Effective Graph Context Representation for Document-level Machine Translation

Kehai Chen[1] , **Muyun Yang**[2] , **Masao Utiyama**[3] , **Eiichiro Sumita**[3] , **Rui Wang**[4] , **Min Zhang**[1*]
[1]Harbin Institute of Technology, Shenzhen, China
[2]Harbin Institute of Technology, Harbin, China
[3]National Institute of Information and Communications Technology, Kyoto, Japan
[4]Shanghai Jiao Tong University, Shanghai, China
{chenkehai, yangmuyun, zhangmin2021}@hit.edu.cn,
{mutiyama, eiichiro.sumita}@nict.go.jp, wangrui12@sjtu.edu.cn

Abstract

Document-level neural machine translation (DocNMT) universally encodes several local sentences or the entire document. Thus, DocNMT does not consider the relevance of document-level contextual information, for example, some context (i.e., content words, logical order, and co-occurrence relation) is more effective than another auxiliary context (i.e., functional and auxiliary words). To address this issue, we first utilize the word frequency information to recognize content words in the input document, and then use heuristical relations to summarize content words and sentences as a graph structure without relying on external syntactic knowledge. Furthermore, we apply graph attention networks to this graph structure to learn its feature representation, which allows DocNMT to more effectively capture the document-level context. Experimental results on several widely-used document-level benchmarks demonstrated the effectiveness of the proposed approach.

1 Introduction

Recently, document-level neural machine translation (DocNMT) has attracted much attention in the machine translation community [Zhang *et al.*, 2018; Yang *et al.*, 2019; Ma *et al.*, 2020]. Typically, DocNMT considers the document-level contextual information, for example, several previous sentences or all sentences in the entire document. All words within the document-level context are then encoded as an additional source representation one-by-one for enhancing the translation of the input document. As a result, the DocNMT models greatly alleviated translation errors in the document-level translation scenario [Tan *et al.*, 2019; Voita *et al.*, 2019; Kang *et al.*, 2020; Bao *et al.*, 2021].

Different from a single sentence, multiple sentences within one input document are often associated with effective items like content words, logical order, co-occurrence, and cohesion/coherence relations, to introduce this document-level context. As illustrated by the English input document in Figure 1, there are obvious multiple content words (e.g., "*owns*" and "*FBI agents*") in green, including co-occurrence words (e.g., "*Austin*" and "*Florid*") in blue, related to the document-level context. Furthermore, the three sentences describe people and things related to "*a scuba diving shop*", and introduced the matter progressively and coherently, for example, "*Austin owns a scuba diving shop, FBI agent will visit his shop, but he did not receive any notification*". Meanwhile, there are also many functional or auxiliary contexts (e.g., not highlighted words) to assist in the expression of this matter. Most existing DocNMT methods universally encode all words within the input document regardless of whether these words are related to this effective document-level context. Thus, those functional and auxiliary context words disperse the attention of DocNMT for this effective document-level context. Particularly, the advanced DocNMT systems [Tu *et al.*, 2018; Maruf *et al.*, 2019] often use several previous sentences instead of the entire document, which further hinders the representation of the global document-level context. Therefore, learning a representation for the effective document-level context is more crucial for the advanced DocNMT.

In this paper, we propose an effective graph context representation (**EGCR**) method to learn the representation of the contextual information from the input document. To this end, we first recognize content words from the input document by term frequency-inverse document frequency and design four heuristic relations to connect the recognized content words as a graph structure without relying on external syntactic knowledge. Graph attention networks (GATs) [Guo *et al.*, 2019] are then used to learn the feature representation of this graph structure. As a result, the proposed EGCR allows DocNMT to effectively make use of the document-level context to enhance the translation of the input document. Experimental results on several widely-used document-level translation benchmarks demonstrated that our DocNMT model with EGCR gained significant improvement over strong baselines. Further quantitative and qualitative analysis verified the effectiveness of the proposed EGCR.

2 Methodology

2.1 Heuristic Graph Structure

To effectively represent the input document, we first recognize content words from an input document M_m by term frequency-inverse document frequency (TF-IDF). Formally, the TF-IDF

* Corresponding author

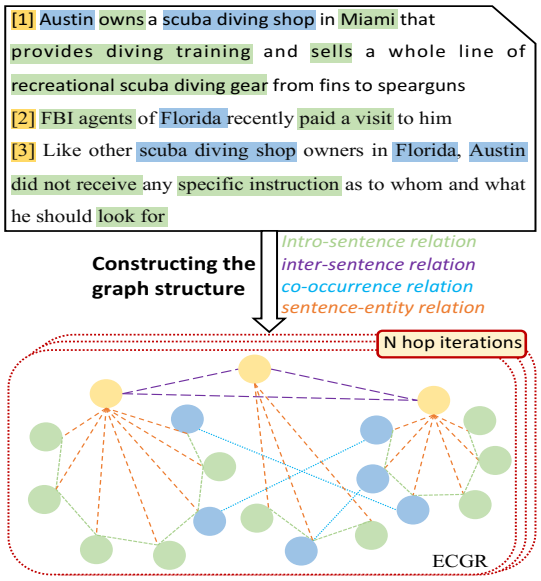

Figure 1: An English input document with three sentences and the proposed EGCR. Words within blue and green boxes denote content and co-occurrence words respectively while orange boxes are for sentences. Note: the dotted lines in green indicate sequential relation between content words in the current sentence; the dotted lines in blue indicate the relation between co-occurrence words; the dotted lines in purple indicate the inter-sentence relation; and the dotted lines in orange indicate the relation between sentence and word.

