Efficient Context-Aware Neural Machine Translation with Layer-Wise Weighting and Input-Aware Gating

Hongfei Xu[1,2], **Deyi Xiong**[3*], **Josef van Genabith**[1,2] and **Qiuhui Liu**[4]
[1]Saarland University, Germany
[2]German Research Center for Artificial Intelligence, Germany
[3]Tianjin University, China
[4]China Mobile Online Services, China
hfxunlp@foxmail.com, dyxiong@tju.edu.cn, Josef.Van_Genabith@dfki.de, liuqhano@foxmail.com

Abstract

Existing Neural Machine Translation (NMT) systems are generally trained on a large amount of sentence-level parallel data, and during prediction sentences are independently translated, ignoring cross-sentence contextual information. This leads to inconsistency between translated sentences. In order to address this issue, context-aware models have been proposed. However, document-level parallel data constitutes only a small part of the parallel data available, and many approaches build context-aware models based on a pre-trained frozen sentence-level translation model in a two-step training manner. The computational cost of these approaches is usually high. In this paper, we propose to make the most of layers pre-trained on sentence-level data in contextual representation learning, reusing representations from the sentence-level Transformer and significantly reducing the cost of incorporating contexts in translation. We find that representations from shallow layers of a pre-trained sentence-level encoder play a vital role in source context encoding, and propose to perform source context encoding upon weighted combinations of pre-trained encoder layers' outputs. Instead of separately performing source context and input encoding, we propose to iteratively and jointly encode the source input and its contexts and to generate input-aware context representations with a cross-attention layer and a gating mechanism, which resets irrelevant information in context encoding. Our context-aware Transformer model outperforms the recent CADec [Voita *et al.*, 2019c] on the English-Russian subtitle data and is about twice as fast in training and decoding.

1 Introduction

NMT has achieved great success in the last few years, and the Transformer [Vaswani *et al.*, 2017] which has outperformed previous RNN/CNN based translation models, is based on multi-layer self-attention networks and can be trained in parallel very efficiently.

Despite the great success of these sequence-to-sequence models, they translate in a sentence-by-sentence manner, utilizing a large amount of sentence-level parallel data, while totally ignoring extra-sentential context information and inter-sentence consistency. This issue has attracted wide attention to context-aware translation recently, and many context-aware translation approaches [Wang *et al.*, 2017; Tiedemann and Scherrer, 2017; Bawden *et al.*, 2018; Voita *et al.*, 2018; Maruf and Haffari, 2018; Kuang *et al.*, 2018; Kuang and Xiong, 2018; Zhang *et al.*, 2018; Läubli *et al.*, 2018; Miculicich *et al.*, 2018; Tu *et al.*, 2018; Voita *et al.*, 2019c; Voita *et al.*, 2019b; Xiong *et al.*, 2019; Tan *et al.*, 2019] are proposed.

However, document-level parallel data are often scarce in practical scenarios, especially compared to the large amount of sentence-level parallel data available. As a result, Zhang *et al.* [2018], Voita *et al.* [2018] and Voita *et al.* [2019c] follow a two-step training procedure which first trains a sentence-level translation model, then freezes those pre-trained parameters and augments the model with extra components which extract and incorporate cross-sentence contexts. To be specific, Zhang *et al.* [2018] introduce an additional context encoder, together with cross-attention networks inserted into each encoder layer and decoder layer to attend context representations. Voita *et al.* [2019c] add a second Context-Aware Decoder (CADec) which attends both the sentence to be translated and its adjacent contexts and generates context-aware translations.

Unfortunately, the cost of such previous approaches for gathering context information is relatively high. In detail, Zhang *et al.* [2018] additionally introduce 1 source context encoder together with 36 cross-attention networks (39 blocks in total) to attend the source context representations when using 3 context sentences. The additional CADec machinery [Voita *et al.*, 2019c] consists of 6 layers (24 blocks) to perform context-aware translation from scratch with word embeddings.

