CoDoNMT: Modeling Cohesion Devices for Document-Level Neural Machine Translation

Yikun Lei, Yuqi Ren, Deyi Xiong *
College of Intelligence and Computing, Tianjin University, Tianjin, China
{yikunlei,ryq20,dyxiong}@tju.edu.cn

Abstract

Cohesion devices, e.g., reiteration, coreference, are crucial for building cohesion links across sentences. In this paper, we propose a document-level neural machine translation framework, CoDoNMT, which models cohesion devices from two perspectives: Cohesion Device Masking (CoDM) and Cohesion Attention Focusing (CoAF). In CoDM, we mask cohesion devices in the current sentence and force NMT to predict them with inter-sentential context information. A prediction task is also introduced to be jointly trained with NMT. In CoAF, we attempt to guide the model to pay exclusive attention to relevant cohesion devices in the context when translating cohesion devices in the current sentence. Such a cohesion attention focusing strategy is softly applied to the self-attention layer. Experiments on three benchmark datasets demonstrate that our approach outperforms state-of-the-art document-level neural machine translation baselines. Further linguistic evaluation validates the effectiveness of the proposed model in producing cohesive translations.

1 Introduction

Neural Machine Translation (NMT) has become the dominant approach for machine translation and achieved substantial progress in comparison to statistical machine translation. Some studies even claim that NMT has reached human parity (Hassan et al., 2018). Despite this, most NMT models are at the sentence level, which translate documents sentence by sentence, ignoring inter-sentential dependencies. Documents translated in this way are usually incoherent and inconsistent across sentences.

In order to address this issue, a wide range of efforts have been made to leverage inter-sentential context information for document-level NMT (Tiedemann and Scherrer, 2017; Zhang et al., 2018; Voita et al., 2019; Tan et al., 2019; Maruf et al., 2019; Xu et al., 2020b; Zhang et al., 2021). Most efforts have been dedicated to modeling local or global context via additional encoders, attention, cache, concatenating inputs, etc (Maruf et al., 2021). However, these approaches normally focus on the way of integrating context into translation, rather than the context itself. The basic assumption behind this is that models are able to detect relevant contextual information. However, Kim et al. (2019) and Li et al. (2020) find that most of the improvements obtained by these approaches cannot be explained as leveraging the right context. We suggest that it could be not a good choice to treat contextual words equally and rely on models to learn contextual clues in an implicit way. This is because it is usually difficult for NMT to capture key information from a long context through itself (Yin et al., 2021). Hence, pinpointing semantically or grammatically relevant context words in an explicit way is desirable for document-level NMT.

In this paper, different from the aforementioned context modeling schemes, we model contextual information for document-level NMT in an explicit way via cohesion devices. Cohesion devices (e.g., reiteration, co-reference) are widely acknowledged as important linguistic items that chain sentences into cohesive discourse (Halliday and Hasan, 1976). Moreover, the interpretation of one cohesion device depends on the corresponding device that is paired to it (Halliday and Hasan, 1976). Consider the following text "*Amy went to the party. She sat with Sara.*". The interpretation of the cohesion device *she* is deeply related to the cohesion device *Amy*. Therefore, explicitly modeling these cohesion devices may guide document-level NMT to actively explore contextual clues, so as to yield cohesive translations.

Inspired by this, we propose **CoDoNMT**, as shown in Fgiure 1, to explore **Co**hesion devices for **Do**cument-level **N**eural **M**achine **T**ranslation.

*corresponding author

Proceedings of the 29th International Conference on Computational Linguistics, pages 5205–5216
October 12–17, 2022.

