Multi-Hop Transformer for Document-Level Machine Translation

Long Zhang[1,2,†], **Tong Zhang**[1,2,†], **Haibo Zhang**[3], **Baosong Yang**[3],
Wei Ye[1,*], **Shikun Zhang**[1]
[1] National Engineering Research Center for Software Engineering, Peking University
[2] School of Software and Microelectronics, Peking University
[3] Alibaba Group
{zhanglong418, zhangtong17, wye, zhangsk}@pku.edu.cn
{zhanhui.zhb, yangbaosong.ybs}@alibaba-inc.com

Abstract

Document-level neural machine translation (NMT) has proven to be of profound value for its effectiveness on capturing contextual information. Nevertheless, existing approaches 1) simply introduce the representations of context sentences without explicitly characterizing the inter-sentence reasoning process; and 2) feed ground-truth target contexts as extra inputs at the training time, thus facing the problem of exposure bias. We approach these problems with an inspiration from human behavior – human translators ordinarily emerge a translation draft in their mind and progressively revise it according to the reasoning in discourse. To this end, we propose a novel Multi-Hop Transformer (MHT) which offers NMT abilities to explicitly model the human-like draft-editing and reasoning process. Specifically, our model serves the sentence-level translation as a draft and properly refines its representations by attending to multiple antecedent sentences iteratively. Experiments on four widely used document translation tasks demonstrate that our method can significantly improve document-level translation performance and can tackle discourse phenomena, such as coreference error and the problem of polysemy.

1 Introduction

Neural machine translation (NMT) employs an end-to-end framework (Sutskever et al., 2014) and has advanced promising results on various sentence-level translation tasks (Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017; Wan et al., 2020). However, most of NMT models handle sentences independently, regardless of the linguistic context that may appear outside the current sentence (Tiedemann and Scherrer, 2017a). This makes NMT insufficient to fully resolve the typical context-dependent phenomena problematic, e.g. coreference (Guillou, 2016), lexical cohesion (Carpuat, 2009), as well as lexical disambiguation (Gonzales et al., 2017).

Recent studies (Tu et al., 2018; Maruf and Haffari, 2018; Maruf et al., 2019; Tan et al., 2019; Kim et al., 2019; Zheng et al., 2020; Chen et al., 2020; Sun et al., 2020; Ma et al., 2020) have proven to be effective on tackling discourse phenomena via feeding NMT with contextual information, e.g. source-side (Wang et al., 2017; Voita et al., 2018; Zhang et al., 2018) or target-side context sentences (Bawden et al., 2018; Miculicich et al., 2018). Despite their successes, these methods simply merge the representations of context sentences together, lacking a mechanism to explicitly characterize the inter-sentence reasoning upon the context. Another shortage in existing document-level NMT is the problem of exposure bias. Most of methods utilized the ground-truth target context for training but the generated translations for inference, leading to inconsistent inputs at training and testing time (Ranzato et al., 2015; Koehn and Knowles, 2017).

Intuitively, human translators tend to acquire useful context information from the reasoning process among sentences, thus figuring out the correct meaning when they encounter ambiguity during translation. Sukhbaatar et al. (2015) and Shen et al. (2017) empirically verified that modeling multi-hop reasoning among sentences benefits to the language understanding task, e.g text comprehension. Voita et al. (2019) showed that document-level NMT model can profit from relative positions with respect to context sentences, which to some extent confirms the importance of the relationship among sentences. Meanwhile, Xia et al. (2017) demonstrated that sentence-level NMT could be improved by a two-pass draft-editing process, of which the second-pass decoder refines the target sentence generated by a first-pass standard decoder.

Accordingly, we propose to improve document-

[†]These authors contributed equally to this work.
[*]Corresponding author.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3953–3963
June 6–11, 2021.

level NMT using a novel framework – Multi-Hop Transformer, which imitates draft-editing and reasoning process of human translators. Specifically, we implement an explicit reasoning process by exploiting source and target antecedent sentences with concurrently stacked attention layers, thus performing the progressive refinement on the representations of the current sentence and its translation. Besides, we leverage the draft to present context information on the target side during both training and testing, alleviating the problem of exposure bias.

