Context-Aware Machine Translation with Source Coreference Explanation

Huy Hien Vu **Hidetaka Kamigaito** **Taro Watanabe**
Nara Institute of Science and Technology, Japan
{vu.huy_hien.va9, kamigaito.h, taro}@is.naist.jp

Abstract

Despite significant improvements in enhancing the quality of translation, context-aware machine translation (MT) models underperform in many cases. One of the main reasons is that they fail to utilize the correct features from context when the context is too long or their models are overly complex. This can lead to the explain-away effect, wherein the models only consider features easier to explain predictions, resulting in inaccurate translations. To address this issue, we propose a model that explains the decisions made for translation by predicting coreference features in the input. We construct a model for input coreference by exploiting contextual features from both the input and translation output representations on top of an existing MT model. We evaluate and analyze our method in the WMT document-level translation task of English-German dataset, the English-Russian dataset, and the multilingual TED talk dataset, demonstrating an improvement of over 1.0 BLEU score when compared with other context-aware models.

1 Introduction

With the rapid development of machine learning techniques, the Machine Translation (MT) field has witnessed changes from exclusively probabilistic models (Brown et al., 1990; Koehn et al., 2003) to neural network based models, such as simplistic Recurrent Neural Network (RNN) based encoder-decoder models (Sutskever et al., 2014) or higher-level attention-based models (Bahdanau et al., 2015; Luong et al., 2015), and finally turn to the current state-of-the-art Transformer model (Vaswani et al., 2017) and its variations.

The quality of MT models, including RNN-based, attention-based, and Transformer models, has been improved by incorporating contextual information (Voita et al., 2018; Wang et al., 2017; and others), or linguistic knowledge (Bugliarello and Okazaki, 2020; Sennrich and Haddow, 2016; and others). In the former context-aware methods, many successful approaches focus on context selection from previous sentences (Jean et al., 2017; Wang et al., 2017) using multiple steps of translation, including additional module to refine translations produced by context-agnostic MT system, to utilize contextual information (Voita et al., 2019; Xiong et al., 2019), and encoding all context information as end-to-end frameworks (Zhang et al., 2020; Bao et al., 2021). Although they have demonstrated improved performance, there are still many cases in which their models perform incorrectly for handling, i.e., the ellipsis phenomenon in a long paragraph. One of the reasons is that their models are still unable to select the right features from context when the context is long, or the model is overly complex. Therefore, the model will easily suffer from an explain-away effect (Klein and Manning, 2002; Yu et al., 2017; Shah et al., 2020; Refinetti et al., 2023) in which a model learns to use only features which are easily exploited for prediction by discarding most of the input features.

In order to resolve the problem of selecting the right context features in the context-aware MT, we propose a model which *explains decisions of translation by predicting input features*. The input prediction model employs the representations of translation outputs as additional features to predict contextual features in the inputs. In this work, we employ coreference as the prediction task since it captures the relation of mentions that are necessary for the context-aware model. The prediction model is constructed on top of an existing MT model without modification in the same manner as done in multi-task learning, but it fuses information from representations used for the decisions of translation in the MT model.

Under the same settings of the English-Russian (En-Ru) dataset and the WMT document-level translation task of the English-German (En-De)

Transactions of the Association for Computational Linguistics, vol. 12, pp. 856–874, 2024. https://doi.org/10.1162/tacl_a_00677
Action Editor: Emily Pitler. Submission batch: 8/2023; Revision batch: 2/2024; Published 7/2024.

dataset, our proposed technique outperforms the standard transformer-based neural machine translation (NMT) model in both sentence and context-aware models, as well as the state-of-the-art context-aware model measured by BLEU (Post, 2018), BARTScore (Yuan et al., 2021), and COMET (Rei et al., 2020), and the human-annotated test set in a paragraph (Voita et al., 2019). Additionally, in the multilingual experiments, our method shows consistent results, paralleling those in the En-Ru and En-De datasets, and proving its versatility across languages.

