Towards Discourse-Aware Document-Level Neural Machine Translation

Xin Tan, **Longyin Zhang**, **Fang Kong** and **Guodong Zhou***
School of Computer Science and Technology, Soochow University, China
{xtan9, lyzhang9}@stu.suda.edu.cn, {kongfang, gdzhou}@suda.edu.cn

Abstract

Current document-level neural machine translation (NMT) systems have achieved remarkable progress with document context. Nevertheless, discourse information that has been proven effective in many NLP tasks is ignored in most previous work. In this work, we aim at incorporating the coherence information hidden within the RST-style discourse structure into machine translation. To achieve it, we propose a document-level NMT system enhanced with the discourse-aware document context, which is named Disco2NMT. Specifically, Disco2NMT models document context based on the discourse dependency structures through a hierarchical architecture. We first convert the RST tree of an article into a dependency structure and then build the graph convolutional network (GCN) upon the segmented EDUs under the guidance of RST dependencies to capture the discourse-aware context for NMT incorporation. We conduct experiments on the document-level English-German and English-Chinese translation tasks with three domains (TED, News, and Europarl). Experimental results show that our Disco2NMT model significantly surpasses both context-agnostic and context-aware baseline systems on multiple evaluation indicators.

1 Introduction

With the maturity of sentence-level neural machine translation (NMT), document-level NMT has been drawing more attention and achieved significant progress in recent years. Reviewing previous work on document-level NMT, context-aware models have made substantial progress by extracting contextual dependencies within sentences to help translate each entire article [Zhang *et al.*, 2018; Voita *et al.*, 2018; Miculicich *et al.*, 2018; Maruf *et al.*, 2019; Tan *et al.*, 2019; Ma *et al.*, 2020; Sugiyama and Yoshinaga, 2021]. However, existing studies are prone to model broad-brush document context from the plain sentences in documents to capture the relatively shallow cohesion information, while the deep coherence information hidden within each article is ignored.

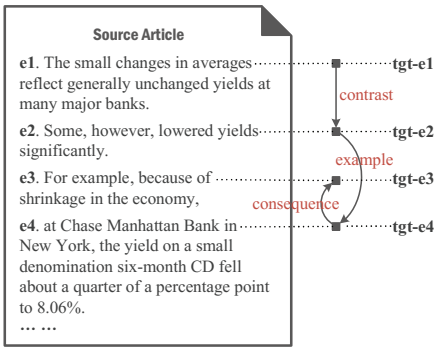

Figure 1: An illustration of the commonality that bridges the source and its corresponding target articles.

Rhetorical Structure Theory (RST) [Mann and Thompson, 1988] is known to be one of the most common and influential discourse theories, which describes the coherence structure of a document and has the ability to facilitate the logical relationship between clauses and sentences in a text [Hobbs, 1979]. As a universal theory about discourse structure, RST analysis has been applied to many different languages like English [Carlson *et al.*, 2001], German [Stede and Neumann, 2014], Chinese [Li *et al.*, 2014b], and so on. In view of this, the rhetorical structure between two different languages serves as an essential commonality between the two languages. Fig. 1 illustrates the importance of the rhetorical structure for document-level NMT. It shows that this commonality serves as a bridge between the source and target languages in document-level machine translation, and the coherence in the source can be used to promote the coherence of the target. Although previous studies have explored various ways of leveraging document context to improve document-level machine translation, the research on discourse structure is rare. To the best of our knowledge, [Chen *et al.*, 2020] is the first attempt to incorporate the discourse information into document-level NMT. However, they only employ the embedded RST tree paths for sentence representation enhancement, limiting the discourse information to shallow features, while the more direct rhetorical relation between two discourse units is ignored to some extent.

In this research, we explore applying RST discourse struc-

*Corresponding author

ture to document-level NMT. To achieve this, we propose the Disco2NMT model to incorporate RST graphs into the document-level NMT model based on a hierarchical attention network. In order to model both intra- and inter-sentence discourse relations, we segment sentences in each article into the finer granularity of elementary discourse units (EDUs) and model document context based on the EDUs. For discourse structure, we convert the tree structure to dependency graphs to avoid the long-range error propagation problem of the original RST constituency tree. On this basis, we encode the discourse dependencies using the graph convolutional network (GCN) [Kipf and Welling, 2016] to capture intra- and inter-sentence interactions. Furthermore, we incorporate the extracted discourse-aware context to the Transformer-based document-level NMT system, guiding the system to generate more coherent translations. It is worth mentioning that this work is the first attempt to prove the validity of the discourse dependency structure in document-level NMT.

