Context-Interactive Pre-Training for Document Machine Translation

Pengcheng Yang, Pei Zhang, Boxing Chen, Jun Xie, Weihua Luo
Machine Intelligence Technology Lab
Alibaba Group
Hangzhou, China
{mingyang.ypc, xiaoyi.zp, boxing.cbx, qingjing.xj, weihua.luowh}@alibaba-inc.com

Abstract

Document machine translation aims to translate the source sentence into the target language in the presence of additional contextual information. However, it typically suffers from a lack of doc-level bilingual data. To remedy this, here we propose a simple yet effective context-interactive pre-training approach, which targets benefiting from external large-scale corpora. The proposed model performs inter sentence generation to capture the cross-sentence dependency within the target document, and cross sentence translation to make better use of valuable contextual information. Comprehensive experiments illustrate that our approach can achieve state-of-the-art performance on three benchmark datasets, which significantly outperforms a variety of baselines.

1 Introduction

Document machine translation (Doc-MT) aims at utilizing the surrounding contexts of the source sentence to tackle some linguistic consistency problems (e.g., deixis, ellipsis, and lexical cohesion) in translation (Tiedemann and Scherrer, 2017). However, due to the introduction of extra contexts, it also presents several intractable challenges:

(1) *Data scarcity of document-level bilingual corpora.* Since most bilingual corpora are preserved by sentence, well-aligned document-level data is relatively scarce (Zhang et al., 2018), especially for low-resource languages or domains. Such a data sparsity not only impairs the effective training of neural machine translation (NMT) models, but also tends to result in potential overfitting.

(2) *Effective utilization of valuable information contained in extra contexts.* Although some efforts (Wang et al., 2017; Tu et al., 2018) have strived to incorporate contextual information via various architectures, they only observe minor performance gains compared with traditional sentence machine translation (Sent-MT). Recent work (Li et al., 2020) also reveals that contextual information cannot be fully leveraged by some existing approaches, where the source contexts tend to act as the data noise enriching the training signals.

(3) *Modeling of cross-sentence dependency within the target document.* Since the input of Doc-MT focuses on documents consisting of multiple sentences, the decoder should be able to deal with some discourse phenomena like coreference resolution, lexical cohesion, and lexical disambiguation. (Voita et al., 2019b). This goal requires the modeling of cross-sentence dependency within the target document.

To tackle the above three challenges, here we propose a simple yet effective context-interactive pre-training approach for Doc-MT. The proposal consists of three pre-training tasks, whose sketch is presented in Figure 1. Specifically, the cross sentence translation task (CST in Figure 1 (A)) strives to generate the target sentence in the absence of the source sentence and only based on the source contexts. With such a goal, the model is encouraged to maximize the utilization of extra contexts. To capture interactions between multiple sentences in the target document so that the discourse phenomena can be modeled, we conduct inter sentence generation (ISG in Figure 1 (B)) that aims to predict the inter sentence based on the target surrounding contexts. This task can be regarded as discourse language modeling that injects the cross-sentence dependency within the target document into the decoder of the translation model. We also introduce parallel sentence translation (PST in Figure 1 (C)) to alleviate the lack of doc-level bilingual corpora and achieve knowledge transfer from abundant sent-level parallel data to limited doc-level parallel data. In order to avoid the catastrophic forgetting of pre-trained model in downstream fine-tuning, elastic weight consolidation (EWC) regularization is introduced to further enhance the model performance.

We perform the evaluation on three benchmark

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3589–3595
June 6–11, 2021. ©2021 Association for Computational Linguistics

