Context-Aware Non-Autoregressive Document-Level Translation with Sentence-Aligned Connectionist Temporal Classification

Hao Yu[1], **Kaiyu Huang**[2*], **Anqi Zhao**[1], **Junpeng Liu**[1], **Degen Huang**[1]
[1]School of Computer Science and Technology, Dalian University of Technology, China
[2]Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
yuhao_dlut@mail.dlut.edu.cn, kyhuang@bjtu.edu.cn

Abstract

Previous studies employ the autoregressive translation (AT) paradigm in the document-to-document neural machine translation. These methods extend the translation unit from a single sentence to a pseudo-document and encode the full pseudo-document, avoiding the redundant computation problem in context. However, the AT methods cannot parallelize decoding and struggle with error accumulation, especially when the length of sentences increases. In this work, we propose a context-aware non-autoregressive framework with the sentence-aligned connectionist temporal classification (SA-CTC) loss for document-level neural machine translation. In particular, the SA-CTC loss reduces the search space of the decoding path by fixing the positions of the beginning and end tokens for each sentence in the document. Meanwhile, the context-aware architecture introduces preset nodes to represent sentence-level information and utilizes a hierarchical attention structure to regulate the attention hypothesis space. Experimental results show that our proposed method can achieve competitive performance compared with several strong baselines. Our method implements non-autoregressive modeling in Doc-to-Doc translation manner, achieving an average 46X decoding speedup compared to the document-level AT baselines on three benchmarks.

Keywords: Document-Level Machine Translation, Non-Autoregressive, Connectionist Temporal Classification

1. Introduction

Document-level neural machine translation (NMT) has attracted increasing attention in the past years (Wang et al., 2017; Voita et al., 2019b; Zhang et al., 2020; Ma et al., 2020; Bao et al., 2021; Zhang et al., 2022, 2023). Due to the success of sentence-level NMT (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Xia et al., 2017; Vaswani et al., 2017), a branch of studies encodes every sentence and its context separately in a sentence-to-sentence manner (Tiedemann and Scherrer, 2017; Wang et al., 2017; Zhang et al., 2018; Maruf and Haffari, 2018; Kuang et al., 2018; Miculicich et al., 2018; Maruf et al., 2019; Zheng et al., 2020; Ma et al., 2020). To reduce the redundant calculation of the context, another branch of studies extends the translation unit from single-sentence to multi-sentences, encoding both the context and the current sentence in a document-to-document (Doc-to-Doc) manner (Zhang et al., 2020; Junczys-Dowmunt, 2019; Liu et al., 2020; Bao et al., 2021).

These works adopt the autoregressive translation (AT) paradigm for document-level NMT and decode token-by-token. However, the AT methods will face two major challenges in the Doc-to-Doc scenario: (1) decoding speed slowly and (2) the accumulation of errors. The former is caused by token-by-token decoding of AT methods, while the latter is caused by exposure bias.

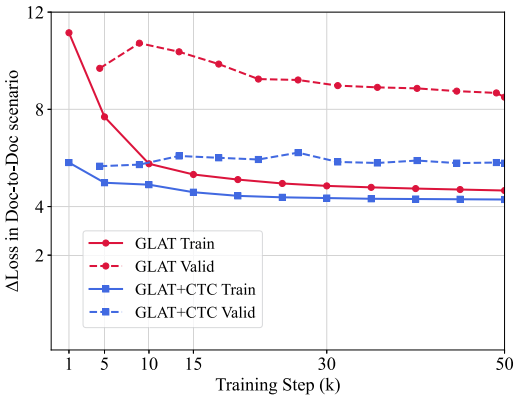

Figure 1: Failure loss curve of the NAT models in Doc-to-Doc scenario.

Fortunately, the Transformer-based non-autoregressive translation (NAT) models (Saharia et al., 2020; Qian et al., 2021; Bao et al., 2022) can alleviate the issues of error accumulation and slow decoding speed. Unlike AT models, NAT models impose conditional independence assumptions to support the parallel decoding of sentences during inference. And some studies (Gu et al., 2018; Qian et al., 2021; Bao et al., 2022) use additional components to explicitly predict

*Corresponding author

LREC-COLING 2024, pages 3812–3824

the length of the target sequence, while others (Libovický and Helcl, 2018; Saharia et al., 2020) implicitly predict the length of the target sequence by upsampling. However, in the Doc-to-Doc translation scenarios, both explicit and implicit NAT methods for predicting the length of the target sequence cannot effectively predict the length of every sentence in the document, resulting in sentence misalignment.

As shown in Figure 1, existing NAT models will fail to train, and the loss falls into a local minimum value that can not continue to decrease. In the implicit prediction NAT methods, the sentence misalignment results in excessively large **search space of decoding path** for the model, which hinders the practical training of the model. Moreover, existing mainstream NAT models adopt the transformer-based architecture, and its attention sublayers struggle with the excessively large **attention hypothesis space** when processing long target sequences (Bao et al., 2021).