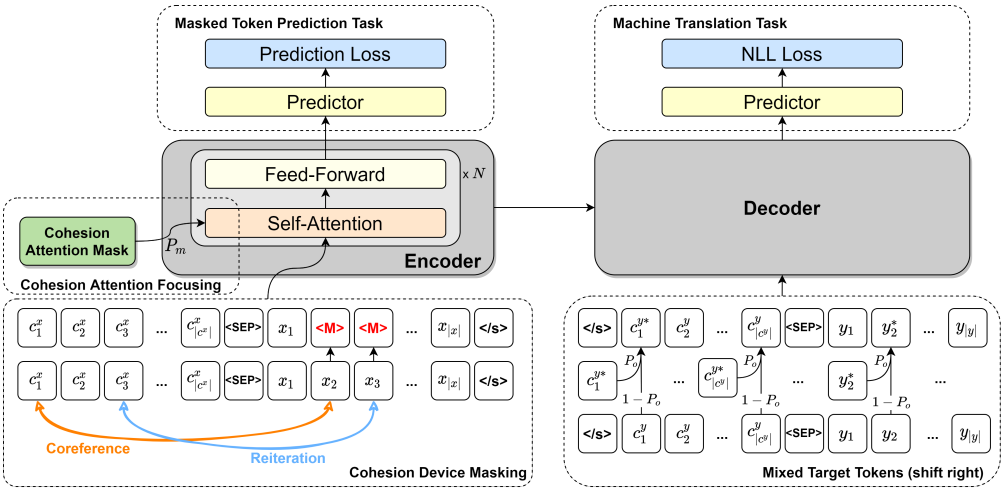

Figure 1: The diagram of CoDoNMT. We prepend the preceding three sentences to each current sentence as its context on both the source and target side. CoDM is applied on the source side of the current sentence while CoAF is softly used in each self-attention layer of the encoder. A masked token prediction task corresponding to CoDM is employed as an auxiliary task to the primary translation task, where the concatenated input (source sentence + context) is translated into the target language. Exposure bias mitigation is applied during training: $c_{\cdot}^{y*}$ and $y_{\cdot}^{*}$ indicate the predicted words in translations of the context and the current sentence, respectively.

Particularly, we present Cohesion Device Masking (CoDM) and Cohesion Attention Focusing (CoAF) as two essential components for CoDoNMT, in an attempt to force NMT to explicitly predict masked cohesion devices and to focus its attention exclusively on related cohesion devices.

Cohesion Device Masking We concatenate the previous context to the current sentence on both the source and target side. Cohesion devices of the current sentence on the source side are masked. In doing so, we force NMT to actively explore previous context to predict the masked cohesion devices, which may teach document-level NMT to pinpoint relevant linguistic context for translation.

Cohesion Attention Focusing As mentioned above, in order to correctly interpret and translate cohesion devices in the current sentence, we need to capture their paired cohesion devices in the concatenated context. For this, we force NMT to pay exclusive attention to previous cohesion clues in the context with attention masks when translating cohesion devices in the current sentence. In this way, we narrow the range of context and enable the model to evade irrelevant contextual information.

In a nutshell, our contributions are three-fold.

- We propose CoDoNMT for document-level NMT, which explicitly explores cohesion devices to capture context information.
- We introduce CoDM and CoAF in CoDoNMT to force NMT to predict masked cohesion devices with context information and to attend to only cohesion devices in the context for translating cohesion devices in the current sentence, respectively.
- We conduct experiments on three widely-used datasets and a linguistic contrastive test set. Results of both automatic and linguistic evaluation demonstrate that our methods are able to significantly improve translation quality over previous state-of-the-art document-level NMT models.

2 Cohesion Devices

A discourse is cohesive when sentences are properly linked by cohesion devices. From the linguistic perspective, cohesion devices can be divided into two categories: lexical cohesion devices and grammatical cohesion devices (Halliday and Hasan, 1976). In this paper, we consider reiteration, synonym and super-subordinate for lexical cohesion devices, and co-reference for grammatical cohesion devices. We choose these devices because they are common and can be annotated automatically.

Reiteration: Reiteration refers to the repetition of the same words in a discourse. This is a common phenomenon in discourse (Church, 2000). And it is easy to detect. Note that we exclude stop words when detecting any type of cohesion devices.

Synonym: We use WordNet (Fellbaum, 2000) to define synonyms, which is a large lexical database

of English. Nouns, verbs, adjectives and adverbs are grouped into sets of semantic groups called *synsets*. We denote $synset(w)$ as a set that includes synonyms grouped in the same synset as word w in WordNet.

Super-subordinate: Superordinate and subordinate are formed by words with an is-a semantic relationship, such as *apple* and *fruit* (hypernym), *furniture* and *cupboard* (hyponym), and so on. As the super-subordinate relation is also encoded in WordNet, we still use WordNet to detect hypernyms and hyponyms. Let $hypset(w)$ be a set that includes both hypernyms and hyponyms in WordNet for word w.

Co-reference: Co-reference is a relationship between two words or phrases in which both refer to the same person or thing and one is a linguistic antecedent of the other. We use CoreNLP (Manning et al., 2014) to parse co-reference relations between the current sentence and its context sentences.

3 CoDoNMT

Figure 1 illustrates the diagram of the proposed CoDoNMT, which uses the standard Transformer (Vaswani et al., 2017) as its backbone. We prepend the previous three sentences to the current sentence on both the source and target side, separated by a special token (i.e., <SEP>). We apply CoDM to the source-side input and CoAF to the self-attention layer of the encoder.

3.1 Cohesion Device Masking

The key of CoDM is to mask cohesion devices in the current sentence and force the model to predict those masked tokens using inter-sentential context. Predicting cohesion devices might offer the model the ability to establish cohesion links, so as to make translation cohesive.

Obtaining Cohesion Devices We denote $\boldsymbol{x}$ as the source side of the current sentence and $\boldsymbol{c^x}$ as the preceding context of $\boldsymbol{x}$. $|\boldsymbol{x}|$ and $|\boldsymbol{c^x}|$ denote the length of $\boldsymbol{x}$ and $\boldsymbol{c^x}$, respectively. Correspondingly, $\boldsymbol{y}$ indicates the target side of the current sentence and its preceding context is $\boldsymbol{c^y}$. $|\boldsymbol{y}|$ and $|\boldsymbol{c^y}|$ are the length of $\boldsymbol{y}$ and $\boldsymbol{c^y}$, respectively.