We conduct experiments on four widely used document translation tasks: English-German and Chinese-English TED, English-Russian Opensubtitles, as well as English-German Europarl-7 datasets. Experimental results demonstrate that our method significantly outperforms both context-agnostic and context-aware methods. The qualitative analysis confirms the effectiveness of the proposed multi-hop reasoning mechanism on resolving many linguistic phenomena, such as word sense disambiguation and coreference resolution. Our contributions are mainly in:

- We propose the Multi-Hop Transformer. To the best of our knowledge, this is the first pioneer investigation that introduces multi-hop reasoning into document-level NMT.
- The proposed model takes target context drafts into account at the training time, which devotes to avoid the training-generation discrepancy.
- Our approach significantly improves document-level translation performance on four document-level translation tasks in terms of BLEU scores and solves some context-dependent phenomena, such as coreference error and polysemy.

2 Preliminary

Transformer NMT is an end-to-end framework to build translation models. Vaswani et al. (2017) propose a new architecture called Transformer which adopts self-attention network for both encoding and decoding. Both its encoder and decoder consist of multiple layers, each of which includes a multi-head self-attention and a feed-forward sub-layer. Additionally, each layer of the decoder applys a multi-head cross attention to capture information from the encoder. Transformer has shown superiority in a variety of NLP tasks. Therefore, we construct our models upon this advanced architecture.

Document-level NMT In order to correctly translate the sentence with discourse phenomena, NMT models need to look beyond the current sentence and integrate contextual sentences as auxiliary inputs. Formally, let $\mathbf{X} = (\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^I)$ be a source-language document composed of I sentences, where $\mathbf{x}^i = (x_1^i, x_2^i, ..., x_N^i)$ denotes the i^{th} sentence containing N words. Correspondingly, the target-language document also consists of I sentences, $\mathbf{Y} = (\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^I)$, where $\mathbf{y}^i = (y_1^i, y_2^i, ..., y_M^i)$ denotes the i^{th} sentence involving M words. Document-level NMT incorporates contextual information from both source side and target side to autoregressively generate the best translation result that has highest probability:

$$P_\theta(\mathbf{y}^i|\mathbf{x}^i) = \prod_{m=1}^{M} P_\theta(y_m^i|y_{<m}^i, \mathbf{x}^i, \mathbf{X}^{-i}, \mathbf{Y}^{-i}) \quad (1)$$

where $y_{<m}^i$ is the sequence of proceeding tokens before position m. $\mathbf{X}^{-i}$ and $\mathbf{Y}^{-i}$ denote the context sentences of the i^{th} sentence.

Related Work Several studies have explored multi-input models to leverage the contextual information from source-side (Jean et al., 2017; Kuang and Xiong, 2018) or target-side sentences (Kuang et al., 2018; Miculicich et al., 2018). For the former, Zhang et al. (2018) propose a new encoder to represent document-level context from previous source-side sentences . Tiedemann and Scherrer (2017b) and Junczys-Dowmunt (2019) utilize the concatenation of previous source-side sentences as input, while Voita et al. (2018) make use of gate mechanism to balance the weight between current source sentence and its context. For the latter, Miculicich et al. (2018) propose a hierarchical attention (HAN) framework to capture the target contextual information in the decoder. Bawden et al. (2018), Maruf and Haffari (2018) and Maruf et al. (2019) take both source-side and target-side context into account.

Motivation As seen, both of the existing methods simply introduce the context sentences without explicitly characterizing the inter-sentence reasoning. Intuitively, when humans have difficulty in translation like encountering ambiguity phenomenon, they could acquire more information

