle$ … $\mathbf{y}_{k-1} = \langle \langle \text{bos} \rangle, \cdots, y_{k-1,t}, \cdots \rangle$
$\mathbf{y}_k = \langle \langle \text{bos} \rangle, \cdots, y_{k,t}, \cdots, y_{k,T} \rangle$

Figure 2: Illustration of the proposed model. The local encoding is complete and independent, which also allows context-agnostic generation.

- Given a source document, the encoder builds local context for each individual sentence (local encoding) and then retrieves global context from the entire source document to understand the inter-sentential dependencies (global encoding) and form hybrid contextual representations (context fusion). For single sentence generation, the global encoding will be dynamically disabled and the local context can directly flow through to the decoder to dominate translation. (Section 4.1)
- Once the local and global understanding of the source document is constructed, the decoder generates target document by sentence basis, based on source representations of the current sentence as well as target global context from previous translated history and local context from the partial translation so far. (Section 4.2)

This general-purpose modeling allows the proposed model to fully utilize bilingual and entire document context and go beyond the restricted scenario where models must have document context as additional input streams and fail to translate single sentences. These two advantages meet our expectation of a unified and general NMT framework.

4.1 Encoder

Lexical and Positional Encoding

The source input will be transformed to representations by lexical and positional encoding. We use word position embedding in Transformer [Vaswani *et al.*, 2017] to represent the ordering of words. Note that we reset the word position for each sentence, i.e., the i-th word in each sentence shares the word position embedding E_i^w. Besides, we introduce segment embedding E_k^s to represent the k-th sentence. Therefore, the representation of i-th word in k-th sentence is given by $\tilde{x}_{k,i} = E[x_{k,i}] + E_k^s + E_i^w$, where $E[x_{k,i}]$ means word embedding of $x_{k,i}$.

Local Context Encoding

We construct the local context for each sentence with a stack of standard transformer layers [Vaswani *et al.*, 2017]. Take the k-th source sentence x_k as an example. The local encoder leverages $N-1$ stacked identical layers to map the sentence into corresponding encoded representations.

$$\hat{\mathbf{h}}_k^l = \texttt{MultiHead}(\texttt{SelfAttn}(\mathbf{h}_k^{l-1}, \mathbf{h}_k^{l-1}, \mathbf{h}_k^{l-1})),$$
$$\mathbf{h}_k^l = \texttt{LayerNorm}(\texttt{FeedForward}(\hat{\mathbf{h}}_k^l) + \hat{\mathbf{h}}_k^l),$$

where $\texttt{SelfAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes self-attention, while $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ indicate *queries, keys*, and *values*, respectively. $\texttt{MultiHead}(\cdot)$ means the attention is performed in a multiheaded fashion [Vaswani *et al.*, 2017]. We let the input representations $\tilde{\mathbf{x}}_k$ to be the 0-th layer representations $\mathbf{h}_k^0$, while we denote the $(N-1)$-th layer of the local encoder as the local context for each sentence, i.e., $\mathbf{h}_k^L = \mathbf{h}_k^{N-1}$.

Global Context Encoding

We add an additional layer on the top of the local context encoding layers, which retrieves global context from the entire document by a *segment-level relative attention*, and outputs final representations based on hybrid local and global context by *gated context fusion* mechanism.

Segment-level Relative Attention Given the local representations of each sentences, we propose to extend the relative attention [Shaw *et al.*, 2018] from token-level to segment-level to model the inter-sentence global context:

$$\mathbf{h}^G = \texttt{MultiHead}(\texttt{Seg-Attn}(\mathbf{h}^L, \mathbf{h}^L, \mathbf{h}^L)),$$

where $\texttt{Seg-Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes the proposed segment-level relative attention. Let us take $x_{k,i}$ as query as an example, its the contextual representations $z_{k,i}$ by the proposed attention is computed over all words (e.g., $x_{\kappa,j}$) in the document regarding the sentence (segment) they belong to:

$$z_{k,i} = \sum_{\kappa=0}^{n} \sum_{j=1}^{|\mathbf{x}_\kappa|} \alpha_{k,i}^{\kappa,j}(x_{\kappa,j}W^V + \gamma_{k-\kappa}^V),$$
$$\alpha_{k,i}^{\kappa,j} = \texttt{softmax}(e_{k,i}^{\kappa,j}),$$

where $\alpha_{k,i}^{\kappa,j}$ is the attention weight of $x_{k,i}$ to $x_{\kappa,j}$. The corresponding attention logit $e_{k,i}^{\kappa,j}$ can be computed with respect to relative sentence distance by:

$$e_{k,i}^{\kappa,j} = \frac{(x_{k,i}W^Q)(x_{\kappa,j}W^K + \gamma_{k-\kappa}^K)^\top}{\sqrt{d_z}}, \tag{1}$$

where $\gamma_{k-\kappa}^*$ is a parameter vector corresponding to the relative distance between the k-th and κ-th sentences, providing inter-sentential clues. W^Q, W^K, and W^V are linear projection matrices for the queries, keys and values, respectively.

Gated Context Fusion After the global context is retrieved, we adopt a gating mechanism to obtain the final encoder representations $\mathbf{h}$ by fusing local and global context:

$$\mathbf{g} = \sigma(W_g[\mathbf{h}^L; \mathbf{h}^G]),$$
$$\mathbf{h} = \texttt{LayerNorm}\big((1-\mathbf{g}) \odot \mathbf{h}^L + \mathbf{g} \odot \mathbf{h}^G\big),$$

where W_g is a learnable linear transformation. $[\cdot;\cdot]$ denotes concatenation operation. $\sigma(\cdot)$ is sigmoid activation which leads the value of the fusion gate to be between 0 to 1. $\odot$ indicates element-wise multiplication.

4.2 Decoder

The goal of the decoder is to generate translations sentence by sentence by considering the generated previous sentences as target global context. A natural idea is to store the hidden states of previous target translations and allow the self attentions of the decoder to access to these hidden states as extended history context.

To that purpose, we leverage and extend Transformer-XL [Dai *et al.*, 2019] as the decoder. Transformer-XL is a novel Transformer variant, which is designed to cache and reuse the previous computed hidden states in the last segment as an extended context, so that long-term dependency information occurs many words back could propagate through the recurrence connections between segments, which just meets our requirement of generating document long text. We cast each sentence as a "segment" in translation tasks and equip the Transformer-XL based decoder with cross-attention to retrieve time-dependent source context for the current sentence. Formally, given two consecutive sentences, y_k and y_{k-1}, the l-th layer of our decoder first employs self-attention over the extended history context:

$$\tilde{\mathbf{s}}_k^{l-1} = [\texttt{SG}(\mathbf{s}_{k-1}^{l-1}); \mathbf{s}_k^{l-1}],$$
$$\bar{\mathbf{s}}_k^{l} = \texttt{MultiHead}(\texttt{Rel-SelfAttn}(\mathbf{s}_k^{l-1}, \tilde{\mathbf{s}}_k^{l-1}, \tilde{\mathbf{s}}_k^{l-1})),$$
$$\bar{\mathbf{s}}_k^{l} = \texttt{LayerNorm}(\bar{\mathbf{s}}_k^{l} + \mathbf{s}_k^{l-1}),$$

where the function $\texttt{SG}(\cdot)$ stands for stop-gradient. $\texttt{Rel-SelfAttn}(\mathbf{Q},\mathbf{K},\mathbf{V})$ is a variant of self-attention with word-level relative position encoding. For more specific details, please refer to [Dai *et al.*, 2019]. After that, the cross-attention module fetching the source context from encoder representation $\mathbf{h}_k$ is computed as:

$$\hat{\mathbf{s}}^{l} = \texttt{MultiHead}(\texttt{CrossAttn}(\bar{\mathbf{s}}_k^{l}, \mathbf{h}_k, \mathbf{h}_k)),$$
$$\mathbf{s}_k^{l} = \texttt{LayerNorm}(\texttt{FeedForward}(\hat{\mathbf{s}}_k^{l}) + \hat{\mathbf{s}}_k^{l}).$$