We perform several experiments on document-level translation tasks with different languages (English-German and English-Chinese) and domains (TED, News, and Europarl). Experiments show that our proposed Disco2NMT outperforms current competitive document-level NMT systems and can generate more coherent translations.

2 Background

Document-level NMT. Different from sentence-level NMT systems that translate sentences separately, document-level NMT systems generally translate each source sentence $X_i = (x_1, x_2, ..., x_n)$ within a document $\mathcal{D} = (X_1, ..., X_N)$ into the target sentence $Y_i = (y_1, y_2, ..., y_n)$ with the consideration of contextual information C. The training criterion for the document-level NMT model is to maximize the conditional log-likelihood as:

$$\mathcal{L}(\mathcal{D};\theta) = \sum_{i=1}^{N}\sum_{j=1}^{n} \log p(y_j^i | y_{<j}^i, x^i, C; \theta) \qquad (1)$$

where $y_{<j}^i$ denotes the generated target hypothesis before y_j^i.

RST Discourse Structure. According to the RST theory, an article can be segmented into a series of clause-like, contiguous, and non-overlapping units, namely elementary discourse units (EDUs). Through certain rhetorical relations, two adjacent underlying discourse units are merged to form the upper-layer nodes recursively. Besides, each two sibling nodes in the tree also maintain a nuclearity relation, where the node labeled with nucleus (N) is more central than the node labeled with satellite (S) in the discourse structure. Fig. 2 shows the hierarchical structure of an RST tree with four EDUs. In the example, the rhetorical relationship between the leaf node e1 and the text span (e2, e4) is `contrast`. Both the two sibling nodes are annotated as the nucleus, which means they are equally crucial to the RST structure.

3 Disco2NMT

In this section, we first introduce how to convert a discourse constituency tree into the corresponding discourse dependency structure. After that, we further introduce the proposed Disco2NMT model integrated with the discourse dependency structure.

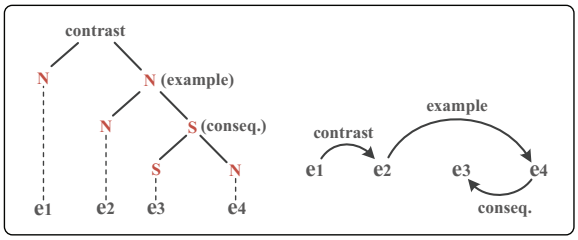

Figure 2: Conversion between constituency and dependency trees.

3.1 Discourse Dependency Structure Conversion

In the literature, Hirao *et al*. [2013] and Li *et al*. [2014a] have provided two different approaches to convert RST trees to dependency graphs. In the dependency tree of [Hirao *et al.*, 2013], different EDUs within a sentence could have multiple heads outside the sentence. Unlike Hirao *et al.*'s method, Li *et al*. [2014a] frame each sentence as a single-rooted dependency tree, which vastly reduces the complexity of the dependency graphs. Since machine translation usually takes each sentence as the elementary translation unit, the method of Li *et al*. [2014a] seems to be more compatible with machine translation. Concretely, the conversion method we use is detailed as follows:

- For each tree node N, we take the head node of its leftmost **N**ucleus child as its head node; if no child is Nucleus, we take the head of the leftmost child as the head of N.
- For each non-leaf node, if it yields a multi-nucleus relation, we follow the principle of leftmost priority and treat the left child as the only **N**ucleus node.
- For each leaf node, we pick the nearest Satellite on the path from the leaf node to the root and define the head of the Satellite node's parent as its head. The leaf node is the root of the graph when there exists no such Satellite node.

Following the above rules, the RST tree shown in Fig. 2 is finally converted into a dependency tree on the right.