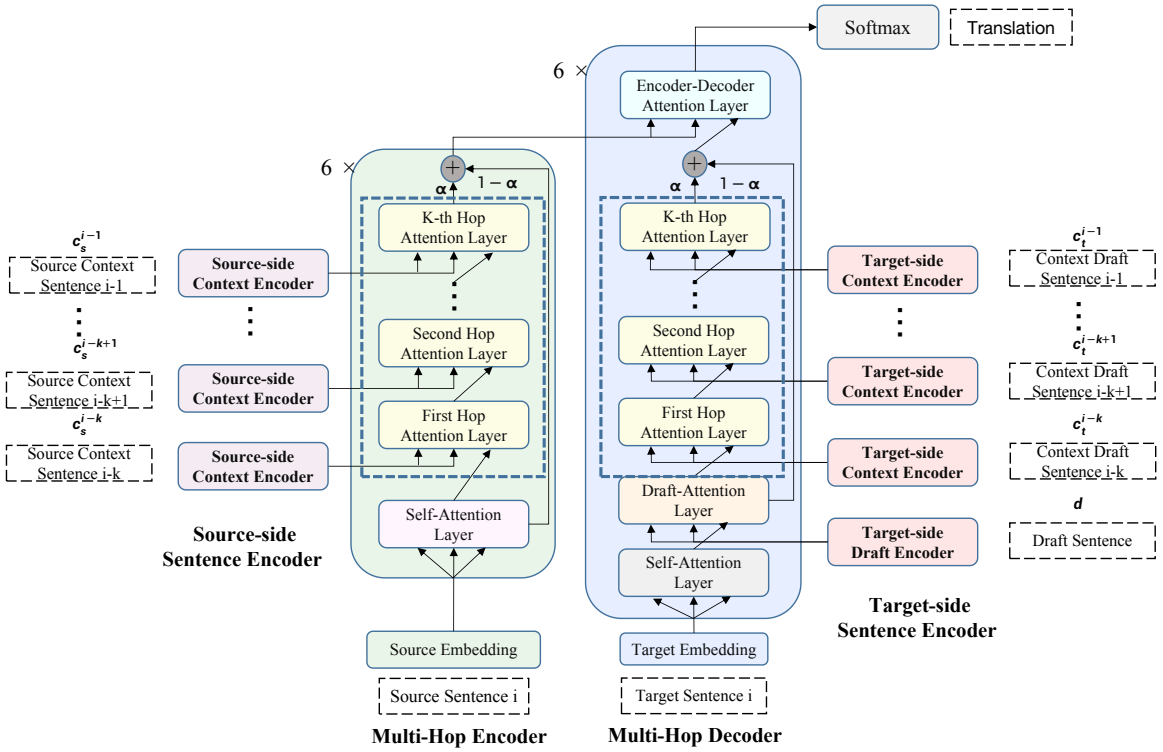

Figure 1: Illustration of Multi-Hop Transformer. c_s^{i-j} and c_t^{i-j} indicate the j^{th} previous sentence in the source side and target side respectively. d denotes the draft of current source sentence. All drafts are generated by a pre-trained sentence-level NMT model. The modules inside dashed box are the proposed multi-hop attention layers, which gradually refine the representation of current sentence. Finally, the context gate α is used to control the contextual information.

from the contexts sentence by sentence and then perform reasoning to figure out the exact meaning. We attribute that such reasoning process is also beneficial to machine translation task. Recent successes in text comprehension communities have to some extent supported our hypothesis (Hill et al., 2015; Kumar et al., 2016). For example, Sukhbaatar et al. (2015) propose a multi-hop end-to-end memory network, which can renew the query representation with multiple computational steps (which they term "hops"). Dhingra et al. (2016) extend an attention-sum reader to multi-turn reasoning with a gating mechanism. In addition, Shen et al. (2017) introduce multi-hop attention, which used multiple turns to effectively exploit and reason over the relation among queries and documents.

In this paper, we propose to bring the idea of multi-hop into document translation and aim at mimicking the multi-step comprehension and revising process of human translators. Contrast with those models for text comprehension which scan the query and document for multiple passes, our model iteratively focuses on different context sentences, which captures the inter-sentence reasoning semantics of contextual sentences to incrementally refine the representation of current sentence.

3 Multi-Hop Transformer

With this mind, we propose a novel method called Multi-Hop Transformer, which models the reasoning process among multiple contextual sentences in both source side and target side. The source-side contexts are directly acquired from the document. The target-side contexts, called target-side drafts in this paper, are generated by a sentence-level NMT model. These contexts are fed into the Multi-Hop Transformer with pre-trained encoders. The overall architecture of our proposed model is illustrated in Figure 1, which consists of three components:

- **Sentence Encoder:** This component contains two pre-trained encoders, one of which is called source-side sentence encoder and the other is called target-side sentence encoder. These encoders generate representations for

source-side contexts and target-side drafts respectively.

- **Multi-Hop Encoder:** We extend the original Transformer encoder with a novel multi-hop encoder to efficiently perform sentence-by-sentence reasoning on source-side contexts and generate the representation for the current sentence.

- **Multi-Hop Decoder:** Similarly, a multi-hop decoder is proposed to acquire information from the target-side drafts and models the translation probability distribution.

