Dynamic Context Selection for Document-level Neural Machine Translation via Reinforcement Learning

Xiaomian Kang[1,2], **Yang Zhao**[1,2], **Jiajun Zhang**[1,2,3], and **Chengqing Zong**[1,2,4]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Beijing Academy of Artificial Intelligence, Beijing, China
[4]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{xiaomian.kang, yang.zhao, jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

Document-level neural machine translation has yielded attractive improvements. However, majority of existing methods roughly use all context sentences in a fixed scope. They neglect the fact that different source sentences need different sizes of context. To address this problem, we propose an effective approach to select dynamic context so that the document-level translation model can utilize the more useful selected context sentences to produce better translations. Specifically, we introduce a selection module that is independent of the translation module to score each candidate context sentence. Then, we propose two strategies to explicitly select a variable number of context sentences and feed them into the translation module. We train the two modules end-to-end via reinforcement learning. A novel reward is proposed to encourage the selection and utilization of dynamic context sentences. Experiments demonstrate that our approach can select adaptive context sentences for different source sentences, and significantly improves the performance of document-level translation methods.

1 Introduction

Although neural machine translation (NMT) has achieved great progress in recent years (Cho et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017), when fed an entire document, standard NMT systems translate sentences in isolation without considering the cross-sentence dependencies. Consequently, document-level neural machine translation (DocNMT) methods are proposed to utilize source-side or target-side inter-sentence contextual information to improve translation quality over sentences in a document (Jean et al., 2017; Wang et al., 2017; Tiedemann and Scherrer, 2017; Tu et al., 2018; Kuang et al., 2018; Junczys-Dowmunt, 2019; Ma et al., 2020).

#	Test Context Settings	Model1	Model2
1	previous 2 sentences	<u>20.84</u>	20.94
2	previous 6 sentences	20.90	<u>21.15</u>
3	select 2 from previous 6	22.03	22.14
4	dynamic size from previous 6	22.90	22.74

Table 1: The BLEU (%) scores with different context settings. "Model1" and "Model2" are trained with previous 2 and 6 context sentences, respectively. Underlined results indicate that training and test context settings are consistent.

More recently, researchers of DocNMT mainly focus on exploring various attention-based networks to leverage the cross-sentence context efficiently, and evaluate the special discourse phenomena (Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019b; Jwalapuram et al., 2019). However, there is still an issue that has received less attention: *which context sentences should be used when translating a source sentence?*

We conduct an experiment to verify an intuition: the translation of different source sentences requires different context. As shown in Table 1, we train two DocNMT models and test them using various context settings[1]. During the test, we obtain dynamic context sentences that achieve the best BLEU scores by traversing all the context combinations for each source sentence. Compared with the fixed size context (row 1 and 2), dynamic context (row 3 and 4) can significantly improve translation quality. Although row 2 uses more context, redundant information may hurt the results. Experiments indicate that only the limited context sentences are really useful, and they change with source sentences.

Majority of existing DocNMT models set the context size or scope to be fixed. They utilize all of

[1]We apply a typical DocNMT method (Zhang et al., 2018) to train models on Zh→En TED, and select 1,000 sentences to test. The BLEU of sentence-level baseline is 20.06.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2242–2254,
November 16–20, 2020.

the previous k context sentences (Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Voita et al., 2019b; Yang et al., 2019; Xu et al., 2020), or the full context in the entire document (Maruf and Haffari, 2018; Tan et al., 2019; Xiong et al., 2019; Zheng et al., 2020). As a result, the inadequacy or redundancy of contextual information is almost inevitable. From this viewpoint, Maruf et al. (2019) propose a selective attention approach that uses the *sparsemax* function (Martins and Astudillo, 2016) instead of the softmax to normalize the attention weights. The sparsemax assigns the low probability in softmax to zero so that the model can focus on the sentences with high probability. However, the learning of attention weights lacks guidance, and they cannot handle the situation where the source sentences achieve the best translation results without relying on any context, which happens in about 39.4% of sentences in the experiment.

To address the problem, we propose an effective approach to select contextual sentences **dynamically** for each source sentence in the document-level translation. Specifically, we propose a *Context Scorer* to score each candidate context sentence according to the currently translated source sentence. Then, we utilize two selection strategies to select useful context sentences for the translation module. The size of selected context is variable for different sentences. A core challenge of our approach is that the selection process is non-differentiable. Therefore, we leverage the reinforcement learning (RL) method to train the selection and DocNMT modules together. We design a novel reward to encourage the model to be aware of different context sentences and select more appropriate context to improve translation quality.

In this paper, we make the following contributions:

- Our approach can measure the contribution of each context sentence to the source, and select dynamic context for the translation of different source sentences. Independent of the translation network, our approach is easily adaptable to existing DocNMT models.
- We bridge the training of context selection and context-aware translation via reinforcement learning. Experiments show that our approach can significantly improve the performance of DocNMT models with the selected dynamic context sentences.

