Document-level Neural Machine Translation Using BERT as Context Encoder

Zhiyu Guo
Japan Advanced Institute
of Science and Technology
guozhiyu@jaist.ac.jp

Minh Le Nguyen
Japan Advanced Institute
of Science and Technology
nguyenml@jaist.ac.jp

Abstract

Large-scale pre-trained representations such as BERT have been widely used in many natural language understanding tasks. The methods of incorporating BERT into document-level machine translation are still being explored. BERT is able to understand sentence relationships (Devlin et al., 2019) since BERT is pre-trained using the next sentence prediction task. In our work, we leverage this property to improve document-level machine translation. In our proposed model, BERT performs as a context encoder to achieve document-level contextual information, which is then integrated into both the encoder and decoder. Experiment results show that our proposed method can significantly outperform strong document-level machine translation baselines on BLEU score. Moreover, the ablation study shows our method can capture document-level context information to boost translation performance.

1 Introduction

Recent years have witnessed the great success of neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017). NMT systems have even achieved human parity on resource-rich language pairs (Hassan et al., 2018). However, standard NMT systems perform translation only at the sentence level, which ignores the dependencies among sentences when translating entire documents. To address the above challenges, various document-level NMT models, have been proposed to extract contextual information from surrounding sentences and have achieved substantial improvements in generating consistent translations (Voita et al., 2018; Zhang et al., 2018; Werlen et al., 2018; Maruf et al., 2019; Ma et al., 2020).

Large-scale pre-trained text representations like GPT-2 (Radford et al., 2018, 2019), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), have been widely used in many natural language understanding tasks. Among them, BERT is one of the most effective representations that has inspired many other representations such as RoBERTa, ALBERT. It significantly boosts the performance of many natural language understanding tasks, including text classification, question answering, etc (Devlin et al., 2019). There have been few recent attempts to incorporate BERT into NMT models (Xiong et al., 2019; Zhu et al., 2020; Weng et al., 2019; Chen et al., 2020).

Intuitively, as one of BERT's pre-training tasks is the binarized next sentence prediction (NSP) task, a natural assumption is that the NSP task has enabled the model to understand the relationship between two sentences, the relationship information is helpful to model the context information for document-level machine translation.

In this work, we propose to extend the Transformer model to take advantage of BERT document-level contextual representation. We use the pre-trained BERT as a context encoder to achieve document-level representation, which is then integrated into both the encoder and the decoder of Transformer NMT model. We use a multi-head attention mechanism and context gate to control how each layer interacts with BERT context representations adaptively.

We conducted experiments on two document-level machine translation datasets. Experimental results show that our proposed model can outperform Transformer baselines and previous state-of-the-art document-level NMT models on BLEU score. Also, we perform an ablation study showing that the BERT context encoder can capture document-level context representation to improve translation performance.