To address these problems, we propose CASA, a **c**ontext-**a**ware framework in a non-autoregressive paradigm with the **s**entence-**a**ligned connectionist temporal classification (SA-CTC) loss for document-level neural machine translation. The SA-CTC loss reduces the excessively large search space of the decoding path based on the intuitive sentence-aligned assumption that the sentence number of the source document is equal to the target document and the positions correspond. By explicitly fixing the position of the beginning and end tokens for each sentence in the document, we effectively remove the wrong path in the decoding path's search space, which does not conform to the sentence-aligned assumption. Meanwhile, the CA architecture regulates the excessively large attention hypothesis space based on a hierarchical attention structure that uses preset nodes to represent sentence-level information. To sum up, our contributions are as follows:

- We investigate the document-level NAT model in a Doc-to-Doc manner and propose a novel framework that consists of the SA-CTC loss and the CA architecture.
- Our proposed method can effectively alleviate the issue of NAT models that fail to train in the Doc-to-Doc scenarios and achieve an average decoding speed of 46 times compared with the document-level AT method on three benchmarks.
- The experimental results demonstrate that our method performs competitively compared to several strong baselines in both AT and NAT manners.

2. Background

2.1. Doc-to-Doc Autoregressive Translation

For the sequence-to-sequence task of document-level machine translation, given a source language document $X = \{X_1, X_2, .., X_n\}$ consisting of a series of sentences and predict a target language document $Y = \{(Y_1, Y_2, ..., Y_n\}$, where the X_i and Y_i represents the i-th sentence of document X and Y. And $Y_i = \{y_{[i,1]}, y_{[i,2]}, ..., y_{[i,m]}\}$, which $y_{[i,j]}$ represents the j-th token of the i-th sentence Y_i.

In Doc-to-Doc translation scenarios, traditional autoregressive factorization factorizes $\mathrm{P_{AT}}(Y|X)$ with a series of conditional probabilities:

$$\begin{aligned}\mathrm{P_{AT}}(Y|X) &= \prod_{i=1}^{n} \mathrm{P}(Y_i|Y_{<i}, X_{[1:n]}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{m} \mathrm{p}(y_{[i,j]}|y_{[i,<j]}, Y_{[1:i-1]}, X_{[1:n]})\end{aligned} \tag{1}$$

where $y_{i,<j} = (y_{[i,1]}, y_{[i,2]}, ..., y_{[i,j-1]})$.

2.2. Sentence-level Non-Autoregressive Translation

The autoregressive translation is predicted based on prefix words for inference, which suffers from error accumulation and slow decoding. To tackle the above problems, Gu et al. (2018) first proposes a sentence-level non-autoregressive translation, introducing a non-autoregressive factorization as:

$$\begin{aligned}\mathrm{P_{NAT}}(Y|X) &= \sum_{i=1}^{n} \mathrm{P}(Y_i|X_i) \\ &= \sum_{i=1}^{n}\prod_{j=1}^{m} \mathrm{p}(y_{[i,j]}|X_i)\end{aligned} \tag{2}$$

where each word $y_{[i,j]}$ are modeled independently.

During inference, the NAT model can simultaneously decode all tokens of a target language by:

$$y_{[i,j]} = \mathrm{argmax}\ \mathrm{p}(y_{[i,j]}|X_i) \tag{3}$$

which significantly improves decoding efficiency.

3. Methodology

3.1. Doc-to-Doc Non-Autoregressive Translation

Non-autoregressive machine translation methods are widely studied in sentence-level scenarios, ignoring inter-sentence relations and global context information. Therefore, we propose a non-autoregressive framework in a Doc-to-Doc translation manner, where the model's input is the entire

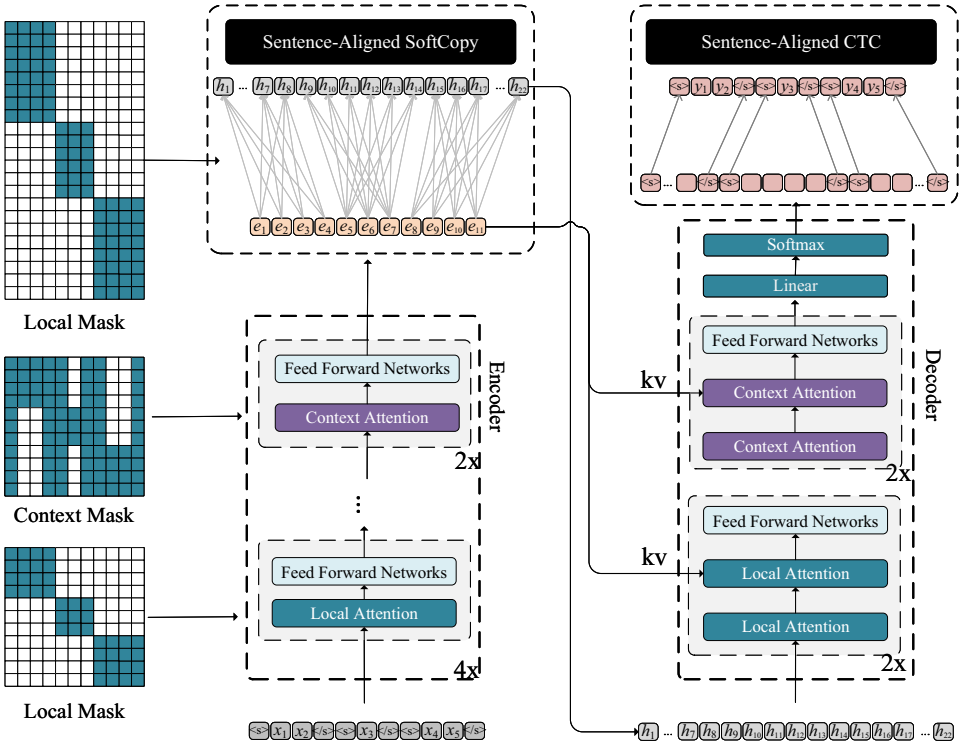

Figure 2: Illustration of our proposed framework (CASA). It consists of the SA-CTC loss and CA architecture.