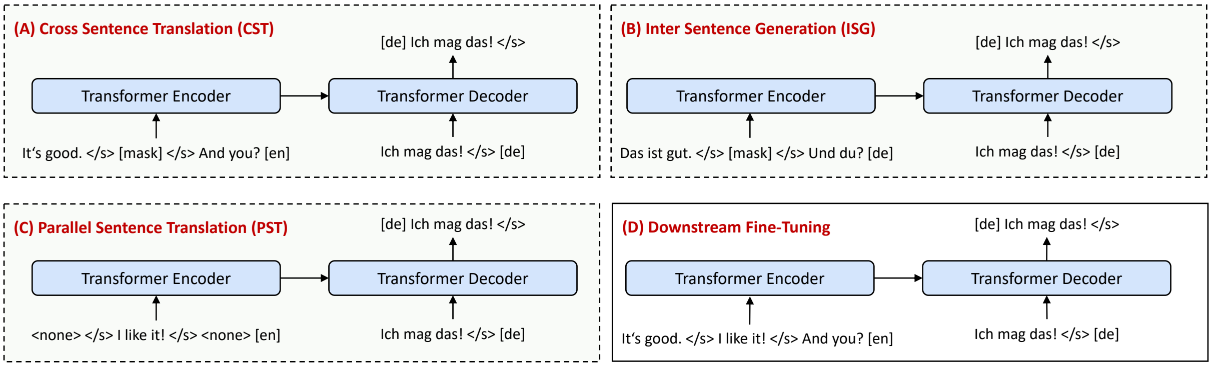

Figure 1: The sketch of our proposed context-interactive pre-training for Doc-MT. The pre-training tasks consist of: (A) CST, (B) ISG, and (C) PST. The lower-right sub-figure (D) shows the illustration of downstream fine-tuning.

datasets and results illustrate that our approach can achieve state-of-the-art performance, which is able to outperform a variety of baselines.

2 Related Work

Document machine translation (Doc-MT) aims to translate the source sentence into another different language in the presence of additional contextual information. The mainstream advances of this research field can be divided into three lines: *uni-encoder structure*, *dual-encoder structure*, and *pre-trained models*.

Uni-encoder structure. This line of research aims at performing Doc-MT based on a universal Transformer, which takes the concatenation of the additional contexts and the source sentence as the input. Tiedemann and Scherrer (2017) explores multiple different concatenation strategies and proves that the translation with extended source achieves the best performance. Bawden et al. (2018) presents several new discourse testsets, which aims to evaluate the ability of the models to exploit previous source and target sentences. Kuang et al. (2018) utilizes dynamic or topic cache to model coherence for Doc-MT by capturing contextual information either from recently translated sentences or the entire document. Going a step further, they (Kuang and Xiong, 2018) presents an inter-sentence gate model to encode two adjacent sentences and controls the amount of information flowing from the preceding sentence to the translation of the current sentence with an inter-sentence gate. Tu et al. (2018) augments translation model with a cache-like memory network that stores recent hidden representations as translation history. Yang et al. (2019) introduce a query-guided capsule networks into document-level translation to capture high-level capsules related to the current source sentence. Ma et al. (2020) proposes a unified encoder to process the concatenated source information that only attends to the source sentence at the top of encoder blocks.

Dual-encoder structure. This line of work tends to adopt two encoders or another components to model the source sentences and the document-level contexts. Wang et al. (2017) summarize the source history in a hierarchical way and then integrate the historical representation into translation model with multiple strategies. Maruf and Haffari (2018) takes both source and target document context into account using memory networks, which modeling Doc-MT as a structured prediction problem with inter-dependencies among the observed and hidden variables. Zhang et al. (2018) introduces a light context encoder to represent source context and performs information fusion with the unidirectional multi-head attention. Werlen et al. (2018) uses a hierarchical attention network (HAN) with two levels of abstraction: word level abstraction allows attention to words in previous sentences, and sentence level abstraction allows access to relevant previous sentences. Source and target context both can be exploited. Voita et al. (2019b) introduces a two-pass framework that first translates each sentence with a context-agnostic model, and then refines it using context of several previous sentences. Furthermore,Voita et al. (2019a) presents a monolingual Doc-Repair model that performs automatic post-editing on a sequence of sentence-level translations to correct inconsistencies among them. Li et al. (2020) investigates multi-encoder approaches in Doc-MT and find that the context encoder does not only encode the surrounding sentences but also behaves as a noise generator. Maruf et al. (2019) presents a hierarchical context-aware translation model, which selectively focus on relevant sentences in the document context and then attends to key words in those sentences.

