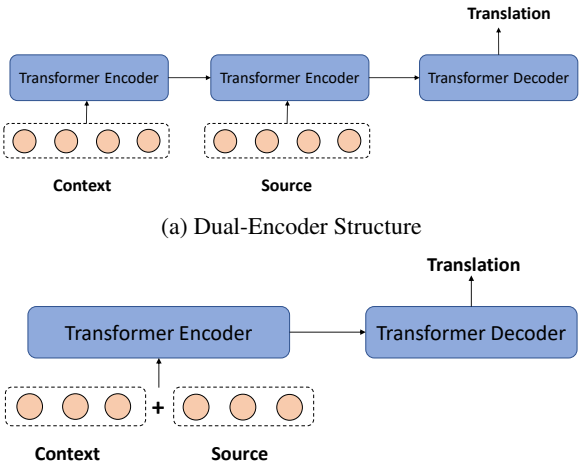
A Simple and Effective Unified Encoder for Document-Level Machine Translation

Shuming Ma, Dongdong Zhang, Ming Zhou
Microsoft Research Asia
{shumma, dozhang, mingzhou}@microsoft.com

Abstract

Most of the existing models for document-level machine translation adopt dual-encoder structures. The representation of the source sentences and the document-level contexts[1] are modeled with two separate encoders. Although these models can make use of the document-level contexts, they do not fully model the interaction between the contexts and the source sentences, and can not directly adapt to the recent pre-training models (e.g., BERT) which encodes multiple sentences with a single encoder. In this work, we propose a simple and effective unified encoder that can outperform the baseline models of dual-encoder models in terms of BLEU and METEOR scores. Moreover, the pre-training models can further boost the performance of our proposed model.



(a) Dual-Encoder Structure

(b) Uni-Encoder Structure

Figure 1: The overview of the dual-encoder structure and the uni-encoder structure for document-level machine translation.

1 Introduction

Thanks to the development of the deep learning methods, the machine translation systems have achieved good performance that is even comparable with human translation in the news domain (Hassan et al., 2018). However, there are still some problems with machine translation in the document-level context (Läubli et al., 2018). Therefore, more recent work (Jean et al., 2017; Wang et al., 2017; Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Bawden et al., 2018; Voita et al., 2019a; Junczys-Dowmunt, 2019) is focusing on the document-level machine translation.

Most of the existing models (Zhang et al., 2018; Maruf et al., 2019; Werlen et al., 2018) for document-level machine translation use two encoders to model the source sentences and the document-level contexts. Figure 1a illustrates the structure of these models. They extend the standard Transformer model with a new context encoder, and the encoder for source sentences is conditioned on this context encoder. However, they do not fully model the interaction between the contexts and the source sentences because the self-attention layers are performed inside each encoder separately. Moreover, it cannot be directly adapted to the recent pre-training models (Devlin et al., 2019; Peters et al., 2018; Radford et al., 2019; Dong et al., 2019; Song et al., 2019; Lample and Conneau, 2019), which encodes multiple sentences with a single encoder.

Different from the dual-encoder structure, the uni-encoder structure takes the concatenation of contexts and source sentences as the input (as shown in Figure 1b). Therefore, when modeling the contexts, it can make full use of the interaction between the source sentences and the contexts, while the dual-encoder model fails to exploit this information. Moreover, the uni-encoder structure is identical to the recent pre-training models (e.g.,

[1]In this work, document-level contexts denote the surrounding sentences of the current source sentence.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3505–3511
July 5 - 10, 2020.

BERT). However, the previous uni structure suffers from two problems for document-level machine translation. First, the attention is distracted due to longer sequences. Second, the source sentences and the contexts are modeled equally, which is contrary to the fact that the translation is more related to the current source sentences.

To address these problems, we propose a novel flat structure with a unified encoder called Flat-Transformer. It separates the encoder of standard Transformers into two parts so that the attention can concentrate at both the global level and the local level. At the bottom of the encoder blocks, the self-attention is applied to the whole sequence. At the top of the blocks, it is only implemented at the position of the source sentences. We evaluate this model on three document-level machine translation datasets. Experiments show that it can achieve better performance than the baseline models of dual-encoder structures in terms of BLEU and METEOR scores. Moreover, the pre-training models can further boost the performance of the proposed structure.

2 Flat-Transformer

In this section, we introduce our proposed flat structured model, which we denote as **Flat-Transformer**.