source document. The target translation probability can be expressed as:

$$\begin{aligned} \mathrm{P}_{\mathrm{NAT}}(Y|X) &= \prod_{i=1}^{n} \mathrm{P}(Y_i|X_{[1:n]}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{m} \mathrm{P}(Y_{[i,j]}|X_{[1:n]}) \end{aligned} \tag{4}$$

During inference, the NAT model can simultaneously decode all tokens of a target document by Equation 3.

3.2. CASA Framework

The non-autoregressive framework, CASA is proposed with a sentence-aligned connectionist temporal classification loss and a context-aware architecture for the Doc-to-Doc scenarios. The framework is illustrated in Figure 2.

3.2.1. Sentence-Aligned Connectionist Temporal Classification (SA-CTC)

Based on the conditional independence assumption, the connectionist temporal classification (CTC) (Graves et al., 2006) implements frame-level alignment between input X and output Y, predicting conditional probabilities $\mathrm{P}_{\mathrm{CTC}}(Y|X)$ for all tokens based on all possible paths:

$$\begin{aligned} \mathcal{L}_{CTC} =& \mathrm{P}_{\mathrm{CTC}}(\mathrm{Y}|\mathrm{X}) \\ =& \sum_{A\in\beta^{-1}(Y)} \mathrm{P}(A|X) \\ =& \sum_{A\in\beta^{-1}(Y)} \prod_{t=1}^{T} \mathrm{P}(\alpha_t|X) \end{aligned} \tag{5}$$

$$A = \{\alpha_t \in \nu \cup \varepsilon | t = 1, 2, \cdots, T\} \tag{6}$$

where ν is the vocabulary, ε is the $< blank >$ token, the A is a path consisting of multiple vocabulary tokens and the $< blank >$ token, and the $\beta^{-1}(Y)$ is the set of all possible paths A.

Existing CTC-based methods are unsuitable for documents in Doc-to-Doc translation scenarios. Since the prediction length is too large, the decoding path's search space of CTC-based methods increases exponentially. We propose SA-CTC loss that automatically interferes with predicted token probabilities, leveraging sentence-level alignment information in documents to reduce the decoding path's search space. The SA-CTC loss prespecifies the positions of the beginning and end tokens in the decoding sequence according to the length of each sentence in the source document. Predicting conditional probabilities $\mathrm{P}_{\mathrm{SA-CTC}}(Y|X)$ for all tokens based on all sentence-aligned possi-

ble paths SA, the target probability can be formulated as below:

$$\begin{aligned}\mathcal{L}_{SA-CTC} &= \mathrm{P}_{\mathrm{SA-CTC}}(Y|X) \\ &= \sum_{SA\in\hat{\beta}^{-1}(Y)} \mathrm{P}(SA|X) \\ &= \sum_{SA\in\hat{\beta}^{-1}(Y)} \prod_{t=1}^{T} \mathrm{P}(\alpha_t|X)\end{aligned} \tag{7}$$

where $\hat{\beta}^{-1}(Y)$ is a sentence-aligned path subset of all possible alignment paths $\beta^{-1}(Y)$.

$$\hat{\beta}^{-1}(Y) \subseteq \beta^{-1}(Y) \tag{8}$$

Given source language documents X and corresponding translations Y, the document sequence can be obtained by concated the sentences in the document, expressed as follows:

$$\begin{aligned}X &= B + X_1 + E + ... + B + X_n + E \\ Y &= B + Y_1 + E + ... + B + Y_n + E\end{aligned} \tag{9}$$

The B/E is $< BOS >/< EOS >$ token, which means the beginning and end of the sentence. The length of the target sequence SA is empirically set to twice the length of the source sequence, $|SA| = 2|X|$. The conditional probability $P \in \mathbb{R}^{2|X|\times Vocabsize}$ of SA is:

$$P = \mathrm{Decoder}(\mathrm{SoftCopy}(\mathrm{Encoder}(X))) \tag{10}$$

We assume that the length of each sentence in the SA sequence is also twice that of each sentence in the source sequence. Therefore, the position of each sentence's B/E token in the SA sequence is defined as follows:

$$\begin{aligned}IndexB = \{I_i = 1 \; if \; X[(i-1)/2] = B \\ else \; 0\}|_{i=1}^{2|X|} \\ IndexE = \{I_i = 1 \; if \; X[i/2] = E \\ else \; 0\}|_{i=1}^{2|X|}\end{aligned} \tag{11}$$

"1" in $IndexB/IndexE$ indicates that the current position in the SA sequence of the B/E token. Set the probability of token B/E at the position of $IndexB/IndexE$ with conditional probability P to be positive infinity, thereby fixing the position of token B/E in the target sequence SA. Meanwhile, the tokens of each sentence will be fixed between the B/E tokens of the corresponding sentences.

