Dynamic-FACT: A Dynamic Framework for Adaptive Context-Aware Translation

Linqing Chen, Weilei Wang*
(PatSnap Co., LTD. Suzhou, Jiangsu 215000)
{chenlinqing, wangweilei}@patsnap.com

Abstract

Document-level neural machine translation (NMT) has garnered considerable attention since the emergence of various context-aware NMT models. However, these static NMT models are trained on fixed parallel datasets, thus lacking awareness of the target document during inference. In order to alleviate this limitation, we propose a dynamic adapter-translator framework for context-aware NMT, which adapts the trained NMT model to the input document prior to translation. Specifically, the document adapter reconstructs the scrambled portion of the original document from a deliberately corrupted version, thereby reducing the performance disparity between training and inference. To achieve this, we employ an adaptation process in both the training and inference stages. Our experimental results on document-level translation benchmarks demonstrate significant enhancements in translation performance, underscoring the necessity of dynamic adaptation for context-aware translation and the efficacy of our methodologies.

1 Introduction

Numerous recent studies have introduced a variety of context-aware models aiming to effectively harness document-level context either from the source side (Maruf and Haffari, 2018; Zhang et al., 2018; Miculicich et al., 2018; Tan et al., 2019; Zheng et al., 2020; Kang et al., 2020), target side (Xiong et al., 2019; Yu et al., 2020; Sugiyama and Yoshinaga, 2021), or both (Kuang et al., 2018; Tu et al., 2018; Maruf et al., 2019; Chen et al., 2020; Chen et al., 2022). In the prevailing practice, a context-aware model remains fixed after training and is then employed for every testing document. Nonetheless, this approach presents a potential challenge, as the model is required to encapsulate all translation knowledge, particularly from diverse domains, within a predefined set of parameters. Accomplishing this task within the confines of reality poses a formidable undertaking.

The "one sentence one model" approach for sentence-level NMT, as proposed by (Li et al., 2018), aims to familiarize the model with each sentence in the test dataset by fine-tuning the NMT model for every testing sentence. However, acquiring suitable fine-tuning sentences for a given testing sentence proves to be highly time-consuming, as they require meticulous extraction from the bilingual training data through similarity search. This presents a significant challenge when attempting to replicate their methodology by seeking similar documents from the bilingual document-level training data. Moreover, this approach assesses sentence similarity solely based on the Levenshtein distance, thereby disregarding the document-level context of these sentences extracted from distinct documents.

To address the potential challenge of employing a fixed, trained model for all testing documents, we propose the "one document one model" approach in this paper. This alternative approach aims to achieve the objective by introducing the *document adapter*. Unlike other methods, the adapter relies solely on the input document itself and does not require additional input forms. Its primary function is to reconstruct the original document from a deliberately corrupted version, thereby enabling the model to familiarize itself with the task of document-level translation. Notably, this approach differs from previous methods where the input and output are in different languages, as opposed to the same language. Following adaptation, this modified model is utilized to translate the document.

*Corresponding author

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

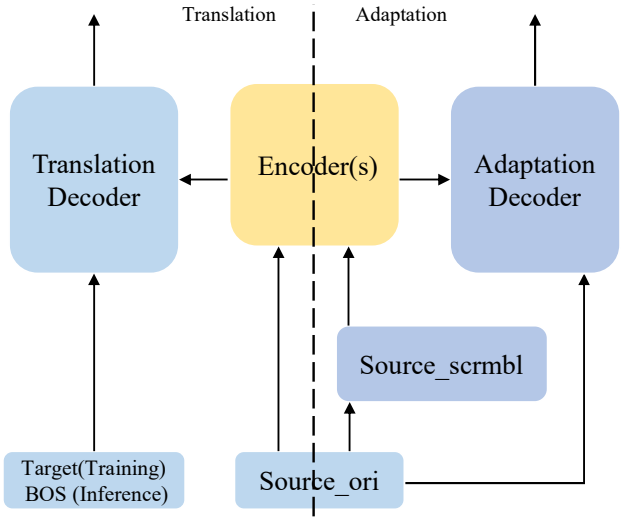

Figure 1: The figure presented in this section depicts the adapter-translator architecture designed for context-aware neural machine translation. In this architecture, the encoder(s) are shared between the adapter denoted as ϕ and the translator denoted as ψ. It is important to note that the translator and adapter constitute two distinct stages within the same model, rather than being treated as separate models.

