Hierarchical Modeling of Global Context for Document-Level Neural Machine Translation

Xin Tan, Longyin Zhang, Deyi Xiong, Guodong Zhou*
School of Computer Science and Technology, Soochow University, Suzhou, China
{annieT.x,zzlynx}@outlook.com
{dyxiong,gdzhou}@suda.edu.cn

Abstract

Document-level machine translation (MT) remains challenging due to the difficulty in efficiently using document context for translation. In this paper, we propose a hierarchical model to learn the global context for document-level neural machine translation (NMT). This is done through a sentence encoder to capture intra-sentence dependencies and a document encoder to model document-level inter-sentence consistency and coherence. With this hierarchical architecture, we feedback the extracted global document context to each word in a top-down fashion to distinguish different translations of a word according to its specific surrounding context. In addition, since large-scale in-domain document-level parallel corpora are usually unavailable, we use a two-step training strategy to take advantage of a large-scale corpus with out-of-domain parallel sentence pairs and a small-scale corpus with in-domain parallel document pairs to achieve the domain adaptability. Experimental results on several benchmark corpora show that our proposed model can significantly improve document-level translation performance over several strong NMT baselines.

1 Introduction

Due to its flexibility and much less demand of manual efforts for feature engineering, neural machine translation (NMT) has achieved remarkable progress and become the *de-facto* standard choice in machine translation. During the last few years, a variety of NMT models have been proposed to reduce the quality gap between human translation and machine translation (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017). Among them, the Transformer model (Vaswani et al., 2017) has achieved the state-of-the-art performance in sentence-level translation and results on news benchmark test sets have shown its "translation quality at human parity when compared to professional human translators" (Hassan et al., 2018). However, when turning to document-level translation, even the Transformer model yields a low performance as it translates each sentence in the document independently and suffers from the problem of ignoring document context.

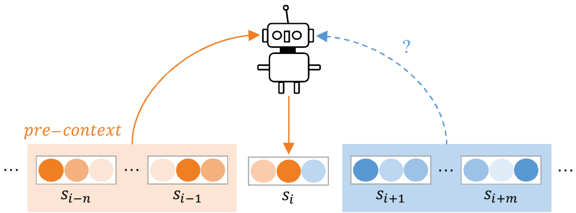

Figure 1: An illustration of document-level translation under the guidance of context.

To address above challenge, various extraction-based methods (Maruf and Haffari, 2018; Wang et al., 2017; Zhang et al., 2018; Miculicich et al., 2018) have been proposed to extract previous context (*pre-context*) to guide the translation of the current sentence s_i, as shown in Figure 1. However, when there exists a huge gap between the *pre-context* and the context after the current sentence s_i, the guidance from *pre-context* is not sufficient for the NMT model to fully disambiguate the sentence s_i. On the one hand, the translation of the current sentence s_i may be inaccurate due to the one-sidedness of partial context. On the other hand, translating the succeeding sentences in the document may much suffer from the semantic bias due to the transmissibility of the improper *pre-context*.

To alleviate the aforementioned issues, we propose to improve document-level translation with the aid of global context, which is hierarchically extracted from the entire document with a sen-

[*] Corresponding author.

*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing*, pages 1576–1585,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