Based on the final representations of the last decoder layer $\mathbf{s}_k^N$, the output probability of current target sentence y_k are computed as:

$$p(y_k|y_{<k}, x_k, x_{-k}) = \prod_t p(y_{k,t}|y_{k,\le t}, y_{<k}, x_k, x_{-k})$$
$$= \prod_t \texttt{softmax}(E[y_{k,t}]^\top s_{k,t}^N).$$

5 Experiment

We experiment on four widely used document-level parallel datasets in two language pairs for machine translation:

- TED (ZH-EN/EN-DE). The Chinese-English and English-German TED datasets are from IWSLT 2015 and 2017 evaluation campaigns respectively. We mainly explore and develop our approach on TED ZH-EN, where we take dev2010 as development set and tst2010-2013 as testset. For TED EN-DE, we use tst2016-2017 as our testset and the rest as development set.
- News (EN-DE). We take News Commentary v11 as our training set. The WMT newstest2015 and newstest2016 are used for development and testsets respectively.
- Europarl (EN-DE). The corpus are extracted from the Europarl v7 according to the method mentioned in Maruf *et al.* [2019].[1]

We applied byte pair encoding [Sennrich *et al.*, 2016, BPE] to segment all sentences with 32K merge operations. We splited each document by 20 sentences to alleviate memory consumption in the training of our proposed models. We used the Transformer architecture as our sentence-level, context-agnostic baseline and develop our proposed model on the top of it. For models on TED ZH-EN, we used a configuration smaller than `transformer_base` [Vaswani *et al.*, 2017] with model dimension $d_z = 256$, dimension $d_{\text{ffn}} = 512$ and number of layers $N = 4$. As for models on the rest datasets, we change the dimensions to 512/2048. We used the Adam optimizer [Kingma and Ba, 2014] and the same learning rate schedule strategy as [Vaswani *et al.*, 2017] with 8,000 warmup steps. The training batch consisted of approximately 2048 source tokens and 2048 target tokens. Label smoothing of value 0.1 [Szegedy *et al.*, 2016] was used for training. For inference, we used beam search with a width of 5 with a length penalty of 0.6. The evaluation metric is BLEU [Papineni *et al.*, 2002][2]. We did not apply checkpoint averaging [Vaswani *et al.*, 2017] on the parameters for evaluation.

5.1 Main Results

Document-level Translation We list the results of our experiments in Table 1, comparing four context-aware NMT models, i.e., Document-aware Transformer [Zhang *et al.*, 2018, DocT], Hierarchical Attention NMT [Miculicich *et al.*, 2018, HAN], Selective Attention NMT [Maruf *et al.*, 2019, SAN] and Query-guided Capsule Network [Yang *et al.*, 2019, QCN].

[1] The last two corpora are from Maruf *et al.* [2019]

[2] https://github.com/mjpost/sacreBLEU

Model	$\Delta\lvert\boldsymbol{\theta}\rvert$	v_{train}	v_{test}	Zh-En	En-De			
				TED	TED	News	Europarl	avg.
SentNmt [Vaswani *et al.*, 2017]	0.0m	1.0×	1.0×	17.0	23.10	22.40	29.40	24.96
DocT [Zhang *et al.*, 2018]	9.5m	0.65×	0.98×	n/a	24.00	23.08	29.32	25.46
HAN [Miculicich *et al.*, 2018]	4.8m	0.32×	0.89×	17.9	24.58	**25.03**	28.60	26.07
SAN [Maruf *et al.*, 2019]	4.2m	0.51×	0.86×	n/a	24.42	24.84	29.75	26.33
QCN [Yang *et al.*, 2019]	n/a	n/a	n/a	n/a	**25.19**	22.37	29.82	25.79
Final	4.7m	0.22×	1.08×	**19.1**	25.10	24.91	**30.40**	**26.80**