3.2 Discourse-Aware Context Modeling

In document-level NMT, the context of an article plays an essential role in generating smooth translations. To our knowledge, most previous studies model document context from sentences of a document through hierarchical architectures and have achieved certain success. However, the context information modeled from the sentences of an article captures shallow and broad-brush context information, while the deeper coherence information like the rhetorical relation between text units is widely ignored. Considering this status quo, we propose to model discourse-aware contextual information based on the RST graph for more coherent translations. As stated before, RST segments each article into several contiguous, adjacent, and non-overlapping EDUs and builds rhetorical coherence between these units. In view of

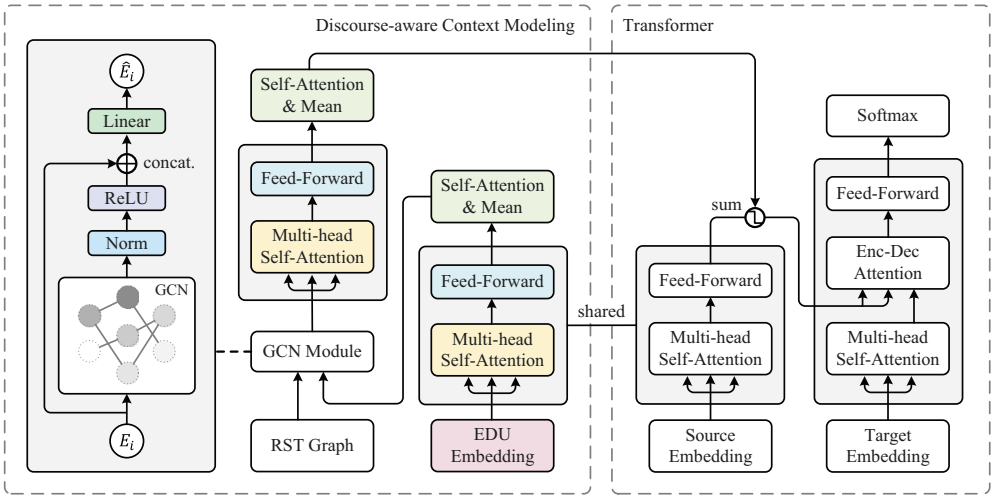

Figure 3: Architecture of the proposed discourse-aware document-level neural machine translation model (Disco2NMT).

this, we take EDUs[1] as inputs to Disco2NMT and model the discourse-aware document context through a three-level encoder structure as follows:

Sentence-Level Encoder. We employ a multi-head attention mechanism [Vaswani *et al.*, 2017] to construct the sentence-level encoder. Formally, given the EDUs of the input document, $(e_1, e_2, ..., e_n)$, we model the correlations between EDUs and obtain the EDU representation as:

$$\begin{aligned} H_i &= \mathrm{MultiHead}(h_i, h_i, h_i) \\ E_i &= \mathrm{Mean}_{\mathrm{sent}}(\text{Self-Attn}(H_i)) \end{aligned} \tag{2}$$

where $\mathrm{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes a standard multi-head self-attention with multiple stacked layers, **Q**, **K**, **V** refer to queries, keys, and values respectively, h_i contains the representations of the n word units in the i-th EDU, $\mathrm{Mean}_{\mathrm{sent}}$ means averaging all the word units in the EDU, and E_i is the EDU representation after the averaging operation.

RST-Graph Encoder. To well utilize the coherence information hidden within the RST dependency graph, we incorporate the RST dependencies into the EDU representations through a graph convolutional network (GCN) [Kipf and Welling, 2016], as shown in Fig. 3. More formally, given the converted RST graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a document, each node $e_i \in \mathcal{V}$ represents an independent EDU, and the edge set $\mathcal{E}(e_i)$ contains the arcs between e_i and all the units connected to e_i, including e_i itself. Moreover, the directions of edges (along, opposite, and self-loop) and rhetorical relations in RST graphs are also attached to the edges. Referring to Fig. 3, we incorporate the discourse structures as follows:

$$\begin{aligned} G_i &= \mathrm{ReLU}((\sum_{e_j \in \mathcal{E}(e_i)} W E_j + b)/|\mathcal{E}(e_i)|) \\ \hat{E}_i &= \mathrm{Linear}(\mathrm{dropout}(G_i) \oplus E_i) \end{aligned} \tag{3}$$

[1]Following [Lin *et al.*, 2019], we employed a discourse segmenter (95.1 F1) based on the pointer nets to obtain EDUs.

where $\oplus$ denotes the concatenation function, and $\hat{E}_i$ denotes the refined representation of the EDU e_i. Notably, we divide the summed vectors by the number of the nodes connected to e_i in Eq. 3 aiming to avoid gradient explosion.