Further analysis shows that our coreference explanation sub-model consistently enhances the quality of translation, regardless of type of dataset size. Notably, the model demonstrates consistent improvement when additional context is incorporated, highlighting its effectiveness in handling larger context sizes. Additionally, the analysis highlights a strong correlation between the self-attention heat map and coreference clusters, underscoring the significance of our coreference prediction sub-model in capturing coreference information during the translation process. Moreover, our proposed training method proves to be effective in the coreference prediction task. We also provide a suggestion to finetune the contribution of the sub-model to optimize its impact within the overall MT system. We release our code and hyperparameters at *https://github.com/hienvuhuy/TransCOREF*.

2 Background

2.1 Transformer-based NMT

Given an input single sentence $\boldsymbol{x} = (x_1, \ldots, x_{|\boldsymbol{x}|})$ and its corresponding translation $\boldsymbol{y} = (y_1, \ldots, y_{|\boldsymbol{y}|})$, an MT system directly models the translation probability

$$p(\boldsymbol{y}|\boldsymbol{x};\theta) = \prod_{t=1}^{|\boldsymbol{y}|} p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta), \quad (1)$$

where t is the index of target tokens, $\boldsymbol{y}_{<t}$ is the partial translation before y_t, and θ is the model parameter. At inference time, the model will find the most likely translation $\hat{\boldsymbol{y}}$ for a given source input

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \prod_{t=1}^{|\boldsymbol{y}|} p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta). \quad (2)$$

To model the translation conditional probability $p(\boldsymbol{y}|\boldsymbol{x};\theta)$, many encoder-decoder architectures have been proposed based either on CNNs (Gehring et al., 2017) or self-attention (Vaswani et al., 2017), and we focus on the Transformer (Vaswani et al., 2017) as our building block, given its superior ability to model long-term dependencies and capture context information features. The encoder of the Transformer comprises l_e stacked layers which transforms the input $\boldsymbol{x}$ into hidden representations $\mathbf{H}^{l_e}_{enc} \in \mathbb{R}^{|\boldsymbol{x}|\times d}$ where d is a dimension for hidden vector representation. Similarly, the decoder of the Transformer comprises l_d stacked layers which consumes the translation prefix $\boldsymbol{y}_{<t}$ and $\mathbf{H}^{l_e}_{enc}$ to yield the final representation $\mathbf{H}^{l_d}_{dec} \in \mathbb{R}^{|\boldsymbol{y}|\times d}$. The two processes can be formally denoted as

$$\mathbf{H}^{i}_{enc} = \textsc{Enc}(\mathbf{H}^{i-1}_{enc}) \quad (3)$$

$$\mathbf{H}^{i}_{dec} = \textsc{Dec}(\mathbf{H}^{i-1}_{dec}, \mathbf{H}^{l_e}_{enc}). \quad (4)$$

Note that $\mathbf{H}^{0}_{enc}$ is the representation of $\boldsymbol{x}$ from the embedding layer, and $\mathbf{H}^{0}_{dec}$ is the representation of $\boldsymbol{y}$ after embedding lookup with shifting by the begin-of-sentence token. $\textsc{Enc}(\cdot)$ and $\textsc{Dec}(\cdot)$ denote the function of the single Transformer encoder and decoder layer, respectively.

The output target sequence is predicted based on the output hidden state $\mathbf{H}^{l_d}_{dec}$ from the top layer of the decoder

$$p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta) = \textsc{Softmax}(\mathbf{W}_{dec}\mathbf{H}^{l_d}_{dec}[t])[y_t] \quad (5)$$

where $\mathbf{W}_{dec} \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the projection weight matrix which maps the hidden state to the probability in the output vocabulary space $\mathcal{V}$, and $[\cdot]$ denotes an index/slice to a vector/matrix.