3 Methodology

Following prior work (Ma et al., 2020), we translate the i-th source sentence $\boldsymbol{x}_i$ into the i-th target sentence $\boldsymbol{y}_i$ in the presence of extra source contexts $\boldsymbol{c} = (\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1})$, where $\boldsymbol{x}_{i-1}$ and $\boldsymbol{x}_{i+1}$ refer to the predecessor and successor of $\boldsymbol{x}_i$ respectively. We adopt Transformer as the model architecture of pre-training and machine translation. The model is trained by minimizing the negative log-likelihood of target sequence $\mathbf{y}$ conditioned on the source sequence $\mathbf{x}$, i.e., $\mathcal{L} = -\mathrm{log}p(\mathbf{y}|\mathbf{x})$. Readers can refer to Vaswani et al. (2017) for more details. We introduce our approach based on EN→DE Doc-MT.

3.1 Pre-Training Tasks

Figure 1 shows the sketch of our context-interactive pre-training approach, elaborated on as follows.

Cross Sentence Translation (CST) When translating the i-th sentence $\boldsymbol{x}_i$ and the source context $\boldsymbol{c} = (\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1})$ into the i-th target sentence $\boldsymbol{y}_i$, prior approaches tend to pay most attention on $\boldsymbol{x}_i$ (Li et al., 2019), resulting in the neglect of $\boldsymbol{c}$. To maximize the use of the source context $\boldsymbol{c}$, we propose cross sentence translation (CST) to encourage the model to more effectively utilize the valuable information contained in $\boldsymbol{c}$. We mask the whole source sentence $\boldsymbol{x}_i$ in the model input, and enforce the model to generate the target sentence $\boldsymbol{y}_i$ only based on $\boldsymbol{c} = (\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1})$. To be specific, we pack both the source context $\boldsymbol{c}$ and the mask token `[mask]` as a continue span, and employ a special token `</s>` to indict the end of each sentence. To distinguish texts from different languages, we add language identifier (e.g., `<en>` for English and `<de>` for German) to the ends of both the source input and target output. Figure 1(A) presents the illustration of this task on EN-DE translation, where the input of Transformer is the concatenation of $(\boldsymbol{x}_{i-1}, \texttt{<mask>}, \boldsymbol{x}_{i+1})$ and the target output is $\boldsymbol{y}_i$.

Inter Sentence Generation (ISG) Voita et al. (2019b) has demonstrated that the cross-sentence dependency within the target document can effectively improve the translation quality. Transformer decoder should be able to model the corresponding historical information to improve coherence or lexical cohesion and other aspects during translation. Motivated by this, here we propose inter sentence generation (ISG) to capture the cross-sentence dependency among the target output. The ISG task aims to predict the inter sentence $\boldsymbol{y}_i$ based on its surrounding predecessor $\boldsymbol{y}_{i-1}$ and successor $\boldsymbol{y}_{i+1}$. In this way, the model is trained to capture the interactions between the sentences in the target document. Besides, the training of ISG only requires the monolingual document corpora of the target language, which effectively alleviates the lack of doc-level parallel data in Doc-MT. Figure 1(B) presents the detailed illustration, where the model input is the concatenation of $(\boldsymbol{y}_{i-1}, \texttt{<mask>}, \boldsymbol{y}_{i+1})$ and the target output is $\boldsymbol{y}_i$. Both source and target language identifiers are `<de>`.

Parallel Sentence Translation (PST) In practice, the available sent-level parallel corpora usually present larger scale than doc-level parallel corpora. Thus, here we introduce parallel sentence translation (PST) performing context-agnostic sentence translation, which only requires sent-level parallel data. This further alleviates the lack of the doc-level parallel data in Doc-MT. The illustration of PST is presented in Figure 1(C), where the input is the concatenation of $(\texttt{<none>}, \boldsymbol{x}_i, \texttt{<none>})$ and the target output is $\boldsymbol{y}_i$.[1] The source and target language identifiers are `<en>` and `<de>`, respectively.

EWC-Based Fine-Tuning. After finishing the pre-training, the pre-trained transformer is used as the model initialization for subsequent fine-tuning on downstream datasets. As shown in Figure 1, the input of Transformer in this scenario is $(\boldsymbol{x}_{i-1}, \boldsymbol{x}_i, \boldsymbol{x}_{i+1})$, i.e., the concatenation of the i-th source sentence $\boldsymbol{x}_i$ and its surrounding context $\boldsymbol{c} = (\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1})$. The desired output is the i-th target sentence $\boldsymbol{y}_i$. The source and target language identifiers are same as PST. However, obvious catastrophic forgetting has been observed during fine-tuning. As fine-tuning continues, the model performance exhibits degradation. Due to large-scale model capacity and limited downstream datasets, pre-trained models usually suffer from overfitting. To remedy this, here we introduce Elastic Weight Consolidation (EWC) regularization (Kirkpatrick et al., 2016). EWC regularizes

[1] We use `<none>` to represent the unavailable content.