Document-Level Encoder. After embedding RST graphs into EDU representations, we further model document context based on the refined EDU representations through another multi-head attention [Vaswani *et al.*, 2017] which shares the same structure as the sentence-level encoder. The document-level encoder is formulated as:

$$\begin{aligned} D_i &= \mathrm{MultiHead}(\hat{E}_i, \hat{E}_i, \hat{E}_i) \\ D &= \mathrm{Mean}_{\mathrm{doc}}(\text{Self-Attn}(D_i)) \end{aligned} \tag{4}$$

where $\mathrm{Mean}_{\mathrm{doc}}$ means averaging all the obtained context-aware EDU representations and D denotes the discourse-aware document context.

3.3 Document-Level NMT Enhanced with RST Dependency Graph

To the best of our knowledge, most of the previous context-aware document-level NMT systems usually design integrated systems that use context information to enhance the encoder and decoder of their systems through various manners. In this work, we aim to keep the overall system as simple as possible to reduce the interference caused by introducing more parameters and revealing the effects of discourse structure purely. Therefore, we integrate the above-mentioned discourse-aware context into the standard Transformer by enhancing the sentence representations, S, with the obtained discourse-aware context, D, as follows:

$$\begin{aligned} S &= \mathrm{MultiHead}(s_i, s_i, s_i) \\ \hat{S} &= S + D \end{aligned} \tag{5}$$

where s_i denotes the word embeddings of the i-th sentence, S denotes sentence representation from the Transformer encoder that shares the same parameters as the EDU encoder,

Model	TED				News				Europarl	
	En-De		En-Zh		En-De		En-Zh		En-De	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
SentNMT	23.32	45.01	11.44	27.15	22.39	42.33	11.00	26.56	28.98	47.47
DocNMT	24.19	44.30	11.64	**27.44**	22.64	42.38	11.16	26.43	29.09	47.27
Ours	24.40	**45.18**	**11.75**	**27.44**	22.83	42.85	**11.65**	**26.56**	**29.39**	**47.94**
Ours (w/ RR)	**24.60**	44.75	11.67	27.36	**23.25**	**42.99**	11.57	26.54	29.36	47.73
DocNMT∗	26.08	47.12	12.99	29.10	27.22	47.54	12.22	28.39	29.32	47.61
Ours∗	26.34	47.47	**13.62**	**30.23**	**28.11**	47.73	12.53	**28.55**	**29.66**	47.87
Ours (w/ RR)∗	**26.99**	**47.48**	13.58	29.89	27.95	**47.84**	**12.61**	28.46	29.62	**47.93**

Table 1: Overall results on English-German and English-Chinese translation tasks. Sign "*" means the pre-training strategy is utilized.

and $\hat{S}$ denotes the enhanced sentence representation with the discourse-aware context. After that, the obtained sentence representations are further input to the vanilla Transformer decoder for language generation and model learning.

4 Experimentation

Datasets. We conduct several experiments on the English-German and English-Chinese translation tasks with corpora from the following three domains: **TED (En-De/En-Zh):** For English-German, we use TED talks from IWSLT2017 [Cettolo *et al.*, 2012] evaluation campaigns[2] as the training corpus, tst2016-2017 as the test corpus, and the rest set as the development corpus. For English-Chinese, we use TED talks from IWSLT2015 [Cettolo *et al.*, 2012] evaluation campaigns as the training corpus, tst2010-2013 as the test corpus, and dev2010 as the development corpus. **News (En-De/En-Zh):** For English-German, we take News Commentary V11 as our training corpus, the WMT newstest2015 our development corpus, and newstest2016 as the test corpus. For English-Chinese, we take the News Commentary V14 as our training corpus, newstest2017 as the development corpus, and newstest2018 as the test corpus. **Europarl (En-De):** The corpus are extracted from the Europarl V7 according to Maruf and Haffari [2018]. We obtain all the above corpora from Maruf *et al*. [2019][3], and the statistics of these corpora are provided in the Appendix. We tokenize and truecase all the datasets with the scripts of Moses Toolkit[4]. We also apply byte pair encoding[5] to segment sentences with 30K merge operations. We employ the current state-of-the-art RST parser [Zhang *et al.*, 2021] that was trained on RST-DT [Carlson *et al.*, 2001] to get discourse trees of the segmented documents for experimentation.