The standard training objective is to minimize the cross-entropy loss function

$$\mathcal{L}_{\text{MT}} = -\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \sum_{t=1}^{|\boldsymbol{y}|} \log p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta) \quad (6)$$

given a parallel corpus $\mathcal{D} = \{(\boldsymbol{x}^w, \boldsymbol{y}^w)\}_{w=1}^{|\mathcal{D}|}$ which contains $|\mathcal{D}|$ pairs of single sentence and its corresponding translation.

2.2 Context-Aware Transformer-base NMT

A context-aware MT model can be regarded as a model which takes a document, i.e., multiple

sentences, as an input and generates multiple sentences as its corresponding translation. We assume that each sentence is translated into a single sentence, and define the source document $\underline{\boldsymbol{x}} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n)$ with n sentences and its corresponding target language document $\underline{\boldsymbol{y}} = (\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n)$. A context-aware MT system directly models the translation probability

$$p(\underline{\boldsymbol{y}}|\underline{\boldsymbol{x}}; \theta) = \sum_{k=1}^{n} p(\boldsymbol{y}^k|\boldsymbol{y}^{<k}, \underline{\boldsymbol{x}}; \theta), \quad (7)$$

where k is an index to a sentence in $\underline{\boldsymbol{y}}$, $\boldsymbol{y}^{<k}$ is the partial translation before the sentence $\boldsymbol{y}^k$. In this model, we assume that $\langle \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}} \rangle$ constitute a parallel document and each $\langle \boldsymbol{x}^k, \boldsymbol{y}^k \rangle$ forms a parallel sentence.

Several approaches can be used to produce a translated document, i.e., keeping a sliding window of size m (Tiedemann and Scherrer, 2017), joining these m sentences as a single input, translating these m sentences and selecting the last sentence as an output (m-to-m) (Zhang et al., 2020), or joining whole sentences in a document as a very long sequence and translating this sequence (Bao et al., 2021), among other methods. To simplify the definition of the context-aware NMT model, we opt for the m-to-m method and use a special character (*_eos*) between sentences when feeding these m sentences to the model. In this way, the context-aware translation model can still be defined as a standard sentence-wise translation in §2.1.

2.3 Coreference Resolution Task

Coreference Resolution is the task of identifying and grouping all the mentions or references of a particular entity within a given text into a cluster, i.e., a set of spans. This task has progressed significantly from its earlier approaches, which were based on hand-crafted feature systems (McCarthy and Lehnert, 1995; Aone and William, 1995), to more advanced and effective deep learning approaches based on span-ranking (Lee et al., 2017, 2018; Kirstain et al., 2021) and for multilingual languages (Zheng et al., 2023).

It is typically formulated as explicitly identifying an antecedent span to the left of a mention span in the same cluster. More formally, a set of clusters $\boldsymbol{\mathcal{C}} = \{..., \mathcal{C}_k, ...\}$ is predicted for an input sequence $\boldsymbol{x}$, either a document or a single sentence, with each cluster comprising a set of non-overlapping spans $\mathcal{C}_k = \{(i, j) : 1 \leq i \leq j \leq |\boldsymbol{x}|\}$ $(1 \leq k \leq |\boldsymbol{\mathcal{C}}|)$. We introduce an alternative view using a variable $\mathcal{A}$ which represents mapping for all possible mention spans $\mathcal{S} = \{(i, j) : \forall 1 \leq i \leq j \leq |\boldsymbol{x}|\}$ of $\boldsymbol{x}$ to its antecedent span within the sample cluster $\mathcal{C}_k$, i.e., $\mathcal{A} = \{..., s \rightarrow c, ...\}$, where $c \in \mathcal{C}_k$ is an antecedent to the left of $s \in \mathcal{C}_k$ and $c = \epsilon$, i.e., an empty span, when s is not a member of any clusters $\mathcal{C}_k$. Note that we can derive a unique $\boldsymbol{\mathcal{C}}$ given a single derivation of $\mathcal{A}$ by forming a cluster of spans connected by antecedent links, but there are multiple derivations of $\mathcal{A}$ for $\boldsymbol{\mathcal{C}}$ when there exists a cluster $|\mathcal{C}_k| > 2$. The task is modeled by the conditional probability distribution of independently predicting any possible antecedents of a mention span in the same cluster