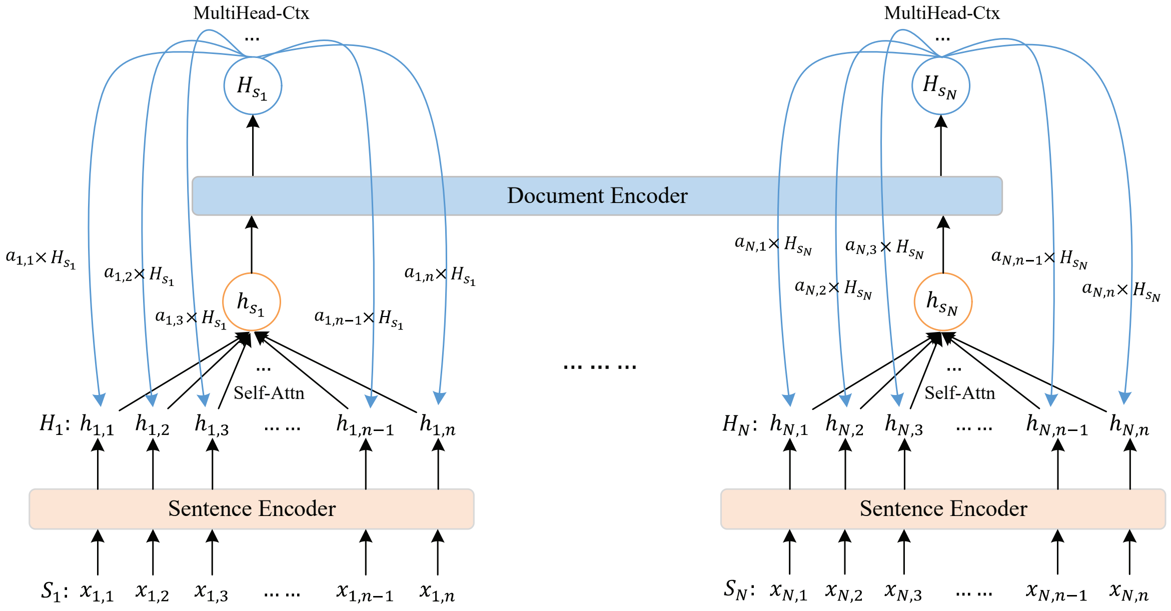

Figure 2: Diagram of the proposed hierarchical modeling of global document context (HM-GDC).

tence encoder modeling intra-sentence dependencies and a document encoder modeling document-level inter-sentence context. To avoid the issue of translation bias propagation caused by improper *pre-context*, we propose to extract global context from all sentences of a document once for all. Additionally, we propose a novel method to feed back the extracted global document context to each word in a top-to-down manner to clarify the translation of words in specific surrounding contexts. In this way, the proposed model can better translate each sentence under the guidance of the global context, thus effectively avoiding the transmissibility of improper *pre-context*. Furthermore, motivated by Zhang *et al.* (2018) and Miculicich *et al.* (2018) who exploit a large amount of sentence-level parallel pairs to improve the performance of document-level translation, we employ a two-step training strategy in taking advantage of a large-scale corpus of out-of-domain sentence-level parallel pairs to pre-train the model and a small-scale corpus of in-domain document-level parallel pairs to fine-tune the pretrained model.

We conduct experiments on both the traditional RNNSearch model and the state-of-the-art Transformer model. Experimental results on Chinese-English and German-English translation show that our proposed model can achieve the state-of-the-art performance due to its ability in well capturing global document context. It is also inferential to notice that our proposed model can explore a large dataset of out-of-domain sentence-level parallel pairs and a small dataset of in-domain document-level parallel pairs to achieve domain adaptability.

2 NMT with Hierarchical Modeling of Global Document Context

In this work, our ultimate goal is to incorporate the global document context into NMT to improve the performance of document-level translation. This is first achieved with the hierarchical modeling of global document context (HM-GDC) based on sentence-level hidden representation and document-level consistency and coherence modeling. Then, we integrate the proposed HM-GDC model into NMT models to help improve the performance of document-level translation.

2.1 Hierarchically Modeling Global Context

To avoid the one-sidedness of partial context and the transmissibility of the improper *pre-context* in previous studies, we take all sentences of the document into account and extract the global context once for all. Inspired by Sordoni *et al.* (2015), we build our HM-GDC model in a hierarchical way which contains two levels of encoder structure, i.e., the bottom sentence encoder layer to capture intra-sentence dependencies and the upper document encoder layer to capture document-level context. In this way, the global document context is captured for NMT. In order to make the translation of each word in specific surrounding context more robust, we propose to equip each word with global document context. This is done by backpropagating the extracted global context to each word in a top-down fashion, as shown in Figure 2. The following is the detailed description of the proposed HM-GDC model.