Model Settings and Evaluation Metrics. We took the vanilla Transformer [Vaswani *et al.*, 2017] as our baseline system (SentNMT) and applied the same system configuration as the Transformer in our experiments. Specifically, the hidden size and filter size were set to 512 and 2048, respectively. Both encoder and decoder were composed of 6 hidden layers. The source and target vocabulary size were set to 30K. The beam size and dropout rate were set to 5 and 0.1, respectively. We used the Adam optimizer to train our model and evaluated the system with case-insensitive BLEU (multi-bleu) [Papineni *et al.*, 2002] and METEOR [Denkowski and Lavie, 2014].

4.1 Overall Results

Compared with Baseline Systems

In this work, we compare the performance of our model with the context-agnostic baseline system (SentNMT) [Vaswani *et al.*, 2017] and the context-aware baseline system (DocNMT)[6]. We implement the Disco2NMT model under two different discourse settings according to whether using the rhetorical relations for discourse incorporation (w/ RR) or not. Table 1 reports the overall results on different language pairs and domains.

The top half of Table 1 reports the results of our system and the baseline systems without using the pre-training strategy. The results show that all the best scores are yielded by our proposed Disco2NMT system with or without the relation labels used. And having rhetorical relations attached to the graph edges does not bring better results necessarily. Although the performance improvements on the context-aware DocNMT system are insignificant in some cases, the overall results indicate that discourse structures like the RST graph are effective in document-level machine translation.

To our knowledge, document-level NMT usually suffers from the issue of corpus size limitation. Therefore, most existing document-level NMT systems usually employ large-scale external sentence-level data for model pre-training. Following these works, we also apply the pre-training strategy to the context-aware baseline system DocNMT and our Disco2NMT for further comparison. Specifically, we mix the training sets of corpora from different domains as our training corpus to pre-train the model in sentence-level translation (only the sentence-level encoder and decoder are tuned in this period). After that, the entire model (sentence- and document-level encoder and decoder) is fine-tuned on in-domain document-level corpora. The results are shown in the under part of Table 1. Comparing the results with and without the pre-training strategy, we find that using out-of-domain corpora can boost the document-level translation performance. Besides, with the pre-training strategy used, our

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

[3]https://github.com/sameenmaruf/selective-attn/tree/master/data

[4]https://github.com/moses-smt/mosesdecoder/tree/master/scripts

[5]https://pypi.org/project/subword-nmt

[6]We achieve the DocNMT model by modeling document context based on sentences and incorporate the contextual information into the sentence representation as described in Section 3.3

Model	BLEU	TER
HAN [2018]*	24.45	56.9
HAN-DS [2020]*	24.84 (+0.39)	56.4 (-0.5)
Disco2NMT	25.19 (+0.74)	56.1 (-0.8)

Table 2: Performance comparison with previous studies. Sign "*" denotes the results are borrowed from their published papers.

Disco2NMT model significantly surpasses the DocNMT system by 0.91 BLEU and 1.13 METEOR. The results show that having Disco2NMT learn general translation knowledge from large-scale external data may strengthen its ability to absorb the discourse structure, thus improving its performance in document-level NMT.

Comparing lines 3 and 4 and lines 6 and 7 in Table 1, we find that attaching the predicted rhetorical relations to arcs between EDUs does not work positively in most cases. And the possible reasons are as follows: On the one hand, the annotated relation labels in RST-DT are very imbalanced, making existing RST parsers perform weakly on relations with fewer examples. Therefore, harnessing the pre-trained RST parser to produce relations for translation corpora will bring in much noise. On the other hand, the results show that Disco2NMT w/ `RR` is more probable to outperform Disco2NMT on the News corpora than the other two fields because the RST parser we use was trained on news articles. This shows that domain inadaptability is another reason for the poor performance of the model with the setting of "w/ `RR`". In general, the error propagation problem generated during the discourse fusion process remains challenging.