In this paper, we propose to reduce the cost for incorporating contexts by reusing the rich and functional representations produced during computation of the pre-trained sentence-level layers for context encoding. We propose a lightweight yet efficient context encoding model for document-level

* Corresponding author.

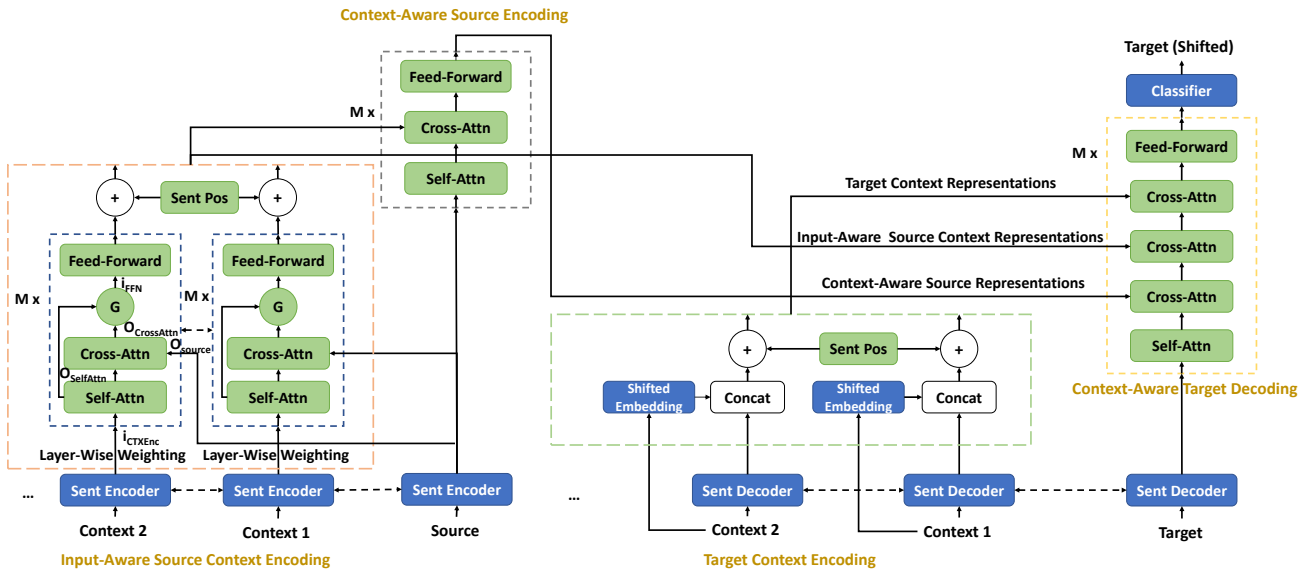

Figure 1: Efficient Context-Aware Transformer with Layer-Wise Weighting and Input-Context Gating. Blue: Parameters from the standard sentence-level Transformer pre-trained on parallel sentence pairs; Green: New parameters for context-level translation. Token Dropout is applied on the shifted embedding of the target context encoding module. Sharing Sent Encoder/Decoder and context encoding layers enables parallel computing of contexts and source/target. Cross-attention networks are surrounded by gated residual connections. Cross-attention networks in Sent Decoder layers attend corresponding outputs of Sent Encoders. Residual connections and layer normalization are omitted for simplicity. Sentence positional embeddings are trained by backpropagation. We use $M = 1$ by default in our experiments. Layer-wise weighting is shown in Figure 2 for better view.