Sentence encoder. Given an input document with N sentences $(S_1, S_2, ..., S_N)$, the sentence encoder maps each word $x_{i,k}$ in the sentence S_i into the corresponding hidden state $h_{i,k}$, obtaining:

$$H_i = \text{SentEnc}(S_i) \tag{1}$$

where $S_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n})$ is the i^{th} sentence in the document, SentEnc is the sentence encoder function (corresponding to Bi-RNNs and multi-head self-attention (Vaswani et al., 2017) for the RNNSearch model and the Transformer model respectively), and $H_i = (h_{i,1}, h_{i,2}, \dots, h_{i,n}) \in \mathbb{R}^{D\times n}$ is the output hidden state.

Document encoder. Following the Transformer model (Vaswani et al., 2017), we employ the multi-head self-attention mechanism to determine the relative importance of different H_i. The model architecture of the document encoder here is the same as the sentence-level encoding stated before. And the document context is formulated as:

$$h_{S_i} = \text{MultiHead-Self}(H_i, H_i, H_i) \tag{2}$$

$$\tilde{h}_{S_i} = \sum\nolimits_{h \in h_{S_i}} h \tag{3}$$

$$H_S = \text{DocEnc}(\tilde{h}_S) \tag{4}$$

where MultiHead-Self$(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ is the multi-head self-attention mechanism corresponding to Self-Attn in Figure 2, $h_{S_i} \in \mathbb{R}^{D\times n}$, $\tilde{h}_{S_i} \in \mathbb{R}^{D\times 1}$, $\tilde{h}_S = (\tilde{h}_{S_1}, \tilde{h}_{S_2}, \dots, \tilde{h}_{S_N}) \in \mathbb{R}^{D\times N}$, DocEnc is the document-level encoding function (corresponding to Bi-RNNs and multi-head self-attention for the RNNSearch model and the Transformer model respectively), and $H_S = (H_{S_1}, H_{S_2}, \dots, H_{S_N}) \in \mathbb{R}^{D\times N}$ is the global document context.

Backpropagation of global context. To equip each word with global document context, we propose to assign the context information to each word in the sentence using another multi-head attention (Vaswani et al., 2017), which we refer to as the multi-head context attention (MultiHead-Ctx in Figure 2). And the context information assigned to the j^{th} word in the sentence S_i is detailed as:

$$\alpha_{i,j} = \text{MultiHead-Ctx}(h_{i,j}, h_{i,j}, H_{S_i}) \tag{5}$$

$$d\_ctx_{i,j} = \alpha_{i,j} H_{S_i} \tag{6}$$

where $\alpha_{i,j}$ is the attention weight assigned to the word and $d\_ctx_{i,j}$ is the corresponding context information distributed to the word.

2.2 Integrating the HM-GDC model into NMT

Different from previous works, we equip the representation of each word with global document context. The word representations are sequential in format, which makes it easy to integrate our proposed HM-GDC model into sequential models like RNNSearch and Transformer. In this section, we mainly introduce the process of integrating HM-GDC into the state-of-the-art Transformer model.

Integration in the Encoding Phase

Consider that the global document context is first extracted during the encoding phase and then distributed to each word in the document, as stated in Section 2.1. On this basis, we propose to employ the residual connection function (He et al., 2016) to incorporate the extracted global context information into the word representation. And the integrated representation of the j^{th} word in the i^{th} sentence is detailed as:

$$h\_ctx_{i,j} = h_{i,j} + \text{ResidualDrop}(d\_ctx_{i,j}, P) \tag{7}$$

where ResidualDrop is the residual connection function, $P = 0.1$ is the rate of residual dropout, $h_{i,j}$ is the corresponding hidden state of the word during the sentence encoding phase, $d\_ctx_{i,j}$ is the global document context assigned to the word, and $h\_ctx_{i,j}$ is the integrated representation of the word.