3.1 Sentence Encoder

We use multi-layer and multi-head self-attention architecture (Vaswani et al., 2017) to obtain the representations for source-side contexts and target-side drafts. Similar to the encoder of Transformer, sentence encoder contains a stack of six identical layers, each of which consists of two sub-layers. The first sub-layer is a multi-head attention(Q,K,V), which takes a query Q, a key K and a value V as inputs. The second sub-layer is a fully connected feed-forward network (FFN).

Source-Side Sentence Encoder. This encoder is utilized to generate the representations for source-side contexts, as shown in Figure 1.

For the current sentence $s = \mathbf{x}^i$ to be translated, we use the previous sentences $\mathbf{X}^{-i} = (\mathbf{x}^{i-k}, \mathbf{x}^{i-k+1}, ..., \mathbf{x}^{i-1})$ in the same document as the source-side context, specially denoted as c_s^{i-k}, c_s^{i-k+1}, ..., c_s^{i-1} for clarity. k is the context window size. For the j^{th} context, we obtain the $A^{(n)}_{c_s^{i-j}}$ which denotes the n^{th} hidden layer representation of c_s^{i-j} as follows:

$$A^{(n)}_{c_s^{i-j}} = \mathrm{MHA}(H^{(n-1)}_{c_s^{i-j}}, H^{(n-1)}_{c_s^{i-j}}, H^{(n-1)}_{c_s^{i-j}}) \quad (2)$$

where $n = 1, 2, ..., 6$. MHA represents the standard Multi-Head Attention function (Vaswani et al., 2017). j denotes the distance between the context sentence and current sentence.

Target-Side Sentence Encoder. Most existing works use ground-truth target-side contexts as the input of decoder during training (Voita et al., 2019). However, the target contexts at training and testing are drawn from different distributions, leading to the inconsistency between training and testing. To alleviate this problem, we instead make use of target-side context drafts generated from a pre-trained sentence-level translation model. Similar to source-side sentence encoder, this target-side context draft encoder is used to obtain the context representation $A^{(n)}_{c_t^{i-j}}$ of the j^{th} target-side draft c_t^{i-j}. Besides, we obtain a draft translation d of the current sentence from the pre-trained sentence-level translation model and use a target-side draft encoder to obtain the representation $A^{(n)}_d$.

3.2 Multi-Hop Encoder

The multi-hop encoder contains a stack of 6 identical layers, each of which contains the following sub-layers:

Self-Attention Layer. The first sub-layer makes use of multi-head self-attention to encode the information of current source sentence s and obtains the representation $A^{(n)}_s$.

Multi-Hop Attention Layer. The second sub-layer uses a multi-hop attention to perform sentence-by-sentence reasoning on c_s in sentence order as shown in Figure 1. Each reasoning step, also called a **hop**, is implemented by a multi-head attention layer. The first hop takes representation $A^{(n)}_s$ as the query and the representation $A^{(n)}_{c_s^{i-k}}$ of the previous k^{th} sentence as the key and value.

$$B^{(n)}_{s^{i-k}} = \mathrm{MHA}(A^{(n)}_s, A^{(n)}_{c_s^{i-k}}, A^{(n)}_{c_s^{i-k}}) \quad (3)$$

The other hops are implemented:

$$B^{(n)}_{s^{i-j}} = \mathrm{MHA}(B^{(n)}_{s^{i-j-1}}, A^{(n)}_{c_s^{i-j}}, A^{(n)}_{c_s^{i-j}}) \quad (4)$$

where $j = k-1, k-2, ..., 1$. j denotes the distance between the context sentence and current sentence.

Context Gating. The information of current source sentence is crucial in translation while the contextual information is auxiliary. In order to avoid excessive utilization of contextual information, a context gating mechanism (Tu et al., 2017; Yang et al., 2017, 2019) is introduced to dynamically control the weight between context sentences and current sentence:

$$\alpha = \sigma(W_a A^{(n)}_s + W_b B^{(n)}_{s^{i-1}}), \quad (5)$$

where σ is the logistic sigmoid function and α is the context gate. W_a and W_b denote the weight matrices of $A^{(n)}_s$ and $B^{(n)}_{s^{i-1}}$, respectively.

$$H^{(n)}_s = \alpha \odot A^{(n)}_s + (1 - \alpha) \odot B^{(n)}_{s^{i-1}} \quad (6)$$

Finally, we obtain the representation $Enc_s = H_s^{(6)}$ as the final output of the multi-hop encoder.

3.3 Multi-Hop Decoder

Similarly, the multi-hop decoder involves a stack of 6 identical layers. Each of them contains five sub-layers.