score of each word w in M_m is:

$$TI_w = \frac{\mathbb{J}_w}{\mathbb{J}} \times \log \frac{|M|}{1 + M_w|}, \tag{1}$$

where $\mathbb{J}_w$ represents the number of occurrences of word w in M_m and $\mathbb{J}$ is the total number of word w in all source documents M; $|M|$ is the total number of source documents; and M_w is the number of document including word w.[1] Note that multiple consecutive content words are as a phrase.

Then, we propose four heuristic relation edges to connect the recognized content words as a graph structure: **Inter-sentence edge** (ISEdg) denotes the relation edges between all sentences in the input document; **Sentence-word edge** (SWEdg) denotes the relationship between one sentence and its content words; **Intra-sentence edge** (IASEdg) used to retain order between content words within the same sentence; **Co-occurrence edge** (CoEdg) denotes the co-occurrence of the same word in different sentences from the input document. Formally, these recognized content words from the input document are denoted as a set of nodes $\mathbb{V}=\{v_1, v_2, \cdots, v_U\}$, where U is the number of nodes and the representation of each node v_u is the average of the word embeddings in the entity or verb. Additionally, an adjacency matrix $\mathbb{E}\in\mathbb{R}^{U\times U}$ denotes connections between two nodes using the aforementioned four kinds of relation edges:

$$\mathbb{E}[i][j] = \begin{cases} 1, \text{an edge between } v_i \text{ and } v_j, \\ 0, \text{no edge between } v_i \text{ and } v_j. \end{cases} \tag{2}$$

[1]This paper empirically select 25% words with higher TF-IDF scores in the input document as content words.

Finally, the constructed graph, which is undirected, is formally denoted as $G=\{\mathbb{V}, \mathbb{E}\}$ as shown in Figure 1.

2.2 Representation of Effective Context

Formally, GATs are applied to the constructed graph $G=\{\mathbb{V}, \mathbb{E}\}$ to learn the feature representation of each node using the multi-hop mechanism [Xu *et al.*, 2021b]. Formally, given the outputs of all previous hop iterations $\{\mathbf{s}_m^1, \mathbf{s}_m^2, \cdots, \mathbf{s}_m^{n-1}\}$, we concatenate them, then transform them into a fixed dimensional vector as the input of the n-th hop operation to learn an intermediate feature representation $\mathbf{z}_m^n$:

$$\mathbf{z}_m^n = \mathbf{W}_d^n \cdot [\mathbf{v}_m : \mathbf{s}_m^1 : \mathbf{s}_m^2 : \cdots : \mathbf{s}_m^{n-1}], \tag{3}$$

where $\mathbf{s}_m^{n-1}\in\mathbb{R}^{d_{model}}$ and $\mathbf{W}_d^n\in\mathbb{R}^{d_{model}\times(n\times d_{model})}$, and d_{model} is the word embedding dimension. According to edge matrix $\mathbb{E}[n][b_i]$ $(0\le b_i<I)$, all I direct adjacent nodes of $\mathbf{v}_n$ are $\{\mathbf{z}_{b_1}^n, \cdots, \mathbf{z}_{b_I}^n\}$ when $\mathbb{E}[n][b_i]$ is equal to one according to Eq.(2). Particularly, $\mathbf{W}_d^n$ aims to summarize direct connections introduced from all its preceding layers. We use the self-attention module [Vaswani *et al.*, 2017] to learn the n-th feature representation $\mathbf{s}_m^n$ using $\mathbf{z}_m^n$ and $\{\mathbf{z}_{b_1}^n, \cdots, \mathbf{z}_{b_I}^n\}$:

$$\mathbf{s}_m^n = \text{ATT}(\mathbf{z}_m^n, \mathbf{K}_d, \mathbf{V}_d), \tag{4}$$

where the direct adjacent feature representations of nodes $\{\mathbf{z}_{b_1}^n, \mathbf{z}_{b_2}^n, \cdots, \mathbf{z}_{b_I}^n\}$ are packed into key matrix $\mathbf{K}_d$ and value matrix $\mathbf{V}_d$. After we perform N hop iterations, there is a sequence of annotations $\{\mathbf{s}_m^1, \mathbf{s}_m^2, \cdots, \mathbf{s}_m^N\}$ to encode the feature information of the m-th node. Finally, we apply another nonlinear layer to integrate the feature information $\{\mathbf{s}_m^1, \mathbf{s}_m^2, \cdots, \mathbf{s}_m^N\}$ and the node vector $\mathbf{v}_m$ in $\mathbf{V}$:

$$\mathbf{g}_u = \text{Relu}(\mathbf{W}_o^d \cdot [\mathbf{v}_m : \mathbf{s}_m^1 : \cdots : \mathbf{s}_m^N]), \tag{5}$$

where $\mathbf{W}_o^d\in\mathbb{R}^{d_{model}\times(d_{model}\times(N+1))}$, $\mathbf{g}_u\in\mathbb{R}^{d_{model}}$. As a result, the proposed EGCR is represented as $\mathbf{G}=\{\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_U\}$ (see Figure 1).