For the ith word x_i in $\boldsymbol{x}$, we consider x_i as a lexical cohesion device if there exists a context word that is the same as x_i, or in the $synset(x_i)$ or $hypset(x_i)$ in $\boldsymbol{c^x}$. We use CoreNLP to parse each concatenated input to obtain co-reference links between $\boldsymbol{x}$ and $\boldsymbol{c^x}$, and refer to words occurring in the detected co-reference links as grammatical cohesion devices. We use $\mathcal{D}$ to denote the set of both lexical and grammatical cohesion devices in $\boldsymbol{x}$.

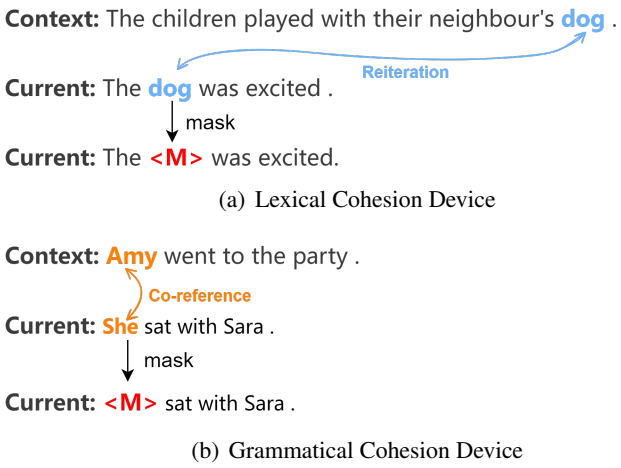

Figure 2: Examples of cohesion device masking. The same words "dog" in Figure 2(a) are reiteration devices. "Amy" and "She" in Figure 2(b) are co-reference devices.

Masking Strategy We mask all cohesion devices in $\mathcal{D}$ to explore contextual dependencies established by these devices as many as possible. Figure 2 shows examples of CoDM.

As there are not many cohesion devices sometimes, only masking cohesion devices is not sufficient. We hence use a masking ratio r as a threshold to mask other words (randomly selected) in addition to cohesion devices. The total number of words being masked in $\boldsymbol{x}$ is $\lceil |\boldsymbol{x}| \times r \rceil$ where $\lceil \cdot \rceil$ indicates the upward rounding operation. We denote the set of masked tokens as $\mathcal{M}$ and the masked version of $\boldsymbol{x}$ as $\acute{\boldsymbol{x}}$ where tokens in $\mathcal{M}$ are substituted by a speical symbol (e.g., <M>). We concatenate $\boldsymbol{c^x}$ and $\acute{\boldsymbol{x}}$ as the input fed into the encoder and use the corresponding hidden states of the last encoder layer to predict the masked tokens in $\mathcal{M}$. The loss of predicting the masked tokens is calculated as follows:

$$\mathcal{L}_{\text{mask}}(\mathcal{M}|\acute{\boldsymbol{x}}, \boldsymbol{c^x}) = -\sum_{i=1}^{|\mathcal{M}|} \log P(\mathcal{M}_i|\acute{\boldsymbol{x}}, \boldsymbol{c^x}) \quad (1)$$

where $\mathcal{M}_i$ is the ith token in $\mathcal{M}$ and $P(\mathcal{M}_i|\acute{\boldsymbol{x}}, \boldsymbol{c^x})$ represents the probability that the model predicts $\mathcal{M}_i$ given the current sentence and its context.

3.2 Cohesion Attention Focusing

Not all information in the context is useful for translating the current sentence. We hence want to guide

the model to attend to only cohesion devices in the context when we translate cohesion devices in the current sentence as they are linguistically linked to each other. We achieve this via a cohesion attention mask.

Constructing the Cohesion Attention Mask We use a key-value pair (x_i, l_{x_i}) to store cohesion devices linked to x_i, where l_{x_i} is a list whose elements are the cohesion devices related to x_i. Specifically, for x_i, if the jth word c_j^x in $\boldsymbol{c^x}$ is the same as x_i, or in the $synset(x_i)$ or $hypset(x_i)$, c_j^x is cohesively linked to x_i. We hence add x_i and c_j^x into $l_{c_j^x}$ and l_{x_i}, respectively. Through CoreNLP, we are able to directly obtain co-reference links between x_i and $\boldsymbol{c^x}$. For each word in each co-reference link, we first find its corresponding cohesion device list l, and then add the remaining words in the co-reference link to l. We denote the collection of all key-value pairs as L.