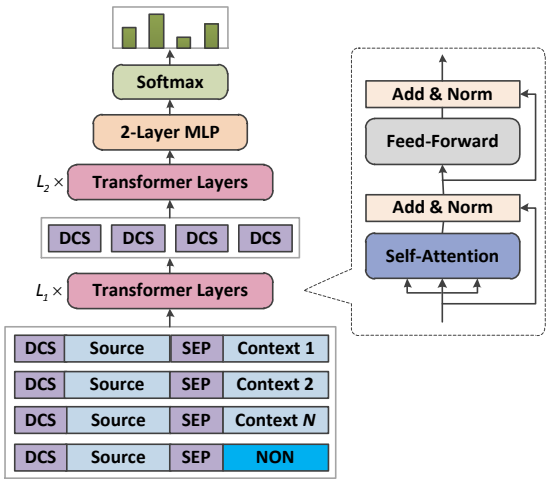

Figure 1: The architecture of context scorer. We add a special empty context sentence "NON" to help the decision of selection strategies. The details of Transformer layers are shown in the right dotted box.

2 Document-level Machine Translation

A standard DocNMT system generally translates a source sentence $X = \{x_1, \cdots, x_I\}$ to a target sentence $Y = \{y_1, \cdots, y_T\}$ with the aid of contextual information $\mathcal{Z}$ that is usually a subset of the candidate context set $\mathbb{Z}$. The model is trained to minimize the negative log-likelihood as:

$$\mathcal{L}_{mle} = -\sum_{t=1}^{T} \log P\left(y_t \mid y_{<t}, X, \mathcal{Z}; \theta\right) \quad (1)$$

Different granularity (word or sentence) and different sources (source-side or target-side) of contextual information $\mathcal{Z}$ have been explored. Maruf et al. (2019) divide the candidate context set $\mathbb{Z}$ into two cases: *offline* where the context comes from the entire document, and *online* that only uses the past context. In this paper, we mainly focus on a general scenario, where DocNMT translates sentences with the online source-side context sentences.

3 Dynamic Context Selection

Our approach translates a source sentence X in the document in two steps. First, we select the appropriate context sentences for the translation of X via the selection module. Independent of DocNMT module, this step is conducted before the context encoding in DocNMT module. The core component is a *Context Scorer* that calculates the contribution of each context sentence $z \in \mathbb{Z}$ to the translation of X (sub-section 3.1). According to the context scores, we propose two strategies to choose the useful context sentences (sub-section 3.2). Sec-

ond, we feed the selected context sentences into a DocNMT module to generate the translation.

To overcome the non-differentiable behavior of the context selection and the lack of direct supervision when training the context scorer, we connect the two steps through the reinforcement learning strategy. We propose an effective reward that is related to the translation quality to guide the dynamic selection of context sentences and the optimization of parameters in DocNMT model (sub-section 3.3).

3.1 Context Scorer

As Figure 1 shows, we obtain the representation of context sentences for scoring. Inspired by the popular pre-training language models such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2019), we produce one instance by concatenating the source sentence with a context sentence, and adding a special symbol "$\langle DCS \rangle$" at the beginning and a separator token "$\langle SEP \rangle$" in between. The instance is fed into a stack of L_1 Transformer encoder layers. We believe the special symbol "$\langle DCS \rangle$" can encode the information of source-context sentence pairs well by the self-attention.

For a candidate context sentence $z \in \mathbb{Z}$, its hidden state of "$\langle DCS \rangle$" after L_1 layers is extracted as the input to L_2 Transformer encoder layers to model the dependencies among context sentences. We denote the hidden state after L_2 layers as $h_z \in \mathbb{R}^{d_1}$. After that, we adopt a two-layer linear scorer network to measure the score as follows:

$$Score_z = \sigma\left(W_2\left(W_1 h_z + b_1\right) + b_2\right) \quad (2)$$

where $W_1 \in \mathbb{R}^{d_1 \times d_2}$, and $W_2 \in \mathbb{R}^{d_2 \times 1}$. σ stands for the logistic sigmoid function.

Considering the sampling operation during training process, we normalize all scores of context sentences in candidate set $\mathbb{Z}$ as a probability distribution:

$$\mathcal{P}_{select} = \text{softmax}\left(\left[Score_1; \cdots; Score_{|\mathbb{Z}|}\right]\right) \quad (3)$$

where $[\cdot;\cdot]$ concatenates elements into a vector.

3.2 Selection Strategies

According to the selection probability in $\mathcal{P}_{select}$, we can obtain useful context sentences for the translation task. To select context dynamically, we add a special empty sentence "$\langle NON \rangle$" into the candidate context set, which stands for the situation that translates a source sentence without any context. As a result, we select those context sentences whose probability is higher than "$\langle NON \rangle$". If the probability of "$\langle NON \rangle$" is the highest, context size is zero. We call this strategy as **probability-first**. The selected context sentences change dynamically with the change of source sentences, and the context size can range from 0 to $|\mathbb{Z}|$.