Integration in the Decoding Phase

With the help of the multi-head attention sub-layer in the decoder, the Transformer model is capable of well employing the information obtained from the encoder. Inspired by this, we introduce an additional sub-layer into the decoder that performs multi-head attention over the output of the document encoder, which we refer to as DocEnc-Decoder attention (shown in Figure 3). Different from (Vaswani et al., 2017), the keys and values of our DocEnc-Decoder attention come from the output of the document encoder. In this way, the global document context is well employed to supervise the decoding process. And Specially,

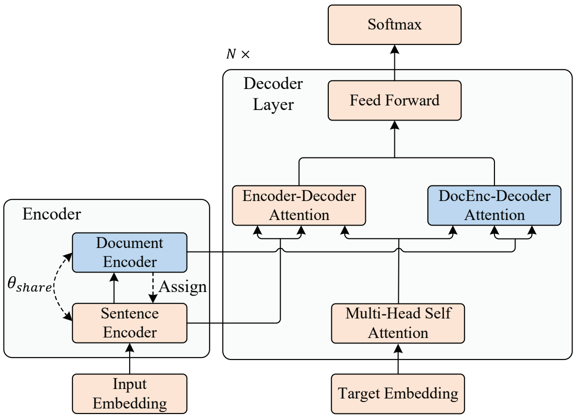

Figure 3: Integration of global document context into the decoder of the Transformer model.

the additional DocEnc-Decoder attention is formulated as:

$$C = [h\_ctx_1; h\_ctx_2; ...; h\_ctx_N] \quad (8)$$

$$G^{(n)} = \text{MultiHead-Attn}(T^{(n)}, C^{(n)}, C^{(n)}) \quad (9)$$

where h_ctx_i denotes the integrated representation of the i^{th} sentence, $C^{(0)} = C$ is the concatenated global document context, $T^{(n)}$ is the output of the multi-head self-attention sub-layer in the decoder. On this basis, we combine the outputs of both the Encoder-Decoder attention sub-layer and the DocEnc-Decoder attention sub-layer into one single output $H^{(n)}$:

$$H^{(n)} = E^{(n)} + G^{(n)} \quad (10)$$

where $E^{(n)}$ is the output of the Encoder-Decoder attention sub-layer, and $G^{(n)}$ is the output of the DocEnc-Decoder attention sub-layer.

2.3 Model Training

In document-level translation, the standard training objective is to maximize the log-likelihood of the document-level parallel corpus. However, due to the size limitation of document-level parallel corpora, previous studies (Zhang et al., 2018; Miculicich et al., 2018; Shen et al., 2016) use two-step training strategies to take advantage of large-scale sentence-level parallel pairs. Following their studies, we also take a two-step training strategy to train our model. Specially, we borrow a large-scale corpus with out-of-domain sentence-level parallel pairs D_s to pre-train our model first, and then use a small-scale corpus with in-domain document-level parallel pairs D_d to fine-tune it.

In this work, we follow Voita *et al.* (2018) to make the sentence and document encoders share the same model parameters. For the RNNSearch model, we share the parameters in the hidden layers of Bi-RNNs in the sentence and document encoders. For the Transformer model, we share the parameters of the multi-head self-attention layers in the sentence and document encoders.