After obtaining L, we construct a cohesion attention mask $\mathrm{M} \in \mathbb{R}^{N \times N}$ and initialize each item with 0. $N = |\boldsymbol{x}| + |\boldsymbol{x^c}|$ indicates the length of the concatenated source input. We use L to set value for each item in the cohesion attention mask matrix. For each key-value pair (x_i, l_{x_i}), we obtain the positions of x_i (p_{x_i}) and words in l_{x_i} ($p_{w \in l_{x_i}}$) in the concatenated source input. Then, at the p_{x_i} row, we mask out all items whose column positions are not $\in \{p_{w \in l_{x_i}}\}$ by setting their values to $-\infty$. This is similarly done for each column $\in \{p_{w \in l_{x_i}}\}$. Note that we do not mask out (inter/intra-sentential) interactions between ordinary words that are not cohesion devices and intra-sentential interactions among words even when they are not cohesion devices. We only force cohesion devices to exclusively attend to cohesion devices that are linked to them in other sentences.

Figure 3 illustrates the cohesion attention mask. For words (e.g., "from", "home") that are not cohesion devices, they can only attend to words in the same sentence. For cohesion devices, they can attend not only to their intra-sentential context words but also to associated cohesion devices across sentences.

Applying the Cohesion Attention Mask In order to make the model not lose the ability to capture important contextual information, we apply the constructed cohesion attention mask softly by using a probability threshold P_m to control whether to apply the cohesion attention mask to self-attention. Thus the model is trained with and without the cohesion attention mask, and we assume that the model is able to acquire the ability to capture cohesion information autonomously with this training strategy. The attention $\boldsymbol{A^l}$ in l-th self-attention can be calculate as follows.

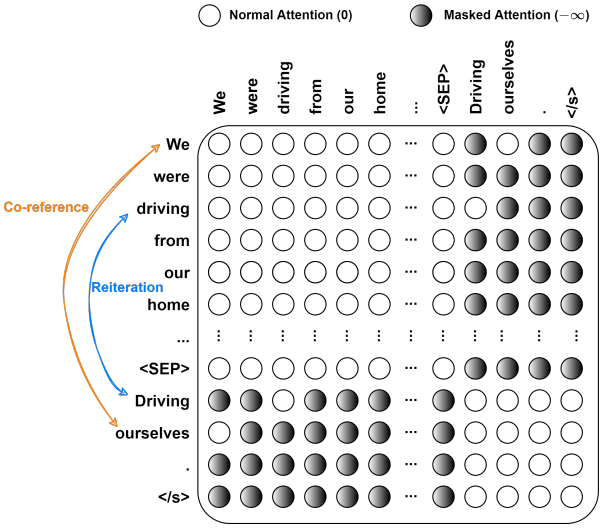

Figure 3: An example of the cohesion attention mask. For brevity, we omit some words in the sentence. The complete sentence is "We were driving from our home to a little farm . <SEP> Driving ourselves . </s>". White positions are set to 0 while black positions are set to $-\infty$.

$$\boldsymbol{A^l} = \begin{cases} \text{Softmax}(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d/h}} + \mathrm{M}) & , \text{if } \varepsilon > P_m \\ \text{Softmax}(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d/h}}) & , \text{otherwise} \end{cases} \quad (2)$$

where the matrices $\boldsymbol{Q}, \boldsymbol{K}$ represent queries and keys in self-attention. d and h indicate the dimension of hidden states and the number of heads, respectively. ε is sampled from $U \backsim (0, 1)$. Only when the sampled ε is larger than the threshold P_m, the cohesion attention mask is applied.

3.3 Training

Our model performs both masking prediction and translation in a multi-task learning fashion. Hence the training objective of our model is composed of the traditional negative log-likelihood (NLL) and the masked token prediction loss $\mathcal{L}_{mask}$.

During training, we use the ground-truth target context (Teacher-Forcing) while we use the previously decoded output tokens as the target context during inference. As a result, our model suffers from exposure bias. Inspired by Zhang et al. (2019), we mix the ground-truth words with the predicted

words as the decoder input during training to alleviate the exposure bias. We refer to a word predicted by the model as the predicted word. For each word in golden context sentences and current sentence, we use the probability P_o to control whether to replace the ground-truth word with its corresponding predicted word. Following Zhang et al. (2019), we gradually decrease P_o from 1 according to the following decay function:

$$P_o = \frac{\mu}{\mu + \exp\left(e/\mu\right)} \tag{3}$$

where μ is a hyper-parameter that controls the decay rate. e is the index of training epochs starting from 0.

For each step, the predicted words are obtained through word-level greedy search. We denote the mixed target context and current sentence as $\boldsymbol{c^{\acute{y}}}$ and $\boldsymbol{\acute{y}}$, respectively. Thus, the NLL is reformulated as follows:

$$\mathcal{L}_{\text{NLL}}(\boldsymbol{y}|\boldsymbol{\acute{x}}, \boldsymbol{c^{x}}, \boldsymbol{c^{\acute{y}}}) = -\sum_{i} \log P(\boldsymbol{y}_i|\boldsymbol{\acute{y}}_{<i}, \boldsymbol{\acute{x}}, \boldsymbol{c^{x}}, \boldsymbol{c^{\acute{y}}}) \tag{4}$$

Please note that we feed the masked version of the current sentence into the encoder for translation during training. In order to correctly translate tokens in masked positions, our model has to utilize previous context, not only for the masked token prediction task, but also for the translation task.

