CONTEXT-ADAPTIVE DOCUMENT-LEVEL NEURAL MACHINE TRANSLATION

Linlin Zhang[⋆] *Zhirui Zhang*[†] *Boxing Chen*[†] *Weihua Luo*[†] *Luo Si*[†]

[⋆] Zhejiang University
[†]Alibaba DAMO Academy
China

ABSTRACT

Document-level translation models are still far from perfect. Most existing document-level neural machine translation (NMT) models leverage a fixed number of the previous or all global sentences to handle the context-independent problem in standard NMT. However, the translating of each source sentence benefits from various sizes of context. And study shows that inappropriate redundant context will increase model burden but not improve the translation performance. This work introduces a data-adaptive method that enables the model to adopt the necessary and helpful context. Specifically, we introduce a light predictor into two document-level translation models to select the explicit context. Experiments demonstrate the proposed approach can significantly improve the performance over the previous methods with a gain up to 1.99 BLEU points.

Index Terms— Machine translation, document-level, context, adaptive

1. INTRODUCTION

Neural machine translation (NMT) based on the encoder-decoder framework has advanced translation performance in recent years [1, 2, 3]. Instead of translating sentences in isolation, document-level machine translation (DocMT) methods are proposed to capture discourse dependencies across sentences by considering a document as a whole.

Current DocMT systems usually leverage a fixed amount of source or target context sentences while translating [4, 5, 6]. It is observed in Kang et al [7], which evaluates some previous DocMT models using different contexts, and the results show that less context can get a higher BLEU than a fixed previous context sometimes. Thus, the translation model may need a more flexible context instead of a fixed static context. Choosing a proper context becomes vital in DocMT systems [8, 7].

To tackle this problem, Maruf et al [8] proposed a selective attention approach that normalizes the attention weights via the sparsemax function instead of the softmax. In their model, the sparsemax converts the low probability in softmax to zero, only keeping the sentences with high probability. However, this method focuses on selective attention weights and cannot handle cases where the source sentence achieves the best translation result without using any context.

In this paper, we propose a novel framework to predict the most appropriate context for model translation. To achieve this goal, we directly utilize a lightweight predictor, which takes the encoder outputs as the inputs, to predict the probabilities of different context options. We take their corresponding classification losses as the training signals to update this lightweight predictor. Based on this, the best prospective context for each source sentence can be selected when inference with only introducing little time cost. Also, the candidate contexts are limited in the most relevant pre-sentence and post-sentence, without searching from the enormous scope like the previous work. Experimental results prove that our proposed approach can significantly outperform the previous baselines with a margin up to 1.99 BLEU points.

Our main contributions can be summarized as follows:

- Our method can select the appropriate context by introducing a lightweight predictor. The predictor and the DocMT model are trained jointly with a few additional parameters.
- Our method is applied to two basic DocMT models where one utilizes source context information, and another uses both source and target. The two models gain significant improvements with our proposed method.

2. CONTEXT-ADAPTIVE DOCMT

In this section, we introduce our Context-adaptive DocNMT by adding a smart context predictor.

2.1. Document-Level Translation

Compared with sentence-level MT that translates a sentence $X = \{x_1, \cdots x_S\}$ into a target sentence $Y = \{y_1, \cdots y_T\}$, DocMT takes advantage of the contextual information C of the document. The DocMT model can be trained to minimize the negative log-likelihood loss as:

$$\mathcal{L}_{DMT} = -\sum_{t=1}^{T} \log P\left(y_t \mid y < t, X, C; \theta\right) \qquad (1)$$