During training, we first optimize the sentence-level parameters θ_s (colored in wheat in Figure 3) with the large-scale sentence-level parallel pairs D_s:

$$\hat{\theta_s} = \arg\max_{\theta_s} \sum_{<x,y>\in D_s} \log P(y|x;\theta_s) \quad (11)$$

Then, we optimize the document-level parameters θ_d (colored in pale blue in Figure 3) with the document-level parallel corpus D_d and fine-tune the pre-trained sentence-level parameters $\hat{\theta_s}$ as follows:

$$\hat{\theta_d} = \arg\max_{\theta_d} \sum_{<x,y>\in D_d} \log P(y|x;\theta_d,\hat{\theta_s}) \quad (12)$$

3 Experimentation

To examine the effect of our proposed HM-GDC model, we conduct experiments on both Chinese-English and German-English translation.

3.1 Experimental Settings

Datasets

For Chinese-English translation, we carry out experiments with sentence- and document-level corpora on two different domains: news and talks. For the sentence-level corpus, we use 2.8M sentence pairs from news corpora LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and LDC2004T08 (Hongkong Hansards/Laws/News). We use the Ted talks corpus from the IWSLT 2017 (Cettolo et al., 2012) evaluation campaigns[1] as the document-level parallel corpus, including 1,906 documents with 226K sentence pairs. We use *dev2010* which contains 8 documents with 879 sentence pairs for development and *tst2012-tst2015* which contain 62 documents with 5566 sentence pairs for testing.

For German-English translation, we use the document-level parallel Ted talks corpus from the IWSLT 2014 (Cettolo et al., 2012) evaluation campaigns, which contain 1,361 documents with 172K

[1] https://wit3.fbk.eu

Method	*pre-training*	tst12	tst13	tst14	tst15	Avg
RNNSearch (Bahdanau et al., 2015)	×	15.50	16.40	14.69	16.88	15.87
Wang *et al.* (2017)	×	15.90	16.99	14.50	17.33	16.18
Ours	×	16.33	17.52	15.80	18.10	**16.94**
Transformer (Vaswani et al., 2017)	×	15.87	16.51	14.67	17.27	16.08
Zhang *et al.* (2018)	×	11.31	12.58	10.22	13.35	11.87
Ours	×	16.58	17.03	15.22	17.96	**16.70**
Transformer (Vaswani et al., 2017)	✓	14.63	16.72	14.43	16.25	15.51
Zhang *et al.* (2018)	✓	16.46	17.80	15.85	18.24	17.09
Ours	✓	16.94	18.31	16.21	19.07	**17.63**

Table 1: Performance (BLEU scores) comparison with the four baseline models on Chinese-English document-level translation. The first three rows are based on RNNSearch and the remaining rows are on top of Transformer. And the *p-value* between Ours and the other models are all less than 0.01.

sentence pairs as training data. We use *dev2012* which contains 7 documents with 1172 sentence pairs for development and *tst2013-tst2014* which contain 31 documents with 2329 sentence pairs for testing.

Model Settings

We integrate our proposed HM-GDC into the original Transformer model implemented by OpenNMT (Klein et al., 2017). Following the Transformer model (Vaswani et al., 2017), the hidden size and filter size are set to 512 and 2048 respectively. The numbers of layers in encoder and decoder are all set to 6. The multi-head attention mechanism of each layer contains 8 individual attention heads. We set both the source and target vocabulary size as 50K and each batch contains 4096 tokens. The beam size and dropout (Srivastava et al., 2014) rate are set to 5 and 0.1 respectively. Other settings with the Adam (Kingma and Ba, 2015) optimization and regularization methods are the same as the default Transformer model.

To comprehensively evaluate the performance of our proposed HM-GDC model, we integrate the HM-GDC into the standard RNNSearch model to serve as a supplementary experiment. For the RNNSearch network, we borrow the implementation from OpenNMT (Klein et al., 2017). The encoder and decoder layers are all set to 2, the size of the hidden layer is set to 500, and the batch size is set to 64. Same as the Transformer model, we use the most frequent 50K words for both source and target vocabularies. We borrow other settings from (Bahdanau et al., 2015). The evaluation metric for both tasks is case-insensitive BLEU (multi-bleu) (Papineni et al., 2002).