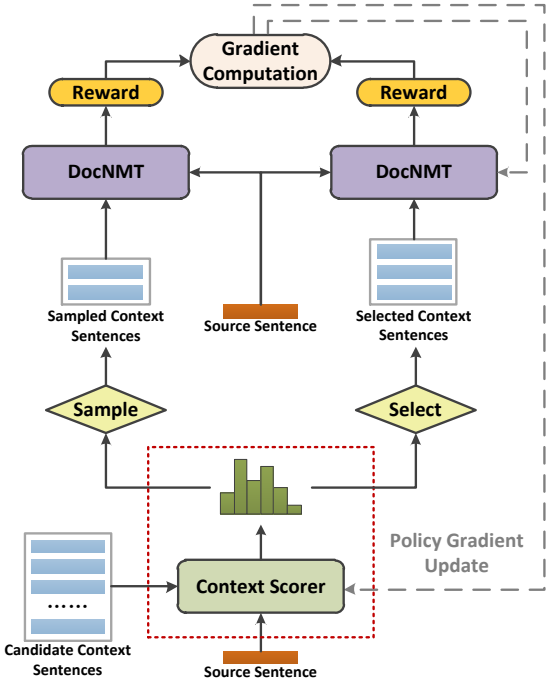

Figure 2: Reinforced training of the context selection and context-aware translation. The two DocNMT models share parameters.

In order to make a fair comparison with existing DocNMT models setting fixed context size, we also propose a **size-first** strategy that selects the certain number of context sentences with the highest probability except "$\langle NON \rangle$". Despite of the fixed size, the context is still dynamic because selected sentences can be anywhere and discontinuous in the document.

3.3 Model Learning

Our strategies perform a non-differentiable hard selection, and it is difficult to decide which context sentences are helpful for the translation. It makes the training quite intractable. Therefore, we apply the policy gradient method to train the selection module and the document-level translation module in an end-to-end fashion through a novel reward. The reward encourages the model to select more useful context to improve the generation probability of the ground truth translations. Figure 2 shows the reinforcement-guided training process.

3.3.1 Modules Initialization

It is well known that a fine initialization of network is important to optimize the parameters in reinforcement learning.

For DocNMT module that is usually trained in two stages (Tu et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Maruf et al., 2019), we load the parameters of standard sentence-level NMT model to initialize the network.

For the selection module, we simplify the initialization of context scorer as a binary classification task without considering the dependencies among context sentences. Its initialization contains two steps. First, we create pseudo labels for candidate context sentences. Each context sentence is labeled as `1` or `0`. The score in Eq. 2 is treated as the probability to predict label `1`. Specifically, pseudo labels are generated by an extra DocNMT model trained with a single random context sentence. We feed different candidate context sentences to the trained model to translate the same source sentence. Candidate context sentences with higher BLEU than "$\langle NON \rangle$" are labeled as `1`, while those with lower BLEU are labeled as `0`. Second, we train the context scorer to predict the pseudo labels. We share the parameters of embedding layer with initialized DocNMT model. The initial scorer is trained to minimize the cross-entropy loss.

3.3.2 Reward

Given that our goal of context selection is to improve translation quality, we propose a reward that can measure translation quality and is sensitive to the context changes [2].

For a decoding time t, we calculate the cost of generating ground truth target word y_t correctly as follows:

$$g_t = \log P_{\tilde{y}_t^{1st}} - \log P_{y_t} + \log P_{\tilde{y}_t^{1st}} \Big/ (P_{\tilde{y}_t^{1st}} - P_{\tilde{y}_t^{2nd}}) \quad (4)$$

where the first two items calculate the gap between the logarithmic probabilities of ground truth target word y_t and the best word $\tilde{y}_t^{1st}$ whose probability is the top one in the prediction probability distribution. And the last item is a regularization that indicates the difference of probabilities between $\tilde{y}_t^{1st}$ and the word $\tilde{y}_t^{2nd}$ with the second-highest probability. The bigger difference means the higher confidence on the prediction.

[2] In our preliminary experiment, we try BLEU as reward but it is not sensitive enough to distinguish different context. Also, decoding a sequence to calculate BLEU is time-consuming.

We obtain the average cost (whose value > 0) of generating the ground truth sentence $Y = \{y_1, \cdots, y_T\}$, and utilize a monotone decreasing function to get the final reward bounded in $0 \sim 1$ as follows:

$$r(g) = e^{-g} = e^{-\frac{1}{T}\sum_{t=1}^{T} g_t} \quad (5)$$

A high value of the reward means that it is easy to generate the ground truth. Therefore, the selected context sentences should be encouraged. Conversely, if a reward is low, generating the ground truth with the selected context would cost a lot, so the selection is discouraged.

3.3.3 Self-Critical Training

We train the whole model with the self-critical training method (Rennie et al., 2017; Bai et al., 2018). The goal of RL training is to minimize the negative expected reward. And in practice, the loss is usually approximated with a single sample u from the policy $\mathcal{P}$ as follows:

$$\mathcal{L}_{rl} = -\mathbb{E}_{u\sim\mathcal{P}}[r(u)] \approx -r(u), u \sim \mathcal{P} \quad (6)$$