2.1 Document-Level Translation

Formally, we denote $X = \{x_1, x_2, \cdots, x_N\}$ as the source document with N sentences, and $Y = \{y_1, y_2, \cdots, y_M\}$ as the target document with M sentences. We assume that $N = M$ because the sentence mismatches can be fixed by merging sentences with sentence alignment algorithms (Sennrich and Volk, 2011). Therefore, we can assume that (x_i, y_i) is a parallel sentence pair.

Following Zhang et al. (2018), $y_{<i}$ can be omitted because $x_{<i}$ and $y_{<i}$ conveys the same information. As a result, the probability can be approximated as:

$$P(Y|X) \approx \prod_{i=1}^{N} P(y_i|x_i; x_{<i}; x_{>i}) \quad (1)$$

where x_i is the source sentence aligned to y_i, and $(x_{<i}, x_{>i})$ is the document-level context used to translate y_i.

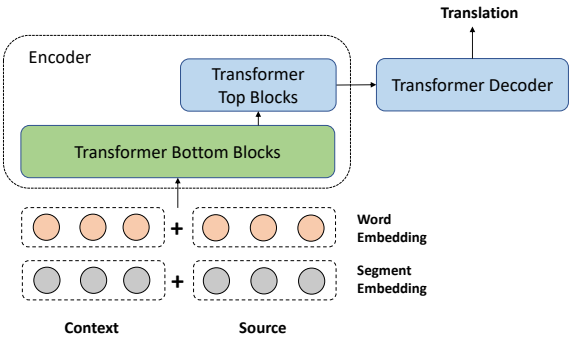


Figure 2: The architecture of the proposed Flat-Transformer model.

2.2 Segment Embedding

The flat structure adopts a unified encoder that does not distinguish the context sentences and the source sentences. Therefore, we introduce the segment embedding to identify these two types of inputs. Formally, given the source input of the surrounding context $\boldsymbol{c}$ and the current sentence $\boldsymbol{x}$, we project them into word embedding and segment embedding. Then, we perform a concatenation operation to unify them into a single input:

$$\boldsymbol{e} = [E(\boldsymbol{c}); E(\boldsymbol{x})] \quad (2)$$

$$\boldsymbol{s} = [S(\boldsymbol{c}); S(\boldsymbol{x})] \quad (3)$$

where $[;]$ denotes the concatenation operation, E is the word embedding matrix, and S is the segment embedding matrix. Finally, we add $\boldsymbol{e}$ and $\boldsymbol{s}$ as the input of the encoder.

2.3 Unified Flat Encoder

Given the document context, the input sequences of Flat-Transformer are much longer than the standard Transformer, which brings additional challenges. First, the attention is distracted, and its weights become much smaller after the normalization operation. Second, the memory consumption and the computation cost increase, so it is difficult to enlarge the model size, which hinders the adaptation to the pre-training model.

To address this problem, we introduce a unified flat encoder. As shown in Figure 2, at the bottom of the encoder blocks, we apply self-attention and the feed-forward layer to the concatenated sequence of the contexts and the current sentence:

$$h_1 = \text{Transformer}(\boldsymbol{e} + \boldsymbol{s}; \theta) \quad (4)$$

where θ is the parameter of the Transformer blocks. At the top of encoder blocks, each self-attention and feed-forward layer is only implemented on the position of the current sentences:

$$h_2 = \text{Transformer}(h_1[s:t]; \theta) \quad (5)$$

where s and t are the starting and ending positions of the source sentences in the concatenation sequence. In this way, the attention can focus more on the current sentences, while the contexts are served as the supplemental semantics for the current sentences. It is noted that the total number of the bottom blocks and the top blocks is equal to the number of standard Transformer's blocks, so there is no more parameter than that of the standard Transformer.

2.4 Training and Decoding

The training of Flat-Transformer is consistent with that of standard Transformer, using the cross entropy loss:

$$L = -\sum_{i=1}^{n} \log P(\boldsymbol{Y_i}|\boldsymbol{X_i}) \quad (6)$$

At the decoding step, it translates the document sentence-by-sentence. When translating each sentences, it predicts the target sequence with the highest probability given the current sentence x_i and the surrounding contexts $x_{<i}, x_{>i}$:

$$\hat{y}_i = \arg\max_{y_i \in \mathcal{V}} P(y_i|x_i; x_{<i}; x_{>i}) \quad (7)$$

2.5 Comparison with Existing Models

Here, we summarize some significant differences compared with the existing models for document-level machine translation:

1. Compared with the dual-encoder models, our model uses a unified encoder. To combine the representation of two encoders for the decoder, these dual-encoder models should add a layer inside the encoders. Flat-Transformer does not put any layer on top of the standard Transformer, so it is consistent with the recent pre-training models.