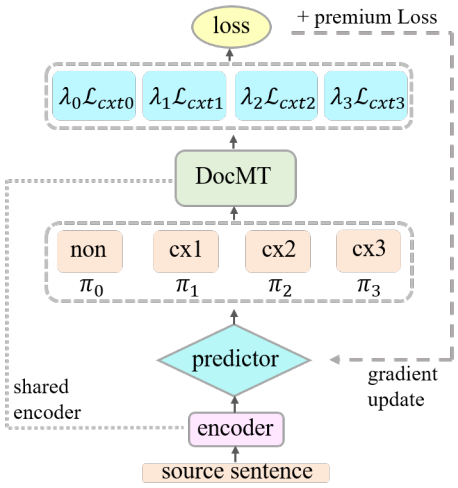

Fig. 1. The framework for Context-adaptive DocMT. π is the context option probability from the predictor. λ and premium losses are detailed in section 2.2.1.

2.2. Context Predictor

To select the appropriate context, we predict context based on the current inputs. As shown in Figure 1, the source sentence is processed by the encoder of the DocMT model. Then we feed the corresponding encoder output to the predictor, to calculate the probability π_i of the i-th choice.

Considering there are N different options for context to choose, including not adopting any context (empty context). Each source sentence has its preference for some context-added options. We leverage the cross-entropy loss as training signals to learn probability. Therefore, for each context-added option cxt_i, we calculate the probability λ_i of cxt_i, and then weigh the corresponding translation loss $\mathcal{L}^i_{MT}$ with the selected context.

2.2.1. *Training loss*

For each source sentence $X = \{x_1, \cdots x_S\}$, where S is the length of the input, $H = \{h_1, \cdots h_S\}$ is the corresponding output of DocMT encoder. The light predictor or classifier leverages the averaged encoder outputs $H_s = \frac{1}{s}\sum_{k=1}^{S} h_{\mathrm{k}}$ to predict the possibility of the context-selected options. Then, uses softmax to calculate the possibility:

$$\pi = \mathrm{softmax}\left(H_s \times W + b\right) \tag{2}$$

where $\mathrm{W} \in \mathbb{R}^{d\times N}$ is the projection weight matrix. The weight λ_i of $\mathcal{L}^i_{MT}$ for each context option cxt_i is as:

$$\lambda_i = \frac{\exp\left(\left(\log\left(\pi_i\right) + g_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\left(\log\left(\pi_j\right) + g_j\right)/\tau\right)} \tag{3}$$

where noise g_i is sampled from the Gumbel distribution, and τ is a constant temperature. The cross-entropy loss of training is a weighted combination of the confidence λ_i and the corresponding loss $\mathcal{L}^i_{MT}$, formulated as:

$$\mathcal{L}_{MT} = \sum_{i=1}^{N}\lambda_i\mathcal{L}^i_{MT} \tag{4}$$

Premium loss

As mentioned above, the predictor is trained in an unsupervised way, which quickly trained to prefer one specific option. To explore the diverse capability of all options, we incorporate the KL-divergence by adding a diversity loss $\mathcal{L}_{div}$:

$$\begin{aligned}\mathcal{L}_{div} &= KL(\mathbb{U}\|\mathbb{E}[\pi]) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log\left(\mathbb{E}\left[\pi_i\right]\right) - \log N\end{aligned} \tag{5}$$

But in the inference time, we hope the predictor can make an unambiguous selection. So we expect the probability to be far away from the uniform distribution $\mathbb{U}$ via adding the $\mathcal{L}_{uni}$ loss:

$$\begin{aligned}\mathcal{L}_{uni} &= -\mathbb{E}\left[KL(\mathbb{U}\|\pi)\right] \\ &= -\mathbb{E}\left[-\frac{1}{N}\sum_{i=1}^{N}\log\pi_i - \log N\right]\end{aligned} \tag{6}$$

Inspired by the BERT [9], we also use segment embedding and mask strategy on our DocMT models. In this way, the final training objective is to minimize the loss as:

$$\mathcal{L} = \mathcal{L}_{MT} + \beta_1\mathcal{L}_{div} + \beta_2\mathcal{L}_{uni} + \beta_3\mathcal{L}_{mask} \tag{7}$$

where $\mathcal{L}_{mask}$ is the token masked loss of source sentence, β_1, β_2, and β_3 are the hyper-parameters described in Table 4.