Self-Attention Layer. The first sub-layer utilizes multi-head self-attention to encode the information of current target sentence t and obtains the representation $A_t^{(n)}$.

Draft-Attention Layer. Inspired by Xia et al. (2017), we introduce the complete draft d translated from current source sentence by a sentence-level NMT. Then this draft representation $A_d^{(n)}$ is encoded by the target-side draft encoder in Section 3.1. The draft attention is achieved by multi-head attention:

$$F_t^{(n)} = \mathrm{MHA}(A_t^{(n)}, A_d^{(n)}, A_d^{(n)}). \quad (7)$$

Multi-Hop Attention Layer. Similar to the encoder, a multi-hop reasoning process is performed on the target-side contexts. The target-side drafts are generated from corresponding source sentences by a pre-trained sentence-level NMT model. The first hop takes representation $F_t^{(n)}$ as the query and the representation $A_{c_t^{i-k}}^{(n)}$ of the previous k^{th} draft as the key and value.

$$B_{t^{i-k}}^{(n)} = \mathrm{MHA}(F_t^{(n)}, A_{c_t^{i-k}}^{(n)}, A_{c_t^{i-k}}^{(n)}) \quad (8)$$

The other hops are achieved:

$$B_{t^{i-j}}^{(n)} = \mathrm{MHA}(B_{t^{i-j-1}}^{(n)}, A_{c_t^{i-j}}^{(n)}, A_{c_t^{i-j}}^{(n)}) \quad (9)$$

where $j = k-1, k-2, ..., 1$. j denotes the distance between the context draft and current target draft.

Context Gating. Same as the multi-hop encoder, the final output of multi-hop decoder is computed as:

$$G_t^{(n)} = \alpha \odot F_t^{(n)} + (1-\alpha) \odot B_{t^{i-1}}^{(n)} \quad (10)$$

where α is used to regulate the weight of target-side contextual information.

Encoder-Decoder Attention Layer. Finally, we use an encoder-decoder attention layer to integrate the output of multi-hop encoder Enc_s with the current target representation $G_t^{(n)}$.

$$H_t^{(n)} = \mathrm{MHA}(G_t^{(n)}, Enc_s, Enc_s) \quad (11)$$

where $H_t^{(n)}$ represents the final representation of decoder.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of the proposed MHT, we conduct experiments on four widely used document translation tasks, including the TED Talk (Cettolo et al., 2012) with two language pairs, Opensubtitles (Maruf et al., 2018) and Europarl7 (Maruf et al., 2018). All datasets are tokenized and truecased with the Moses toolkit (Koehn et al., 2007), and splited into sub-word units with a joint BPE model (Sennrich et al., 2016) with 30K merge operations. The datasets are described as follows:

- **TED Talk (English-German)**: We use the dataset of IWSLT 2017 MT English-German track for training, which contains transcripts of TED talks aligned at sentence level. *dev2010* is used for development and *tst2016-2017* for evaluation. Statistically, there are 0.21M sentences in the training set, 9K sentences in the development set, and 2.3K sentences in the test set.
- **TED Talk (Chinese-English)**: We use the corpus consisting of 0.2M sentence pairs extracted from IWSLT 2014 and 2015 Chinese-English track for training. *dev2010* involves 0.8K sentences for development and *tst2010-2013* contains 5.5K sentences for test.
- **Opensubtitles (English-Russian)**: We make use of the parallel corpus from Maruf et al. (2018). The training set includes 0.3M sentence pairs. There are 6K sentence pairs in development set, and 9K in test set.
- **Europarl7 (English-German)**: The raw Europarl v7 corpus (Koehn, 2005) contains SPEAKER and LANGUAGE tags where the latter indicates the language the speaker was actually using. We process the raw data and extract the parallel corpus as same as Maruf

Method	TED		Opensubtitles	Europarl7	Params	AVG
	En → De	Zh → En	En → Ru	En → De		
Transformer*	24.55	18.36	19.46	30.18	50M	23.14
CA-Transformer†	25.04	18.77	20.21	30.67	72M	23.67
(Maruf et al., 2018)†	-	-	19.13$^{\diamond}$	26.49$^{\diamond}$	-	-
CA-HAN†	25.70	18.79	20.08	26.61	70M	22.79
(Maruf et al., 2019)†	24.62$^{\diamond}$	-	-	-	54M$^{\diamond}$	-
CADec†	26.08	19.01	19.46	30.36	91M	23.98
MHT (Ours)†	**26.22**	**19.52**	**20.46**	**31.25**	80M	**24.36**