NMT, shown in Figure 1. Instead of training a multiple-layer context encoder on a usually small amount of document-level data from scratch, we attempt to train lightweight context modules (e.g., $M = 1$ in the green parts of Figure 1), efficiently incorporating information from a sentence-level Transformer trained on large-scale sentence-level data into context modules. First, we empirically find that representations from shallow pre-trained encoder layers are important in source context encoding, and propose to perform source context encoding upon the weighted combination of pre-trained encoder layers' outputs, instead of only the last layer. Second, previous document-level approaches usually separately perform context encoding and context-aware input encoding, which totally ignores the source input in context encoding and may lead to the inclusion of information irrelevant to the translation of the source input in context representations. To address this, we propose (i) to perform input-aware context encoding with a cross-attention network attending the source input and (ii) an input-aware gating mechanism to reset information in the context representations that is unrelated to the translation of the source.

Our proposed architecture outperforms CADec while being around twice as fast in our experiments with only 3 layers (11 blocks) newly introduced for utilizing contexts in translation. The contributions of this paper are as follows:

- We propose a new efficient document-level NMT model which reuses the rich and functional representations learned in pre-trained sentence-level layers in context-aware encoding and decoding and reduces the additional efforts for incorporating contexts into translation;
- We find that shallow encoder layers extract important feature representations for the source context encoding, and propose to build the input-aware source context encoder layer upon a layer-wise weighted combination of pre-trained encoder layers' outputs;
- We introduce a cross-attention network attending the source input and a gating mechanism to reset information irrelevant to source input translation in context representations for source context encoding;
- Our proposed model outperforms previous approaches [Zhang *et al.*, 2018; Voita *et al.*, 2019c] on the English-Russian subtitle dataset in both BLEU and linguistic evaluations, and is about twice as fast as the CADec.

2 The Proposed Model

We separate context-aware translation into two steps: 1, using a standard sentence-level Transformer to produce context-agnostic translations; 2, building context-aware layers which incorporate inter-sentence context information upon the pre-trained context-agnostic model to make full use of rich and functional representations from context-agnostic layers pre-trained on a large amount of sentence-level parallel data, instead of generating context-aware representations from scratch with word embeddings. Our proposed model involves: a standard sentence-level Transformer, input-aware source context encoding layers, context-aware input encoding layers, target context encoding and context-aware decoding layers. The architecture is shown in Figure 1.

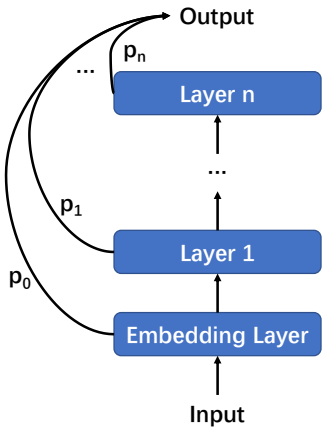

Figure 2: Layer-Wise Weighting.

2.1 Input-Aware Source Context Encoding

Zhang *et al.* [2018] use the standard Transformer encoder layer including a self-attention network followed by a feed forward network as the context encoding layer and encoding from scratch from embeddings. In our work, we propose to use the weighted aggregation of pre-trained encoder layers' outputs as the input of context encoding layers and augment the context encoding layer with a gating mechanism to generate source context representations aware of the source input sentence.

Layer-Wise Weighting

A single source context encoder layer results in the best performance in Zhang *et al.* [2018]. Peters *et al.* [2018] show that different encoder layers capture linguistic properties at different levels, while Voita *et al.* [2019a] show that the Transformer encoder gradually forgets its inputs while going deep. Therefore we conjecture that the representations produced by the last pre-trained encoder layer may not be the best choice for source context encoding. Instead we propose to aggregate individually *weighted* outputs of all encoder layers as input to context-aware encoder layers, as illustrated in Figure 2.

Let w_i be the ith element of a weight vector w, we first normalize w into a probability distribution p with softmax:

$$p_i = \frac{e^{w_i}}{\sum_{j=0}^{n} e^{w_j}} \tag{1}$$

where n is the number of encoder layers and 0 corresponds to the word embedding layer.