The self-critical training introduces a baseline reward $r(u')$ to reduce the variance of the gradient, where u' is obtained by the inference algorithm at test time. The final gradient is estimated by:

$$\nabla\mathcal{L}_{rl} = (r(u) - r(u'))\nabla \log P(u) \quad (7)$$

Specifically, we denote the trainable parameters of the context scorer and DocNMT by ω and θ, respectively. For each source sentence X, we select a set of context sentences $\mathcal{Z}^*$ by our selection strategies in sub-section 3.2. Meanwhile, another set of context sentences $\hat{\mathcal{Z}}$ with the same size of $\mathcal{Z}^*$ is sampled according to $\mathcal{P}_{select}$ in equation 3. Two sets of context sentences are fed into the same DocNMT module to obtain the rewards $r(\mathcal{Z}^*)$ and $r(\hat{\mathcal{Z}})$, respectively. Therefore, referring to equation 7, the final gradient of the context scorer is calculated by:

$$\nabla_\omega\mathcal{L}(\omega) = (r(\hat{\mathcal{Z}}) - r(\mathcal{Z}^*))\nabla_\omega \log P_\omega(\hat{\mathcal{Z}}) \quad (8)$$

where $P_\omega(\hat{\mathcal{Z}})$ is the probability of sampling $\hat{\mathcal{Z}}$ from $\mathcal{P}_{select}$. With the baseline reward $r(\mathcal{Z}^*)$ obtained by the current best policy (i.e., learned selection strategies), the method encourages model to explore more useful context (i.e., sampled context) that yields higher reward than the current best (i.e., selected context).

Datasets		Training	Dev	Test
Zh→En	TED	0.23M	0.88K	4.68K
	News	0.31M	2.00K	3.98K
En→De	TED	0.21M	0.89K	4.70K
	News	0.33M	3.00K	3.00K
	Europarl	1.67M	3.59K	5.14K

Table 2: Dataset statistics in the number of sentences.

For DocNMT module, we can combine the MLE objective (Eq. 1) and RL objective (Eq. 6) together to stabilize the training procedure (Wu et al., 2018) through a balance factor α as follows:

$$\mathcal{L}(\theta) = \alpha * \mathcal{L}_{mle}(Y \mid X, \hat{\mathcal{Z}}, \theta) + (1-\alpha) * \mathcal{L}_{rl}(\theta) \quad (9)$$

We introduce the RL objective into DocNMT module so that the model can make better use of the selected context. The final RL gradient of DocNMT is calculated by:

$$\nabla_\theta \mathcal{L}_{rl}(\theta) = (r(\hat{\mathcal{Z}}) - r(\mathcal{Z}^*)) \nabla_\theta \log P_\theta(\hat{Y} \mid X, \hat{\mathcal{Z}}) \quad (10)$$

where $\hat{Y}$ is a sequence generated by current DocNMT model with the sampled context $\hat{\mathcal{Z}}$.

4 Experiment

4.1 Datasets

We evaluate our approach on different domains of Chinese-English (Zh→En) and English-German (En→De) datasets. The corpora statistics are listed in Table 9. For TED Talks in IWSLT17[3], we use *dev-2010* as the development set, and *tst-2010∼2013* as the test set for both Zh→En and En→De language pairs. For News-Commentary v14[4], we use the *newstest2017* for development and *newstest2018* for testing. Europarl is a large scale corpus extracted from Europarl v7, and we use the same training, development and test sets as Maruf et al. (2019).

4.2 Models

We compare our approach with the following methods: **1) SENTNMT** (Vaswani et al., 2017) is a standard sentence-level Transformer model using the "base" version parameters. **2) TDNMT** (Zhang et al., 2018) introduces the contextual information by adding attention sub-layers at each encoder and decoder layer. We use 2 previous consecutive context sentences as they suggested. **3) HAN** (Miculicich et al., 2018) uses 3 previous sentences as context. We adopt the "HAN encoder + HAN decoder" strategy that adds a hierarchical network on the top of the last encoder and decoder layer to model sentence-level and word-level contextual information. **4) SAN** (Maruf et al., 2019) utilizes all context in the entire document by calculating the sentence-level and word-level weights. It focuses on relevant context sentences through the sparsemax function. We choose the model that integrates the online context into encoder with "sparse-soft H-Attention".

We implement our approach and baseline methods based on the toolkit THUMT (Zhang et al., 2017). The parameters are the "base" version of the original Transformer (Vaswani et al., 2017). The d_1 and d_2 in Eq. 2 are 512 and 256, respectively. We set the layers of $L_1 = 2$ and $L_2 = 2$. The effect of layer depth of context scorer and more implementation details are shown in the appendix.

5 Results and Analysis

5.1 Main Results

We use BLEU (Papineni et al., 2002) score to evaluate the translation quality. Considering the memory limitation and complex sampling space, we select dynamic context from previous six sentences. Table 10 shows the performance of models utilizing different context settings. We always keep the same setting for training and test.

Comparison with Fixed Context Methods. The performance of DocNMT models with fixed context is shown in row 2∼5. Row 2 and 3 follow the context settings in the published papers. It can be found that using more context sentences indiscriminately (row 4 and 5) does not bring significant BLEU improvement. Instead, it increases computational cost.