Table 1: BLEU scores on TED Talk, Opensubtitles and Europarl7 tasks. $\star$ mark indicates context-agnostic NMT models and † mark indicates context-aware NMT models. AVG indicates the average BLEU scores on test sets. $\diamond$ denotes that the value is reported by the corresponding paper. Our MHT model achieves better performance than both context-agnostic and context-aware strong baseline on four examined tasks. The significance tests are conducted for testing the robustness of approaches, and the results are statistically significant with $p < 0.05$.

et al. (2018). 0.1M sentence pairs are used for training, 3K sentence pairs for development, and 5K sentence pairs for evaluation.

4.2 Baselines

We compare our model against four NMT systems as follows:

- **Transformer**: The state-of-the-art context-agnostic NMT model (Vaswani et al., 2017).
- **CA-Transformer**: A context-aware transformer model (CA-Transformer) with an additional context encoder to incorporate document contextual information into model (Zhang et al., 2018).
- **CA-HAN**: A context-aware hierarchical attention networks (CA-HAN) which integrate document contextual information from both source side and target side (Miculicich et al., 2018).
- **CADec**: A two-pass machine translation model (Context-Aware Decoder, CADec) which first produces a draft translation of the current sentence, then corrects it using context (Voita et al., 2019).

4.3 Implementation Details

Our model is implemented on the open-source toolkit **Thumt** (Zhang et al., 2017). Adam optimizer (Kingma and Ba, 2014) is applied with an initial learning rate 0.1. The size of hidden dimension and feed-forward layer are set to 512 and 2048 respectively. Encoder and decoder have 6 layers with 8 heads multi-head attention. Dropout is 0.1 and batch size is set to 4096. Beam size is 4 for inference. Translation quality is evaluated by the traditional metric BLEU (Papineni et al., 2002) on tokenized text. Context window size is set to 3, consistent with the experiments in Section 5.2.

To initialize the source-side sentence encoder in Section 3.1, a sentence-level NMT model is trained from source language to target language using the corresponding datasets without additional corpus. The encoder of this trained model is used to initialize the source-side context encoder. Also, we utilize the trained model to translate the source-side sentences and obtain the target-side drafts. Similarly, we train a sentence-level model from target language to source language to initialize the target-side encoders in Section 3.1. In order to reduce the computational overhead, we share the parameters among the sentence encoders on the same side. The settings of these two sentence-level NMT models are consistent with our baseline Transformer model.

4.4 Results

Table 1 summarizes the BLEU scores of different systems on four tasks. As seen, our baseline and re-implemented existing methods outperform the reported results on the same data, which we believe makes the evaluation convincing.

Clearly, our model MHT significantly improves translation quality in terms of BLEU on these tasks, and obtains the best average results that gain 0.38, 0.69 and 1.57 BLEU points over CADec, CA-Transformer and CA-HAN respectively. These results demonstrate the universality and effectiveness of the proposed approach. Moreover, without in-

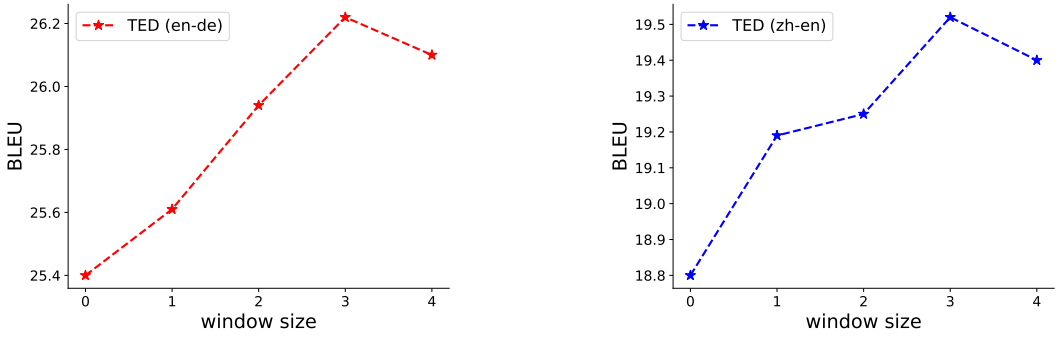

Figure 2: The performance of the MHT model on TED (En-De) and TED (Zh-En) translation task using different context window sizes.

troducing large-scale pre-trained language models, our translation systems achieve new state-of-the-art translation qualities across three examined translation tasks, which are TED (En-De), Opensubtitles (En-Ru) and Europarl7 (En-De). Overall, our experiments indicate the following two points: 1) explicitly modeling underlying reasoning semantics by a multi-hop mechanism indeed benefits neural machine translation, and 2) the improvements of our model are not from enlarging the network.