Table 1: Experiment results of our model in comparison with several baselines, including increments of the number of parameters over Transformer baseline ($\Delta|\boldsymbol{\theta}|$), training/testing speeds (v_{train}/v_{test}, some of them are derived from Maruf *et al.* [2019]), and translation results of the testsets in BLEU score.

Model	Test
SentNMT	17.0
DocNMT (documents as input/output)	14.2
HAN [Miculicich *et al.*, 2018]	15.6
Final	17.8

Table 2: Results of sentence-level translation on TED Zh-En.

As shown in Table 1, by leveraging document context, our proposed model obtains 2.1, 2.0, 2.5, and 1.0 gains over sentence-level Transformer baselines in terms of BLEU score on TED Zh-En, TED En-De, News and Europarl datasets, respectively. Among them, our model archives new state-of-the-art results on TED Zh-En and Europarl, showing the superiority of exploiting the whole document context. Though our model is not the best on TED En-De and News tasks, it is still comparable with QCN and HAN and achieves the best average performance on English-German benchmarks by at least 0.47 BLEU score over the best previous model. We suggest this could probably because we did not apply the two-stage training scheme used in Miculicich *et al.* [2018] or regularizations introduced in Yang *et al.* [2019]. In addition, while sacrificing training speed, the parameter increment and decoding speed could be manageable.

Sentence-level Translation We compare the performance on single sentence translation in Table 2, which demonstrates the good compatibility of our proposed model on both document and sentence translation, whereas the performance of other approaches greatly leg behind the sentence-level baseline. The reason is while our proposed model does not, the previous approaches require document context as a separate input stream. This difference ensures the feasibility in both document and sentence-level translation in this unified framework. Therefore, our proposed model can be directly used in general translation tasks with any input text of any number of sentences, which is more deployment-friendly.

5.2 Analysis and Discussion

Does Bilingual Context Really Matter? Yes. To investigate how important the bilingual context is and corresponding contributions of each component, we summary the ablation study in Table 3. First of all, using the entire document as input and output directly cannot even generate document translation with the same number of sentences as source document, which is much worse than sentence-level baseline and our model in terms of document-level BLEU. For source context modeling, only casting the whole source document as an input sequence (Doc2Sent) does not work. Meanwhile, reset word positions and introducing segment embedding for each sentence alleviate this problem, which verifies one of our motivations that we should focus more on local sentences. Moreover, the gains by the segment-level relative attention and gated context fusion mechanism demonstrate retrieving and integrating source global context are useful for document translation. As for target context, employing Transformer-XL decoder to exploit target historically global context also leads to better performance on document translation. This is somewhat contrasted to [Zhang *et al.*, 2018] claiming that using target context leading to error propagation. In the end, by jointly modeling both source and target contexts, our final model can obtain the best performance.

Model	BLEU (BLEU$_{\mathrm{doc}}$)
SentNmt [Vaswani *et al.*, 2017]	11.4 (21.0)
DocNmt (documents as input/output)	n/a (17.0)
Modeling source context	
Doc2Sent	6.8
+ reset word positions for each sentence	10.0
+ segment embedding	10.5
+ segment-level relative attention	12.2
+ context fusion gate	12.4
Modeling target context	
Transformer-XL decoder [Sent2Doc]	12.4
Final	12.9 (24.4)

Table 3: Ablation study on modeling context on TED Zh-En development set. "Doc" means using a entire document as a sequence for input or output. BLEU$_{\mathrm{doc}}$ indicates the document-level BLEU score calculated on the concatenation of all output sentences.