Both the adapter model and the NMT model employed in our study are context-aware and utilize shared encoder(s), while each having its dedicated decoder. In this paper, we present a training methodology that aims to adapt a pre-trained NMT model to a specific document through a process of alternating document reconstruction and document translation for each document batch. This approach is employed during both the training and inference stages. To evaluate the effectiveness of our proposed approach, we conducted experiments on three English-to-German document-level translation tasks. The results reveal significant enhancements in translation performance, providing strong evidence for the necessity of employing a one document one model approach and the efficacy of our proposed methodology.

Overall, we make the following contributions.

- We present an enhanced context-aware document-level auto-encoder task to facilitate dynamic adaptation of translation models.
- We propose an adapter-translator framework for context-aware NMT. To the best of our knowledge, this is the first study that investigates the one-document-one-model approach specifically for document-level NMT.

2 Adapter-Translator Architecture

The Adapter-Translator architecture entails an iterative procedure involving an adaptation process denoted as ϕ and a translation process denoted as ψ. Figure 1 presents a visual representation of the proposed architecture. The translator ψ, which is a context-aware NMT model, comprises context-aware encoder(s) and a decoder specific to translation.[1] The adapter shares the encoder(s) with the translator while possessing a decoder specifically designed for adaptation. Given a source document $\mathcal{X}$, the corpus processing script generates a deliberately corrupted version $\widehat{\mathcal{X}}$ of the document. This corrupted version is then utilized to optimize the adapter in order to reconstruct the scrambled segments of the original document $\mathcal{X}$. As the encoder(s) are shared between the adapter and the translator, the capability to capture context during document adaptation can also be harnessed during the document translation process. The translation component of this architecture resembles that of other document-level translation models.

[1]It is worth noting that while not all context-aware NMT models possess an additional context encoder (Ma et al., 2020), the adapter-translator architecture can still be adapted to accommodate these models.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

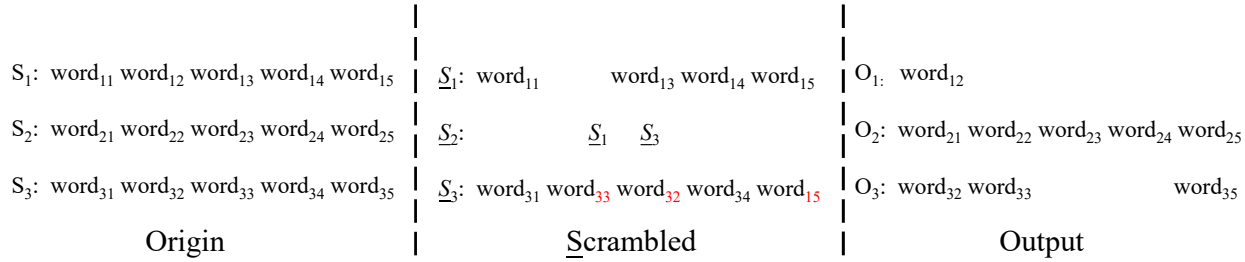

Figure 2: Illustration of document reconstruction task.

Due to its straightforward yet impactful architecture, the proposed method can be employed with diverse document-level translation models.

2.1 Document Adapter

Motivated by the work of (He et al., 2022), we present an adapter-based methodology to restore the scrambled segments of an input document. To be more precise, we adopt a strategy where sentences or words are randomly omitted from the original document, and the adapter is trained to reconstruct these scrambled portions by minimizing the cross-entropy reconstruction loss between the output of its decoder and the corresponding correct part of the original document.