Compared with Existing Discourse-Aware Studies
In the literature, Chen *et al*. [2020] is the first to demonstrate the usefulness of discourse structure in document-level NMT. They achieve discourse incorporation by enriching the word representation with the embedded discourse tree path from each leaf node to the root. Different from them, instead of using the discourse paths as additional features, we borrow RST structures directly to guide the context modeling process. Notably, besides the naked tree structure, we also harness the nuclearity and rhetorical relation information for discourse incorporation. Table 2 presents the results of our system and HAN-DS [Chen *et al.*, 2020] using the same data and model settings, e.g., using the pre-training strategy. Since HAN-DS derives from the context-aware model of Miculicich *et al*. [2018] (HAN), we also report the results of HAN for reference.

The results in table 2 show that both Disco2NMT and HAN-DS outperform the non-discourse-aware HAN, which proves the efficacy of discourse structure in document-level NMT. Besides, our Disco2NMT model further outperforms HAN-DS to a certain extent on both BLEU and TER [Snover *et al.*, 2006] metrics, suggesting that employing discourse structure in a deeper way ensures the NMT model better absorbs the rhetorical information for document-level NMT.

4.2 Analysis

Statistics on Conjuctions. To intuitively investigate the effects brought by our Disco2NMT system on machine translation, we count four kinds of conjunctions that appear most frequently in the translation results. Specifically, we focus on coordinating conjunctions like "and" and "or", adversative conjunctions like "but" and "however", causal conjunctions like "because" and "therefore", and concessive conjunctions like "although". For performance comparison, we perform statistics on the translations produced by the context-agnostic SentNMT, the context-aware DocNMT, and our discourse-aware Disco2NMT. We report the total number of conjunctions mentioned above in Fig. 4. The results show that our method tends to use more explicit connectives to connect text units within an article. And this can improve the coherence of the translation text to a certain extent. More detailed statistics are presented in the Appendix.

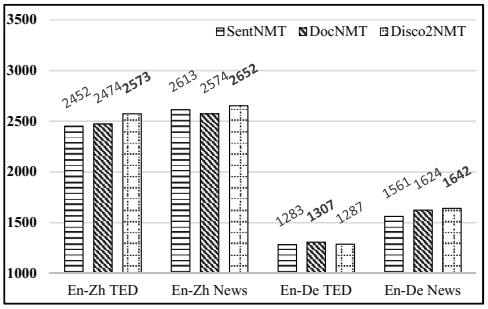

Figure 4: Statistics of conjunctions that appear in the translations.

Model	En-De			En-Zh	
	TED	News	Europ.	TED	News
SentNMT	0.382	0.384	0.413	0.394	0.379
DocNMT	0.385	0.384	0.414	0.393	0.392
Disco2NMT	**0.391**	**0.385**	**0.416**	**0.398**	**0.400**

Table 3: Discourse coherence evaluation of model translations.

Discourse Coherence Evaluation. Following [Lapata *et al.*, 2005], we evaluate the discourse coherence of translations by measuring the cosine similarity between two adjacent sentences' representations in an article. According to their method, the sentence representation is obtained by averaging the distributed word vectors in the sentence, and the coherence score is achieved by averaging all the cosine similarity scores of the article. In this work, we use the pre-trained 300-dimension Fasttext word vectors[7] to generate sentence representations for English-German and English-Chinese translation tasks to evaluate discourse coherence. We report the averaged coherence score of translated articles in each corpus, shown in Table 3. Compared with the context-agnostic SentNMT model, the context-aware model, DocNMT, can effectively improve the discourse coherence of translations. In Particular, our Disco2NMT system can further enhance the improvement through discourse-aware context modeling. Whether it is from the translation of explicit conjunctions above or from the implicit similarity between sentences, the experimental results show that our method can improve the coherence of machine translation to a certain extent.