2. Compared with the previous uni-encoder models, our model limits the top transformer layers to only model the source sentences. In this way, our model has an inductive bias of modeling on more current sentences than the contexts, because the translation is more related to the current sentences.

Dataset	#Sent	Avg. #Sent
TED	0.21M/9K/2.3K	121/96/99
News	0.24M/2K/3K	39/27/19
Europarl	1.67M/3.6K/5.1K	14/15/14

Table 1: Statistics of three document-level machine translation datasets.

3. There are also some alternative approaches to limit the use of context vectors. For example, we can limit only the top attention layers to attend to the source sentence while keeping the feed-forward layers the same. Compared with this approach, our model does not feed the output vectors of the context encoder to the decoder, so that the decoder attention is not distracted by the contexts. The context vectors in our model is only to help encode a better representation for current source sentences.

3 Experiments

We evaluate the proposed model and several state-of-the-art models on three document-level machine translation benchmarks. We denote the proposed model as **Flat-Transformer**.

3.1 Datasets

Following the previous work (Maruf et al., 2019), we use three English-German datasets as the benchmark datasets, which are TED, News, and Europarl. The statistic of these datasets can be found in Table 1. We obtain the processed datasets from Maruf et al. (2019)[2], so that our results can be compared with theirs reported in Maruf et al. (2019). We use the scripts of Moses toolkit[3] to tokenize the sentences. We also split the words into subword units (Sennrich et al., 2016) with 30K merge-operations. The evaluation metrics are BLEU (Papineni et al., 2002) and Meteor (Banerjee and Lavie, 2005).

3.2 Implementation Details

The batch size is limited to $4,000$ tokens for all models. We set the hidden units of the multi-head component and the feed-forward layer as 512 and 1024. The embedding size is 512, the number of heads is 4, and the dropout rate (Srivastava et al., 2014) is 0.3. The number of Transformer blocks

[2]https://github.com/sameenmaruf/selective-attn
[3]https://github.com/moses-smt/mosesdecoder

	Model	TED		News		Europarl	
		BLEU	METR	BLEU	METR	BLEU	METR
Dual	HAN (Werlen et al., 2018)	24.58	45.48	25.03	44.02	29.58	46.91
	SAN (Maruf et al., 2019)	24.62	45.32	24.84	44.27	29.90	47.11
	QCN (Yang et al., 2019)	25.19	45.91	22.37	41.88	29.82	47.86
	Transformer (Zhang et al., 2018)	24.01	45.30	22.42	42.30	29.93	48.16
	+BERT	23.19	45.25	22.06	42.25	30.72	48.62
Uni	RNN (Bahdanau et al., 2015)	19.24	40.81	16.51	36.79	26.26	44.14
	Transformer (Vaswani et al., 2017)	23.28	44.17	22.78	42.19	28.72	46.22
	Our Flat-Transformer	24.87	47.05	23.55	43.97	30.09	48.56
	+BERT	26.61	48.53	24.52	45.40	31.99	49.76

Table 2: Results on three document-level machine translation benchmarks ("Dual" denotes dual-encoder, while "Uni" means uni-encoder).

TED	BLEU	METEOR
Flat-Transformer	24.87	47.05
w/o Segment	24.36	46.20
w/o Unified	23.28	44.17

Table 3: Ablation study on the TED dataset.

for the top encoder is 5, while that for the bottom encoder is 1. When fine-tuning on the pre-training BERT, we adopt the base setting, and the hidden size, the feed-forward dimension, and the number of heads are 768, 3072, 12. To balance the accuracy and the computation cost, we use one previous sentence and one next sentence as the surrounding contexts.

We use the Adam (Kingma and Ba, 2014) optimizer to train the models. For the hyper-parameters of Adam optimizer, we set two momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.98$, and $\epsilon = 1 \times 10^{-8}$. The learning rate linearly increases from 0 to 5×10^{-4} for the first $4,000$ warming-up steps and then decreases proportional to the inverse square root of the update numbers. We also apply label smoothing to the cross-entropy loss, and the smoothing rate is 0.1. We implement the early stopping mechanism with patience that the loss on the validation set does not fall in 10 epochs.

3.3 Baselines

We compare our models with two categories of baseline models: the dual-encoder models and the uni-encoder models.