As a result, the DocMT model and the predictor can be trained jointly.

2.2.2. *Inference prediction*

When inferring, we use the trained predictor to choose the most suitable context option according to the averaged encoder output H_s of source X:

$$cxt_n = \mathrm{argmax}\left(H_s \times W + b\right) \tag{8}$$

In this way, the framework can dynamically select the appropriate context by the predictor. All options use the same DocMT model.

3. EXPERIMENTS

3.1. Datasets and Settings

For a fair comparison with the previous works, we conducted experiments on four widely used document-level parallel datasets of two language pairs:

(1) **TED (ZH-EN)**: The Chinese-English TED dataset are from IWSLT 2015 evaluation campaigns. For TED ZH-EN, we take dev2010 as the valid set and tst2010-2013 as the test set.

(2) For the **EN-DE** language pair, we directly use the 3 prepared EN-DE corpora extracted by Maruf et al [8].

	ZH-EH	EN-DN			
	TED	TED	News	Europarl	$\Delta\|\theta\|$
Sentence Transformer [2]	17.56	23.10	22.40	29.40	
Sentence Transformer (our implementation)	18.38	25.08	24.32	29.98	0.0m
HAN [4]	17.9	24.58	25.03	28.60	4.8m
SAN [8]	n/a	24.42	24.84	29.75	4.2m
QCN [10]	n/a	25.19	22.37	29.82	n/a
Flat-Transformer [11]	n/a	24.87	23.55	30.09	0.0m
MCN [12]	19.1	25.10	24.91	30.40	4.7m
G-Transformer [13]	n/a	25.12	25.52	**32.39**	n/a
Context-unit (our implementation)	19.12	25.75	24.89	30.41	7.9m
Context-unit + Our predictor	19.81	26.29	25.41	30.84	9.5m
Concatenate (our implementation)	18.69	25.36	24.56	30.19	0.0m
Concatenate + Our predictor	**20.68**	**26.77**	**25.81**	31.13	2.1m

Table 1. The translation results of the test sets in BLEU score and the increments of the number of parameters over Transformer baseline ($\Delta|\theta|$), when compared with several baselines.

	concatenate	num	percentage
1	non‖source‖non	336	14.80%
2	pre‖source‖non	578	25.45%
3	non‖source‖pos	322	14.18%
4	pre‖source‖pos	1035	45.57%

Table 2. Statistics of the predictor's selections of concatenate model on the TED EN-DE test set.

model	all tokens	all time
sentence-level	48833	659.5s
concatenate	148469	1962.3s
our model	108622	1878.8s

Table 3. Statistics of the models' inference time of batch size one on the TED EN-DE test set(total 2271 sentences).

For a fair comparison, we use the same model configuration and training settings as Ma et al [11], and implement our experiments on Fairseq [1], detailed in the appendix materials.

3.2. Two Basic DocMT

We apply the above predictor on two prevalent and straightforward basic DocMT models with mini changes. All options share one DocMT model.

3.2.1. Context-Unit model

Many previous DocMT models use two encoders [14, 4, 10], one is to process the source sentence, and another is for the context. In a standard Transformer, each layer unit is composed of Multi-head Attention and a point-wise feed-forward network (FFN). The output F_i of the i-th layer can be calculated from the input X_i as:

$$S_i^{src} = \mathrm{SelfAttn}_i^{src}\left(X_i\right) + X_i \tag{9}$$

$$F_i^{src}\left(X_i\right) = \mathrm{FFN}_i^{src}\left(S_i\right) + S_i \tag{10}$$