3 EGCR for DocNMT

3.1 Background: Transformer-based NMT

In Transformer-based NMT [Vaswani *et al.*, 2017], given an input sentence of length J, $\mathbf{X}=\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_J\}$, The positional embedding $\mathbf{pe}_j$ of each word is added to the corresponding word embedding $\mathbf{x}_j$ as a combined embedding $\mathbf{h}_j$. Thus, there is an input representation $\mathbf{H}_e^0=\{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_J\}$ which will be fed into the stacked encoder to learn the source representation $\mathbf{H}_e^L$:

$$\begin{bmatrix} \mathbf{C}_e^l = \text{LN}(\text{ATT}^l(\mathbf{H}_e^{l-1}) + \mathbf{H}_e^{l-1}) \\ \mathbf{H}_e^l = \text{LN}(\text{FFN}^l(\mathbf{C}_e^l) + \mathbf{C}_e^l) \end{bmatrix}_L, \tag{6}$$

where $\text{ATT}^l(\cdot)$, $\text{LN}(\cdot)$, and $\text{FNN}^l(\cdot)$ are the self-attention module, layer normalization, and feed-forward neural network for the l-th identical layer, respectively. $[\cdots]_L$ denotes the stack of L identical layers. Compared with the stacked encoder, there is an additional encoder-decoder attention sub-layer that performs the dependent-time attention module for the top output of the stacked encoder $\mathbf{H}_e^L$.

$$\begin{bmatrix} \mathbf{T}_i^l = \text{LN}(\text{ATT}_t^l(\mathbf{Q}_i^{l-1}, \mathbf{K}_i^{l-1}, \mathbf{V}_i^{l-1}) + \mathbf{S}_i^{l-1}) \\ \mathbf{B}_i^l = \text{LN}(\text{ATT}_c^l(\mathbf{T}_i^l, \mathbf{K}_e^L, \mathbf{V}_e^L) + \mathbf{T}_i^l) \\ \mathbf{S}_i^l = \text{LN}(\text{FNN}_d^l(\mathbf{B}_i^l) + \mathbf{B}_i^l) \end{bmatrix}_L, \tag{7}$$

where $\mathbf{Q}_i^{l-1}, \mathbf{K}_i^{l-1}$, and $\mathbf{V}_i^{l-1}$ are the query, key, and value matrices that are transformed from the (l-1)-th layer $\mathbf{S}^{l-1}$ in the current time step i and $\{\mathbf{K}_e^L, \mathbf{V}_e^L\}$ are transformed from the top output $\mathbf{H}_e^L$ of the stacked encoder in Eq.(6). $\mathbf{T}_i^l$ represents the generated target fragment and is therefore used to compute the dependent-time context vector $\mathbf{B}_i^l$.

Finally, the top output of the stacked decoder $\mathbf{S}_i^L$ is used to generate the next target word y_i using a linear, potentially multi-layered function:

$$P(y_i|y_{<i}, \mathbf{X}) \propto \exp(\boldsymbol{W}_o^t \tanh(\boldsymbol{W}_w^t \mathbf{S}_i^L), \tag{8}$$

where $\boldsymbol{W}_o^t$ and $\boldsymbol{W}_w^t$ are projection matrices. To obtain the NMT model θ, the training objection maximizes the conditional translation probability over the training dataset $\{[\mathbf{X}, \mathbf{Y}]\}$:

$$\mathcal{J}(\theta) = P(\mathbf{Y}|\mathbf{X}; \theta). \tag{9}$$

3.2 DocNMT with EGCR

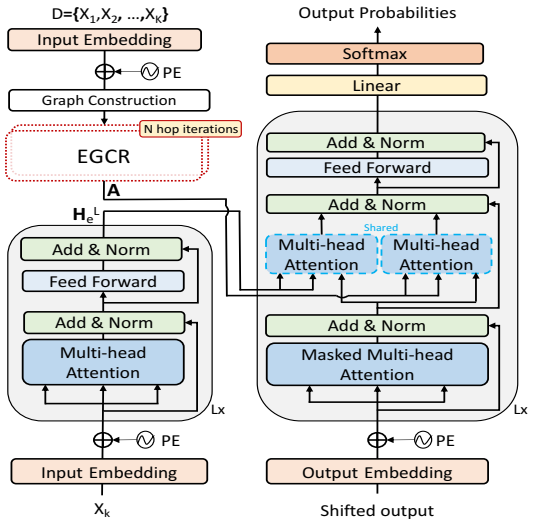

Figure 2: The proposed DocNMT with EGCR. Note that two boxes denotes that their parameters are shared.