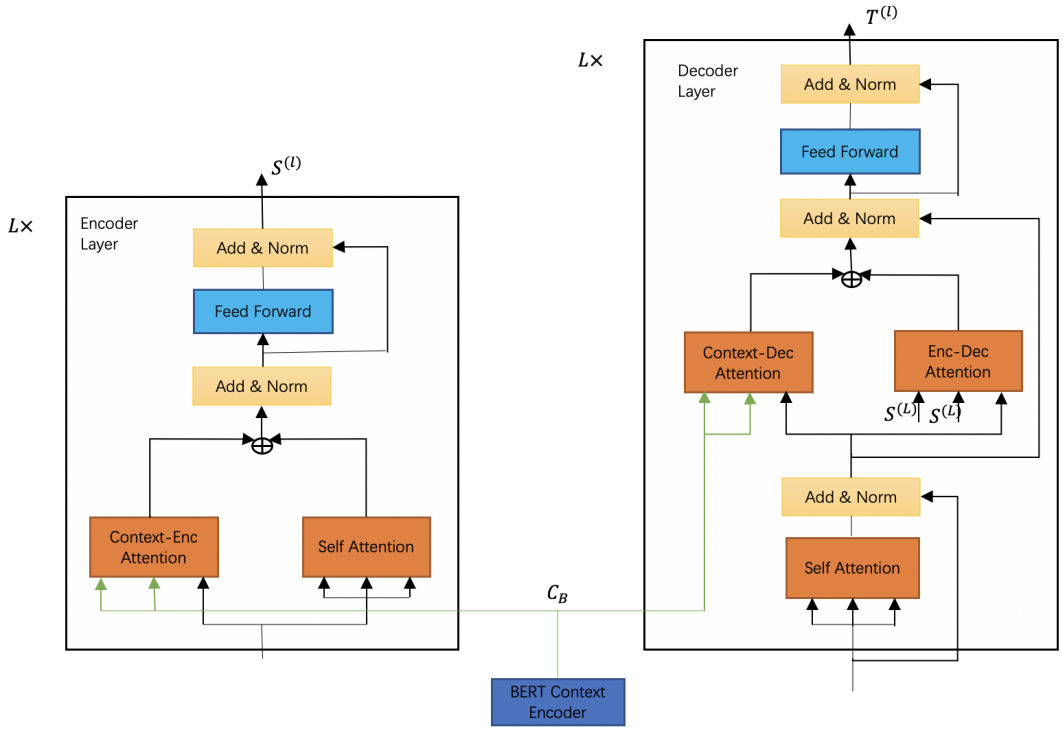


Figure 1: Illustration of using BERT as context encoder for document-level NMT model. C_B denote the output of BERT context encoder, $S^{(L)}$ denote the last layer output of Transformer encoder

2 Approach

2.1 Problem Statement

Formally, denote $\mathbf{X} = x_1, x_2, \ldots, x_N$ as a source-language document with N source sentences. The corresponding target-language document is denoted by $\mathbf{Y} = y_1, y_2, \ldots, y_N$. We assume that (x_i, y_i) is a parallel sentence pair.

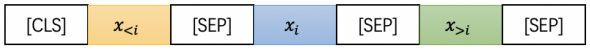
Following (Zhang et al., 2018), we omit the target-side document-level context $y_{<i}$ because of the translation error propagation problem, and source side document-level context $x_{<i}$ conveys the same information with $y_{<i}$. Therefore, the probability can be approximated as:

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

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

2.2 BERT Context Encoder

The context encoder is a BERT model. The input x_{ctx} of BERT is the concatenation of current sentence x_i and document-level context sentences $(x_{<i}, x_{>i})$ as follows:



Where [CLS] and [SEP] are special tokens for BERT. The context input x_{ctx} is encoded by BERT into document-level context representation $\mathbf{C}_B = BERT(x)$. $\mathbf{C}_B$ is the output from the last layer of BERT.

2.3 BERT Context Representation Integration

Inspired by (Zhang et al., 2018; Zhu et al., 2020), we use multi-head attention to integrate BERT context representation $\mathbf{C}_B$ into both the encoder and the decoder of Transformer NMT model.

2.3.1 Integration into the Encoder

As shown in Figure 1, we follow Vaswani et al. (2017) using a stack of L identical layers to encode x_i. Every layer consists of two attention models with different parameters. The first attention model is a multi-head self-attention:

$$\mathbf{B}^{(l)} = MultiHead\left(\mathbf{S}^{(l-1)}, \mathbf{S}^{(l-1)}, \mathbf{S}^{(l-1)}\right) \quad (2)$$

where $\mathbf{S}^{(0)}$ denotes the word embedding of sequence x_i. The second attention model is context

attention that integrate BERT document-level context into the encoder:

$$\mathbf{D}^{(l)} = MultiHead\left(\mathbf{S}^{(l-1)}, \mathbf{C}_B, \mathbf{C}_B\right) \quad (3)$$

If we directly combine the outputs of the two attention mechanisms, the influence of document-level context will be enhanced in an uncontrolled way as the context information will be added to every layer. Also, different source sentences require different amount of context information for translation. Inspired by context gate in Werlen et al. (2018); Zhang et al. (2018), we propose to use context gate to combine the output of the two attention mechanisms.