Given a document $\mathcal{X} = (X_i)\,|_{i=1}^{N}$ consisting of N sentences, we apply token substitution, insertion, and deletion operations to each sentence X_i. Following the approach of BERT (Devlin et al., 2019), we randomly select 15% of the tokens. However, unlike BERT, we do not replace these tokens with [MASK] tokens. Instead, the adapter is responsible for identifying the positions that require correct inputs. Furthermore, we do not preserve 10% of the selected tokens unchanged, as our method does not rely on the [MASK] token. In our experiments, we observed that compared to generating the entire original document, generating only the corrected scrambled part significantly reduced the computational time. Nevertheless, this modification did not significantly compromise the model's ability to capture context and become familiar with the document to be translated.

Figure 2 depicts an example involving 3 sentences in the original document. In this example, the first sentence undergoes a word deletion operation, while the third sentence experiences word scrambling and replacement operations. The scrambled preceding and succeeding sentences serve as context for the second sentence. The adapter produces the missing words in the first sentence, the corrected words in the third sentence, and the complete second sentence. While our document reconstruction task draws inspiration from the similar proposal in (Devlin et al., 2019), there exist two significant distinctions. Firstly, instead of substituting selected words with `[MASK]` tokens, we introduce contextual document corruption by allowing token substitution, insertion, and deletion. Secondly, in contrast to BERT, our training objective simultaneously considers the utilization of both sentence-level and document-level context.

To summarize, we define the document adaptation task by employing the following two sub-tasks:

- Sentence-level: The adapter generates corrected words based on a deliberately scrambled version of the original sentence.
- Document-level: The adapter utilizes the concatenated context sentences to generate the original sentences.

2.2 Context-Aware Translator

The context-aware translation model in our framework is designed as a relatively independent model, which shares the encoder(s) with the adaptation model while having a dedicated decoder for translation. This design ensures the flexibility of the framework, allowing it to be easily integrated with different translation models by simply incorporating the adapter model's encoder.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

#	Model	TED		News		Europarl		Average	
		BLEU	Meteor	BLEU	Meteor	BLEU	Meteor	BLEU	Meteor
1	DocT (Zhang et al., 2018)	24.00	44.69	23.08	42.40	29.32	46.72	25.47	44.60
2	+ Adapter	24.70	45.20	23.68	43.01	29.84	47.15	26.07	45.12
3	HAN (Miculicich et al., 2018)	24.58	45.48	25.03	44.02	28.60	46.09	26.07	45.20
4	+ Adapter	24.90	45.89	25.51	44.38	29.07	46.61	26.49	45.63
5	SAN (Maruf et al., 2019)	24.42	45.26	24.84	44.17	29.75	47.22	26.34	45.55
6	+ Adapter	24.80	45.69	25.24	44.63	30.11	48.20	26.72	46.17
7	QCN (Yang et al., 2020)	25.19	46.09	22.37	41.88	29.82	47.86	25.79	45.28
8	+ Adapter	25.83	46.80	22.89	42.40	30.32	48.35	26.35	45.85
9	GCNMT (Chen et al., 2022)	25.81	46.33	25.32	44.35	29.80	47.77	26.98	46.15
10	+ Adapter	26.50	46.96	25.71	44.83	30.43	48.46	27.55	46.75
11	Transformer (Vaswani et al., 2017)	23.02	43.66	22.03	41.37	28.65	45.83	24.57	43.62

Table 1: Performance on test sets. + Adapter indicates we use our proposed context-aware adapter to guidance the context-aware encoder. Significance test (Koehn, 2004) shows that the improvement achieved by our approach is significant at 0.05 on almost all of the above models.

From a structural perspective, this approach facilitates the applicability of the framework to a wide range of translation models. However, in terms of translation performance, there are significant differences between the output of the adaptation phase and the translation phase. Sharing the decoder between these two phases may introduce bias towards shorter output text during translation, given the relatively short length of the corrected scrambled part produced in the adaptation phase. Furthermore, sharing the decoders may increase the vocabulary size of the translation model decoding end and the dimension of the vector, thereby increase the computational cost of training and inference. Additionally, changes in the decoder's vocabulary may alter the semantic space of the translation model, necessitating retraining even if a well-trained translation model is available.

As discussed earlier, the adapter model's decoder only generates the corrected part of the original document. Therefore, employing two different decoders does not significantly impact the time required during the translation inference phase.