3.2 Experimental Results

In this paper, we compare our model with four strong baselines as shown in Table 1. Among them, the RNNSearch (Bahdanau et al., 2015) is a traditional RNN-based encoder-decoder model. Wang *et al.* (2017) propose to use a hierarchical model to extract partial document context based on the RNNSearch model. To compare with these two RNN-based works, we integrate our proposed HM-GDC model into the encoder of the RNNSearch model using the same method in Section 2.2 and keep other settings the same as the basic RNNSearch model. Different from RNN-based works, Vaswani *et al.* (2017) propose the Transformer model, which achieves the state-of-the-art performance in sentence-level translation with only attention mechanism. On this basis, Zhang *et al.* (2018) add an additional multi-head attention to extract partial document context to improve the Transformer model in document-level translation. In particular, Zhang *et al.* (2018) use a two-step training strategy in their work, so we also report the performance comparison with respect to the training strategy. For the RNNSearch and Transformer models, we run them with their default settings. And we reimplement the models of Wang *et al.* (2017) and Zhang *et al.* (2018) to conduct experiments on our datasets.

As shown in Table 1, we divide the results into two main groups, i.e., in the framework of RNNSearch (the first three rows) and Transformer (the remaining rows). The results in the first group reveal that our proposed model can significantly improve the RNNSearch model and can further improve the model of Wang *et al.* (2017) in document-level translation by 0.76 BLEU points. In addition, the results in the second group is further split into two parts depending on whether the *pre-training* strategy is used. For the first part, we train our model and the two baselines with only the small-scale document-level parallel corpus without *pre-training* (the first three rows in the second group). From the results, the model of Zhang *et al.* (2018) achieves much worse results (-4.21 BLEU points) when compared with the standard Transformer model, which is consistent with what they state in their paper. By contrast, our proposed model can achieve 0.62 BLEU points over the Transformer model, which indicates the robustness of our model. For the second part, to further compare with (Zhang et al., 2018), we use the *pre-training* strategy to take advantage of large-scale sentence-level parallel corpus D_s for these models (the last three rows). From the results, our proposed HM-GDC can significantly improve the performance of the Transformer model by 2.12 BLEU points and can further improve the performance of (Zhang et al., 2018) by 0.54 BLEU points.

From the overall results, it's not difficult to find that using a large-scale corpus with out-of-domain parallel pairs D_s to pre-train the Transformer model results in worse performance due to domain inadaptability (the first and the fourth row in the second group). By contrast, our proposed model can effectively eliminate this domain inadaptability (the third and sixth row in the second group). In general, our proposed HM-GDC model is robust when integrated into frameworks like RNNSearch and Transformer and it can help improve the performance of document-level translation.

4 Analysis and Discussion

To further demonstrate the effectiveness of our proposed HM-GDC model, we illustrate several experimental results in this section and give our analysis on them.

4.1 The Effect of HM-GDC Integration

As the state-of-the-art Transformer model is well-designed in the model structure, an analysis of the integration in Transformer is thus necessary. Therefore, we perform experiments on Chinese-English document-level translation for analyzing. Table 2 illustrates the effectiveness of integrating our proposed HM-GDC model into the encoder, decoder and both sides of the Transformer model with respect to the number of layers in the multi-head self-attention in the document encoder (see Section 2.1).

N	Encoder	Decoder	Both
1	17.34	**17.54**	17.49
2	17.30	17.43	17.56
3	**17.39**	**17.54**	17.45
4	17.27	17.45	17.55
5	17.31	17.49	**17.63**
6	17.25	17.40	17.58

Table 2: The effect of integrating HM-GDC into Transformer with respect to the layer number (N) of the self-attention in the document encoder. The results here refer to the average BLEU scores of test sets.

Model	tst13	tst14	Avg
Baseline	27.89	23.75	25.82
Ours	28.58	24.85	**26.72**

Table 3: Comparison with the Transformer model on German-English document-level translation. And the *p-value* between Ours and Baseline is less than 0.01.