This section will introduce the proposed DocNMT with EGCR, as shown in Figure 2. Formally, we feed the learned EGCR $\mathbf{A}$ into Eq.(7) together with the original representation of source sentence $\mathbf{H}_e^L$ to learn a dependent-time context vector:

$$\begin{bmatrix} \boldsymbol{\mathcal{T}}_i^l = \mathrm{LN}(\mathrm{ATT}_t^l(\mathbf{Q}_i^{l-1}, \mathbf{K}_i^{l-1}, \mathbf{V}_i^{l-1}) + \boldsymbol{\mathcal{S}}_i^{l-1}) \\ \mathbf{B}_i^l = \mathrm{LN}(\mathrm{ATT}_c^l(\boldsymbol{\mathcal{T}}_i^l, \mathbf{K}_e^L, \mathbf{V}_e^L) + \boldsymbol{\mathcal{T}}_i^l) \\ \mathbf{D}_i^l = \mathrm{LN}(\mathrm{ATT}_c^l(\boldsymbol{\mathcal{T}}_i^l, \mathbf{K}_\mathrm{A}, \mathbf{V}_\mathrm{A}) + \boldsymbol{\mathcal{T}}_i^l) \\ \boldsymbol{\mathcal{S}}_i^l = \mathrm{LN}(\mathrm{FNN}_d^l(\mathbf{B}_i^l + \mathbf{D}_i^l) + \mathbf{B}_i^l) \end{bmatrix}_L, \tag{10}$$

where $\mathbf{K}_\mathrm{A}$=$\mathbf{V}_\mathrm{A}$ is the matrix representation of the proposed EGCR $\mathbf{G}$=$\{\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_U\}$. Note that both the original sentence-level context vector $\mathbf{B}_i^l$ and the proposed document-level context vector $\mathbf{D}_i^l$ are learned by a shared crossing attention module ATT_c^l.

We use the top output of decoder $\boldsymbol{\mathcal{S}}_i^L$, which is enhanced by the extracted effective document-level context, as input to Eq.(8) to compute the conditional probability of next target word y_i:

$$P(y_i|y_{<i}, \mathbf{x}) \propto \exp(\boldsymbol{W}_o^t \tanh(\boldsymbol{W}_w^t \boldsymbol{\mathcal{S}}_i^L). \tag{11}$$

Finally, the training objective maximizes the probability over the training dataset $\{[\mathbf{X}, \mathbf{Y}]\}$:

$$\mathcal{L}(\theta, \phi) = P(\mathbf{Y}|\mathbf{X}; (\theta, \phi)), \tag{12}$$

where θ is a parameter set related to the processing of the SentNMT model and ϕ is another parameter set related to the learning of EGCR.

4 Experiments

4.1 Datasets and Setup

We evaluated our approach on several benchmarks for document-level machine translation. For *TED* Talks in IWSLT17, we used *dev-2010* as the development set, and test-2010/2011/2012/2013 as the test sets for both Chinese-English (Zh-En) and English-German (En-De) language pairs. For News-Commentary v14 (*News*), we use the newstest2017 for development and newstest2018 to test both Zh-En and En-De language pairs. Our approach was also evaluated on a large scale corpus *Euro* extracted from Europarl v7 [Maruf *et al.*, 2019]. Table 1 showed the corpora statistics. We adopted the

Datasets		Training	Dev	Test
Zh-En	*TED*	0.23M	0.88K	4.68K
	News	0.31M	2.00K	3.98K
En-De	*TED*	0.21M	0.89K	4.70K
	News	0.33M	3.00K	3.00K
	Euro	1.66M	3.58K	5.13K

Table 1: Corpora statistics (K and M represent thousands and million sentence pairs, respectively).

BPE algorithm [Sennrich *et al.*, 2016], and set the vocabulary size to 32K. We set the dimension of all input and output layers to 512, the dimension of the inner feedforward neural network layer to 1024, and the total heads of all multi-head modules to 8 in both the encoder and decoder layers. The number of multi-hop reasoning N was set to 2 empirically. Each training batch consisted of a set of sentence pairs that contained approximately 4000×8 source tokens and 4000×8 target tokens. The value of label smoothing was set to 0.1, and the attention dropout and residual dropout were 0.1. We varied the learning rate under a warm-up strategy with warmup steps of 8,000. Following the training of 100,000 batches, we used a single model obtained by averaging the last five checkpoints, which validated the model with an interval of 2,000 batches on the dev set. We trained all models on eight V100 GPUs and evaluated them on a single V100 GPU. The multi-bleu.perl script was used as the evaluation metric.