$$\begin{aligned} g^l &= \sigma\left(W_g^l\left[\mathbf{B}^{(l)}, \mathbf{D}^{(l)}\right] + b_g^l\right) \\ \mathbf{A}^{(l)} &= g^l \odot \mathbf{B}^{(l)} + \left(1 - g^l\right) \odot \mathbf{D}^{(l)} \end{aligned} \quad (4)$$

Where σ is a sigmoid function. Then the combination is further processed by a position-wise fully connected feed-forward neural network $FFN(.)$:

$$\mathbf{S}^{(l)} = FFN(\mathbf{A}^{(l)}) \quad (5)$$

$\mathbf{S}^{(l)}$ is the representation for the source sentence x_i and its context at the l-th layer.

2.3.2 Integration into the decoder

Similar to the encoder layer, we use context gate and attention mechanism to integrate the BERT document-level context representation into standard Transformer decoder. In the l-th layer,

$$\begin{aligned} \mathbf{E}^{(l)} &= MultiHead\left(\mathbf{T}^{(l-1)}, \mathbf{T}^{(l-1)}, \mathbf{T}^{(l-1)}\right) \\ \mathbf{F}^{(l)} &= MultiHead\left(\mathbf{E}^{(l)}, \mathbf{C}_B, \mathbf{C}_B\right) \\ \mathbf{G}^{(l)} &= MultiHead\left(\mathbf{E}^{(l)}, \mathbf{S}^{(L)}, \mathbf{S}^{(L)}\right) \\ d^l &= \sigma\left(W_d^l\left[\mathbf{F}^{(l)}, \mathbf{G}^{(l)}\right] + b_d^l\right) \\ \mathbf{H}^{(l)} &= d^l \odot \mathbf{F}^{(l)} + \left(1 - d^l\right) \odot \mathbf{G}^{(l)} \\ \mathbf{T}^{(l)} &= FFN(\mathbf{H}^{(l)}) \end{aligned} \quad (6)$$

After achieving the final representations of the last decoder layer $\mathbf{T}^{(L)}$, the output probability of the current target sentence y_i are computed as:

$$\begin{aligned} & p\left(y_i \mid x_i, x_{<i}, x_{>i}\right) \\ &= \textstyle\prod_t p\left(y_{i,t} \mid y_{i,\leq t}, x_i, x_{<i}, x_{>i}\right) \\ &= \textstyle\prod_t softmax\left(E\left[y_{i,t}\right]^\top \mathbf{T}_{i,t}^L\right) \end{aligned} \quad (7)$$

Dataset	Sent No.	Doc len avg
TED	0.21M / 9K / 2.3K	121.4 / 96.4 / 98.7
News	0.24M / 2.2K / 3K	38.9 / 26.8 / 19.4

Table 1: Statistics of the train/valid/test corpora.

3 Experiments

3.1 Dataset

We use two English-German datasets as the benchmark datasets, which are TED and News. The corpora statistics are shown in Table 1.

- **TED**: This corpus is from the IWSLT 2017 MT track (Cettolo et al., 2012) aligned at the sentence level. Every TED talk is treated as a document.
- **News Commentary**: This corpus is from document-delimited News Commentary v11 [1] aligned at the sentence level.

We obtain the processed datasets from Maruf et al. (2019)[2]. We use the same train/valid/test datasets with Maruf et al. (2019), so that our results can be compared with previous work. We use the script of Moses toolkit[3] to tokenize the sentence. We use byte pair encoding (Sennrich et al., 2016) to segment all sentences with 30K merge operations. The evaluation metrics is BLEU (Papineni et al., 2002).

3.2 Implementation Details

Firstly, we train a Transformer sentence-level NMT model until convergence, then use the obtained model to initialize our proposed document-level model. The context encoder attention and context decoder attention are randomly initialized. The pre-trained BERT model is *bert_base_uncased*. When training our proposed document-level model, the parameter of the BERT encoder is not trainable. To balance the accuracy and the computation cost, we only use one previous sentence as the context.