2.3 Training and Inference

During the model training phase, the framework follows different procedures based on whether it is built upon a pre-trained translation model or trained from scratch. When using a pre-trained model, the parameters of the translator are frozen, and only the decoder part of the adapter is trained. In the case of training from scratch, parallel corpora are employed as input and output for the translator, while the source corpus and its scrambled versions are used as input and output for the adapter. Training is performed iteratively, alternating between the translation and reconstruction tasks.

In the framework's inference phase, the decoders of both the translator and adapter are frozen for two primary reasons. Firstly, these decoders have undergone sufficient training during the training phase. Secondly, freezing them saves computational time during inference. Similarly, a certain percentage (P%) of the context-aware encoder parameters are also frozen for similar reasons. This not only reduces computational overhead but also facilitates multi-round learning by utilizing multiple scrambled versions of the same document, enabling the translation model to become familiar with the document to be translated. By freezing most of the encoder parameters and increasing the dropout rate, overfitting on a single document is mitigated, preventing potential performance degradation on other documents in the test set.

Specifically, the document restoration process consists of the following steps:

1. Expansion of $\mathcal{X}$: We expand the original document $\mathcal{X}$ by creating K copies, where K is the expansion ratio. Each copy is processed independently, forming instances for the document restoration task.

2. Freezing of Translator and Adapter Parameters: We freeze a portion of the parameters in both the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

#	Model	MT06	MT02	MT03	MT04	MT05	MT08	All		
		BLEU	BLEU	BLEU	BLEU	BLEU	BLEU	BLEU	Meteor	d-BLEU
1	DocT (Zhang et al., 2018)	37.08	43.40	43.83	41.51	41.79	32.47	40.35	27.45	42.91
2	+ Adapter	38.65	44.57	44.17	42.80	43.19	33.75	41.52	28.66	44.07
3	HAN (Miculicich et al., 2018)	37.20	42.96	44.53	41.89	42.31	32.57	40.83	28.00	43.28
4	+ Adapter	38.11	43.62	45.99	43.51	43.03	33.91	42.47	29.49	45.10
5	SAN (Maruf et al., 2019)	37.40	43.28	44.82	41.99	42.60	32.46	41.01	28.19	43.54
6	+ Adapter	**39.62**	**45.37**	46.72	**43.91**	43.59	**34.48**	**42.93**	**30.01**	**45.38**
7	GCNMT (Chen et al., 2022)	38.39	44.33	46.43	42.92	43.60	33.41	41.51	28.73	44.08
8	+ Adapter	39.51	45.28	**47.26**	43.70	**44.56**	34.27	42.43	29.50	44.96
9	Transformer (Vaswani et al., 2017)	36.27	42.71	43.51	41.25	41.07	31.54	39.64	26.70	42.16

Table 2: Performance on ZH-EN test sets with and without the context-aware adapter is presented in this Table. The "+Adapter" indicates that our proposed context-aware adapter was used to guide the context-aware encoder. Significance testing (Koehn, 2004) demonstrates that the improvements achieved by our approach are statistically significant at the 0.05 level for almost all of the aforementioned models.

translator and adapter. The dropout rate is set to 0.2, while P% (the percentage of frozen context-aware encoder parameters) is set to 99%.

3. Training the Context-Aware Model: We utilize the corrupted instances to train the context-aware model, which follows the adapter-translator architecture, with the aim of familiarizing it with the document. This involves updating part of the parameters in the context-aware encoder(s). During adaptation, the learning rate is set to 0.1.

4. Document Translation: We employ the adapted model to translate the original document $\mathcal{X}$. This entails utilizing the updated parameters in the context encoder and the sentence encoder to encode the source sentences, and employing the translator decoder to decode the target sentences.

3 Application to various Document-level NMT Model

To evaluate the effectiveness of our proposed framework in context-aware NMT, we select the following five representative NMT models:

- DocT (Zhang et al., 2018): This model considers two previous sentences as context. It employs a document-aware transformer that incorporates context representations into both the sentence encoder and decoder.

- HAN (Miculicich et al., 2018): HAN leverages all previous source and target sentences as context and introduces a hierarchical attention network to capture structured and dynamic context. The context representations are then fed into the decoder.