Uni-encoder: RNNSearch (Bahdanau et al., 2015) is an RNN-based sequence-to-sequence model with the attention mechanism. Transformer (Vaswani et al., 2017) is a popular model for machine translation, based solely on attention mechanisms. For a fair comparison, we use the same hyper-parameters as our model's, which is described in Section 3.2.

Dual-encoder: Zhang et al. (2018) extends the Transformer model with a new context encoder to represent the contexts. HAN (Werlen et al., 2018) is the first to use a hierarchical attention model to capture the context in a structured and dynamic manner. SAN (Maruf et al., 2019) proposes a new selective attention model that uses sparse attention to focus on relevant sentences in the document context. QCN (Yang et al., 2019) proposes a query-guided capsule networks to cluster context information into different perspectives.

3.4 Results

We compare our Flat-Transformer model with the above baselines. Table 2 summarizes the results of these models. It shows that our Flat-Transformer can obtain scores of 24.87/23.55/30.09 on three datasets in terms of BLEU, and 47.05/43.97/48.56 in terms of METEOR, which significantly outperforms the previous flat models (RNNSearch and Transformer).

By fine-tuning on BERT, Flat-Transformer can achieve improvements of +1.74/+0.97/+1.90 BLEU scores as well as +1.48/+1.43/+1.20 METEOR scores. It proves that Flat-Transformer can be compatible with the pre-training BERT model. Except for the BLEU score on the News dataset, the Flat-Transformer can significantly outperform the dual-encoder models, achieving state-of-the-

art performance in terms of both BLEU and METEOR scores. On the contrary, the dual-encoder Transformer is not compatible with BERT. It gets slightly worse performance on two datasets, mainly because the model size becomes larger to adapt the setting of BERT. Still, BERT does not provide a good prior initialization for modeling the uni-directional relationship from contexts to source sentences.

3.5 Ablation Study

To analyze the effect of each component of Flat-Transformer, we conduct an ablation study by removing them from our models on the TED dataset. Table 3 summarizes the results of the ablation study. We remove the segment embedding but reserve the unified structure. It concludes that the segment embedding contributes to an improvement of 0.51 BLEU score and 0.85 METEOR score, showing the importance of explicitly identifying the contexts and the source sentences. After further removing the unified structure of Flat-Transformer, the model becomes a standard Transformer. It shows that the unified structures contribute a gain of 1.08 in terms of BLEU and 2.03 in terms of METEOR. The reason is that the unified structures encourage the model to focus more on the source sentences, while the contexts can be regarded as the semantic supplements.

4 Related Work

Here we summarize the recent advances in document-level neural machine translation. Some work focuses on improving the architectures of the document machine translation models. Tiedemann and Scherrer (2017) and Wang et al. (2017) explore possible solutions to exploit the cross-sentence contexts for neural machine translation. Zhang et al. (2018) extends the Transformer model with a new context encoder to represent document-level context. Werlen et al. (2018) and (Maruf et al., 2019) propose two different hierarchical attention models to model the contexts. Yang et al. (2019) introduces a capsule network to improve these hierarchical structures. There are also some works analyzing the contextual errors (Voita et al., 2018, 2019b; Bawden et al., 2018) and providing the test suites (Müller et al., 2018). More recently, Voita et al. (2019a) explores the approaches to incorporate the mono-lingual data to augment the document-level bi-lingual dataset. Different from these works, this paper mainly discusses the comparison between dual-encoder models and uni-encoder models and proposes a novel method to improve the uni-encoder structure.

5 Conclusions

In this work, we explore the solutions to improve the uni-encoder structures for document-level machine translation. We propose a Flat-Transformer model with a unified encoder, which is simple and can model the bi-directional relationship between the contexts and the source sentences. Besides, our Flat-Transformer is compatible with the pre-training model, yielding a better performance than both the existing uni-encoder models and the dual-encoder models on two datasets.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable suggestions and comments. We appreciate Sameen Maruf providing the same processed document data as in their work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72.

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

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language

model pre-training for natural language understanding and generation. *CoRR*, abs/1905.03197.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

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

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4791–4796.

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

Sameen Maruf, André 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.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL 2018*, pages 2227–2237.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Rico Sennrich and Martin Volk. 2011. Iterative, mt-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics, NODALIDA 2011, May 11-13, 2011, Riga, Latvia*, pages 175–182.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *ICML 2019*, pages 5926–5936.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

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

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)*, pages 876–885. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings*

of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

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

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

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.

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. *CoRR*, abs/1909.00564.

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.