The two loss functions are integrated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}}(\boldsymbol{y}|\boldsymbol{\acute{x}}, \boldsymbol{c^{x}}, \boldsymbol{c^{\acute{y}}}) + \lambda \mathcal{L}_{\text{mask}}(\mathcal{M}|\boldsymbol{\acute{x}}, \boldsymbol{c^{x}}) \tag{5}$$

where λ is a hyper-parameter that balances the contribution from the masked token prediction task.

4 Experiments

To examine the effectiveness of our proposed approaches, we carried out experiments on three widely-used datasets and linguistic evaluation on a contrastive test set.

4.1 Data and Settings

Following previous work (Zhang et al., 2021), we used three datasets on two different language pairs as the benchmark datasets, which are TED (Cettolo et al., 2012), Opensubtitles (Maruf et al., 2018) and Europarl7 (Maruf et al., 2018).

Dataset	Language	#Sentences	#Documents
		train/dev/test	train/dev/test
TED	En-De	0.2M/09k/2.2k	1.7k/7/22
Opensubtitles	En-Ru	0.3M/6k/9k	23k/461/693
Europarl7	En-De	0.1M/2k/3.3k	3.6k/69/107

Table 1: Statistics of the used datasets on different language pairs.

- **TED (English-German)**: The corpus contains transcriptions of TED talks from IWSLT 2017. Each talk is used as a document, aligned at the sentence level. *dev2010* was used as our development set and *tst2016-tst2017* for testing.
- **Opensubtitles (English-Russian)**: This corpus is extracted from the OpenSubtitles2016 corpus (Maruf et al., 2018), where sentences are segmented and aligned using additional information.
- **Europarl7 (English-German)**: Following (Maruf et al., 2018), we used the same method to preprocess the raw Europarl v7 Corpus (Koehn, 2005) and extract the parallel coprus.

The statistics of these corpora are shown in Table 1. We used scripts from MOSES (Koehn et al., 2007) to tokenize and truecase sentences. We applied BPE (Sennrich et al., 2016) with 30K merge operations for each language in the datasets. Translation quality was evaluated by BLEU (Papineni et al., 2002). [1]

We followed the same Transformer base setting used in (Vaswani et al., 2017) and trained all models on 4 GeForce RTX 2080 Ti GPU. The dropout was set to 0.1. The masking ratio r was set to 0.15. The weight of the masked token predction loss λ was set to 0.5. The probability threshold P_m was set to 0.5. For TED, we set μ as 10. For Europarl7 and Opensubtitles, we set μ as 12. For inference, we set the beam size to 4. The source code is available at https://github.com/codeboy311/CoDoNMT.

4.2 Baselines

We used five baselines to compare against our model. The sentence-level baseline Sent is the standard Transformer (Vaswani et al., 2017) trained on sentence-level parallel data. The rest four baselines are all document-level NMT, including: 1) DocT

[1]We use sacrebleu to calculate BLEU score for each dataset, and the signature of sacrebleu we used is "BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.1".

Model	TED En-De	Opensubtitles En-Ru	Europarl7 En-De	AVG	#Param
Sent (Vaswani et al., 2017)	24.30	19.50	30.70	24.83	50M
DocT (Zhang et al., 2018) †	25.04	20.21	30.67	25.31	72M
HAN (Miculicich et al., 2018) †	25.70	20.08	26.61	24.13	70M
CADec (Voita et al., 2019) †	26.08	19.46	30.36	25.30	91M
MHT (Zhang et al., 2021) †	26.22	20.46	31.25	25.97	80M
Our Model	**26.89**	**21.49**	**32.07**	**26.82**	50M

Table 2: Overall results on the TED, OpenSubtitles and Europarl translation tasks. †indicates that the results of baselines are reported from corresponding papers.

Model	Deixis	Lexical Cohesion
Sent (Voita et al., 2019)	50.0%	45.9%
CADec (Voita et al., 2019)	81.6%	58.1%
Concat (Voita et al., 2019)	83.5%	47.5%
Our Model	**89.4%**	**62.3%**

Table 3: Linguistic evaluation results on the contrastive test set (accuracy).

(Zhang et al., 2018) that employs an additional encoder for context, 2) HAN (Miculicich et al., 2018) which integrates document contextual information from both the source and target side through context-aware hierarchical attention networks, 3) CADec (Voita et al., 2019) that explores contextual information to refine sentence-level translation, and 4) MHT (Zhang et al., 2021) that applies a multi-hop mechanism to imitate reasoning process.