- SAN (Maruf et al., 2019): SAN extends the context coverage to the entire document. It adopts sparse attention to selectively attend to relevant sentences and focuses on key words within those sentences.

- MCN (Zheng et al., 2020): MCN employs an encoder to generate local and global contexts from the entire document, enabling the model to understand inter-sentential dependencies and maximize the utilization of contextual information.

- GCNMT (Chen et al., 2022): GCNMT comprises a global context encoder, a sentence encoder, and a sentence decoder. It incorporates two types of global context to enhance translation performance.

All of these models utilize a context encoder to encode global or local contexts, thereby improving document-level translation performance. To apply our proposed adapter-translator architecture to these models, we introduce an adapter decoder.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Set	TED		News	
	#SubDoc	#Sent	#SubDoc	#Sent
Training	7,491	206,126	10,552	236,287
Dev	326	8,967	112	2,169
Test	87	2,271	184	2,999
Set	**Europarl**			
	#SubDoc	#Sent		
Training	132,721	1,666,904		
Dev	273	3,587		
Test	415	5,134		

Table 3: Statistics of the training, development, and test sets of the three translation tasks.

K	**BLEU**	**Meteor**
0	26.98	46.15
1	27.33	46.50
5	27.55	46.75
10	27.41	46.50
15	27.13	46.32

Table 4: Averaged performance with respect to different data expansion ratio in inferring stage.

4 Experimentation

4.1 Settings

Datasets and Evaluation Metrics. We conduct experiments on English-to-German (EN→DE) translation tasks in three different domains: talks, news, and speeches. Additionally, we evaluate our proposed framework for the Chinese-to-English translation task.

- TED: This dataset is obtained from the IWSLT 2017 MT track (Cettolo et al., 2012). We combine test2016 and test2017 as our test set, while the remaining data is used as the development set.
- News: This dataset is derived from the News Commentary v11 corpus. We use news-test2015 and news-test2016 as the development set and test set, respectively.
- Europarl: This dataset is extracted from the Europarl v7 corpus. We randomly split the corpus to obtain the training, development, and test sets.
- For ZH-EN: The training set consists of 41K documents with 780K sentence pairs.[2] We use the NIST MT 2006 dataset as the development set and the NIST MT 02, 03, 04, 05, and 08 datasets as the test sets. The Chinese sentences are segmented using Jieba, while the English sentences are tokenized and converted to lowercase using Moses scripts.

We obtained the three document-level translation datasets from (Maruf et al., 2019). [3] For the source-side English sentences, we segmented them using the corresponding BPE model trained on the training data. Meanwhile, for the target-side German sentences, we used the BPE model with 25K operations trained on the corresponding target-side data. Table 3 provides a summary of the statistics for the three translation tasks. It should be noted that we divided long documents into sub-documents containing at most 30 sentences to enable efficient training.

Model Settings. For all translation models, we have set the hidden size to 512 and the filter size to 2048. The number of heads in the multi-head attention mechanism is 8, and the dropout rate is 0.1. During the training phase, we train the models for 100K steps using four A100 GPUs, with a batch size of 40960 tokens. We employ the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a learning rate of 1, incorporating a warm-up step of 16K. As for the fine-tuning stage, we fine-tune

[2]It consists of LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, LDC2010T03.
[3]https://github.com/sameenmaruf/selective-attn/tree/master/data

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

the models for 40K steps on a single A100 GPU, with a batch size of 40960 tokens, a learning rate of 0.3, and a warm-up step of 4K. During the inference phase, we set the beam size to 5.

4.2 Experimental Results

We utilize two evaluation metrics, BLEU (Papineni et al., 2002) and Meteor (Lavie and Agarwal, 2007), to assess the quality of translation. The results, presented in Table 1, demonstrate that our proposed approach consistently achieves state-of-the-art performance, outperforming previous context-aware NMT models on average. We observe significant improvements across all datasets by adapting the NMT model to the characteristics of each input document. Of particular note is the comparison between models #9 and #10, where our approach demonstrates a notable improvement with a gain of +0.57 in BLEU and +0.60 in Meteor.