3.2.2. Context-Aware (CA) Architecture

Our model adopts the transformer architecture (Vaswani et al., 2017) with an encoder, decoder, and an additional predictor in Latent-GLAT (Bao et al., 2022). The encoder is composed of a stack of $N = 6$ identical blocks. Each block has two sub-layers, the multi-head self-attention and the position-wise fully connected feed-forward network (FFN). The decoder and predictor comprise a stack of $N = 4$ identical blocks.

Local Attention Following the previous work (Bao et al., 2021), we introduce the group tag to construct the local mask, formally expressed as:

$$\begin{aligned}G_Q &= \{g_p = t \; if \; Q_p \in sent_t^Q\}|_{p=1}^{|Q|} \\ G_K &= \{g_p = t \; if \; K_p \in sent_t^K\}|_{p=1}^{|K|} \\ LocalMask_{ij} &\propto 1 \;\; if \;\; (G_Q[i] = G_K[j]) \\ &\quad else \;\; 0|_{i,j=1,1}^{|Q|,|K|}\end{aligned} \tag{12}$$

The G_Q and G_K are the group tags of a set of corresponding query vectors and key vectors, representing the position of the corresponding sentence in the document. And the $LocalMask \in \mathbb{R}^{|Q|\times|K|}$ represents the masking matrix of attention. $\mathrm{I}(G_Q)$ are constant vectors of 1 with the same dimension as G_Q.

$$\begin{aligned}&\mathrm{LocalAttention}(Q, K, V) \\ &= \mathrm{Softmax}(\frac{QK^T}{\sqrt{d_K}} + LocalMask \cdot \gamma)V\end{aligned} \tag{13}$$

The constant value γ can typically be $-1e8$ (negative infinity).

Context Attention We set the token at the beginning and end of each sentence in the document as the sentence-level node and the remaining nodes as token-level nodes. The token-level node can observe the sentence-level node to capture the context information under the premise of a small attention hypothesis space.

$$\begin{aligned}C_Q &= \{c_p = 1 \;\; if \; Q_p \in \{B, E\} \; else \; 2\}|_{p=1}^{|Q|} \\ C_K &= \{c_p = 1 \;\; if \; K_p \in \{B, E\} \; else \; 2\}|_{p=1}^{|K|} \\ ContextMask_{ij} &\propto 1 \\ &\quad if \; (G_Q[i] = G_K[j] \; or \; C_Q[i] = 1) \\ &\quad else \; 0|_{i\in\{1:|Q|\} \; j\in\{1:|K|\}}\end{aligned} \tag{14}$$

The C_Q and C_K represent the category tags of the tokens corresponding to the query vector Q and the key vector K in the document, where "1" represents a sentence-level node, and "2" represents a token-level node. The context attention is formally expressed as:

$$\begin{aligned}&\mathrm{ContextAttention}(Q, K, V) \\ &= \mathrm{Softmax}(\frac{QK^T}{\sqrt{d_K}} + ContextMask \cdot \gamma)V\end{aligned} \tag{15}$$

Dataset		Sentences	Documents	Instances
TED	Train	0.21M	1.70K	11.19K
	Valid	9.06K	93	483
	Test	2.29K	23	123
News	Train	0.24M	6.07K	18.46K
	Valid	2.25K	81	172
	Test	3.15K	155	263
Europarl	Train	1.78M	0.12M	0.16M
	Valid	3.83K	240	346
	Test	5.49K	360	498

Table 1: Statistical results of document En-De datasets.

Sentence-Aligned SoftCopy Following the most common practices in NAT models (Wei et al., 2019; Li et al., 2019; Bao et al., 2022), we apply the softcopy method to initialize the decoder and predictor input $H = \{h_1, h_2, ..., h_t\}$ with the encoder output $E = \{e_1, e_2, \cdots, e_s\}$. Based on the sentence alignment assumption, we propose the sentence-aligned softcopy method. The formulaic representation is as follows:

$$A = \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,s} \\ \vdots & \ddots & \vdots \\ \alpha_{t,1} & \cdots & \alpha_{t,s} \end{pmatrix} \quad (16)$$
$$\alpha_{ij} \propto \exp[-(i - j \cdot \frac{s}{t})^2]$$

$$H = \mathrm{Softmax}(A + LocalMask \cdot \gamma)E \quad (17)$$

Hierarchical Attention Structure Based on hierarchical modeling, we apply local attention in the bottom sublayer to aggregate sentence-level information and context attention in the top sublayer of the model to capture document-level dependency information, thereby reducing the excessively large hypothesis space of the standard attention mechanism in Doc-to-Doc translation scenarios.

4. Experiments

4.1. Datasets and Settings

Datasets We evaluate the CASA against widely adopted benchmark datasets (Maruf et al., 2019), including three English-German (En-De) translation domains: TED, News, and Europarl. En-De benchmark data statistics are shown in Table 1. We also conducted experiments in three translation directions of IWSLT17. We preprocess the document and split it into instances of approximately 512 tokens. The sentences were tokenized and truecased using the MOSES (Koehn et al., 2007) tool. Applying BPE (Sennrich et al., 2016) with 30,000 merging operations to encode words to subwords.