We re-implemented several advanced document-level NMT systems based on the sentence-level Transformer NMT model (SentNMT) [Vaswani *et al.*, 2017] by using the *fairseq* toolkit: **CTX** [Zhang *et al.*, 2018] uses an additional encoder to learn context representations, which are then integrated by cross-attention mechanisms. **HAN** [Miculicich *et al.*, 2018] uses a hierarchical attention mechanism with two levels

(word and sentence) of abstraction to incorporate context information from both the source and target documents. **Selective** [Maruf *et al.*, 2019] uses all contexts in the entire document by calculating the sentence-level and word-level weights. **HierAtt** [Tan *et al.*, 2019] learns a global context representation using a hierarchical attention mechanism. **DynS** [Tu *et al.*, 2018] introduces a cache to memorize previous hidden states as dynamically updated context during decoding. Additionally, we reported the results of two recent DocNMT models: **Unified** [Ma *et al.*, 2020] employed the first encoder layer of Transformer to encode the current sentence with context information and fed it into the decoder; **DGraph** [Xu *et al.*, 2021a] applied syntactic parser to both source and target documents, and incorporated their graphs into the Transformer with graph convolutional networks.

4.2 Main Results

Methods	*TED*		*News*		*Euro*	#Speed.	#Param.
	Zh-En	En-De	Zh-En	En-De	En-De		
SentNMT	19.27	28.79	13.83	26.26	29.23	13.8k	69.16M
+DGraph*	20.46	N/A	N/A	N/A	N/A	N/A	N/A
+Unified*	N/A	N/A	N/A	N/A	30.09	N/A	N/A
+HierAtt	20.11	29.93	14.79	27.38	30.19	10.6k	77.14M
+Selective	20.17	30.01	14.56	27.79	30.43	8.4k	77.65M
+CTX	19.89	29.51	14.11	26.62	29.66	10.2k	88.59M
+HAN	20.06	29.89	14.62	27.12	30.23	9.61k	76.63M
+DynS	20.58	30.39	14.95	27.76	30.76	12.6k	76.14M
+EGCR	**20.95**†	**30.93**†	**15.46**†	**28.32**†	**31.31**†	13.2k	72.64M

Table 2: Main results (BLEU) of five test sets on TED, News, and Euro tasks. Note that: Results of methods with "$*$" were reported in the original papers. "†" denotes a statistical significant improvement over the best baseline DynS model with the proposed EGCR on each task at $p<0.05$ [Collins *et al.*, 2005]. "#Speed." denotes training speeds (tok/sec) and "#Param." denotes the size of model parameters (M is million and k is thousand). Bold results indicate this method has the best performance on the test set.

Table 2 shows the overall results on all translation tasks, where the hyperparameter N in Eq.(5) is set to 2 for the Zh-En and En-De tasks (the relation between N and BLEU scores is as shown in Appendix **??**). The results demonstrated that all the document-level NMT models performed better than the SentNMT model, which indicates that document-level contextual information is beneficial for NMT. +EGCR achieved the best performance among all document-level NMT models on five test sets. This means that the proposed EGCR can leverage the document-level context to improve translation performance. Particularly, +EGCR outperformed the comparison DocGraph by 0.49 BLEU scores on the same TED Zh-En test set. We think that the proposed EGCR summarized more effective context from the input document while DocGraph constructed all words within one document as a graph that contains other redundant contextual information (i.e., functional and auxiliary words). Additionally, +EGCR had higher efficiency than comparison methods in terms of model parameters and training speeds.

4.3 Ablation of EGCR

As mentioned in Section 1, those functional and auxiliary context words within the input document disperse the attention of DocNMT for this document-level context. To evaluate the relevance of our constructed document-level context, Table 3 shows results of our SentNMT+EGCR model when the different percentage of remaining words from the input document is introduced into EGCR (the +EGCR in Table 2 contained about 25% words in the input document). BLEU scores began to decrease as the increase of the remaining words (e.g., functional and auxiliary context words) but the EGCR with the least amount of content and co-occurrence words gained the highest BLEU scores. This indicates that these functional and auxiliary context words hindered the learning of the document-level context.

EGCR	*TED*		*News*		*Euro*
	Zh-En	En-De	Zh-En	En-De	En-De
+25%	20.95	30.63	15.46	28.32	31.31
+50%	20.76	30.53	15.16	27.94	31.08
+75%	20.28	30.17	15.01	27.78	30.89
+100%	20.12	29.93	14.94	27.67	30.71

Table 3: Ablation results (BLEU) of SentNMT+EGCR when EGCR encodes different percentage of remaining words in the input document.

4.4 Evaluating GATs for Learning EGCR

Methods	*TED*		*News*		*Euro*
	Zh-En	En-De	Zh-En	En-De	En-De
SentNMT	19.27	28.79	13.83	26.26	29.23
+EGCR(RNNs)	20.17	29.52	14.76	27.12	30.22
+EGCR(CNNs)	20.05	29.83	14.83	27.58	30.67
+EGCR(SANs)	20.46	30.15	15.11	28.01	31.11
+EGCR(GATs)	**20.95**	**30.63**	**15.46**	**28.32**	**31.31**

Table 4: Ablation results (BLEU scores) for our SentNMT+EGCR model on the five test sets when replacing GATs with different neural networks, including RNNs, CNNs, and SANs.