Table 2 showcases the performance results for Chinese-English translation. The table presents the BLEU scores for each sub-test set and the average Meteor score across all sets. The results demonstrate that our proposed adapter-translator framework consistently achieves state-of-the-art performance when compared to the original versions of previous context-aware NMT models. Moreover, we consistently observed improvements across all datasets by adapting the trained NMT model to fit each input document. For instance, comparing models #8 and #7, our approach achieves an improvement with a gain of +0.92 in BLEU, +0.77 in Meteor, and +0.88 in d-BLEU.

Effect of Hyper-Parameter K in Dynamic Translation In the inference stage, the expansion ratio is an important hyperparameter for dynamic translation. A low ratio may restrict the effectiveness of adaptation in parameter optimization, whereas a high ratio may lead to overfitting of the model to the document restoration task. As indicated in Table 4, we observe that the optimal performance is attained with a ratio of 5 for the EN-DE translation task using the GCNMT model.

5 Analysis and Discussion

In this section, we employ the Chinese-to-English translation task as a representative to offer additional evidence for the efficacy of our proposed framework. In addition to reporting s-BLEU, we also present case-insensitive document-level BLEU (d-BLEU) scores.

5.1 Effect of Adapting Task

In a previous study (Li et al., 2020), it was suggested that context encoders not only utilize context to guide models but also encode noise. Therefore, the improvement in translation quality can sometimes be attributed to enhanced model robustness. The authors discovered that two context-aware models exhibited superior performance during inference even when the context input was replaced with noise. To ascertain whether our framework genuinely benefits from the document adaptation task, we compare the experimental results with and without an adapter in a Chinese-to-English translation task.

We conducted an investigation on the impact of adapting a document prior to translation. We define **Fake adapting** as the process wherein nonsensical words are employed as the target output during the model's adaptation phase, and **Noisy adapting** as the process wherein the model employs shuffled noisy sentences as input and corrects portions of these sentences as output. The results in Table 5 demonstrate that our proposed framework achieves improvements of +3.12 and +1.67 compared to Fake adapting and Noisy adapting, respectively. Furthermore, a notable performance disparity is observed between the results of Fake adapting and Noisy adapting. The adapter that employs shuffled documents as input achieves a gain of +1.45 compared to Fake adapting, indicating that document adaptation indeed has a positive effect on the translation model.

5.2 Architecture of the Adapter

As elaborated in Section 2 on the **Adapter-Translator Architecture**, our proposed framework employs shared encoder(s) for both the adaptation process and translation process. It is worth noting that some previous context-aware models have utilized multiple encoders. To determine whether this architecture

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Context	s-BLEU	d-BLEU	Meteor
HAN (Miculicich et al., 2018)	40.83	43.28	28.00
Fake adapting	39.35	42.00	26.29
Noisy adapting	40.80	43.55	28.33
ours	**42.47**	**45.10**	**29.49**

Table 5: Performance on ZH-EN test sets of effectiveness of adapting process.

Share	s-BLEU	d-BLEU	Meteor
Sentence encoder	41.59	44.19	28.40
Context encoder	41.55	44.13	28.42
Both	**42.47**	**45.10**	**29.49**

Table 6: Performance on ZH-EN test sets of sharing the sentence encoder, the context encoder, or both.

is the optimal choice for our research objectives, we investigated the impact of the adapter architecture on the translation model's performance.

In our framework, the encoder(s) are shared between the adapter and translator; however, the effectiveness of each encoder remains uncertain. To explore this, we conducted experiments and present the results in Table 6. The table demonstrates that sharing either the sentence encoder, the context decoder, or both leads to significant improvements in translation performance. These findings align with our intuition, and we observe that sharing both encoders yields the best performance, as indicated in the first row of the table. A possible explanation for these results is that sharing both encoders maximizes the preservation and exchange of information acquired during the reconstruction process in the adaptation phase, specifically concerning the test document.