Knowledge Distillation According to previous works (Qian et al., 2021; Saharia et al., 2020), sequential distillation is essential for non-autoregressive model training. Our experiments use both the sentence-level and document-level autoregressive teacher model (SENTNMT (Vaswani et al., 2017)/G-Trans (Bao et al., 2021)) (randinit/finetune) to distill the training dataset from the raw corpus and obtain the sentence-level KD(sent-KD)/document-level KD(doc-KD) (randinit/finetune) corpus.

Parameters Settings Our model is implemented based on Fairseq (Ott et al., 2019). Following the settings of the previous work (Bao et al., 2022): 6 layers for the encoder and 4 for the decoder, 8 attention heads per layer, 512 model dimensions, and 2048 hidden dimensions. We follow the weight initialization schema from BERT (Devlin et al., 2019). For the regularization, we set dropout to 0.3/0.2/0.1 on the TED/News/Europarl7 data set, respectively. We train batches of 16k tokens for our model using Adam (Kingma and Ba, 2015) with $\beta = (0.9, 0.999)$ and $\varepsilon = 10^{-6}$. The learning rate warms up to $5e-4$ within 4k steps and then decays with the inverse square-root schedule. We train all models for 50k steps, measure the validation loss at the end of each epoch, and select the last checkpoints to create the final model. And all models are trained on 2 Nvidia A6000 GPUs.

Evaluation We report the tokenized BLEU (Papineni et al., 2002) of models, as reported in previous NAT work (Qian et al., 2021; Bao et al., 2022) by the ScareBLEU evaluation script (Post, 2018). Following the previous document machine translation works (Liu et al., 2020; Bao et al., 2021), we report the sentence-level BLEU (s-BLEU) and document-level BLEU (d-BLEU), respectively.

Baselines We compare our proposed CASA with various representative methods in both AT and NAT manner for document-level NMT. The baselines can be listed as follows:

SENTNMT (Vaswani et al., 2017): We reproduce the Transformer model on the sentence-level machine translation scenario using random initialization settings.

G-Trans (randinit/finetune) (Bao et al., 2021): We reproduce the document-level G-Transformer model on Doc-to-Doc scenarios using random initialization settings or finetune on sentence-level Transformer (SENTNMT).

GLAT/GLAT+CTC (Qian et al., 2021): We reproduce the GLAT and GLAT+CTC models on Doc-to-Doc scenarios, which introduce a two-step glancing training strategy and sampling partial ground-truth tokens for training NAT.

Method	Data	TED		News		Europarl	
		s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU
autoregressive translation							
SENTNMT (2017)	raw	23.10	-	22.40	-	29.40	-
HAN (2018)	raw	24.58	-	25.03	-	28.60	-
SAN (2019)	raw	24.42	-	24.84	-	29.75	-
Hybrid Context (2020)	raw	25.10	-	24.91	-	30.40	-
Flat-Transformer (2020)	raw	24.87	-	23.55	-	30.09	-
G-Trans (randinit) (2021)	raw	23.53	25.84	23.55	25.23	32.18	33.87
G-Trans (finetune) (2021)	raw	25.12	27.17	25.52	27.11	32.39	34.08
Disco2NMT (2022)	raw	24.60	-	23.25	-	29.36	-
SENTNMT (2017) †	raw	25.00	27.32	25.26	26.78	31.50	33.19
G-Trans (randinit) (2021) †	raw	23.84	26.14	23.44	25.00	31.95	33.65
G-Trans (finetune) (2021) †	raw	24.98	27.17	25.50	27.09	32.54	34.22
non-autoregressive translation							
GLAT (2021)†	sent-KD	-	0.00	-	0.00	-	0.94
GLAT+CTC (2021)†	sent-KD	-	8.05	-	0.00	-	0.00
GLAT-Latent (2022)†	sent-KD	-	0.75	-	0.93	-	16.77
CASA	sent-KD	**24.24**	**26.45**	23.25	24.72	29.50	31.07
CASA-Latent	sent-KD	24.04	26.28	**23.78**	**25.92**	**29.75**	**31.33**
CASA	doc-KD(finetune)	**24.16**	**26.24**	**23.47**	**25.00**	29.49	31.12
CASA-Latent	doc-KD(finetune)	23.88	26.00	23.09	24.68	**29.85**	**31.44**
CASA	raw	22.44	24.61	**19.16**	**20.55**	25.47	27.06
CASA-Latent	raw	**22.50**	**24.78**	18.55	19.94	**26.31**	**27.85**

Table 2: Results on three document benchmark datasets. The "Latent" means introducing discrete latent variables in the CASA method like GLAT-Latent. The better score of each setting is highlighted in **bold**, and the best score of all NAT models is underlined. The "0.00" represents the training failure. † means that we reproduce the model and report the tokenized s-BLEU and d-BLEU.