By contrast, our approach (row 10∼15) can significantly improve translation quality on all datasets. Let us take the TDNMT models on Zh→En TED for example. Row 11 applies the size-first strategy to select context sentences of the same size as original TDNMT model in row 3. The result achieves +0.70 BLEU improvement (20.09 vs. 19.39). Compared with row 5 that uses all context in previous six sentences, our approach can filter some redundant information and focus on fewer selected context sentences to gain +0.64 BLEU scores (20.09 vs. 19.45). On the other hand, even if the context is selected from previous two sentences (row 14), our model utilizing probability-

[3] https://wit3.fbk.eu/mt.php?release=2017-01-trnted

[4] http://data.statmt.org/news-commentary/v14

#	Model	Context Settings			TED		News		Europarl
		Scope	Size	Method	Zh-En	En-De	Zh-En	En-De	En-De
	Baselines with Fixed Context								
1	SENTNMT (Vaswani et al., 2017)	–	–	–	18.67	28.23	13.21	25.85	28.80
2	HAN (Miculicich et al., 2018)	3	3	full	19.54	29.45	13.87	26.81	29.85
3	TDNMT (Zhang et al., 2018)	2	2	full	19.39	29.14	13.51	26.25	29.32
4	HAN	6	6	full	19.33	29.37	13.90	26.90	29.82
5	TDNMT	6	6	full	19.45	29.02	13.54	26.26	29.26
	Baselines with Dynamic Context w/o RL Selection								
6	HAN	6	3	random	19.41	29.45	14.03	26.87	29.78
7	TDNMT	6	2	random	19.40	29.13	13.57	26.28	29.29
8	SAN (Maruf et al., 2019)	all	dyn	attend	19.60	29.41	14.08	26.79	29.81
9	SAN	6	dyn	attend	19.49	29.43	14.11	26.82	29.77
	Our Methods								
10	HAN + *DCS-SF*	6	3	select	20.06	29.92	14.43	27.38	30.40
11	TDNMT + *DCS-SF*	6	2	select	20.09	29.70	14.36	26.93	29.89
12	HAN + *DCS-PF*	3	dyn	select	19.97	29.87	14.37	27.34	30.36
13	HAN + *DCS-PF*	6	dyn	select	20.26	**30.22**	14.48	**27.61**	**30.48**
14	TDNMT + *DCS-PF*	2	dyn	select	19.91	29.50	14.19	26.62	29.64
15	TDNMT + *DCS-PF*	6	dyn	select	**20.34**	30.09	14.65	27.06	30.18
16	SAN + *DCS-PF*	6	dyn	select	20.18	30.13	**14.71**	27.43	30.37

Table 3: Performance of models on BLEU (%) using different context settings. "full" means using all context in the scope. "random", "attend", and "select" stand for selecting sentences randomly, implicitly based on attention weights, and explicitly by our approaches, respectively. "dyn" stands for dynamic size. "DCS-SF" and "DCS-PF" mean dynamic context selection by size-first and probability-first strategies respectively. All results using "DCS" are statistically significantly (p-values < 0.05) better than corresponding original DocNMT models.

first strategy can still improve original TDNMT by +0.52 BLEU (19.91 vs. 19.39). It indicates that useless context sentences still exist in a small scope. Conclusions are similar for other models and datasets.

Comparison with Other Dynamic Methods. Row 6∼16 show the models trained and tested with dynamic context settings. Row 6 and 7 show a lower bound that randomly selects the fixed size context sentences. The results are similar to original models with the fixed size previous sentences (row 2 and 3). In contrast to the random selection, our approach (row 10 and 11) can select the same size of context sentences that are really helpful to generate better translations.

SAN (row 8) implicitly selects context from all previous sentences through sharpening the attention weights. It resets low attention weights to zero to filter out some sentences. For a fair comparison, we also implement SAN in a limited context scope (row 9). Even if the candidate set is limited to previous six sentences, the BLEU does not decrease significantly. Different from SAN, our approach explicitly selects context sentences via reinforced guide. As row 16 shows, when added into SAN (row 9), our approach can obtain +0.69 BLEU gains (20.18 vs. 19.49) on Zh→En TED by picking a more focused context candidate set for SAN. Furthermore, our approach can set the context size to be zero, but SAN cannot deal with this common cases that do not require any context.

#	Training		Balance RL Loss	BLEU
	Scorer	DocNMT		
1	×	×	–	29.04
2	✓	×	–	29.58
3	✓	✓	$\alpha = 1.00$	29.61
4	✓	✓	$\alpha = 0.75$	29.92
5	✓	✓	$\alpha = 0.50$	29.73
6	✓	✓	$\alpha = 0.25$	29.70
7	✓	✓	$\alpha = 0.00$	29.65

Table 4: Effect of training settings for DocNMT models. BLEU scores are measured based on TDNMT on the development set in En→De Europarl. ✓ means training the module while × means not. Row 1 stands for the original TDNMT model.

Comparison of Selection Strategies. We also compare the two selection strategies proposed in section 3.2. Results with probability-first strategy (row 13 and 15) are slightly better than size-first strategy (row 10 and 11). The size-first strategy has to contain some useless sentences because of the fixed size. By contrast, the probability-first strategy allows more flexible context selection of dynamic size. It can achieve +0.72 (20.26 vs. 19.54) and +0.95 (20.34 vs. 19.39) BLEU improvement on Zh→En TED when applied to HAN and TDNMT model, respectively.