We add a context-unit to process the context, the context output of i-th layer as:

$$S_i^{cxt} = \mathrm{SelfAttn}_i^{cxt}\left(C_i\right) + C_i \tag{11}$$

$$F_i^{cxt}\left(C_i\right) = \mathrm{FFN}_i^{cxt}\left(S_i^{cxt}\right) + S_i^{cxt} \tag{12}$$

Then add the context-unit output with a cross-attention weighted parameter α. The final output of i layer as:

$$F_i\left(X_i, C_i\right) = F_i^{src}\left(X_i\right) + \alpha\,\mathrm{CrossAttn}_i\left(F_i^{cxt}\left(C_i\right)\right) \tag{13}$$

In this model, there are $\mathrm{N} = 3$ different context inputs: previous sentence, next sentence, empty context replaced by the source sentence. The values of α and last layer's $CrossAttn$ correspond to different options, while other parameters are the same.

3.2.2. Concatenate model

Concatenating the context and current sentence is a native DocMT model. There are $\mathrm{N} = 4$ context-added options, shown in Table 2.

The target sentence concatenation is the same as the source. Inspired by Li et al [15], we assume that the concatenated input of different lengths corresponds to a different number of model layers. We increase the number of encoder layers by one. In the decoder, corresponding to the above four options in Table 2 in respectively: reduce two layers (exit before the last two layers), reduce one layer, reduce one layer, not reduce. Thus the number of decoder layers are 4, 5, 5, 6 when inference. All options work on one shared DocMT model.

3.3. Results

We list the results of experiments in Table 1, comparing with a standard sentence-level Transformer and six previous DocMT baselines. Our method is at the lower part.

[1]This tool can be accessed via `https://github.com/pytorch/fairseq`

src+cxt	通过演奏音乐，谈论音乐，这个人从一个偏执，不安的，刚才还在洛杉矶大街上晃悠的流浪汉，变成了一个迷人，博学，优秀的受过朱丽亚音乐学院教育的音乐家。音乐是良药，音乐改变着我们。对nathaniel来说，音乐是帮助他开启心智。
ref+cxt	And through playing music and talking about music, this man had transformed from the paranoid, disturbed man that had just come from walking the streets of downtown Los Angeles to the charming, erudite, brilliant, Juilliard-trained musician. Music is medicine, Music changes us. And for Nathaniel, music is sanity.
sys0	Music is medicine, Music changes us. For nathaniel, music is helping him turn on his mind.
sys1	By playing music and talking about music, this person went from being a paranoid, restless, tramp who was just walking on the streets of Los Angeles to a charming, knowledgeable, and outstanding musician who was educated by Juilliard-trained. Music was medicine . Music changed us . And for nathaniel, music was helping him turn on his mind.
sys2	Music is medicine, Music is changes us. And for nathaniel, music is helping him turn on the mind .

Fig. 2. Examples of translation results. sys0: sentence-level transformer. sys1: concatenate baseline. sys2: our context-adaptive DocMT. A unified grammatical tense sometimes reduces the quality of translation.

	ZH-EN	EN-DN		
parameter	TED	TED	News	Europarl
β_1	0.1	0.1	0.1	0.2
β_2	0.01	0.01	0.02	0.02
β_3	0.1	0.1	0.1	0.1

Table 4. The hyper-parameters of training loss.

	BLEU
Our predictor	26.77
w/o $\mathcal{L}_{uni}$	26.63
w/o $\mathcal{L}_{div}$	26.22
w/o Doc tips	25.79

Table 5. Ablation study on the TED dataset.

As shown in Table 1, our proposed method on the context-unit model and the concatenate model both achieved leading results over other DocMT baselines. With our predictor, the performance of two DocMT models has been significantly improved. For the concatenate model, our method receives 2.30, 1.69, 1.49, 1.15 BLEU gains over the sentence-level Transformer, receives 1.99, 1.41, 1.25, 0.94 over the concatenate baseline, on TED ZH-EN, TED EN-DE, News and Europarl datasets, respectively.