$$\begin{aligned} p(\boldsymbol{\mathcal{C}}|\boldsymbol{x}) &= \sum_{\mathcal{A} \in a(\boldsymbol{\mathcal{C}})} p(\mathcal{A}|\boldsymbol{x}) \\ &= \prod_{s \in \mathcal{S}} \sum_{\mathcal{A} \in a(\boldsymbol{\mathcal{C}})} p(\mathcal{A}_s|s, \boldsymbol{x}) \\ &= \prod_{s \in \mathcal{S}} \sum_{\mathcal{A} \in a(\boldsymbol{\mathcal{C}})} \frac{\exp\left(f(\mathcal{A}_s, s; \mathbf{H}_{coref})\right)}{\sum_{c \in \mathcal{M}_s} \exp\left(f(c, s; \mathbf{H}_{coref})\right)} \\ &\triangleq \prod_{s \in \mathcal{S}} \sum_{\mathcal{A} \in a(\boldsymbol{\mathcal{C}})} \textsc{Coref}(\mathcal{A}_s, s; \mathbf{H}_{coref}), \end{aligned} \quad (8)$$

where $a(\cdot)$ is a function that returns all possible derivations for clusters and $\mathcal{M}_s$ is a set of all possible spans to the left of s including ϵ. $f(\cdot, \cdot)$ is a score function (Kirstain et al., 2021) to compute both mention and antecedent scores and $\mathbf{H}_{coref} \in \mathbb{R}^{|\boldsymbol{x}| \times d}$ is contextualized representation of the input sequence $\boldsymbol{x}$, i.e., BERT (Devlin et al., 2019). We denote the final function as $\textsc{Coref}(\cdot, \cdot)$ for brevity.

We adopt the training scheme proposed by Kirstain et al. (2021), which filters spans to avoid the explicit enumeration of all possible mention spans, and represents antecedent relations using only the endpoints of the retained spans with a biaffine transformation. At the training stage, we minimize the negative log-likelihood of predicting clusters

$$\mathcal{L}_{\textsc{Coref}} = -\sum_{(\boldsymbol{\mathcal{C}}, \boldsymbol{x}) \in \mathcal{D}_{\textsc{Coref}}} \log \prod_{s \in \mathcal{S}} \sum_{\mathcal{A} \in a(\boldsymbol{\mathcal{C}})} p(\mathcal{A}_s|s, \boldsymbol{x}), \quad (9)$$

where $\mathcal{D}_{\textsc{Coref}}$ is a training data for coreference resolution.

3 Context-Aware MT with Coreference Information

Our motivation stems from the observation that when translating a paragraph, translators are able to pick up precise words and explain why a particular choice of word is better given the context especially by relying on linguistic cues such as discourse structure, verb equivalence, etc. Thus, instead of modeling a translation by directly relying on an additional conditional variable of coreference clusters $\boldsymbol{\mathcal{C}}$ for $\boldsymbol{x}$, i.e., $p(\boldsymbol{y}|\boldsymbol{\mathcal{C}}, \boldsymbol{x})$, we propose a model that is akin to the noisy channel framework (Yee et al., 2019), to explain the decision $\boldsymbol{y}$ made by the translation model:

$$p(\boldsymbol{y}|\boldsymbol{\mathcal{C}}, \boldsymbol{x}) = \frac{p(\boldsymbol{y}, \boldsymbol{\mathcal{C}}|\boldsymbol{x})}{p(\boldsymbol{\mathcal{C}}|\boldsymbol{x})} = \frac{p(\boldsymbol{y}|\boldsymbol{x})(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x})}{p(\boldsymbol{\mathcal{C}}|\boldsymbol{x})} \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x}), \quad (10)$$