5.2 Effect of DocNMT Training

Our proposed context selection module is independent of the translation module. Therefore, the con-

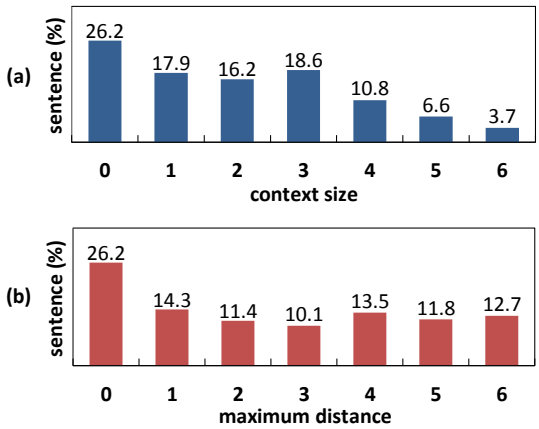

Figure 3: Distribution of dynamic context size and position. (a) shows the ratio of source sentences with different context sizes. (b) counts the maximum distance from the selected context sentences to their corresponding source sentences.

text scorer and DocNMT can be trained separately. As shown in Table 11, we discuss the impact of selected context on the training of DocNMT. In row 2, we only train the context scorer, and keep the original DocNMT model unchanged as a component to calculate rewards. The result shows that our selection module can effectively distinguish between useful and useless context sentences for translation, and achieves +0.54 BLEU gains on the En→De Europarl development set.

We also explore whether the selected context would be helpful for the DocNMT training. We set the balance factor α Eq. 9 to be [0, 0.25, 0.5, 0.75, 1.0] in our experiments. Row 3 shows the model setting $\alpha = 1.0$ that optimizes the standard MLE loss using the selected context sentences. Row 7 sets $\alpha = 0.0$ to fine-tune DocNMT with the RL loss. By contrast, DocNMT models guided by the combination of MLE and RL loss can be learned better. We think the RL loss may make the model more sensitive to the selected context sentences. When $\alpha = 0.75$, DocNMT can obtain the best BLEU score on development set, thus we use the setting in our experiments.

5.3 Distribution of Dynamic Context

Figure 3 shows the distribution of different context sizes and maximum distances in the test sets of Zh→En TED. Our approach selects context sentences whose size can range from zero to six. In Figure 3 (a), 78.9% of source sentences tend to select no more than three context sentences. 26.2% of sentences can be translated well without contextual information. The average context size over the test sets is 2.05 sentences. In Figure 3 (b), we show the maximum distances from the selected context sentences to the currently translated sentence. Except for the cases that need no context (distance 0), the distance distribution is relatively uniform. The total average distance is previous 2.57 sentences.

Precision	Recall	F1
68.46	51.78	58.96

Table 5: Results of empty context prediction on 500 sentences with human annotation as reference.

Model	Ctx-Empty	Ctx-Nonempty
SENTNMT	19.75	20.04
TDNMT	20.49	20.72
+*DCS-PF*	21.51 (+1.02)	21.43 (+0.71)

Table 6: BLEU (%) scores on the context-empty and context-nonempty test sets. "+" stands for the improvement when compared with TDNMT.

5.4 Selection of Empty Context

Our approach has the ability to select empty context for translation, which other models such as SAN (Maruf et al., 2019) cannot do. To evaluate whether the selected empty context is reasonable, we annotate a special test set that contains 500 sentences selected randomly from Zh→En TED test sets. Each sentence is given its previous 6 sentences as context. Two annotators are instructed to mark context-empty sentences that can be translated well without any contextual information. The annotation details and statistics are shown in the appendix. The Cohen's Kappa value (Cohen, 1960) of annotation is 0.72. We gather sentences marked by both annotators as the final context-empty sentences (about 39.4% in 500 sentences). Therefore, the test set is divided into context-empty and context-nonempty subsets. Their sizes are 197 and 303, respectively.

Table 5 shows the performance of our approach (using "TDNMT+*DCS-PF*" model) for predicting empty context on the 500 annotated sentences. For the selection of empty context, our approach can achieve 58.96 F1-score.

Table 6 shows the BLEU scores on the context-empty and context-nonempty subsets. Through our context selection, the improvement of BLEU on context-empty set is higher than context-nonempty set. The analysis indicates that our approach is aware of context-empty sentences, and can select empty context to improve translation quality.

Model	deixis	lex.c.	ell.infl.	ell.VP
TDNMT+*DCS-PF*	60.9	85.1	52.4	81.0
	83.4	89.6	88.2	90.6
CADec+*DCS-PF*	67.0	90.5	50.8	84.4
	85.7	95.4	89.2	91.8

Table 7: Accuracy (%) of context selection on the discourse phenomena test sets. A model contains two rows: upper – exact match, lower – selected context contains the golden answer.

Model	deixis	lex.c.	ell.infl.	ell.VP
SENTNMT	50.0	45.9	53.0	28.4
TDNMT	50.0	46.0	56.4	48.0
CADec	81.6	58.1	72.2	80.0
+*DCS-PF*	79.2	62.0	71.8	80.8

Table 8: Accuracy (%) of discourse phenomena.

5.5 Analysis of Discourse Phenomena

In addition to the selection of empty context, we also want to examine whether our approach can select context sentences that are helpful to improve the translation of discourse phenomena.