DocMT models can be trained in two stages: first, train a sentence-level base model, then finetune from the pre-trained model with document-level data. All our context-adaptive DocMT models adopt two stages training to save training time, as in previous works. Due to the N different options, similarly, the same model is trained by N times. Thus, the training time of every epoch increases, but the number of convergence rounds is reduced. As in Table3, when inference, the prediction of the context is increased, and the total decoding time increases very little. The last column of Table 1 shows that although the training time increases, the model parameters increase very little. Thus the inference speed is controllable. It indicates that the predictor is indeed light-weight.

In Table 2, there are 4 context options of concatenate model ("non" indicates empty context sentence). 14.80% of the source sentences choose not to use any context, indicating most translations need contextual information. 45.57%, less than half prefer to use the context of both the previous and next sentences. As observed from the translation results, contextual information can increase consistency, such as a unified tense, supplementary pronouns, and conjunctions. Nevertheless, as the qualitative example in appendix materials, the DocMT model prefers the previous sentence's tense. In contrast, our model selected the tense of the latter sentence instead of unified all sentences into one tense. The experimental results prove our conjecture that the DocMT model may need a more flexible context instead of a fixed static context.

3.4. Ablation Study

We conduct an ablation study by removing the components of our method from our concatenate model on the TED EN-DE dataset. As shown in Table 5, when we remove the $\mathcal{L}_{uni}$, the gain of the model dropped by 0.14. After removing the $\mathcal{L}_{div}$, the gain dropped by 0.41. Further removing minor DocNMT tips, including segment and adaptive depth, the gain dropped by 0.43. The results show that context predictor helps to improve the translation.

4. CONCLUSION

In this paper, we proposed a data-driven framework on DocMT for adaptive context. The method introduces a lightweight predictor to select the most appropriate context without increasing many parameters. Moreover, it is not limited by the specific circumstances of different contexts: empty context, source context, or target context. Experimental results show that the proposed DocMT framework can achieve significant improvements on two baseline models and various datasets.

5. REFERENCES

[1] Y. Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Q. Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, G. Kurian, Nishant Patil, W. Wang, C. Young, J. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. S. Corrado, Macduff Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *ArXiv*, vol. abs/1609.08144, 2016.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[3] Hany Hassan, Anthony Aue, C. Chen, Vishal Chowdhary, J. Clark, C. Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, W. Lewis, M. Li, Shujie Liu, T. Liu, Renqian Luo, Arul Menezes, Tao Qin, F. Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and M. Zhou, "Achieving human parity on automatic chinese to english news translation," *ArXiv*, vol. abs/1803.05567, 2018.

[4] Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson, "Document-level neural machine translation with hierarchical attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2947–2954.

[5] Elena Voita, Rico Sennrich, and Ivan Titov, "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*, 2019, pp. 1198–1212.

[6] Hongfei Xu, Deyi Xiong, Josef van Genabith, and Qiuhui Liu, "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*, 2020, pp. 3933–3940.

[7] Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong, "Dynamic context selection for document-level neural machine translation via reinforcement learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2242–2254.

[8] Sameen Maruf, André FT Martins, and Gholamreza Haffari, "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)*, 2019, pp. 3092–3102.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou, "Enhancing context modeling with a query-guided capsule network for document-level translation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 1527–1537.

[11] Shuming Ma, Dongdong Zhang, and Ming Zhou, "A simple and effective unified encoder for document-level machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3505–3511.

[12] Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch, "Towards making the most of context in neural machine translation," in *IJCAI*, 2020.

[13] Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo, "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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, Eds. 2021, pp. 3442–3455, Association for Computational Linguistics.

[14] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu, "Improving the transformer translation model with document-level context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 533–542.

[15] Xian Li, Asa Cooper Stickland, Yuqing Tang, and Xiang Kong, "Deep transformers with latent depth," *Advances in Neural Information Processing Systems*, vol. 33, 2020.