To evaluate the effectiveness of GATs, we replaced GATs with recurrent neural networks (RNNs), convolutional neural networks (CNNs), and self-attention networks (SANs). Specifically, we flat the constructed graph in Section 2.1 as a long sequence according to word order in the original document and then feed this long sentence into RNNs, CNNs, and SANs to learn an approximation EGCR. Table 4 showed ablation results on the five translation tasks. First, four SentNMT+EGCR models outperformed the baseline SentNMT model in terms of BLEU scores, confirming the effectiveness of the proposed approach. The EGCR learned by GATs gave more improvement than ones learned by three alternative RNNs, CNNs, and SANs. We believe that GATs can better encode structural and hierarchical translation knowledge in the input document.

4.5 Ablation Study for Different Edges

To evaluate different relation edges of the constructed graph, we report ablation test results in Table 5. Among all relationship edges, removing CoEdg yielded the greatest performance degradation on all test sets. This means that the co-occurrence edge contributed the most to the improvement, which is consistent with the effectiveness of coreference information in the advanced DocNMT systems [Maruf and

Haffari, 2018; Maruf *et al.*, 2019; Ma *et al.*, 2020]. Similarly,

Methods	*TED*		*News*		*Euro*
	Zh-En	En-De	Zh-En	En-De	En-De
SentNMT	19.27	28.79	13.83	26.26	29.23
EGCR	20.95	30.63	15.46	28.32	31.31
-SWEdg	20.61	30.41	15.23	28.21	30.96
-ISEdg	20.26	29.98	14.86	27.84	30.61
-CoEdg	20.01	29.33	14.13	27.57	30.16
-IASEdg	20.16	29.79	14.46	27.76	30.27

Table 5: Ablation results (BLEU scores) over all test sets when removing different relation edges.

removing IASEdg was in second place for the improvement, which indicates that the sequential edge is an important contribution. The performance drop for our model without SWEdg and ISEdg demonstrated that the sentence-entity edges and inter-sentence edges were also useful for DocNMT. In short, the four types of relation edges effectively simulated order dependency in the input document.

4.6 Ablation Study of Different Nodes

Methods	*TED*		*News*		*Euro*
	Zh-En	En-De	Zh-En	En-De	En-De
EGCR	20.95	30.63	15.46	28.32	31.31
-Sentence	20.57	30.18	15.09	28.04	30.71
-Content	20.31	29.82	14.76	27.73	30.59
-Co-occurrence	20.36	30.13	15.31	27.86	30.91

Table 6: Ablation results (BLEU) of Sentence, Content, and Co-occurrence nodes in the EGCR.

To evaluate the effect of sentence, content, and co-occurrence nodes in the EGCR, we report ablation results in Table 6. Removing any of three types of nodes resulted in the performance degradation but outperformed the baseline SentNMT, which indicates that they were useful for representing the document-level context. Moreover, EGCR without sentence or co-occurrence nodes had a lesser degradation in terms of BLEU scores than EGCR without content nodes. By comparison, content nodes can encode more document-level contextual information than sentence and co-occurrence nodes in EGCR.

4.7 Ablation Study of Sentence Order

Multiple sentences within one document often introduce this matter of the input document gradually according to some logical order. The proposed graph structure encoded the document-level contextual information in a hierarchical manner, which may have a strong capability of capturing logical order. Therefore, we simulated a scenario where the order of some sentences was randomly swapped in one input document. For example, "2" indicates that there were two randomly swapped sentences for each document in the test set. Figure 3 showed the results of TED Zh-En and En-De test sets. As the number of sentences whose word order is changed increases, the performance of DocNMT will decrease significantly. Particularly, when Num is greater than two (Zh-En) or three (En-De), BLEU scores of +DynS were even lower than the sentence-level SentNMT model. This means that the

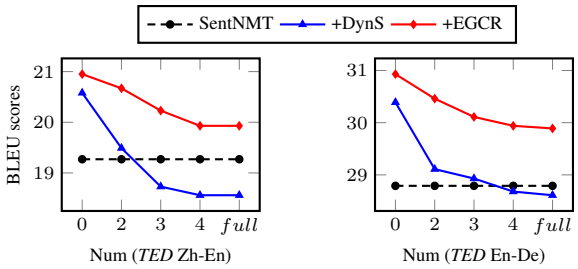

Figure 3: The effect of sentence order where sentence order within input document is partially changed for two TED test sets. "Num" denotes the number of sentences whose orders were changed.

logical order plays a very important role for DocNMT, but the proposed approach does have the strong ability to capture logical order between multiple sentences in one document.

4.8 Analysis of Discourse Phenomena

Methods	BLEU	deixis.	lex.c.	ell.infl.	ell.VP.
SentNMT	31.36	51.2	46.7	65.2	37.5
+DynS	32.61	55.3	51.6	68.3	38.9
+EGCR	**33.93**	**58.8**	**55.7**	**71.8**	**40.9**

Table 7: Accuracy on contrastive test sets. "deixis." is the deistic words or phrases whose denotation depends on the context. "lex.c." focuses on the reiteration of named entities. "ell.infl." aims at the morphological form. "ell.VP." is a test for the ellipsis verb phrase.