We use the same model configuration with the setting of the Maruf et al. (2019). For the Transformer NMT model, the hidden size is 512, and the FFN layer dimension is 2048. The number of layers is 4; the number of attention head is 8. The

[1] http://www.casmacat.eu/corpus/news-commentary.html
[2] https://github.com/sameenmaruf/selective-attn
[3] https://github.com/moses-smt/mosesdecoder

Model	TED	News
HAN (Werlen et al., 2018)	24.58	25.03
SAN (Maruf et al., 2019)	24.62	24.84
QCN (Yang et al., 2019b)	25.19	22.37
Doc-Transformer (Zhang et al., 2018)	24.01	22.42
Transformer (Vaswani et al., 2017)	23.28	22.78
Flat-Transformer (Ma et al., 2020)	24.87	23.55
+BERT	**26.61**	24.52
BERT-fused (Zhu et al., 2020)	25.59	25.05
Our Reproduced Transformer	23.99	22.50
Our Proposed Model	26.23	**26.55**

Table 2: BLEU scores on the two document-level machine translation benchmarks

dropout (Srivastava et al., 2014) is 0.1 for sentence model and 0.2 for document-level model.

We use the Adam (Kingma and Ba, 2014) optimizer and the same learning rate schedule strategy as (Vaswani et al., 2017) with 4000 warmup steps. The batch size is limited to 4000 tokens. We also apply label smoothing to the cross-entropy loss, and the smoothing rate is 0.1. Our Transformer implementation is based on Fairseq (Ott et al., 2019).

3.3 Experimental results

We list the results of our experiments in Table 2, comparing six context-aware NMT models. For Document-aware Transformer (Zhang et al., 2018), Hierarchical Attention NMT (Werlen et al., 2018), Selective Attention NMT (Maruf et al., 2019) and Query-guided Capsule Network (Yang et al., 2019b), Flat-Transformer (Ma et al., 2020), using BERT to initialize the encoder of Flat-Transformer(+BERT). Most of the previous work's results are from Ma et al. (2020), except BERT-fused (Zhu et al., 2020). The result of BERT-fused (Zhu et al., 2020) is my re-implementation using the current sentence and one previous sentence as BERT input. The reproduced Transformer uses the 4-layers setting, which is the same as our proposed model.

As shown in Table 2, by leveraging BERT document-level context representation, our proposed model obtains 2.24 and 4.05 gains over our reproduced sentence-level Transformer baselines in BLEU score on TED and News datasets, respectively. Among them, our model achieves new state-of-the-art results on the News dataset, showing the superiority of exploiting BERT document-level context representation.

Integration	BLEU
none	22.50
encoder	25.65
decoder	25.55
both	26.55

Table 3: Effect of context integration. "none" means no BERT context representation is integrated, "encoder" means BERT context representation is only integrated into the encoder, "decoder" means BERT context representation is only integrated into the decoder, "both" means BERT context representation is integrated into both the encoder and the decoder.

Our model achieved significant improvement on the News dataset, but relatively smaller gains on the TED dataset and haven't achieved state-of-the-art performance. Since BERT is pre-trained on BooksCorpus and Wikipedia, and the document in News dataset is more similar to the pre-training corpus, BERT can better encode context information on News dataset.

3.4 Ablation study

Effect of Context Integration Table 3 shows the effect of integrating BERT context representation into the encoder and the decoder. We can find that integrating BERT context representation into the encoder brings more improvements, it is also beneficial to integrate representation into the decoder. The results indicate that the BERT context representation should be integrated into both encoder and decoder to achieve better performance.

Does the BERT encoder really capture the contextual information? Yes. Li et al. (2020) claims that the improvements of the multi-encoder

News	BLEU
Context	26.55
Random	25.96
Fixed	26.14

Table 4: BLEU scores using three context inputs

document-level NMT approach is not from leveraging of contextual information, instead, it is from the noise generated by the context encoder that can provide richer training signals. To investigate whether the BERT context encoder has captured contextual information, we follow the experimental setting in Li et al. (2020) presenting three types input for the BERT context encoder and make experiments on News dataset.