From the results, the overall performance of integrating HM-GDC into both the encoder and decoder is better than integrating it into the encoder or decoder only. However, the layer number of the multi-head self-attention does not make much difference in our experiments. It shows that when the HM-GDC is integrated into both the encoder and decoder and the layer number equals to 5, the Transformer model can achieve a relatively better performance.

4.2 Different Language Pairs

In this paper, we aim to propose a robust document context extraction model. To achieve this goal, we perform experiments on different language pairs to further illustrate the effectiveness of our proposed HM-GDC model. Table 3 shows the performance of our model on German-English document-level translation and the baseline here refers to the Transformer model. For clarity, we only use the German-English document-level parallel corpus to train these two models without *pre-training*. From the results, our proposed HM-GDC model can help improve the Transformer model on German-English document-level translation by 0.90 BLEU points. The experimental re-

sent	*doc*	tst12	tst13	tst14	tst15	Avg
×	×	16.58	17.03	15.22	17.96	16.70
✓	×	16.94	18.31	16.21	19.07	**17.63**
✓	✓	20.08	21.04	19.48	22.46	**20.77**

Table 4: Results of our model with and without *pre-training* on Chinese-English document-level translation. *sent* refers to the sentence-level parallel corpus and *doc* refers to the document-level parallel corpus. ✓means that the corresponding corpus is used for pre-training while × means not.

	Chinese-English					German-English		
Model	tst12	tst13	tst14	tst15	Avg	tst13	tst14	Avg
Baseline	58.06	55.10	51.71	52.48	54.34	86.47	84.78	85.63
Ours	59.29	55.48	52.95	53.22	**55.24**	87.36	85.64	**86.50**

Table 5: Evaluation on pronoun translation of Chinese-English and German-English document-level translation. The baseline model refers to the Transformer.

sults further validate the robustness of our model in different language pairs.

4.3 The Effect of Pre-training

Due to the size limitation of document-level parallel corpora, previous studies (Zhang et al., 2018; Miculicich et al., 2018; Shen et al., 2016) use two-step training strategies to take advantage of a large-scale corpus with sentence-level parallel pairs. Inspired by them, we use a two-step training strategy to train our model, which we refer to as the *pre-training* strategy (see Section 2.3). In this section, we perform the *pre-training* strategy on the HM-GDC integrated Transformer model to further analysis the ability of our model in utilizing resources of different domains. The results of our model with and without the *pre-training* strategy are shown in Table 4. The first row in the table gives the result of our model without *pre-training*, where only the talks-domain document-level parallel corpus D_d are used to train the model. The remaining rows give the results of our model with the *pre-training* strategy, where we first use the large-scale sentence-level parallel pairs D_s to pre-train the model and then the small-scale talks-domain document-level parallel corpus D_d to fine-tune the entire model.

From the results, the performance of our model is significantly improved by 0.93 BLEU points (the first two rows in Table 4) when the large-scale sentence-level parallel corpus is used for the pre-training process. In particular, when we use the mixed data of both sentence- and document-level parallel corpora[2] to first pre-train our model, the performance of our model is significantly improved by 5.26 BLEU points (the last row in Table 4). The overall results prove that our proposed model is robust and promising. It can significantly improve the performance of document-level translation when a two-step training strategy is used.

4.4 Pronoun & Noun Translation

To intuitively illustrate how the translation performance is improved by our proposed HM-GDC model, we conduct a further analysis on pronoun and noun translation.

For the pronoun translation, we evaluate the coreference and anaphora using the reference-based metric: the accuracy of pronoun translation (Miculicich Werlen and Popescu-Belis, 2017) in Chinese-English and German-English translation as shown in Table 5. From the results, our proposed HM-GDC model can well improve the performance of pronoun translation in both corpora due to the well captured global document context assigned to each word. Correspondingly, we display a translation example in Table 6 to further illustrate this. From the example, given the surrounding context, our proposed HM-GDC model can well infer the latent pronoun *it* and thus improve the translation performance of the Transformer model.