Then we use p_i as the weight of the output of the ith encoder layer o_i, and the input to the source context encoder layers i_{CTXEnc} is:

$$i_{CTXEnc} = \sum_{i=0}^{n} p_i o_i \tag{2}$$

w is trained together with the model by backpropagation to find the best weighted combination of pre-trained layers' outputs for source context encoding during training.

Unlike previous work [Dou *et al.*, 2018; Bapna *et al.*, 2018; Dou *et al.*, 2019] which combines outputs of layers for sentence-level translation, we weigh shallow layers only for source context encoding rather than aggregating for the source input.

Input-Aware Gating

We propose to insert a cross-attention sub-layer together with a reset gate into the context encoding layer to perform input-aware source context encoding. During the encoding of context sentences, the context encoding layer first performs self attention with the input to this layer i_{CTXEnc} to build an intra-sentence contextual representation:

$$O_{SelfAttn} = SelfATTN(i_{CTXEnc}) \tag{3}$$

Then the cross-attention layer attends to the representations of the source input sentence O_{source} to be translated from the pre-trained sentence encoder to gather information about the source sentence:

$$O_{CrossAttn} = CrossATTN(O_{SelfAttn}, O_{source}) \tag{4}$$

With the original representation of the context sentence $O_{SelfAttn}$ and representation $O_{CrossAttn}$ relevant to the source sentence, we compute a gate to reset $O_{SelfAttn}$ as the input to the feed-forward network i_{FFN}:[1]

$$g_{\text{context}} = \sigma(W_1[O_{SelfAttn}|O_{CrossAttn}] + b_1) \tag{5}$$

$$i_{FFN} = O_{SelfAttn} \bullet g_{\text{context}} \tag{6}$$

where "|", "•" and "σ" are vector concatenation, element-wise multiplication and the sigmoid activation function.

Then, the feed-forward network takes the output of the gating mechanism and generates the output of this layer.

At the end of the input-aware context sentence encoding, sentence positional embeddings are added to the output of the last encoding layer to each context sentence representation sequence, allowing us to concatenate all context sentence representation sequences into one.

2.2 Context-Aware Input Encoding

To generate context-aware source sentence input representations, we add context-aware input encoding layers on top of the pre-trained context-agnostic encoder. The context-aware input encoding layer first performs self attention over inputs, followed by cross attention to the input-aware source context representations to incorporate context information and a feed-forward network.

To control the influence of the residual connections of the Transformer after the cross-attention sub-layer attending source context representations, we follow Zhang *et al.* [2018] and use a gated residual connection. The gated residual connection first takes input of the sub-layer $input_{residual}$ and the

[1]Note that $O_{SelfAttn}$ corresponds to the residual connection in the standard encoder layer, and in our case, the residual connection is reset by the gating mechanism.

output of the sub-layer $O_{Sub-layer}$ to compute the residual gate g:

$$g = \sigma(W_2[input_{residual}|O_{Sub-layer}] + b_2) \quad (7)$$

Then we apply the residual gate g to $input_{residual}$ and $O_{Sub-layer}$ to obtain o as the result.

$$o = input_{residual} \bullet g + O_{Sub-layer} \bullet (1 - g) \quad (8)$$

2.3 Target Context Encoding

For the context encoding on the target side, we follow Voita *et al.* [2019c] and use the gold reference translations as context translation on the target side during training, utilizing the concatenation of word embeddings and the outputs of the last pre-trained decoder layer. Instead of replacing 20% tokens of 50% sentences in the gold references with random tokens like Voita *et al.* [2019c], we apply a token dropout of 0.1 randomly masking the whole embeddings of randomly selected tokens with zeros in the embedding matrix. Empirically we find this provides comparable performance while being more efficient than randomly replacing tokens in target contexts, since the use of token dropout avoids looking up embeddings in the new reference matrix with noise for a second time.

After we obtain the target context representation sequences, we add sentence positional embeddings to each of them to indicate their position and concatenate them into 1 sequence as the target context representation sequence.