Effect of Quantity of Context: the More, the Better. We also experiment to show how the quantity of context affects our model in document translation. As shown in Figure 3, we find that providing only one adjacent sentence as context helps performance on document translation, but that the more context is given, the better the translation quality is, although there does seem to be an upper limit of 20 sentences. Successfully incorporating context of this size is something related work has not successfully achieved [Zhang *et al.*, 2018; Miculicich *et al.*, 2018; Yang *et al.*, 2019]. We attribute this advantage to our hierarchical model design which leads to

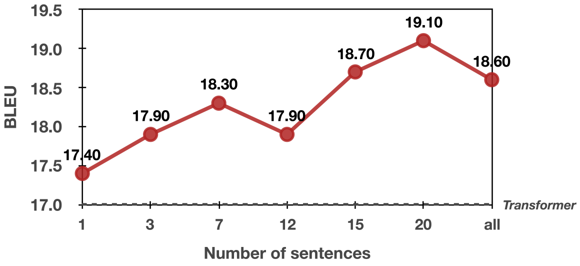

Figure 3: BLEU score w.r.t. #sent. of context on TED ZH-EN.

Model	Dev	Test
Transformer [Vaswani *et al.*, 2017]	11.4	17.0
BERT+MLM [Li *et al.*, 2019]	n/a	20.7
FINAL	12.9	19.1
FINAL + source TL	13.9	19.7
FINAL + source & target TL	14.9	21.3

Table 4: Effect of transfer learning (TL).

more gains than pains from the increasingly noisy global context guided by the well-formed, uncorrupted local context.

Effect of Transfer Learning: Data Hungry Remains a Problem for Document-level Translation. Due to the limitation of document-level parallel data, exploiting sentence-level parallel corpora or monolingual document-level corpora draws more attention. We investigate transfer learning (TL) approaches on TED ZH-EN. We pretrain our model on WMT18 ZH-EN sentence-level parallel corpus with 7m sentence pairs, where every single sentence is regarded as a document. We then continue to finetune the pretrained model on TED ZH-EN document-level parallel data (source & target TL). We also compare to a variant only whose encoder is initialized (source TL). As shown in Table 4, transfer learning approach can help alleviate the need for document level data in source and target languages to some extent. However, the scarcity of document-level parallel data still prevents the document-level NMT from extending at scale.

What Does Model Learns about Context? A Case Study. Furthermore, we are interested in what the proposed model learns about context. In Figure 4, we visualize the sentence-to-sentence attention weights of a source document based on segment-level relative attention. Formally, the weight of the k-th sentence attending to the κ-th sentence are computed by $\alpha_k^\kappa = \frac{1}{|x_k|}\sum_i \sum_j \alpha_{k,i}^{\kappa,j}$, where $\alpha_{k,i}^{\kappa,j}$ is defined by Eq.(1). As shown in Figure 4, we find very interesting patterns (which are also prevalent in other cases): 1) first two sentences (blue frame), which contain the main topic and idea of a document, seem to be a very useful context for all sentences; 2) the previous and subsequent adjacent sentences (red and purple diagonals, respectively) draw dense attention, which indicates the importance of surrounding context; 3) although sounding contexts are crucial, the subsequent sentence significantly outweighs the previous one. This may imply that the lack of target future information but the availability of the past information in the decoder forces the encoder to retrieve more knowledge about the next sentence than the previous one; 4) the model seems to not that care about the current sentence. Probably because the local context can flow through the context fusion gate, the segment-level relative attention just focuses on fetching useful global context; 5) the 6-th sentence also gets attraction by all the others (brown frame), which may play a special role in the inspected document.