In this section, we investigate the contribution of document-level context information in improving the translation of discourse phenomena. The experiment is designed according to [Voita *et al.*, 2019] to evaluate four types of discourse phenomena (i.e., deixis, lexical cohesion, inflection, and VP ellipses). Each test instance consisted of positive and several negative translations with incorrect phenomena. We evaluated the models using accuracy, which is defined as the proportion of times the generation probability of a positive translation was higher than that of a negative translation. We trained the SentNMT, +DynS, and +EGCR models on applying the OpenSubtitles2018 corpus for English and Russian following the settings of [Voita *et al.*, 2019]. The accuracy of the discourse phenomena is shown in Table 7. Both +DynS and our +EGCR comprehensively improved consistency over the document-level context-agnostic SentNMT. Moreover, our +EGCR outperformed +DynS over four types of discourse phenomena. We attribute this to the fact that our document graph contained structural information within the document-level context (i.e., lexical consistency, logical order, and co-reference), thereby directly linking relevant contexts for repeated and deistic words.

4.9 Accuracy of Pronoun/Noun Translations

We followed the [Miculicich Werlen and Popescu-Belis, 2017]'s setting to evaluate coreference and anaphora using the reference-based metric, that is, the accuracy of pronoun translation, which can be extended to nouns. The list of evaluated pronouns is predefined in the metric, while the list of nouns is extracted using POS tagging, as shown in the results

of two Zh-En test sets in Table 8. The results demonstrate that both +DynS and +EGCR achieved accuracy with a significant improvement compared with the SentNMT model, which indicates that the document-level context was beneficial to the translation of discourse nouns and pronouns. Furthermore, the proposed +EGCR gained higher accuracy than the baseline +DynS. This indicates that EGCR more effectively leveraged the document-level contextual information than DynS to improve the translation of discourse nouns and pronouns.

Methods	Noun Translation		Pronoun Translation	
	TED	*News*	*TED*	*News*
SentNMT	41.26	46.23	64.31	50.61
+DynS	42.54	47.68	65.79	52.13
+EGCR	**43.12**	**48.32**	**66.37**	**52.38**

Table 8: Evaluation of discourse noun and pronoun translation over two Zh-En tasks.

4.10 Cohesion and Coherence Evaluation

Following Wong and Kit's setting, we evaluated the lexical cohesion, which is defined as the ratio of the number of repeated and lexical similar content words to the total number of content words in a target document. Meanwhile, for coherence, we used a metric based on latent semantic analysis [Foltz *et al.*, 1998], which is used to obtain sentence representations, then calculated the cosine similarity from one sentence to the next, and averaged the results to obtain a document score. Table 9 shows the average ratio per tested

Methods	Lexicon Cohesion		Coherence	
	TED	News	TED	News
SentNMT	53.78	31.64	0.301	0.282
+DynS	54.89	32.19	0.308	0.286
+EGCR	**55.11**	**32.39**	**0.311**	**0.289**

Table 9: Evaluation of discourse lexical cohesion and coherence for two Zh-En tasks. Lexical cohesion: ratio of repeated and lexical similar words over the number of content words. Coherence: average cosine similarity of consecutive sentences.

document for lexical cohesion and the average coherence score of documents. +DynS and +EGCR consistently obtained better scores for two Zh-En tasks, which is in line with the improvements to lexical cohesion and coherence in [Xiong *et al.*, 2013b; Lin *et al.*, 2015]. In particular, the scores of our +EGCR model were higher than those of +DynS, which indicates that the proposed EGCR captured more document-level contextual information related to discourse lexical cohesion and coherence.

5 Related Work

Typically, document-level context information was used to improve the lexical cohesion of the input document [Xiong *et al.*, 2013a; Ben *et al.*, 2013; Xiong *et al.*, 2013b], where the sentences are connected via syntactic and lexical devices. Document-level context information has also been used to capture coherence for document translation [Xiong and Zhang, 2013; Lin *et al.*, 2015], sentences of which are tied into a meaningfully connected structure.

Recently, document-level contextual information has begun to be introduced into NMT to enhance translation performance. Part of the document-level context has been encoded to assist the translation of the current sentence, for example, a single previous sentence [Tiedemann and Scherrer, 2017; Voita *et al.*, 2018]; one previous sentence both in the source and target [Bawden *et al.*, 2018]; more than one previous source sentence [Zhang *et al.*, 2018; Voita *et al.*, 2019; Kang *et al.*, 2020]; or a few previous source and target sentences [Miculicich *et al.*, 2018], and the contextual information in a fixed scope [Kuang *et al.*, 2018; Wong *et al.*, 2020]. By contrast, in some studies, the full document context has been used effectively for the source or target-side document-level contextual information in NMT [Maruf and Haffari, 2018; Zhang *et al.*, 2018; Maruf *et al.*, 2019; Tan *et al.*, 2019; Ma *et al.*, 2020].