5 Analysis

In this section, to gain further insight, we explore the effectiveness of several factors of our model, including 1) multi-hop attention; 2) context window size; 3) reasoning direction; 4) sides for introducing context; and 5) target contexts. Moreover, we show qualitative analysis on discourse phenomena to better understand the advantage of our model.

5.1 Multi-Hop Attention

To further investigate the effect of multi-hop reasoning, we compare our multi-hop attention with two baseline context modeling methods, including "Concat" and "Hierarchical Attention". Table 2 shows the results of three different context modeling modules on TED, which use same inputs containing original training data and drafts. "Concat" denotes the MHT model simply using the concatenation of the three context sentences representations to get the final context representation. "Hierarchical Attention" denotes the MHT model with a hierarchical attention to model context, which consists of a sentence-level attention and a token-level attention to capture information from the appropriate context sentences and tokens, as in Miculicich et al. (2018). As depicted in Table 2, we replace multi-hop attention with these two baseline modules for experiments. "Hierarchical Attention" slightly outperforms "Concat", while multi-hop attention leads both of them by a much larger margin. The results demonstrate that multi-hop attention is capable of providing a more fine-grained representation of reasoning state over context and consequently capturing context semantic information more accurately.

Method	TED (En-De)	TED (Zh-En)
Concat	25.52	18.53
Hierarchical Attention	25.65	18.71
Multi-Hop Attention	**26.22**	**19.52**

Table 2: Comparison of different context modeling methods.

5.2 Context Window Size

As shown in Figure 2, we conduct experiments with different context window sizes to explore its effect. When the window size is less than 4, the model obtains more information from contexts and achieves better performance as the window size gets larger. However, when window size is increased to 4, we find that the performance doesn't improve further, but decreases slightly. This phenomenon shows that contexts far from the target sentence may be less relevant and cause noise (Kim et al., 2019). Therefore, we choose the window size 3 for our model MHT.

5.3 Reasoning Direction

In Table 3, we conduct an ablation study to investigate the effect of reasoning direction on MHT model. L2R denotes the MHT model with natural reasoning direction, which encodes context sentences from left to right by multi-hop layers, while

Direction	TED (En-De)	TED (Zh-En)
L2R	**26.22**	**19.52**
R2L	25.80	19.18

Table 3: The performance of the MHT model on TED (En-de) and TED (Zh-En) using different reasoning direction. L2R denotes the left to right direction for reasoning in context, while R2L is the opposite reasoning direction.

R2L indicates the MHT model encoding context sentences with an opposite direction. We observe that integrating reasoning processes by multi-hop attention with both direction can improve the effect of Transformer due to the incorporation of extra context information. Besides, MHT model reasoning with natural sentence order outperforms the MHT model with an opposite reasoning direction. This is within our expectation since the L2R reasoning is consistent with the reading and reasoning direction of human being.

5.4 Different Sides for Introducing Context

As shown in Table 4, we conduct an ablation study to explore how MTH model benefits from contexts on source side and target side of MTH model. "None" indicates the MTH model without multi-hop attention module on any side of MHT model, but only the draft of the current sentence. "Source", "Target" and "Source & Target" indicate the MHT models with multi-hop attention module to introducing context on only source side, only target side and both sides respectively. We find that integrating source-side context or target-side context into the model brings improvements over "None" that ignores context on both side. Besides, MHT with context on both sides achieves the best performance, indicating that the beneficial context information captured by multi-hop attention on the source side and the target side are divergent and complementary.

Side	TED (En-De)	TED (Zh-En)
None	25.40	18.80
Source	25.86	19.24
Target	25.73	19.20
Source & Target	**26.22**	**19.52**

Table 4: Comparison of introducing context on different sides of MHT model.