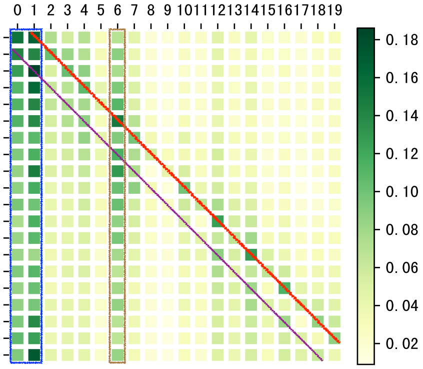

Figure 4: Visualization of sentence-to-sentence attention based on segment-level relative attention. Each row represents a sentence while each column represents another sentence to be attended. The weights of each row sum to 1.

Model	deixis	lex.c.	ell.infl.	ell.VP
SENTNMT	50.0	45.9	52.2	24.2
OURS	61.3	46.1	61.0	35.6
Voita *et al.* [2018]∗	81.6	58.1	72.2	80.0

Table 5: Accuracy (%) of discourse phenomena. ∗ different data and system conditions, only for reference.

Analysis on Discourse Phenomena. We also want to examine whether the proposed model actually learns to utilize document context to resolve discourse inconsistencies that context-agnostic models cannot handle. We use contrastive test sets for the evaluation of discourse phenomena for English-Russian by Voita *et al.* [2018]. There are four test sets in the suite regarding deixis, lexicon consistency, ellipsis (inflection), and ellipsis (verb phrase). Each testset contains groups of contrastive examples consisting of a positive translation with correct discourse phenomenon and negative translations with incorrect phenomena. The goal is to figure out if a model is more likely to generate a correct translation compared to the incorrect variation. We summarize the results in Table 5. Our model is better at resolving discourse consistencies compared to context-agnostic baseline. Voita *et al.* [2018] use a context-agnostic baseline, trained on $4\times$ larger data, to generate first-pass drafts, and perform post-processings, which is not directly comparable, but would be easily incorporated with our model to achieve better results.

6 Conclusion

In this paper, we propose a unified local and global NMT framework, which can successfully exploit context regardless of how many sentence(s) are in the input. Extensive experimentation and analysis show that our model has indeed learned to leverage a larger context. In future work we will investigate the feasibility of extending our approach to other document-level NLP tasks, e.g., summarization.

7 Acknowledgements

Shujian Huang is the corresponding author. This work was supported by National Science Foundation of China (No. U1836221, 61672277) and National Key R&D Program of China (No. 2019QY1806). Zaixiang Zheng was also supported by China Scholarship Council (No. 201906190162). Alexandra Birch was supported by the European Union's Horizon 2020 research and innovation programme under grant agreements No 825299 (GoURMET) and also by the UK EPSRC fellowship grant EP/S001271/1 (MTStretch).

References

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

[Bawden *et al.*, 2018] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In *NAACL-HLT*, 2018.

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

[Hassan *et al.*, 2018] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*, 2018.

[Jean *et al.*, 2017] Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135, 2017.

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

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

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

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

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

[Li *et al.*, 2019] Liangyou Li, Xin Jiang, Qun Liu, Huawei Noah', and Ark Lab. Pretrained Language Models for Document-Level Neural Machine Translation. *arXiv preprint*, 2019.

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

[Maruf *et al.*, 2019] Sameen Maruf, André FT Martins, and Gholamreza Haffari. Selective attention for context-aware neural machine translation. In *NAACL-HLT*, 2019.

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

[Müller *et al.*, 2018] Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *WMT*, 2018.

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

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

[Shaw *et al.*, 2018] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL-HLT*, 2018.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

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

[Tiedemann and Scherrer, 2017] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *DiscoMT*, 2017.

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NIPS*, 2017.

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

[Voita *et al.*, 2019] Elena Voita, Rico Sennrich, and Ivan Titov. Context-aware monolingual repair for neural machine translation. In *EMNLP-IJCNLP*, 2019.

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

[Xiong *et al.*, 2019] Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling coherence for discourse neural machine translation. In *AAAI*, 2019.

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

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