	One Instance				Fully GPU Memory			
	TED	News	Europarl	Avg.	TED	News	Europarl	Avg.
autoregressive translation on raw data								
SENTNMT (2017) †	1.37x	1.36x	1.34x	1.36x	8.03x	8.40x	7.16x	7.86x
G-Trans(randinit) (2021) †	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
Shadow(8+4)	1.27x	1.24x	1.23x	1.25x	1.09x	1.12x	1.14x	1.12x
Shadow(10+2)	1.91x	1.91x	1.87x	1.90x	1.31x	1.39x	1.33x	1.34x
2to2	0.97x	0.90x	0.93x	0.93x	3.15x	3.19x	2.78x	3.04x
non-autoregressive translation on sentence-level KD data								
CASA-Latent	30.27x	29.90x	29.74x	29.97x	14.19x	20.85x	15.01x	16.68x
CASA	**46.67x**	**44.21x**	**47.15x**	**46.01x**	**25.14x**	**32.33x**	**23.00x**	**26.82x**

Table 3: Model accelerated evaluation with one instance setting. We use the decoding speed of the document-level AT model G-Trans (randinit) on a single GPU and a single instance as the benchmark to evaluate our models. The better performance of all models is highlighted in **bold**.

GLAT-Latent (Bao et al., 2022): In document-to-document scenarios, we reproduce the GLAT-Latent model, which has an additional discrete latent variable predictor and a gating component.

4.2. Main Results

Results on Benchmarks As shown in Table 2, we investigate the translation quality of AT and NAT methods in Doc-to-Doc translation scenarios. Results show that existing NAT methods suffer from training failures in Doc-to-Doc scenarios, exhibiting near-zero d-BLEU values on three benchmark datasets. The results also show that our CASA is successfully trained on sentence-level/document-level KD and RAW datasets. Meanwhile, the model trained on the sentence-level/document-level KD dataset achieved competitive performance compared with the document-level AT method (G-Trans(randinit)) on TED and News. On the

Method	Data	Zh-En		Ar-En		Fr-En	
		s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU
autoregressive translation							
SENTNMT (2017) †	raw	24.08	28.47	32.52	36.19	38.92	42.00
G-Trans (randinit) (2021) †	raw	21.88	26.02	31.58	35.32	36.32	38.48
non-autoregressive translation							
GLAT (2021)†	sent-KD	-	0.00	-	0.00	-	0.00
GLAT+CTC (2021)†	sent-KD	-	0.57	-	0.00	-	12.38
GLAT-Latent (2022)†	sent-KD	-	1.74	-	1.36	-	5.82
CASA	raw	14.51	16.45	27.83	30.41	33.97	35.99
	sent-KD	**19.31**	**22.06**	29.62	32.35	36.28	38.42
	doc-KD(finetune)	18.67	21.14	**29.75**	**32.22**	**36.50**	**38.55**

Table 4: Results in IWSLT17 Datasets. The better score of CASA is highlighted in **bold**.

Method	Data	Deixis	E_vp	E_infl	L_coh
autoregressive translation					
SENTNMT (2017) †	raw	50.00	26.20	51.60	45.87
CADec (2019b)	raw	81.60	80.00	72.20	58.10
DocRepair (2019a)	raw	91.80	75.20	86.40	80.60
LSTM-Trans (2020)	raw	90.50	81.00	80.60	73.90
D-LM(PMI) (2021)	raw	96.80	90.60	75.80	97.80
G-Trans (randinit) (2021) †	raw	85.36	76.00	76.00	58.00
G-Trans (finetune) (2021) †	raw	74.48	25.20	50.80	45.87
non-autoregressive translation					
CASA	raw	50.00	33.80	55.20	45.87
CASA-Latent	raw	50.00	38.40	55.00	45.87
CASA	sent-KD	50.00	19.40	50.40	45.87
CASA-Latent	sent-KD	50.00	21.00	51.00	45.87
CASA	doc-KD(randinit)	50.00	**51.80**	59.40	45.87
CASA-Latent	doc-KD(randinit)	50.00	49.60	**60.00**	45.87
CASA	doc-KD(finetune)	**50.60**	36.20	47.80	**46.13**
CASA-Latent	doc-KD(finetune)	50.48	32.80	47.60	45.87

Table 5: Results on discourse phenomena. We only use the 1.5M document pairs from the OpenSubtitles2018 (Lison et al., 2018) dataset training model and testing in the human-labeled evaluation set (Voita et al., 2019b). The SENTNMT and G-Trans (randinit/finetune) are trained on the raw dataset, and our "CASA/CASA-Latent" are trained on both the raw and KD datasets.

Method			TED		News		Europarl		Avg.	
			s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU
	SA	CA	*non-autoregressive translation on sentence-level KD data*							
CASA-Latent	✓	✓	**24.04**	**26.28**	**23.78**	**25.92**	29.75	31.33	**25.85**	**27.84**
	✓	×	0.00	0.00	0.00	0.00	**29.90**	**31.49**	9.97	10.50
	×	✓	23.09	23.61	0.00	0.00	29.81	31.38	17.63	18.33
CASA	✓	✓	**24.24**	**26.45**	**23.25**	**24.72**	29.50	31.07	**25.66**	**27.41**
	✓	×	23.86	26.06	22.76	24.24	**29.83**	**31.44**	25.48	27.24
	×	✓	21.68	21.26	23.14	24.62	29.33	30.92	24.71	25.60

Table 6: Results on different strategies and modules for CASA and CASA-Latent. The better score of each base model is highlighted in **bold**.

random initialization settings, our model CASA-Latent achieves 0.2/0.34 percentage points higher s-BLEU values than G-Trans (randinit) on TED and News datasets. On the benchmark dataset Eu-

Method	TED		News	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU
CASA	**24.24**	**26.45**	**23.25**	**24.72**
w/o context attn	23.98	26.09	22.97	24.39
w/o local attn	23.89	25.96	23.07	24.59
w/o SA softcopy	24.00	26.09	23.04	24.54

Table 7: Results on different strategies for CA modules. The best score is highlighted in **bold**.