5.5 Draft vs. Reference

In training, the context draft sentences can be the drafts from a pre-trained MT system or the context references, while only the generated drafts are accessible during inference. Table 5 shows the BLEU scores of the MHT models using generated drafts and context references during training. We can see that the MHT model using drafts as contexts outperforms the MHT model directly using target-side context references, possibly because using context references faces the problem of exposure bias and the drafts generated from pre-trained translation system can bridge the gap between training and testing data.

Target Contexts	TED (En-De)	TED (Zh-En)
Reference	26.03	19.21
Draft	**26.22**	**19.52**

Table 5: The performance of the MHT models using drafts or context references.

5.6 Qualitative Analysis

We present the translated results from baselines and our model in Table 6 to explore how multi-hop reasoning mitigate the impact of common discourse phenomena in translation process. According to Case 1 in Table 6, the noun "hum" in source sentence is translated to "der Summen" by Transformer and CA-Transformer, which fail to understand the correct coreference. In German, "der" is a masculine article. The correct article is neutral article "das" because the "hum" is from a machine. MHT can perform a reasoning process to leverage the context information effectively and figure out the "hum" is from an engine according to Context 2. Case 2 indicates that MHT can understand the exact meaning of a polysemous word, benefiting from the reasoning process among the contexts. In this case, Transformer, CA-Transformer and CA-HAN all translates the noun "show" into "zeigt", which means "display". The translation is clearly wrong in this context. The correct meaning of "show" is TV shows like "Breaking Bad" according to the Context 1. In contrast, our model can take previous contexts in consideration and reason out the exact meaning of the polysemous word.

6 Conclusion

In this paper, we propose a novel document-level translation model called Multi-Hop Transformer

	Case 1	Case 2
Context3	They were 12, 3, and 1 when the hum stopped.	So if your show gets a rating of nine points ...
Context2	The hum of the engine died.	Then you have a top two percent show.
Context1	I stopped loving work. I couldn't restart the engine.	That's shows like "Breaking Bad," ...
Source	The hum would not come back.	That kind of show .
Reference	Das Summen kam nicht zurück.	diese Art von Show .
Transformer	der Summen kam nicht zurück .	das zeigt .
CA-Transformer	der Summen kam nicht zurück .	das zeigt .
CA-HAN	das Summen würde nicht zurückkommen.	das zeigt irgendwie .
MHT	das Summen kam nicht zurück .	diese Art von Show .

Table 6: Examples of the translation results of the baselines and MHT model.

with an inspiration from human reasoning behavior to explicitly model the human-like draft-editing and reasoning process. Experimental results on four widely used tasks show that our model can achieve better performance than both context-agnostic and context-aware strong baseline. Furthermore, the qualitative analysis shows that the multi-hop reasoning mechanism is capable of solving some discourse phenomena by capturing context semantics more accurately.

Acknowledgement

This work was supported by National Key R&D Program of China (2018YFB1403202). We 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.

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.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27.

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.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. *arXiv preprint arXiv:2006.04721*.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*.

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.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19.

Liane Kirsten Guillou. 2016. Incorporating pronoun function into statistical machine translation.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

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 (Volume 2: Shared Task Papers, Day 1)*, pages 225–233.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source

toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Shaohui Kuang and Deyi Xiong. 2018. 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.

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.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387.

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 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.

Sameen Maruf, André FT Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112.

Sameen Maruf, André FT 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, Volume 1 (Long and Short Papers)*, pages 3092–3102.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

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.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1055. ACM.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2020. Capturing longer context for document-level neural machine translation: A multi-resolutional approach. *arXiv preprint arXiv:2010.08961*.

I Sutskever, O Vinyals, and QV Le. 2014. Sequence to sequence learning with neural networks. *Advances in NIPS*.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. 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.

Jörg Tiedemann and Yves Scherrer. 2017a. Neural machine translation with extended context. *arXiv preprint arXiv:1708.05943*.

Jörg Tiedemann and Yves Scherrer. 2017b. 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.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5(1):87–99.

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, Rico Sennrich, and Ivan Titov. 2019. 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.

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.

Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. Self-Paced Learning for Neural Madchine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*.

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.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*, pages 1784–1794.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 387–394.

Baosong Yang, Derek F Wong, Tong Xiao, Lidia S Chao, and Jingbo Zhu. 2017. Towards bidirectional hierarchical representations for attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1432–1441.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017. Thumt: An open source toolkit for neural machine translation. *arXiv preprint arXiv:1706.06415*.

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. pages 533–542.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *IJCAI*.