2.4 Context-Aware Decoding

To introduce cross-sentence contexts into the Transformer decoder, we follow Voita *et al.* [2019c] to explore contexts from both source side and target side. Rather than using an additional context-aware decoder, we put context-aware decoding layers on top of the sentence-level decoder. Thus the context-aware decoding layer can work directly on the rich and functional representations from the pre-trained layers.

In each context-aware decoding layer, we first perform self attention with subsequent masking, followed by cross attention networks to context-aware representations of the source sentence, input-aware source context representations and target context representations. Finally, a feed-forward network is applied to process the collected information. Cross-attention networks are also wrapped by gated residual connections like in the context-aware source encoding layers.

3 Efficiency Analysis

Compared to Zhang *et al.* [2018], our model separates context-aware layers from the sentence-level Transformer by putting them on top of the pre-trained model and only introducing 1 input-aware source context encoding layer, 1 context-aware source encoding layer and 1 context-aware target decoding layer (i.e., $M = 1$ in Figure 1) instead of inserting cross-attention networks in every pre-trained encoder and decoder layer. Rather than using individual cross-attention networks for context sentences at different positions as in Zhang *et al.* [2018], we use sentence position embedding to distinguish contexts of different distances.

Compared to Voita *et al.* [2019c], our approach reuses representations produced during the forward computation of the sentence-level Transformer instead of introducing a CADec which builds deep representations from scratch from word embeddings.

Since the pre-trained encoder and decoder are shared for both source sentences, target translations and related contexts, our approach can flatten a batch of documents into a larger batch of independent sentences and perform forward computation in parallel effciently in the implementation. Context encoding layers can utilize rich and functional representations extracted directly from the pre-trained sentence-level encoder / decoder without additional cost (e.g., **Context 1** in Figure 1 is the source input of the preceding sentence in a batch flattened by a document of sentences).

The context encoding layers are shared for context sentences at different positions, and context sentences are independently encoded. In practice, context sentences can also be merged into 1 batch of independent sentences and computed in parallel efficiently.

4 Experiments

We conducted experiments to validate the performance and efficiency of our approach. Our approach was implemented based on the Neutron implementation [Xu and Liu, 2019] of the Transformer translation model. To compare with Voita *et al.* [2019c], we used the same corpus which is based on the publicly available OpenSubtitles2018 corpus [Lison *et al.*, 2018] for English and Russian. The corpus consists of 6 million training instances, among which 1.5 million have contexts of three sentences. We also compared our approach with Zhang *et al.* [2018].

In addition to tokenized BLEU, we also performed linguistic evaluations on the contrastive test sets [Voita *et al.*, 2019c] which are specifically designed to test the ability of a system to adapt to contextual information in handling frequent discourse phenomena (i.e., deixis, lexical cohesion, VP and inflection ellipses) in context-aware translation. Each test instance in the contrastive test sets consists of a true example (sequence of sentences and their reference translation from the data) and several contrastive translations which differ from the true one only in the concerned aspect. All contrastive translations are correct plausible translations at the sentence level, and only context reveals the errors introduced. The model is asked to score each candidate example, and the accuracy is defined as the proportion of times the true translation is preferred over the contrastive ones.

4.1 Settings

For fairness, we followed the setting of Voita *et al.* [2019c] to use 3 previous context sentences. Corresponding to randomly masking 20% tokens of 50% sentences with random tokens, we used a token dropout of 0.1. We employed $h = 8$ parallel attention heads. The dimension of input and output (d_{model}) was 512, and the hidden dimension of feed-forward networks was 2048. We used 0.1 as the dropout probability and the label smoothing value. For the Adam optimizer, we used 0.9, 0.98 and 10^{-9} as $\beta 1$, $\beta 2$ and ϵ, and all context-aware mod-