Dropout	TED		News	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU
0.1	23.52	25.80	22.30	23.83
0.2	23.85	26.13	**23.78**	**25.92**
0.3	**24.04**	**26.28**	22.67	24.19
0.4	24.04	26.15	21.97	23.43
0.5	23.78	25.92	21.40	22.84

Table 8: Results on different dropouts. The best score is highlighted in **bold**.

roparl, our method underperforms the AT method.

Model Acceleration Evaluation As is shown in Table 3, we report the decoding speeds of our models and the baseline systems on three benchmark datasets, respectively.

Shadow(8+4/10+2): We reproduce the deep encoder + shadow decoder model on Doc-to-Doc scenarios based on G-Trans(randinit). 8+4 means using an 8-layer encoder + 4-layer decoder.

2to2: We reproduce the context-aware model based on G-Trans(randinit), using a setting of 2 source sentences and 2 target sentences.

Our method CASA/CASA-Latent respectively achieves 46.01/29.97 times the average decoding speed of the document-level AT method in one instance setting and 16.68/26.82 times the average decoding speed of the document-level AT method in full gpu memory setting. The decoding speed of our model is significantly better than the existing document-level AT systems. Meanwhile, the speed of the CASA-Latent model is inferior to CASA, which is caused by the two-step decoding process of GLAT-Latent.

Results in Other Language Directions To verify that the proposed method is also effective in other language directions, we conducted experiments on the IWSLT17 document-level translation dataset in Zh/Ar/Fr-En translation directions, and the experimental results are shown in Table 4. Our model performs better in the document-level KD dataset for Ar-En/Fr-En language directions, with s-BLEU values of 29.75/36.50, respectively, achieving comparable performance to the AT baseline system. Therefore, it can be concluded that our method can be effectively applied to other language directions.

Method	TED		News	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU
CASA-Latent	24.04	26.28	23.78	25.92
w/o source side	23.98	26.24	22.49	24.02
w/o target side	24.06	26.27	22.44	23.99
CASA	24.24	26.45	23.25	24.72
w/o source side	24.20	26.53	23.48	25.01
w/o target side	24.13	26.37	23.01	24.52

Table 9: Impact of Source-side and Target-side Context.

Results on Discourse Phenomena To evaluate the abilities of the AT and the NAT methods on discourse phenomena, we conducted experiments on the English-Russian discourse evaluation dataset (Voita et al., 2019b). We use the sentence-level AT method "SENTNMT" and document-level AT method "G-Trans" (randinit/finetune) as the teacher model to obtain the corresponding KD datasets. And our document-level NAT models are trained on these KD datasets with random initialization. We tested four discourse evaluation metrics Deixis, $\mathrm{E\_vp}$, $\mathrm{E\_infl}$, $\mathrm{L\_coh}$ on this evaluation dataset.

As shown in Table 5, the document-level G-Trans (randinit) model has the best ability to model discourse phenomena. And the G-Trans (finetune) model dropped significantly on four discourse metrics. Meanwhile, the NAT model trained on the document-level KD (finetune) dataset also dropped significantly in two discourse metrics, $\mathrm{E\_vp}$, $\mathrm{E\_infl}$. The result shows that fine-tuning the sentence-level SENTNMT model will decline the abilities of discourse phenomena, lower the quality of the KD dataset, and eventually deteriorate the abilities of NAT for the discourse phenomena. Utilizing the document-level teacher model G-Trans (randinit) for sequence distillation is better.

4.3. Ablation Studies

Effect of Different Components As shown in Table 6, we performed ablation studies with different components on two base models, "CASA-Latent". For the "CASA", eliminating the SA-CTC method leads to an average decrease of 0.18 percentage points in s-BELU and 0.17 percentage points in d-BELU. Meanwhile, ablation of the CA architecture leads to an average decrease of 0.95 percentage points in s-BLEU and 1.81 percentage points in d-BLEU. For the "CASA-Latent", ablating the SA-CTC loss and CA architecture fails model training on the News dataset. It is due to the Latent-GLAT needs to train an additional predictor, which is more susceptible to the problem of excessively

large search space of decoding path and attention hypothesis space. In addition, on the large-scale dataset Europarl, only using the SA-CTC method achieved 29.90/29.83 s-BLEU values, because the larger dataset scale helps the model overcome the excessively large attention hypothesis space. In addition, we ablated the components of the CA module in the basic model CASA. The experimental results are shown in the Table 7. The ablation of the local attn component caused the s-BLEU value to decrease by 0.35 percentage points on the TED data set. The ablation of the context attn component caused the s-BLEU value to decrease by 0.28 percentage points on the News data set. The ablation of the component SA softcopy caused the s-BLEU value to decrease by 0.24/0.21 percentage points on the TED/News data set.

Results on Different Dropouts As shown in Table 8, we investigate the effectiveness of different dropout settings on the "CASA-Latent" in the TED and News datasets with sentence-level kd. The experimental results show that the datasets TED and News achieve the best s-BLEU value and d-BLEU when the dropout equals to 0.3 and 0.2, respectively. When the corpus size is larger, smaller dropout values are more beneficial. In contrast, we can increase the dropout value for a small-scale dataset.