[7]https://fasttext.cc/docs/en/crawl-vectors.html

Model	En-De			En-Zh	
	TED	News	Europ.	TED	News
DocNMT	24.19	22.64	29.09	11.64	11.16
DocNMT-EDU	24.13	22.66	29.12	11.66	11.14
Disco2NMT	**24.60**	**23.25**	**29.39**	**11.75**	**11.65**

Table 4: Performance comparison between modeling contextual information on the sentences and segmented EDUs.

Model	TED	News	Europ.
GCN-BEF	26.99	28.11	29.66
GCN-AFT	26.12	27.54	29.30

Table 5: Results of applying GCN to different system positions.

Effects of Modeling Context Based on EDUs. As stated before, our discourse-aware document context is modeled on segmented EDUs instead of sentences. Therefore, two factors may cause performance improvements in our Disco2NMT system compared with DocNMT, i.e., the contextual information from EDUs and the discourse structure we use. To prove that the performance improvements of our model come from the incorporated discourse information instead of the segmented EDUs, we conduct an additional experiment where we exclude the discourse structures and only use the EDUs for context modeling, noted by DocNMT-EDU. As shown in Table 4, the similar performances of DocNMT and DocNMT-EDU indicate that using the sentences or segmented EDUs for context modeling makes a tiny difference. In particular, the last two lines show that our Disco2NMT system improves the performance of DocNMT-EDU using only discourse information, which suggests the effectiveness of our method.

Ablation Analysis of GCN Application. In this part, we perform an ablation analysis on the application of GCN in discourse-aware context modeling. As shown in Table 5, we conduct experiments on the English-German translation tasks to show the effects of employing the GCN model in different positions, i.e., before and after the document encoder, noted by GCN-BEF and GCN-AFT respectively. The results show that directly applying GCN to sentence-level EDU representations (GCN-BEF) for discourse incorporation performs better than applying it to document-level EDU representations (GCN-AFT). Therefore, we take the scheme of applying GCN before the document-level encoder for discourse incorporation in this work. It is worth mentioning that we also attempt to use multiple GCN layers for encoding in our experiments, but the overall results show that the number of GCN layers has a weak impact on the performance of our approach.

Pronoun Translation. Furthermore, we follow [Miculicich Werlen and Popescu-Belis, 2017] to evaluate the coreference and anaphora of our generated translations using the reference-based metric, namely, Accuracy of Pronoun Translation (APT). We list the results of our system and the baseline systems in Table 6. The results show that although our system achieves better results in En-Zh translation tasks, it performs relatively poorly on TED and Europarl in En-De translation. To some extent, this result is reasonable since RST pays more attention to implicit coherence hidden within texts rather than shallow lexical cohesion in between words;

Model	En-De			En-Zh	
	TED	News	Europ.	TED	News
SentNMT	**77.12**	83.86	74.06	63.09	55.24
DocNMT	76.89	85.24	**74.40**	63.91	55.46
Disco2NMT	76.67	**85.40**	73.92	**64.43**	**55.57**

Table 6: Pronoun translation results (APT) of our Disco2NMT system and the baseline systems on the En-De and En-Zh tasks.

the original RST structure and pronouns do not have an obvious correlation in theory. Besides, the domain inadaptability is also a possible reason since the discourse structure we use is limited to the news domain, and our Disco2NMT model performs well on the news corpora. To our knowledge, recent research [Tan *et al.*, 2021] has proven that a pronoun-targeted method can significantly improve pronoun translation performance. In view of this, it is worthy of in-depth study to enhance Disco2NMT's ability in modeling lexical cohesion for more coherent translations.

5 Related Work

As a new research hotspot in the machine translation community, document-level NMT has made considerable progress in recent years with various context-aware approaches proposed so far. The mainstream of previous context-aware studies employed additional context related modules to capture document context information for the vanilla Transformer [Wang *et al.*, 2017; Voita *et al.*, 2018; Maruf *et al.*, 2019; Sugiyama and Yoshinaga, 2021; Fernandes *et al.*, 2021]. Among these studies, [Miculicich *et al.*, 2018; Zhang *et al.*, 2018; Xu *et al.*, 2021] used local context sentences in their translation system, while [Maruf and Haffari, 2018; Tan *et al.*, 2019; Ma *et al.*, 2020] utilized the entire ones for more comprehensive context modeling. In addition, some studies applied cache/memory-based approaches to store the representations of translated sentences to strengthen their original system [Tu *et al.*, 2018; Kuang *et al.*, 2018]. Although the above work has put forward the study of document-level NMT to a certain extent, the primary discourse information is widely ignored in current document-level NMT research. To our knowledge, [Chen *et al.*, 2020] is the first work that incorporates discourse structures into document-level NMT. This shows that the work on discourse-aware NMT is still limited, and it is worthy of further study in future work.