- *Context*: Concatenation of the previous sentence and the current sentence.
- *Random*: Concatenation of a sentence consisting of words randomly selected from the source vocabulary and the current sentence.
- *Fixed*: Concatenation of a fixed sentence and the current sentence.

As shown in Table 4, the performance of *Random* and *Fixed* decrease due to the incorrect context, which is different from the result in Li et al. (2020). This indicates that our proposed model can really capture the contextual information. Although the performance of *Random* and *Fixed* decreases, they can still outperform the standard Transformer model significantly. This is because current sentence usually plays a more important role in target sentence generation, and our proposed model can leverage the representation of current sentence obtained by BERT as extra representation.

4 Related Work

Document-level NMT Document-level NMT models incorporate the document-level contextual information to generate more consistent and coherent translations compared with sentence-level NMT models. Most of the existing document-level NMT models can be divided into two categories: Uni-encoder models (Tiedemann and Scherrer, 2017; Li et al., 2019; Ma et al., 2020) and dual-encoder models (Voita et al., 2018; Zhang et al., 2018; Werlen et al., 2018; Maruf et al., 2019; Yang et al., 2019b). Uni-encoder models (Tiedemann and Scherrer, 2017; Li et al., 2019; Ma et al., 2020) take the concatenation of contexts and source sentences as the input. Dual-encoder (Voita et al., 2018; Zhang et al., 2018; Werlen et al., 2018; Maruf et al., 2019; Yang et al., 2019b) models integrate an additional encoder to incorporate the contextual information into standard NMT models. Our proposed model can be categorised as a dual-encoder model. More recently, Li et al. (2020) indicates that in dual-encoder document-level NMT models, the BLEU score improvement is not attributed to the use of contextual information. We have shown that our model can really capture the contextual information to improve the BLEU score.

BERT for Neural Machine Translation Recently, some works tried to apply BERT into NMT. Song et al. (2019) proposed MASS pre-training, showing promising results in unsupervised NMT. Yang et al. (2019a); Weng et al. (2019); Chen et al. (2020) leverage knowledge distillation to acquire knowledge from BERT to NMT. Li et al. (2019); Ma et al. (2020) use BERT to initialize parameters of document-level NMT model encoder. BERT-fused model (Zhu et al., 2020) exploits the representation from BERT by integrating it into all layers of Transformer model. BERT-fused model can also be extended to document-level NMT, but our work is different in the modeling and experimental part. While Zhu et al. (2020) are mainly focusing on improving sentence-level machine translation performance, they proposed a drop-net trick to combine the output of BERT encoder and the standard Transformer encoder, our proposed context gate combination can better leverage document-level context information since it is more correspond to the fact that different source sentences require a different amount of context information for translation.

5 Conclusion

We have presented a method for leveraging BERT to capture contextual information for document-level neural machine translation. Experiments on two document-level machine translation tasks demonstrate the effectiveness of our approach. Besides, we have shown that our approach can really capture the context information to improve the translation performance.

For future work, we plan to compress our model into a light version to leverage more context sen-

tences. Also, we plan to do experiments on large-scale datasets and some other language pairs like Chinese-English.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of european association for machine translation*, pages 261–268.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in bert for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7893–7905.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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

Liangyou Li, Xin Jiang, and Qun Liu. 2019. Pretrained language models for document-level neural machine translation. *arXiv preprint arXiv:1911.03110*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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 (Volume 1: Long Papers)*, pages 1715–1725.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2019. Acquiring knowledge from pre-trained model to neural machine translation. *arXiv preprint arXiv:1912.01774*.

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*, pages 2947–2954.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li. 2019a. Towards making the most of bert in neural machine translation. *arXiv preprint arXiv:1908.05672*.

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

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*, pages 533–542.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.