More recently, [Xu *et al.*, 2021a] proposed a graph-based DocNMT method to simulate both source and target document-level context according to inter-sentential and intra-sentential relations. Although the graph-based DocNMT method is closely related to our work, there are two main differences. To begin with, we represent the document-level contextual information as the graph structure constructed by heuristic rules without demanding external syntactic knowledge. This breaks through the constraints of prior syntactic knowledge and provides a general alternative method for researchers to explore the document-level context for DocNMT. Moreover, our graph structure can encode the relevant contextual information from the input document, which allows DocNMT to more effectively make use of the document-level context than the existing universal encoding method.

6 Conclusion

This paper heuristically constructed the graph structure of the input document without external syntactic knowledge to summarize the document-level contextual information. We then applied GATs to the constructed graph to learn its feature representation, and thereby effectively leveraged the document-level context to improve translation performance. The experiment results on several widely-used document-level translation benchmarks demonstrated the effectiveness of the proposed approach. Furthermore, the exhaustive quantitative and qualitative analysis demonstrated that the proposed approach helped DocNMT capture the relevant document-level translation knowledge, for example, logical order, discourse phenomena, and coherence information. In the future, we will explore another efficient document-level translation knowledge for advancing the DocNMT.

Acknowledgments

We are grateful to the anonymous reviewers, area chair, senior area chair, and program committee for their insightful comments and suggestions. Min Zhang was partially supported by the National Natural Science Foundation of China (No. 62036004). Rui Wang is with MT-Lab, Department of Computer Science and Engineering, School of Electronic Information and Electrical Engineering, and also with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200204, China.

References

[Bao *et al.*, 2021] Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. G-transformer for document-level machine translation. In *Proc. of ACL*, 2021.

[Bawden *et al.*, 2018] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In *Proc. of NAACL*, 2018.

[Ben *et al.*, 2013] Guosheng Ben, Deyi Xiong, Zhiyang Teng, Yajuan Lü, and Qun Liu. Bilingual lexical cohesion trigger model for document-level machine translation. In *Proc. of ACL*, 2013.

[Collins *et al.*, 2005] Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *Proc. of ACL*, 2005.

[Foltz *et al.*, 1998] Peter W. Foltz, Walter Kintsch, and Thomas K Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 1998.

[Guo *et al.*, 2019] Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. In *Proc. of ACL*, 2019.

[Kang *et al.*, 2020] Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proc. of EMNLP*, 2020.

[Kuang *et al.*, 2018] Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proc. of COLING*, 2018.

[Lin *et al.*, 2015] Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. Hierarchical recurrent neural network for document modeling. In *Proc. of EMNLP*, 2015.

[Ma *et al.*, 2020] Shuming Ma, Dongdong Zhang, and Ming Zhou. A simple and effective unified encoder for document-level machine translation. In *Proc. of ACL*, 2020.

[Maruf and Haffari, 2018] Sameen Maruf and Gholamreza Haffari. Document context neural machine translation with memory networks. In *Proc. of ACL*, 2018.

[Maruf *et al.*, 2019] Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Selective attention for context-aware neural machine translation. In *Proc. of NAACL*, 2019.

[Miculicich *et al.*, 2018] Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. In *Proc. of EMNLP*, 2018.

[Miculicich Werlen and Popescu-Belis, 2017] Lesly Miculicich Werlen and Andrei Popescu-Belis. Validation of an automatic metric for the accuracy of pronoun translation. In *Workshop*, 2017.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proc. of ACL*, 2016.

[Tan *et al.*, 2019] Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. Hierarchical modeling of global context for document-level neural machine translation. In *Proc. of EMNLP*, 2019.

[Tiedemann and Scherrer, 2017] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *Workshop*, 2017.

[Tu *et al.*, 2018] Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. Learning to remember translation history with a continuous cache. *TACL*, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of NeurIPS*, 2017.

[Voita *et al.*, 2018] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In *Proc. of ACL*, 2018.

[Voita *et al.*, 2019] Elena Voita, Rico Sennrich, and Ivan Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proc. of ACL*, 2019.

[Wong and Kit, 2012] Billy T. M. Wong and Chunyu Kit. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proc. of EMNLP*, 2012.

[Wong *et al.*, 2020] KayYen Wong, Sameen Maruf, and Gholamreza Haffari. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proc. of ACL*, 2020.

[Xiong and Zhang, 2013] Deyi Xiong and Min Zhang. A topic-based coherence model for statistical machine translation. In *Proc. of AAAI*, 2013.

[Xiong *et al.*, 2013a] Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lü, and Qun Liu. Modeling lexical cohesion for document-level machine translation. In *Proc. of IJCAI*, 2013.

[Xiong *et al.*, 2013b] Deyi Xiong, Yang Ding, Min Zhang, and Chew Lim Tan. Lexical chain based cohesion models for document-level statistical machine translation. In *Proc. of EMNLP*, 2013.

[Xu *et al.*, 2021a] Mingzhou Xu, Liangyou Li, Derek F. Wong, Qun Liu, and Lidia S. Chao. Document graph for neural machine translation. In *Proc. of EMNLP*, 2021.

[Xu *et al.*, 2021b] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. In *Proc. of AAAI*, 2021.

[Yang *et al.*, 2019] Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proc. of EMNLP*, 2019.

[Zhang *et al.*, 2018] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the transformer translation model with document-level context. In *Proc. of EMNLP*, 2018.