	BLEU	Train	Decode	Para.
Sent-level Trans.	32.40			76.6
CADec	32.38	2,939	2.46	+34.6
Zhang *et al.* [2018]	32.51	2,566	3.07	+75.8
Our approach	**32.75**	**1,556**	**4.98**	**+17.8**

Table 1: BLEU scores and efficiency measured by training time, decoding speed and additional parameters. Train, Decode and Para. are training time (s) per epoch, the decoding speed (#documents / s) and numbers of parameters (M), respectively. Trans.: Transformer.

		latest relevant context		
	total	1st	2nd	3rd
deixis				
Sent-level Trans.	50.0	50.0	50.0	50.0
CADec	81.6	84.6	84.4	75.9
Zhang *et al.* [2018]	50.0	50.0	50.0	50.0
Our approach	**85.4**	**87.8**	**87.7**	**80.6**
lexical cohesion				
Sent-level Trans.	45.9	46.1	45.9	45.4
CADec	58.1	63.2	52.0	56.7
Zhang *et al.* [2018]	46.0	46.4	45.8	45.4
Our approach	**64.3**	**67.3**	**62.2**	**61.9**

Table 2: Accuracy for deixis and lexical cohesion.

els were trained for $200k$ training steps following Voita *et al.* [2019c].

We kept the other settings the same as Voita *et al.* [2019c]. All models were trained on 2 GTX 1080 Ti GPUs, and translation was performed on 1 GPU. Our model uses only 1 source context encoding layer, 1 context-aware input encoding layer and 1 context-aware decoding layer by default.

4.2 Main Results

We compared our approaches with Voita *et al.* [2019c] and Zhang *et al.* [2018]. Results are shown in Tables 1, 2 and 3.

Like in Voita *et al.* [2019c], none of approaches resulted in statistically significant differences in BLEU, but BLEU is widely acknowledged as a bad metric for discourse phenomena, and contrastive linguistic evaluation is considered to complement it. Table 1 shows that our approach with significantly fewer additional parameters for context modeling is significantly faster than previous approaches in both training and decoding. Particularly, both the training and decoding of our approach are almost twice as fast as the CADec [Voita *et al.*, 2019c].

Table 2 and 3 further show that our approach outperforms previous approaches in the linguistic evaluations even at significantly less additional cost (computation / parameters). Zhang *et al.* [2018] only utilize source contexts, while both Voita *et al.* [2019c] and our approach use target contexts in addition to source contexts. Our approach outperforms the CADec [Voita *et al.*, 2019c] in all tests, potentially because the layer-wise weighting provides more suitable representations for the source context encoding and the input-aware gating resets translation-irrelevant information, while

	ellipsis (infl.)	**ellipsis (VP)**
Sent-level Trans.	53.0	28.4
CADec	72.2	80.0
Zhang *et al.* [2018]	56.4	48.0
Our approach	**77.1**	**81.3**

Table 3: Accuracy on ellipsis test sets.

the CADec directly uses the output of the last pre-trained sentence-level encoder layer.

4.3 Ablation Study

We studied the effects of the input-aware gating and the layer-wise weighting in source context encoding. The results are shown in Table 4.

We observe that both the input-aware gating and the layer-wise weighting play vital roles in our model in the linguistic evaluation. If we disable the layer-wise weighting (i.e., using the last layer of the pre-trained encoder), we lose 2.65 points in accuracy on average. We further show the learned weight w_i for each pre-trained encoder layer in Table 5. It is obvious that shallow layers, e.g., the word embedding layer and the first layer, are almost as important as the last pre-trained encoder layer for the source context encoding.

If we do not use the input-aware gating mechanism, the average accuracy on the four test sets further drops 2.38 points, indicating the advantage of input-aware context encoding over input-agnostic context encoding.

4.4 Analysis on Layer Depth

We studied the effects of layer depth M (Figure 1) in both source context encoding and target decoding modules, and results are shown in Table 6.