Effect of Source-side and Target-side Contexts As shown in Table 9, we ablate the source context and the target separately and study the impact of different contexts on the "CASA-Latent". From the experimental results, removing the target context leads to a decrease of 0.33 percentage points s-BLEU value in the TED dataset and 0.49 percentage points s-BLEU value in the News dataset. Removing the source context decreases the News dataset's 1.1 percentage point s-BELU value. Therefore, both source and target-side contexts positively influence our framework.

5. Related Work

Existing document-level MT works focus on expanding the translation unit, conforming to the Doc-to-Doc paradigm. Kalchbrenner and Blunsom (2013) expands the translation unit from phrases to sentences, which is purely based on a continuous representation of words, phrases, and sentences. Zhang et al. (2020) expands the translation unit from single-sentence to multiple-sentences, changing the sentence-to-sentence translation paradigm of previous context-aware work, which encodes every sentence and its context separately. Bao et al. (2021) further expands the translation unit from multiple-sentences to pseudo-document, which shows stable document-level BLEU scores for inputs containing 512 and 1024 tokens.

The above works adopt the autoregressive paradigm, which has the problem of error accumulation caused by exposure bias and the problem of slow speed caused by token-by-token decoding. Since Gu et al. (2018) introduces a NAT model based on the Transformer network, prior works (Lee et al., 2018; Ghazvininejad et al., 2019, 2020; Qian et al., 2021; Bao et al., 2022) introduce various training strategies to reduce the model burden of dealing with dependencies among output words. Ghazvininejad et al. (2019) adopts an iterative training strategy and uses multiple masking and prediction methods to reduce training difficulty but requires multi-step decoding, resulting in a reduced speedup. Qian et al. (2021) proposes a two-step training strategy, introducing the ground-truth token to help model training and realize single-step decoding. Bao et al. (2022) employs the discrete latent variables to capture word categorical information, alleviating the multi-modality problem. Another branch of non-autoregressive (Libovický and Helcl, 2018; Saharia et al., 2020) implicitly predicts the length of the target sequence by introducing various training losses. Libovický and Helcl (2018) proposes a CTC-based model to predict the implicit alignment of source and target sequences, enabling variable-length predictions.

6. Conclusion

In this work, we investigate the failure of NAT model training in Doc-to-Doc translation scenarios. To address this, we propose CASA, a **c**ontext-**a**ware framework in a non-autoregressive paradigm with the **s**entence-**a**ligned connectionist temporal classification (SA-CTC) loss for the excessively large search space of decoding path and attention hypothesis space in document-level neural machine translation. The SA-CTC loss eases the search space of the decoding path by fixing the position of the beginning and ending tokens of each sentence in a document. Meanwhile, the context-aware architecture represents sentence-level information through preset sentence-level nodes and uses a hierarchical attention structure to regulate the excessively large attention hypothesis space. Experimental results show that our method solves the training failure problem of NAT methods in Doc-to-Doc translation scenarios and achieves competitive performance compared to the document-level AT method on two benchmark datasets. Furthermore, our method achieves fully non-autoregressive decoding, an average of 46 times faster than the document-level AT baseline method.

7. Acknowledgements

We sincerely thank all anonymous reviewers for their comments and suggestions. This work was supported by National Key R&D Plan (Approval Number: 2020AAA0108004) and Key R&D plans Yunnan Province (Approval No.: 202203AA080004).

8. Bibliographical References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Yu Bao, Hao Zhou, Shujian Huang, Dongqi Wang, Lihua Qian, Xinyu Dai, Jiajun Chen, and Lei Li. 2022. latent-GLAT: Glancing at latent variables for parallel text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8398–8409, Dublin, Ireland. Association for Computational Linguistics.

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

Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, and Omer Levy. 2020. Aligned cross entropy for non-autoregressive machine translation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3515–3523. PMLR.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. 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, 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.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.

Zhuohan Li, Zi Lin, Di He, Fei Tian, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Hint-based training for non-autoregressive machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5708–5713, Hong Kong, China. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2018. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3021, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

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.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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, Belgium, Brussels. Association for Computational Linguistics.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2021. Glancing transformer for non-autoregressive neural 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 1993–2003, Online. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1098–1108, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Amane Sugiyama and Naoki Yoshinaga. 2021. Context-aware decoder for neural machine translation using a target-side document-level language model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5781–5791, Online. 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.

Xin Tan, Longyin Zhang, Fang Kong, and Guodong Zhou. 2022. Towards discourse-aware document-level neural machine translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4383–4389. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. 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.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Bingzhen Wei, Mingxuan Wang, Hao Zhou, Junyang Lin, and Xu Sun. 2019. Imitation learning for non-autoregressive neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1304–1312, Florence, Italy. Association for Computational Linguistics.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. 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 1784–1794.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192, Dublin, Ireland. Association for Computational Linguistics.

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

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

Zhen Zhang, Junhui Li, Shimin Tao, and Hao Yang. 2023. Lexical translation inconsistency-aware document-level translation repair. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12492–12505, Toronto, Canada. 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-20*, pages 3983–3989. International Joint Conferences on Artificial Intelligence Organization. Main track.