5.3 Designing of Adapting Task

Masked sentence auto-encoding tasks have been extensively utilized in natural language processing and have consistently shown their effectiveness and generalizability in numerous previous studies. In Table 7, we present the performance of various document adaptation tasks on the Chinese-to-English translation task. Interestingly, we observe a decline in performance when using the translation process itself as a document adaptation task, which aligns with findings from prior research on double-translation. Similarly, the experiment employing the reconstruction of typical masked sentences as an adaptation task also exhibited a similar phenomenon. These findings indicate that our proposed approach effectively assists translation models in capturing valuable information from documents.

5.4 Pronoun Translation

To evaluate coreference and anaphora, we adopt the reference-based metric proposed by Werlen and Belis (2017), following the methodology of Miculicich et al.(2018) and Tan et al.(2019). This metric measures the accuracy of pronoun translation. Table 8 displays the performance results. We observe that our proposed approach significantly improves the translation of pronouns, indicating that pronoun translation benefits from leveraging global context. This finding is consistent with the results reported in related studies (Werlen and Popescu-Belis, 2017; Miculicich et al., 2018; Tan et al., 2019).

5.5 Adapting with Human Feedback

Adapting with human feedback has been widely employed in various natural language models, and its effectiveness and generalization have been demonstrated in numerous prior studies. We sought to investigate whether human feedback could enhance our translator-adapter framework.

Table 9 presents the performance of the adapting task augmented with human feedback on the Chinese-to-English translation task. The term "**Fake feedback**" refers to using the adapter's outputs as simulated human feedback, while "**Real feedback**" denotes the process of reviewing and correcting the adapter's outputs, and using the corrected sequences as target sentences. From the results, we observe that using

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Task	s-BLEU	d-BLEU	Meteor
Translation	41.30	44.00	28.01
Masked sentences	41.98	44.60	28.33
Ours	**42.47**	**45.10**	**29.49**

Table 7: Performance of different document adapting task on ZH-EN translation task.

Model	Pronoun
Transformer	68.68
GCNMT (Chen et al., 2022)	68.77
+ adapter	68.95
SAN (Maruf et al., 2019)	69.37
+ adapter	**69.84**

Table 8: Evaluation on pronoun translations of ZH-EN.

the adapter's output as simulated human feedback leads to a decrease in performance. Additionally, employing human-corrected sentences as feedback incurs a doubling of the adaptation task cost, but only yields marginal improvements in translation performance. One possible assumption is that significant positive impact on translation quality can be achieved only when a substantial amount of high-quality human feedback data is available. Therefore, we did not integrate this method into our adapter-translator framework.

5.6 The Impact of Frozen Encoder Parameters Proportion

We performed preliminary experiments to examine the optimal proportion of frozen encoder parameters during the inference phase of the translator. The results in Table 10 demonstrate that the translator's performance steadily improved as we increased the proportion of frozen encoder parameters, reaching its peak at 99%. However, when we further increased the proportion to 99.5%, the translator's performance started to decline. Consequently, in our experiments, we set the proportion of frozen encoder parameters to 99% during the inference phase of the translator.

6 Related Work

Local context has been extensively investigated in neural machine translation (NMT) models, including both RNN-based RNNSearch and Transformer-based models (Bahdanau et al., 2015; Vaswani et al., 2017). An early attempt in RNN-based NMT was the concatenation method proposed by (Tiedemann and Scherrer, 2017). Subsequently, the adoption of multiple encoders emerged as a promising direction in both RNNSearch and Transformer-based NMT models (Jean et al., 2017; Wang et al., 2017; Zhang et al., 2018; Bawden et al., 2018; Voita et al., 2018; Voita et al., 2019b; Yang et al., 2020). Cache/memory-based approaches (Tu et al., 2018; Kuang et al., 2018; Maruf and Haffari, 2018) also fall under this category, as they utilize a cache to store word/translation information from previous sentences.

An alternative approach in document-level NMT treats the entire document as a unified translation unit and dynamically extracts pertinent global knowledge for each sentence within the document. This global context can be derived either from the source side (Maruf and Haffari, 2018; Mace and Servan, 2019; Maruf et al., 2019; Tan et al., 2019) or the target side (Xiong et al., 2019).