Task	Input	Output	SLI	TLI	Use Mono-Doc	Use Bi-Doc	Use Bi-Sent
CST	$(\boldsymbol{x}_{i-1}, \texttt{<mask>}, \boldsymbol{x}_{i+1})$	$\boldsymbol{y}_i$	<en>	<de>		✓	
ISG	$(\boldsymbol{y}_{i-1}, \texttt{<mask>}, \boldsymbol{y}_{i+1})$	$\boldsymbol{y}_i$	<de>	<de>	✓	✓	
PST	$(\texttt{<none>}, \boldsymbol{x}_i, \texttt{<none>})$	$\boldsymbol{y}_i$	<en>	<de>		✓	✓
Fine-tune	$(\boldsymbol{x}_{i-1}, \boldsymbol{x}_i, \boldsymbol{x}_{i+1})$	$\boldsymbol{y}_i$	<en>	<de>		✓	

Table 1: The detailed illustration of different tasks. "SLI" and "TLI"denotes the source and target language identifier, respectively. "Use Mono-Doc", "Use Bi-Doc" and "Use Bi-Sent" means that the corresponding task can use monolingual doc-level, bilingual doc-level, bilingual sent-level corpora, respectively.

the weights individually based on their importance to that task, which forces the model to remember the original language modeling tasks. Formally, the EWC regularization is computed as:

$$\mathcal{R} = \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta_i^*)^2 \qquad (1)$$

where λ is a hyperparameter weighting the importance of old LM tasks compared to new MT task, and i labels each parameter. The final loss $\mathcal{J}$ for fine-tuning is the sum of negative log-likelihood in all pre-training tasks and newly introduced $\mathcal{R}$, i.e., $\mathcal{J} = \mathcal{L}_{\text{CST}} + \mathcal{L}_{\text{ISG}} + \mathcal{L}_{\text{PST}} + \mathcal{R}$.

We summarize the key information of our approach in Table 1, which also shows the available data of different tasks.

4 Experiments

4.1 Settings

We train Transformer consisting of 12 encoder and 12 decoder layers with 1024 hidden size on 16 heads. We adopt the public mBART.CC25 released by Liu et al. (2020) as the initialization. For CST task, the pre-training data consists of: *TED*, *Europarl*, *News Commentary* and *Rapid* corpus. The monolingual target documents used in ISG task are extracted from Wikipedia. For PST task, we sample bilingual sentences in NewsCrawl utill 2018. We use sentence piece model (Kudo and Richardson, 2018) to tokenize all data. Gradient accumulation is used to simulate the batch size of 128K tokens. We use Adam optimizer with linear learning rate decay. The learning rate and dropout is set to $3e-5$ and 0.1, respectively. We set λ in Eq. 1 to 0.01. We evaluate on three EN-DE Doc-MT datasets provided by Maruf et al. (2019): *TED*, *News*, and *Europarl* and perform limited grid-search of hyperparameter.

4.2 Baselines

Unpretrained models. Transformer (Vaswani et al., 2017) performs context-agnostic sent-level translation and HAN (Werlen et al., 2018) employs hierarchical attention to capture extra contexts. SAN (Maruf et al., 2019) utilizes top-down attention to selectively focus on relevant sentences and QCN (Yang et al., 2019) uses query-guided capsule networks to capture the related capsulese.

Pretrained models. Flat-Transformer (Ma et al., 2020) apply BERT as the initialization of encoder. We also implement the parallel sentence translation-based pre-training with mBART (Liu et al., 2020) initialization as the most comparable baseline.

To have a fair comparison, we adopt multi-BLEU as the evaluation metric. We first conduct SPM-based detoken on the generated texts and then use Moses to re-tokenize all texts like the baselines.

4.3 Main Results

Table 2 shows the performance of different systems. Results first confirm that large-scale pre-training can effectively accomplish model transferring and advance the performance of Doc-MT. Besides, we can observe significant performance gain for our approach compared to the baselines. For instance, it surpasses the mBART initialized model with PST by 0.72 BLEU. With the proposed pre-training tasks, our approach succeeds in acquiring more effective knowledge from external large-scale corpora, leading to better translation quality.