4.3 Main Results

Results on the three translation benchmarks are shown in Table 2. As can be seen, our model achieves the highest BLEU scores on all tasks over both sentence- and document-level baselines. Furthermore, we gain improvements of 0.67, 1.03 and 0.82 BLEU points over the strongest document-level baseline on TED, Opensubtitles and Europarl7, respectively, using fewer parameters. This suggests that modeling cohesion devices is able to benefit document-level NMT and could be better than simply integrating full context into NMT without showing relevant context information explicitly.

4.4 Linguistic Evaluation

To further investigate whether our method is able to improve translation cohesion, we conducted linguistic evaluation on Deixis and Lexical Cohesion using a linguistic contrastive test set (Voita et al., 2019). These two discourse phenomena are relevant to the cohesion devices that we attempt to capture (i.e. co-reference and lexical cohesion devices) in our model.

To make a fair comparison, we follow Voita et al. (2019) to use 6M sentence-level instances to train the sentence-level baseline and then use 1.5M document-level instances to train our document-level model. Results are shown in Table 3. Our model achieves significant improvements on Deixis and Lexical Cohesion compared with sentence-level baseline Sent (Voita et al., 2019) and document-level baselines CADec (Voita et al., 2019) and Concat (Voita et al., 2019). This indicates that our model can make better use of context to deal with discourse phenomena.

4.5 Ablation Study

In order to take a deep look into the improvements gained by our model, we further conducted ablation study to investigate the contributions of the three components in our model: 1) cohesion device masking, 2) cohesion attention focusing and 3) exposure bias mitigation introduced in Section 3.3. Results are shown in Table 4. Without using cohesion device masking, CoDoNMT drops by 0.58, 0.62 and 0.55 BLEU on TED, Opensubtitles and Europarl7, respectively. This demonstrates that forcing the model to predict masked cohesion devices is beneficial for document-level NMT. Similarly, without using the other two techniques, we also see performance drops of 0.2 BLEU for the exposure bias mitigation, 0.34 BLEU for cohesion attention focusing. These results validate the effectiveness of the three methods used in CoDoNMT.

Ablation	TED		Opensubtitles		Europarl7		AVG	
	BLEU	Δ	BLEU	Δ	BLEU	Δ	BLEU	Δ
Our	26.89	-	21.49	-	32.07	-	26.82	-
w/o Exposure Bias Mitigation	26.60	-0.29	21.38	-0.11	31.86	-0.21	26.62	-0.20
w/o CoAF	26.53	-0.36	21.16	-0.33	31.75	-0.32	26.48	-0.34
w/o CoDM	26.31	**-0.58**	20.87	**-0.62**	31.52	**-0.55**	26.23	**-0.59**

Table 4: Ablation study results.

Cohesion Device	TED	Opensubtitles	Europarl7	AVG	Deixis	Lexical Cohesion
Lexical	26.58	21.24	31.99	26.60	84.1%	60.3%
Grammatical	26.64	21.21	31.86	26.57	87.4%	58.2%
Both	**26.89**	**21.49**	**32.07**	**26.82**	**89.4%**	**62.3%**

Table 5: Impact of different cohesion types modeled on translation quality on the TED, Opensubtitles and Europarl7 translation tasks (BLEU) and the constrastive test set (accuracy). "Lexical" and "Grammatical" denote the lexical and grammatical cohesion devices.

5 Analysis

In this section, we analyzed three factors to examine their impact on the performance of the proposed model, including: 1) cohesion device type, 2) masking strategy, 3) cohesion attention mask.

5.1 Cohesion Device Type

In CoDoNMT, we take into account both lexical and grammatical cohesion devices. To further understand the effect of the type of cohesion devices on the model performance, we conducted experiments to model them separately during training.

Results on TED, Opensubtitles and Europarl7 are shown in Table 5. As can be seen, when we model both lexical and grammatical cohesion devices, CoDoNMT achieves the highest BLEU scores over the three translation tasks. This indicates that considering both lexical and grammatical cohesion devices is more beneficial to document-level NMT than only modeling one type of cohesion devices.

The performance differences of "Lexical" and "Grammatical" on the three translation tasks in terms of BLEU are slight. We conjecture that it may be because BLEU is not a good metric for discourse phenomena (Xu et al., 2020a). Therefore, we performed another linguistic evaluation (Voita et al., 2019). As shown in Table 5, "Both" continues to achieve the best performance on the two discourse phenomena. This again suggests that considering both lexical and grammatical cohesion devices are more helpful for the model to deal with discourse phenomena. In addition, the results of different types of cohesion devices demonstrate that modeling the corresponding type of cohesion devices can better solve discourse phenomenon relevant to these devices. In other words, lexical cohesion devices can better solve Lexical Cohesion than grammatical cohesion devices, but it is the other way around on Deixis.