6 Conclusion

In this paper, we proposed Disco2NMT, which models discourse-aware document context through a GCN net and incorporates the context information into an attention-based hierarchical model for document-level NMT. Experimental results show that our method can significantly improve the performance of document-level machine translation.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600 and the National Natural Science Foundation of China (NSFC) via Grant Nos. 62076175 and 61876118.

References

[Carlson *et al.*, 2001] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *SIGDIAL*, 2001.

[Cettolo *et al.*, 2012] Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: Web inventory of transcribed and translated talks. In *EAMT*, 2012.

[Chen *et al.*, 2020] Junxuan Chen, Xiang Li, Jiarui Zhang, et al. Modeling discourse structure for document-level neural machine translation. *arXiv:2006.04721*, 2020.

[Denkowski and Lavie, 2014] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT*, 2014.

[Fernandes *et al.*, 2021] Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. Measuring and increasing context usage in context-aware machine translation. In *ACL-IJCNLP*, 2021.

[Hirao *et al.*, 2013] Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, et al. Single-document summarization as a tree knapsack problem. In *EMNLP*, 2013.

[Hobbs, 1979] Jerry R Hobbs. Coherence and coreference. *Cognitive science*, 3(1), 1979.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907*, 2016.

[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 *COLING*, 2018.

[Lapata *et al.*, 2005] Mirella Lapata, Regina Barzilay, et al. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, volume 5, 2005.

[Li *et al.*, 2014a] Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. Text-level discourse dependency parsing. In *ACL*, 2014.

[Li *et al.*, 2014b] Yancui Li, wenhe Feng, jing Sun, Fang Kong, and Guodong Zhou. Building chinese discourse corpus with connective-driven dependency tree structure. In *EMNLP*, 2014.

[Lin *et al.*, 2019] Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. A unified linear-time framework for sentence-level discourse parsing. In *ACL*, 2019.

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

[Mann and Thompson, 1988] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3), 1988.

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

[Maruf *et al.*, 2019] Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Selective attention for context-aware neural machine translation. In *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 *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 (APT). In *DiscoMT*, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Snover *et al.*, 2006] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *AMTA*, 2006.

[Stede and Neumann, 2014] Manfred Stede and Arne Neumann. Potsdam commentary corpus 2.0: Annotation for discourse research. In *LREC*, 2014.

[Sugiyama and Yoshinaga, 2021] Amane Sugiyama and Naoki Yoshinaga. Context-aware decoder for neural machine translation using a target-side document-level language model. In *NAACL*, 2021.

[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 *EMNLP*, 2019.

[Tan *et al.*, 2021] Xin Tan, Longyin Zhang, and Guodong Zhou. Coupling context modeling with zero pronoun recovering for document-level natural language generation. In *EMNLP*, 2021.

[Tu *et al.*, 2018] Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. Learning to remember translation history with a continuous cache. *TACL*, 6, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *NIPS*, 2017.

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

[Wang *et al.*, 2017] Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploiting cross-sentence context for neural machine translation. In *EMNLP*, 2017.

[Xu *et al.*, 2021] Hongfei Xu, Deyi Xiong, Josef Van Genabith, and Qiuhui Liu. Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *IJCAI*, 2021.

[Zhang *et al.*, 2018] Jiacheng Zhang, Huanbo Luan, Maosong Sun, et al. Improving the transformer translation model with document-level context. In *EMNLP*, 2018.

[Zhang *et al.*, 2021] Longyin Zhang, Fang Kong, and Guodong Zhou. Adversarial learning for discourse rhetorical structure parsing. In *ACL-IJCNLP*, 2021.