Table 6 shows that increasing the depth of context layers hurts performance, and using only 1 context layer seems to be the best choice in terms of both efficiency and performance. This is consistent with Zhang *et al.* [2018], and we conjecture potential reasons might be:

- The limited amount of document-level parallel data has difficulty in supporting the training of increasing numbers of parameters with increasing context layers;
- The pre-trained encoder / decoder layers are already deep, and representations of the pre-trained layers are rich enough, thus stacking more layers on top of the pre-trained layers is less likely to to be beneficial.
- The gated residual connection shrinks the gradients during backpropagation, which hampers convergence while stacking more layers.

We suggest that the fact that simply increasing context layer depth cannot further improve the performance in the current scenario highlights the importance of our proposed layer-wise weighting and input-aware gating approaches.

4.5 Analysis on Using Layer-Wise Weighting in Target Context Encoding

Applying input-aware context encoding to the target context encoding may bypass the subsequent mask and leak the gold

model	deixis	lex. c	ell. infl.	ell. VP
Our approach	85.4	64.3	77.1	81.3
- Layer-wise weighting	85.2	60.9	72.8	78.6
- Input-aware gating	83.7	57.7	69.4	77.2

Table 4: Ablation study.

Layer	0	1	2	3	4	5	6
Weight (%)	22.62	15.17	9.83	8.59	8.72	11.17	23.91

Table 5: Layer weights. 0 stands for the embedding layer.

Source	Target	deixis	lex. c	ell. infl.	ell. VP
1		85.4	64.3	77.1	81.3
1	2	80.9	60.5	76.9	79.9
2	1	82.3	61.8	76.7	80.6
2		79.1	58.9	74.5	80.9

Table 6: Effects of the depth of context layers (M in both source context encoding and target decoding modules).

	deixis	lex. c	ell. infl.	ell. VP
on source context	85.4	64.3	77.1	81.3
+ on target context	83.3	58.7	72.4	78.5

Table 7: Effects of layer-wise weighting on the target context encoding.

references, but will layer-wise weighting help target context encoding? We tried to apply layer-wise weighting into the target context encoding and results are shown in Table 7.

Table 7 shows that applying layer-wise weighting on the target side hurts the performance especially for lexical cohesion and inflection ellipsis. We conjecture this might be because the decoder gradually transforms the representation from current token to the next token, and aggregating outputs of different layers may lead to inconsistency in the representation.

5 Related Work

Analyzing the effects of contexts. Tiedemann and Scherrer [2017] discuss the effect of increasing the segments beyond single translation units, and observe cross-sentential attention patterns that improve textual coherence in translation. Läubli *et al.* [2018] show that human assessment has a stronger preference for human over machine translation when evaluating documents as compared to isolated sentences. Bawden *et al.* [2018] present hand-crafted, discourse test sets to test the models' ability to exploit previous source and target sentences, and highlight the importance of target-side context. Voita *et al.* [2018] introduce a context-aware NMT model which controls the flow of information from the context to the translation model, and show that the model deals with pronoun translation and implicitly captures anaphora. Voita *et al.* [2019c] perform a human study on an English-Russian subtitle dataset and identify deixis, ellipsis and lexical cohesion as three main sources of inconsistency, and create test sets targeting these phenomena.

Document-level modeling of NMT. Wang *et al.* [2017] summarize the history in a hierarchical way, and propose a cross-sentence context-aware approach to integrate the history representation into NMT. Maruf and Haffari [2018] present a document-level NMT model which takes both source and target document context into account using memory networks. Tu *et al.* [2018] augment NMT models with a cache-like memory network, which stores recent hidden representations as translation history. Kuang *et al.* [2018] propose a cache-based approach to modeling coherence for NMT by capturing contextual information either from recently translated sentences or the entire document. Kuang and Xiong [2018] propose an inter-sentence gating model that uses the same encoder to encode adjacent sentences and controls the amount of information flowing from the preceding sentence to the translation of the current sentence. Maruf *et al.* [2019] propose a hierarchical attention approach for context-aware NMT which first uses sparse attention to selectively focus on relevant sentences in the document context and then attends to key words in those sentences.