5.2 Masking Strategy

Due to the scarcity of cohesion devices, we mask not only cohesion devices, but also a part of remaining tokens of the current sentence randomly. In other words, if the current sentence does not include any cohesion devices, CoDM degenerates to random masking. In order to investigate the effect of masking cohesion devices against randomly masking tokens, we conducted experiments to compare our masking strategy with the random masking strategy. We mask both lexical and grammatical cohesion devices in CoDM. For the random masking, we randomly mask a set of tokens in the current sentence according to the masking ratio r. Note that we do not apply CoAF in this analysis for a fair comparison as CoAF may further improve CoDM.

As shown in Table 6, in terms of BLEU, the differences between the two masking strategies are marginal. However, on the linguistic test, CoDM significantly outperforms Random on both linguistic phenomena. This suggests that predicting cohesion devices could be more efficient to produce document cohesion than predicting other words and reconfirms that BLEU is not sensitive on discourse phenomena (Xu et al., 2020a).

Masking Strategy	TED	Opensubtitles	Europarl7	AVG	Deixis	Lexical Cohesion
CoDM	**26.53**	21.16	**31.75**	**26.48**	**83.40%**	**55.70%**
Random	26.48	**21.26**	31.67	26.47	79.08%	46.20%

Table 6: Comparison on masking strategies on the TED, Opensubtitles and Europarl7 translation tasks (BLEU) and the constrastive test set (accuracy). "Random" denotes the random masking strategy that randomly masks a subset of tokens in the current sentence according to the masking ratio r.

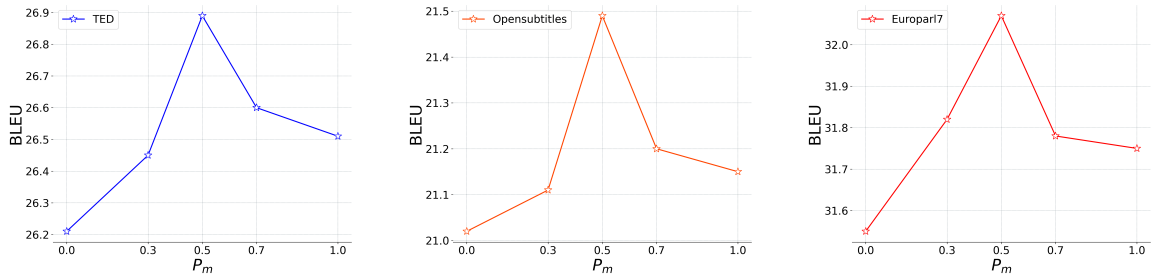

Figure 4: Impact of the probability threshold P_m on the TED, Opensubtitles and Europarl7 translation tasks.

5.3 Cohesion Attention Mask

As shown in Figure 4, we conducted experiments with different probability threshold P_m to explore its effect. As can be seen, when P_m is 0.5, our model achieves the best BLEU score over the three translation tasks. Furthermore, if we rigidly focus the model on cohesion devices (i.e., $P_m = 0.0$, applying the cohesion attention mask in a hard way), the model performance drops significantly. This suggests that a soft application of the cohesion attention mask allows the model to gain the ability to capture cohesion information and maintain the ability to explore other important contextual information.

5.4 Case Study

To better illustrate how our model improves translation quality, we provide an example of Deixis from the linguistic contrastive test set in Table 7, which is translated by both the document-level baseline Concate and our model. The document-level baseline **Concate** is a standard Transformer. We trained **Concate** on the document-level corpus where the current sentence is concatenated to its corresponding previous context by a special token (e.g., <SEP>) on both source- and target-side. As shown in Table 7, according to the translation "твой" in previous context, the word "*you*" in "*Maybe you know.*" should be translated into "ты", instead of "вы". Obviously, **Concate** fails to capture the co-reference information, while our model is able to translate the word correctly.

6 Related Work

Document-level NMT aims to improve translation quality with the aid of contextual information beyond the scope of current sentences. Most previous works focus on the integration of the inter-sentential context into NMT. One typical approach is introducing an addtional encoder to encode context, and then integrate the context representation into the primary encoder and/or the decoder (Zhang et al., 2018; Voita et al., 2018; Kuang and Xiong, 2018; Xu et al., 2020a; Zheng et al., 2020). Miculicich et al. (2018) propose a hierarchical attention model to capture the contextual information from both word and sentence level. Tan et al. (2019) propose a hierarchical model to learn global context for document-level NMT. Voita et al. (2019) introduce a two-pass framework that uses several previous sentences as context to refine the translation generated by a strong sentence-level NMT. Zhang et al. (2021) apply a multi-hop mechanism to document-level NMT to simulate the human-like draft-editing and reasoning process. Liu et al. (2020), Ma et al. (2020) and Bao et al. (2021) combine pre-trained models with document-level NMT. Zhang et al. (2020) pretrain a source context prediction model on a large-scale monolingual document corpus to learn contextualized sentence embeddings.