Moreover, several endeavors have been undertaken to enhance the performance of document-level translation through the utilization of monolingual document data. For instance, in order to improve translation coherence within a document, Voita et al.(2019a) propose DocRepair, which is trained on monolingual target language document corpora to address inconsistencies in sentence-level translations. Similarly, Yu et al.(2020) train a document-level language model to re-evaluate sentence-level translations. In contrast, Dowmunt(2019) harness monolingual source language document corpora to investigate multi-task training using the BERT-objective on the encoder.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Task	s-BLEU	d-BLEU	Meteor
Real feedback	42.64	45.37	29.60
Fake feedback	42.03	44.71	28.50
Ours	**42.47**	**45.10**	**29.49**

Table 9: Performance of human feedback augmented adapting task on ZH-EN translation task.

K	BLEU	Meteor
97.0%	23.57	43.46
98.0%	26.69	45.83
99.0%	27.55	46.75
99.5%	27.50	46.68
99.7%	27.46	46.60

Table 10: The impact of frozen encoder parameters proportion.

7 Conclusion

To enhance the alignment between the trained context-aware NMT model and each input document, we present in this paper an adapter-translator framework, designed to facilitate the model's familiarity with a document prior to translation. Our modification to the NMT model involves incorporating an adapter encoder, which reconstructs the intentionally corrupted portions of the original document. Empirical findings from Chinese-to-English translation tasks and various English-to-German translation tasks demonstrate the considerable performance improvement achieved by our approach compared to several robust baseline models.

Limitations

Our experimental findings and analysis validate the effectiveness of the proposed adapter-translator framework in facilitating model familiarity with documents prior to translation, thereby yielding substantial enhancements across multiple evaluation benchmarks. However, it should be noted that the inclusion of the adapter module may introduce a certain degree of computational overhead to the framework's efficiency. Nevertheless, it is widely recognized that the time-consuming aspect of machine translation during the inference stage primarily stems from the serial decoding process of beam search. In contrast, our approach, as described in this paper, does not employ beam search during the adaptation stage; instead, it leverages parallel attention and mask mechanisms that align with the training stage. The main increase in computational time for this approach arises from the storage of checkpoints after the completion of parameter updates during the adaptation stage.

References

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

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of NAACL*, page 1304–1313.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*, pages 261–268.

Linqing Chen, Junhui Li, and Zhengxian Gong. 2020. Hierarchical global context augmented document-level neural machine translation. In *Proceedings of CCL*, pages 434–445.

Linqing Chen, Junhui Li, Zhengxian Gong, Min Zhang, and Guodong Zhou. 2022. One type context is not enough: Global context-aware neural machine translation. In *TALLIP*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, page 4171–4186.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of CVPR*, pages 16000–16009.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machinetranslation benefit from larger context? *Computing Research Repository*, arXiv:1704.05135.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of WMT*, page 225–233.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of EMNLP*, page 2242–2254.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, page 388–395.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of COLING*, page 596–606.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of WMT*, pages 228–231, June.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2018. One sentence one model for neural machine translation. In *Proceedings of LREC*, pages 910–917.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of ACL*, page 3505–3511.

Valentin Mace and Christophe Servan. 2019. Using whole document context in neural machine translation. In *Proceedings of IWSLT*.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of ACL*, pages 1275–1284.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of NAACL*, pages 3092–3102.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of EMNLP*, pages 2947–2954.

Kishore Papineni, Salim Roukos, Ward Todd, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Amane Sugiyama and Naoki Yoshinaga. 2021. Context-aware decoder for neural machine translation using a target-side document-level language model. In *Proceedings of NAACL*, pages 5781–5791.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of EMNLP-IJCNLP*, page 1576–1585.

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.

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.

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 *Proceedings of NIPS*, pages 5998–6008.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of ACL*, pages 1264–1274.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of EMNLP-IJCNLP*, pages 877–886.

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 ACL*, pages 1198–1212.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of EMNLP*, page 2826–2831.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *Proceedings of Workshop on Discourse in Machine Translation*, pages 17–25.

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

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2020. Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proceedings of EMNLP*, pages 1527–1537.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

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 EMNLP*, pages 533–542.

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 IJCAI*, pages 3983–3989.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.