Two-step training models. Zhang *et al.* [2018] extend the Transformer model with a new context encoder to represent document-level context, which is then incorporated into the original encoder and decoder. Miculicich *et al.* [2018] propose to integrate a hierarchical attention model into the original NMT architecture to capture context. Voita *et al.* [2019c] propose the CADec which demonstrates major gains over a context-agnostic baseline on their benchmarks without sacrificing BLEU. Tan *et al.* [2019] propose a hierarchical model consisting of a sentence encoder to capture intra-sentence dependencies and a document encoder to model document-level information.

Other approaches. Xiong *et al.* [2019] propose to train the model to learn the policy that produces discourse coherent text by a reward teacher. Voita *et al.* [2019b] perform automatic post-editing on a sequence of sentence-level translations with a DocRepair model trained on very large monolingual document-level data in the target language and their round-trip translations of each isolated sentence, and analyze

which discourse phenomena are hard to capture using monolingual data only.

6 Conclusion

We present a model to learn context representations based on the weighted combination of pre-trained layers, and incorporate the source sentence to be translated in the encoding of source contexts with a reset gating mechanism to reduce translation-irrelevant information in context representations.

Our model makes the most of pre-trained sentence-level layers in context-aware translation by adding context-aware layers on top of the pre-trained layers instead of building representations from scratch with word embeddings.

Our approach reduces the additional costs for context-ware translation, and outperforms previous approaches while being more efficient.

Acknowledgments

We thank anonymous reviewers for their insightful comments and helpful advice. Hongfei Xu acknowledges the support of China Scholarship Council ([2018]3101, 201807040056). Deyi Xiong is supported by the National Natural Science Foundation of China (Grant No. 61861130364), the Natural Science Foundation of Tianjin (Grant No. 19JCZDJC31400) and the Royal Society (London) (NAF\R1\180122). Hongfei Xu and Josef van Genabith are supported by the German Federal Ministry of Education and Research (BMBF) under the funding code 01IW17001 (Deeplee).

References

[Bapna *et al.*, 2018] Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Bawden *et al.*, 2018] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Dou *et al.*, 2018] Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang. Exploiting deep representations for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4253–4262, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Dou *et al.*, 2019] Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Longyue Wang, Shuming Shi, and Tong Zhang. Dynamic layer aggregation for neural machine translation with routing-by-agreement. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 86–93. AAAI Press, 2019.

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

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

[Läubli *et al.*, 2018] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Lison *et al.*, 2018] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA).

[Maruf and Haffari, 2018] Sameen Maruf and Gholamreza Haffari. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Maruf *et al.*, 2019] Sameen Maruf, André F. T. 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)*, pages 3092–3102, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Miculicich *et al.*, 2018] Lesly Miculicich, 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*, pages 2947–2954, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Peters *et al.*, 2018] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Tan *et al.*, 2019] Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, Hong Kong, China, November 2019. Association for Computational Linguistics.

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

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

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

[Voita *et al.*, 2019a] Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4395–4405, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Voita *et al.*, 2019b] Elena Voita, Rico Sennrich, and Ivan Titov. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 876–885, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Voita *et al.*, 2019c] 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*, pages 1198–1212, Florence, Italy, July 2019. Association for Computational Linguistics.

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

[Xiong *et al.*, 2019] Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling coherence for discourse neural machine translation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7338–7345. AAAI Press, 2019.

[Xu and Liu, 2019] Hongfei Xu and Qiuhui Liu. Neutron: An Implementation of the Transformer Translation Model and its Variants. *arXiv preprint arXiv:1903.07402*, March 2019.

[Zhang *et al.*, 2018] 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*, pages 533–542, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.