Voita et al. (2019b) construct contrastive test sets for English-Russian to evaluate four types of discourse phenomena (i.e., deixis, lexical cohesion, inflection and VP ellipses). Each test instance consists of a positive and several negative translations with incorrect phenomena. Models are evaluated by the accuracy that is defined as the proportion of times the generation probability of positive translation is higher than negative ones. Meanwhile, each instance has three context sentences. Among them, there is one and only one context sentence that is decisive in resolving the phenomena. It has been marked. Therefore, we can evaluate the accuracy of context selection, taking the marked context sentences as the standard answer.

We use the same datasets as Voita et al. (2019b) to train models. Different from TDNMT (Zhang et al., 2018) that only uses source-side context, CADec (Voita et al., 2019b) is proposed to utilize both source-side and target-side context. Based on CADec, we try to extend our approach in a simple way to select the target-side context. When the context scorer selects a source-side context sentence, the corresponding sentence-level translation is directly selected as target-side context.

Table 7 shows the accuracy of context selection at four test sets. It can be found that our approach can select more than 85% standard context sentences for special phenomena, and achieve more than 80% exact match on lexical cohesion and VP ellipses sets.

The accuracy of discourse phenomena are shown in Table 8. TDNMT does not perform well because it only uses source-side context, which is unchanged in contrastive instances of test sets. Compared with original CADec, our approach can improve the performance of lexical cohesion. Although the simple way of selecting target-side context bears the risk of missing selection, the accuracy of some phenomena does not change significantly. Table 7 has shown that our approach can select useful target-side context in most cases. And the selection mechanism can make the model focus more on the useful context to resolve the discourse phenomena.

6 Related Work

Standard neural machine translation methods usually focus on the sentence-level translation (Cho et al., 2014; Bahdanau et al., 2015; Zhang and Zong, 2015; Luong et al., 2015; Tu et al., 2016; Zhang and Zong, 2016; Vaswani et al., 2017; Wang et al., 2019; Zhou et al., 2019; Zhao et al., 2020). As a contrast, document-level neural machine translation methods mainly pay attention to how to utilize the cross-sentence context. Researchers propose various context-aware networks to utilize contextual information to improve the performance of DocNMT models on the translation quality (Jean et al., 2017; Tu et al., 2018; Kuang et al., 2018) or discourse phenomena (Bawden et al., 2018; Xiong et al., 2019; Voita et al., 2019b,a). However, most methods roughly leverage all context sentences in a fixed size that is tuned on development sets (Wang et al., 2017; Miculicich et al., 2018; Zhang et al., 2018; Yang et al., 2019; Voita et al., 2018; Xu et al., 2020) , or full context in the entire document (Maruf and Haffari, 2018; Tan et al., 2019; Kang and Zong, 2020; Zheng et al., 2020). They ignore the individualized needs for context when translating different source sentences.

Some works have noticed that not all context is useful (Jean and Cho, 2019; Kim et al., 2019). Kimura et al. (2019) explore the context selection in the single-encoder framework (Tiedemann and Scherrer, 2017), and select context sentences that yield highest forced back-translation probability. However, the method cannot optimize DocNMT model at training phase, and requires back-translation model at inference phrase. Maruf et al. (2019) sharpen the attention weights between the source and context sentences through the *sparse-*

max function, and implicitly select context with high attention weights. Nevertheless, the method lacks direct supervision over context selection, and it cannot cover the situation where context is not needed. Inspired by the extractive-abstractive summarization (Chen and Bansal, 2018), our approach is different from above DocNMT methods. Our approach can explicitly select dynamic size (that can be 0) of context sentences for the translation of different source sentences.

7 Conclusion and Future Work

We propose a dynamic selection method to choose variable sizes of context sentences for document-level translation. The candidate context sentences are scored and selected by two proposed strategies. We train the whole model via reinforcement learning, and design a novel reward to encourage the selection of useful context sentences. When applied to existing DocNMT models, our approach can improve translation quality significantly. In the future, we will select context sentences in larger candidate space, and explore more effective ways to extend our approach to select target-side context sentences.

Acknowledgments

We thank anonymous reviewers for their insightful comments and suggestions. The research work described in this paper has been supported by the Natural Science Foundation of China under Grant No. U1836221 and 61673380. The research work in this paper has also been supported by Beijing Advanced Innovation Center for Language Resources and Beijing Academy of Artificial Intelligence (BAAI2019QN0504).

References

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

He Bai, Yu Zhou, Jiajun Zhang, Liang Zhao, Mei-Yuh Hwang, and Chengqing Zong. 2018. Source critical reinforcement learning for transferring spoken language understanding to a new language. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3597–3607, Santa Fe, New Mexico, USA. 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.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*.

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.

Sébastien Jean and Kyunghyun Cho. 2019. Context-aware learning for neural machine translation. *arXiv preprint arXiv:1903.04715*.

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

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

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. 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 2964–2975, Hong Kong, China. Association for Computational Linguistics.

Xiaomian Kang and Chengqing Zong. 2020. Fusion of discourse structural position encoding for neural machine translation. *Chinese Journal of Intelligent Science and Technology*, 2(2):144–152.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Ryuichiro Kimura, Shohei Iida, Hongyi Cui, Po-Hsuan Hung, Takehito Utsuro, and Masaaki Nagata. 2019. Selecting informative context sentence by forced back-translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 162–171, Dublin, Ireland. European Association for Machine Translation.