4.4 Incremental Analysis

Here we perform further incremental analysis. We treat Transformer with mBART initialization as the base model and cumulatively add each pre-training task until the full approach is rebuilt. The results are shown in Table 3. We can observe that the removal of the parallel sentence translation (PST) task results in the largest performance degradation. First, the scale of parallel sentences used for PST far exceeds that for the other two tasks, bringing the significant performance gains; In addition, PST closely resembles the downstream Doc-MT task,

Model	TED	News	Eurporal	Avg
Transformer (Vaswani et al., 2017)	23.28	22.78	28.72	24.93
Doc-Transformer (Zhang et al., 2018)	24.01	22.42	29.93	25.45
HAN (Werlen et al., 2018)	24.58	25.03	29.58	26.40
SAN (Maruf et al., 2019)	24.62	24.84	29.90	26.45
QCN (Yang et al., 2019)	25.19	22.37	29.82	25.79
Flat-Transformer (Ma et al., 2020)	26.61	24.52	31.99	27.71
mBART+PST	27.23	27.18	32.04	28.82
Context-interative pre-training (Ours)	**27.84**	**27.93**	**32.85**	**29.54**

Table 2: The results of different systems. “Avg” denotes the average BLEU score on all datasets.

CST	ISG	PST	BLEU
		✓	27.18
	✓	✓	27.74
✓	✓		26.82
✓	✓	✓	**27.93**

Table 3: The results of incremental analysis on *News* dataset. “✓” represents that the corresponding pre-training task is adopted.

encouraging more effective knowledge transfer. Besides, Table 3 also reveals that other CST and ISG tasks play an active role in improving translation quality. By masking the whole source sentence in the input via CST, the model is encouraged to more effectively extract and utilize valuable information from extra contexts. With the target doc-level language modeling, the cross-sentence dependency within the document is better captured. Both contributes to improving the quality of Doc-MT.

4.5 Effectiveness of EWC Regularization

To avoid the catastrophic forgetting of pre-trained models in downstream fine-tuning, we introduce EWC regularization to force the model to remember the original language modeling task. Table 4 presents the comparison of our approach with or without EWC regularization, demonstrating its effectiveness in improving model performance. Results show that EWC regularization can achieve consistent improvements on various datasets, increasing the average BLEU score from 29.24 to 29.54. By weighing the original LM task and newly introduced NMT task based on the importance of parameters, the overfitting of the pre-trained model on the limited downstream data is effectively alleviated, bringing consistent performance gains.

Model	TED	News	Eurporal	Avg
w/o EWC	27.46	27.78	32.49	29.24
w/ EWC	27.84	27.93	32.85	29.54

Table 4: The comparison of our approach with or without elastic weight consolidation (EWC) regularization.

5 Conclusion

This work presents context-interactive pre-training to benefit document machine translation from external large-scale mono or bi-lingual corpora. The proposed approach strives to capture the cross-sentence dependency within the target document via inter sentence generation, and utilize valuable information contained in the source context via cross sentence translation. Extensive experiments illustrate that our approach can consistently outperform extensive baselines, achieving state-of-the-art performance on various benchmark datasets.

References

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Shaohui Kuang and Deyi Xiong. 2018. Fusing recency into neural machine translation with an inter-sentence gate model. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 607–617. 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, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 596–606. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? A case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3512–3518. Association for Computational Linguistics.

Liangyou Li, Xin Jiang, and Qun Liu. 2019. Pretrained language models for document-level neural machine translation. *CoRR*, abs/1911.03110.

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. *CoRR*, abs/2001.08210.

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, ACL 2020, Online, July 5-10, 2020*, pages 3505–3511. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1275–1284. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3092–3102. 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.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 82–92. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Trans. Assoc. Comput. Linguistics*, 6:407–420.

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, 4-9 December 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 2019, Hong Kong, China, November 3-7, 2019*, pages 877–886. 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1198–1212. 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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2826–2831. Association for Computational Linguistics.

Lesly Miculicich Werlen, 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2947–2954. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9386–9393. AAAI Press.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 1527–1537. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 533–542. Association for Computational Linguistics.