For the analysis of noun translation, we display another example in Table 7. From the example,

[2] We shuffle sentences in D_d to get sentence-level parallel pairs.

Source	dan xifang zhengfu ye tongyang dui tamen ziji zheyang zuo **ta**.
Reference	But western governments are doing **it** to themselves as well.
Baseline	But the western governments do the same for themselves.
Ours	But the western governments do **it** the same for themselves.

Table 6: An example of pronoun translation in the Chinese-English document-level translation. The word "ta" in bold in the source sentence is an omitted pronoun.

Pre-context	··· renhe yige youxiu de **chengxuyuan** dou hui zuo de ···
Rear-context	··· zai xiaweiyi de **IT** bumen shangban de ren kandao le ···
Source	ta xie le yige xiangduilaishuo bijiao xiao de **chengxu**
Reference	he wrote a modest little **app**
Baseline	he wrote a relatively small **procedure**
Ours	he wrote a relative smaller **program**

Table 7: An example of noun translation in the Chinese-English document-level translation. The baseline model here refers to the Transformer.

the word *chengxu* is translated into *procedure* and *program* by the Transformer model and the model integrated with HM-GDC respectively. Comparing with the reference translation, the word *program* translated by our model is more appropriate. Although the words *chengxuyuan* and *IT* in the global context provide essential evidence for an accurate translation of *chengxu*, it is hard for the baseline model to obtain the information. Different from previous works which do not use or use only partial document context, we propose to incorporate the HM-GDC into the NMT model to take global context into consideration and thus it can safely disambiguate those multi-sense words like *chengxu*.

5 Related Work

Recent years have witnessed a variety of approaches proposed for document-level machine translation. Most of existing studies aim to improve overall translation quality with the aid of document context. Among them, Maruf and Haffrai (2018), Wang *et al.* (2017), Zhang *et al.* (2018) and Miculicich *et al.* (2018) use extraction-based models to extract partial document context from previous sentences of the current sentence. In addition, Tu *et al.* (2018) and Kuang *et al.* (2018) employ cache-based models to selectively memorize the most relevant information in the document context. Different from above extraction-based models and cache-based models, there are also some works (Bawden et al., 2018; Voita et al., 2018) that pay much attention to discourse phenomena (Mitkov, 1999) related to document-level translation.

Although these approaches have achieved some progress in document-level machine translation, they still suffer from incomplete document context. Further more, most of previous works are based on the RNNSearch model, and only few exceptions (Zhang et al., 2018; Miculicich et al., 2018) are on top of the state-of-the-art Transformer model.

6 Conclusion

We have presented a hierarchical model to capture the global document context for document-level NMT. The proposed model can be integrated into both the RNNSearch and the state-of-the-art Transformer frameworks. Experiments on two benchmark corpora show that our proposed model can significantly improve document-level translation performance over several strong document-level NMT baselines. Additionally, we observe that pronoun and noun translations are significantly improved by our proposed HM-GDC model. In our future work, we plan to enrich our HM-GDC model to solve discourse phenomena such as (zero) anaphora.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) via Grant Nos. 61751206, 61673290, 61622209 and A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education

Institutions (PAPD). Also, we would like to thank the anonymous reviewers for their insightful comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *In Proceedings of the Third International Conference on Learning Representations (ICLR2015).*

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.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

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

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

Lesly Miculicich, 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 (EMNLP)*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

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

Ruslan Mitkov. 1999. Introduction: special issue on anaphora resolution in machine translation and multilingual nlp. *Machine translation*, 14(3-4):159–161.

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 (ACL).*

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562. ACM.

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.

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, 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 (ACL) (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 (EMNLP)*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.