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.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

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.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA. PMLR.

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 FT 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.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

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.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level 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 1576–1585, Hong Kong, China. Association for Computational Linguistics.

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.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. 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.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019. A compact and language-sensitive multilingual translation method. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223, Florence, Italy. Association for Computational Linguistics.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Hongfei Xu, Deyi Xiong, Josef van Genabith, and Qiuhui Liu. 2020. Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3933–3940. International Joint Conferences on Artificial Intelligence Organization.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level 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 1527–1537, Hong Kong, China. Association for Computational Linguistics.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017. Thumt: An open source toolkit for neural machine translation. *arXiv preprint arXiv:1706.06415*.

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.

Jiajun Zhang and Chengqing Zong. 2015. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, (5):16–25.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graphs enhanced neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4039–4045. International Joint Conferences on Artificial Intelligence Organization.

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.

Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics*, 7:91–105.

A Experimental Setup

A.1 Parameters and Implementation

We implement all models based on the toolkit THUMT[5] with the parameters of the "base" version of Transformer (Vaswani et al., 2017). Specifically, we use 6 layers of encoder and decoder with 8 attention heads. The hidden size and feed-forward layer size are 512 and 2,048, respectively. For Zh→En, Chinese and English vocabulary sizes are 30K and 25K, respectively. For En→De, source-side and target-side share a vocabulary table. The vocabulary size is 30K. Chinese sentences are segmented into words by our in-house toolkit. English and German datasets are tokenized and truecased by the Moses toolkit[6]. Words are segmented by byte-pair encoding (Sennrich et al., 2016).

We introduce a context scorer that is independent of the DocNMT models, which allows our approach to be easily deployed on many baseline DocNMT systems. Compared with original DocNMT models, the amount of additional parameters depends on the number of Transformer encoder layers L_1 and L_2 in the context scorer.

		L_1		
		1	2	4
L_2	0	29.35	29.58	29.64
	1	29.60	29.72	29.77
	2	29.76	**29.92**	29.83

Table 9: BLEU (%) scores on En→De TED development set using different layers of context scorer.

In Table 9, we discuss the effect of layer depth of context scorer (defined in subsection 3.1). Experiments are conducted using "TDNMT+*DCS-PF*" model with the balance factor $\alpha = 0.75$. Our approach achieves the highest BLEU with a context scorer setting $L_1 = 2$ and $L_2 = 2$, which introduces 12.7M extra parameters to any original DocNMT models.

[5] https://github.com/thumt/THUMT

[6] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

A.2 Training and Inference

For training, we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. We employ label smoothing with a value of 0.1 and dropout with a rate of 0.1. The batch size is 3,000 tokens. We employ 4 Titan Xp GPUs to train all models. Compared with original DocNMT (TDNMT), the training and testing speeds are slowed down by an order of 1.61 (mainly because of the generation of $\hat{Y}$ in Eq. 10) and 1.05, respectively.

We use *multi-bleu.perl*[7] to compute the BLEU score. The beam size is set to 4. The significance test is conducted by the script "bootstraphypothesis-difference-significance.pl" in Moses.

B Annotation and Statistics of Empty Context

In this section we describe the annotation process and statistics of the special test set constructed to evaluate the selection of empty context.

B.1 Annotation

We randomly select 500 sentences with previous 6 sentences as context from Chinese-English TED *tst-2010∼2013*. Each example to be annotated contains a source-reference sentence pair and six source-reference contextual sentence pairs. Two annotators proficient in both Chinese and English are instructed to annotate the sentences that can be translated well without any context. The process consists of three steps, and is carried out independently between two annotators.

Step1. Annotators are instructed to read a single source sentence X without any context, and translate it into Y' by themselves.

Step2. The reference Y of the sentence X is shown to annotators. Then, they are instructed to compare Y' with Y word-by-word to answer whether Y' is appropriate.

Step3. Annotators are instructed to read source-reference contxtual sentences, and compare Y' with Y word-by-word again. After that, they are asked to determain whether Y' needs to be modified better.

If a annotator insists that his translation Y' is appropriate at Step2 and needs no modification at Step3, the sentence X is annotated as "context-empty", which means it can be translated well with-

[7] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

		A1	
		Ctx-Empty	Ctx-Nonempty
A2	Ctx-Empty	197	51
	Ctx-Nonempty	20	232

Table 10: Statistics of human annotation for empty context. A1 and A2 stand for two annotators.

		Human Annotation	
		Ctx-Empty	Ctx-Nonempty
Ours	Ctx-Empty	102	47
	Ctx-Nonempty	95	256

Table 11: Statistics of our approach (*Ours*) for empty context prediction.

out relying on any context. Otherwise, the sentence is annotated as "context-nonempty".

Table 10 shows the statistics of annotation. The Cohen's Kappa value is 0.72. 197 context-empty sentences are annotated by both annotators. These sentences are gathered as the final context-empty test set. The other 303 sentences make up the context-nonempty test set.

B.2 Statistics of Empty Context Selection

Taking the human annotation in Table 10 as the golden test set, Table 11 shows the statistics of empty context prediction by our approach in subsection 5.3.