Yet another research strand is to focus on context selection. Maruf and Haffari (2018), Kuang et al.

Source	c^x	Like your boss said, might get you killed.
	c^x	Well, that's what I keep hearing.
	c^x	Nobody wants to share this dangerous entity's idea with me.
	x	Maybe you know.
Reference	c^x	Как сказал твой босс, это может кончиться твоей смертью.
	c^x	Что ж, я это я слышу все время.
	c^x	Никто не хочет делиться со мной личностью той опасной сущности.
	x	Может, ты знаешь.
Concate	c^x	Как сказал твой босс, может убить тебя.
	c^x	Ну, это то, что я продолжаю слышать.
	c^x	Никто не хочет делиться со мной идеей этой опасной сущности.
	x	Может быть, вы знаете.
Our Model	c^x	Как сказал твой босс, тебя могут убить.
	c^x	ну, это то, что я всегда слышал.
	c^x	никто не хочет делиться со мной мыслями об этой опасной сущности.
	x	может быть, ты знаешь.

Table 7: An example of Deixis translation in English-Russian. x and c^x denote the current sentence and its corresponding previous context, respectively. Blue words indicate the correct translations, while red words are the opposite.

(2018) and Tu et al. (2018) use a cache-like memory network to memorize the translation history, and treat it as context to translate future sentences. Maruf et al. (2019) uses sparse attention to selectively focus on relevant sentences in the document context. Kang et al. (2020) adopt reinforcement learning to select dynamic context for document-level NMT.

Recently, approaches have been proposed to leverage discourse information. Xu et al. (2020b) build directed graphs of documents with intra-sentential and inter-sentential relations and use GCN to obtain the document representation. Lyu et al. (2021) use word links to encourage the model to generate more consistent translations.

Different from the above works, we attempt to leverage cohesion devices to enhance the ability of model to capture inter-sentential contextual information to generate cohesive translations.

7 Conclusion

In this paper, we have presented CoDoNMT for document-level NMT, which models cohesion devices with two key methods, CoDM and CoAF. CoDM masks cohesion devices in the current sentence to force the model to actively explore inter-sentential contextual information. CoAF softly guides the model to focus attention on cohesion devices. Both automatic and linguistic evaluations show that our model can significantly improve translation quality in terms of BLEU and lexical and grammatical cohesion accuracy on a discourse-oriented contrastive test set. Further analyses demonstrate the impact of cohesion device type and masking strategy on translation quality.

8 Acknowledgement

The present research was supported by the Key Research and Development Program of Yunnan Province (Grant No. 202203AA080004-2). We would like to thank the anonymous reviewers for their insightful comments.

References

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-Transformer for Document-Level Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Kenneth W. Church. 2000. Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to p/2 than p2. In *COLING 2000 Volume 1: The 18th*

International Conference on Computational Linguistics.

Christiane D. Fellbaum. 2000. WordNet : an Electronic Lexical Database. *Language*, 76:706.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Routledge.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *CoRR*, abs/1803.05567.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic Context Selection for Document-level Neural Machine Translation via Reinforcement Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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, Prague, Czech Republic. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling Coherence for Neural Machine Translation with Dynamic and Topic Caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and changliang Li. 2020. Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8(0):726–742.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging Lexical Translation Consistency for Document-Level Neural Machine Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

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, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, André F. T. 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, Brussels, Belgium. Association for Computational Linguistics.

Sameen Maruf, André F. T. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A Survey on Document-level Neural Machine Translation: Methods and Evaluation. *ACM Computing Surveys*, 54(2):1–36.

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, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

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(0):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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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, Florence, Italy. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Hongfei Xu, Deyi Xiong, Josef van Genabith, and Qiuhui Liu. 2020a. Efficient Context-Aware Neural Machine Translation with Layer-Wise Weighting and Input-Aware Gating. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3933–3940. International Joint Conferences on Artificial Intelligence Organization. Main track.

Mingzhou Xu, Liangyou Li, Derek F. Wai, Qun Liu, and Lidia S. Chao. 2020b. Document Graph for Neural Machine Translation. *arXiv:2012.03477 [cs]*. ArXiv: 2012.03477.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do Context-Aware Translation Models Pay the Right Attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801, Online. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the Transformer Translation Model with Document-Level Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Long Zhang, Tong Zhang, Haibo Zhang, Baosong Yang, Wei Ye, and Shikun Zhang. 2021. Multi-Hop Transformer for Document-Level Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3953–3963, Online. Association for Computational Linguistics.

Pei Zhang, Xu Zhang, Wei Chen, Jian Yu, Yanfeng Wang, and Deyi Xiong. 2020. Learning Contextualized Sentence Representations for Document-Level Neural Machine Translation. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2298–2305. IOS Press.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the Gap between Training and Inference for Neural Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards Making the Most of Context in Neural Machine Translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3983–3989, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.