where $p(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x})$ is a model to predict coreference clusters given both an input sentence and its translation. Note that we can omit the denominator $p(\boldsymbol{\mathcal{C}}|\boldsymbol{x})$ given that it is a constant when predicting $\boldsymbol{y}$, similar to the noisy channel modeling, since both $\boldsymbol{x}$ and $\boldsymbol{\mathcal{C}}$ are input to our model. The direct model $p(\boldsymbol{y}|\boldsymbol{\mathcal{C}}, \boldsymbol{x})$ is prone to ignoring features in $\boldsymbol{\mathcal{C}}$, especially when the context is long, since the information in $\boldsymbol{x}$ has direct correspondence with $\boldsymbol{y}$. In contrast, the model for the coreference resolution task, $p(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x})$, will explain the coreference cluster information in $\boldsymbol{x}$ not only by the features from $\boldsymbol{x}$ but additional features from $\boldsymbol{y}$ and, thus, the higher $p(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x})$, the more likely $\boldsymbol{y}$ will be a translation for $\boldsymbol{x}$. When coupled with the translation model $p(\boldsymbol{y}|\boldsymbol{x})$ especially when jointly trained together, our formulation will be able to capture long-distance relations in coreference clusters since the coreference resolution task needs to predict it given $\boldsymbol{x}$ and $\boldsymbol{y}$.

Architecture The two sub-models, i.e., $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x})$, could be trained separately as done in a noisy channel modeling approach of MT (Yee et al., 2019). This work formulates it as a multi-task setting by predicting two tasks jointly, i.e., translation task and coreference resolution task, by using the representations of the encoder and decoder of Vaswani et al. (2017). More specifically, we do not alter translation task $p(\boldsymbol{y}|\boldsymbol{x})$, but obtain the representation for the coreference task by fusing the representations of the encoder and decoder as follows

$$p(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x}) = \prod_{s \in \mathcal{S}} \sum_{\mathcal{A} \in a(\boldsymbol{\mathcal{C}})} \textsc{Coref}(\mathcal{A}_s, s; \mathbf{H}'_{coref})$$
$$\mathbf{H}'_{coref} = \textsc{Dec}(\boldsymbol{H}^{l_e}_{enc}, \boldsymbol{H}^{l_d}_{dec}). \quad (11)$$

Note that we obtain $\mathbf{H}'_{coref}$ from an additional decoder layer for the encoder representation $\boldsymbol{H}^{l_e}_{enc}$ with cross attention for $\boldsymbol{H}^{l_d}_{dec}$.

Training We jointly train our two sub-models using the label-smoothing variant of the cross-entropy loss function in Equation 6 and the marginal log-likelihood loss function in Equation 9, but using $\mathbf{H}'_{coref}$ in Equation 11 as follows

$$\mathcal{L} = \mathcal{L}_{\text{MT}} + \alpha \mathcal{L}'_{\textsc{Coref}}, \quad (12)$$

where α is a hyperparameter that controls a contribution of the coreference resolution task. During the training step, we feed pairs of sentences together with coreference cluster information generated by an external coreference resolution framework since human annotation is not available in MT tasks.

Inference Inference is complex in that the model for the coreference resolution task has to be evaluated every time a target token is generated by the translation model as done in the noisy channel approach of MT (Yee et al., 2019). We resort to a simpler approach of ignoring the term for coreference clusters, i.e., $p(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x})$, and using only the token prediction, i.e., $p(\boldsymbol{y}|\boldsymbol{x})$; alternatively, we generate a large set of N-best translations from $p(\boldsymbol{y}|\boldsymbol{x})$ and rerank them using the joint probabilities

$$\log p(\boldsymbol{y}|\boldsymbol{x}) + \beta \log p(\boldsymbol{\mathcal{C}}|\boldsymbol{y}, \boldsymbol{x}), \quad (13)$$

where β is a hyperparameter to control the strength of the coreference resolution task.

4 Experiments

4.1 Dataset

We utilized the En-Ru dataset (Voita et al., 2019) and the widely adopted En-De benchmark dataset IWSLT 2017, as used in Maruf et al. (2019), with details provided in Table 1. We also used the multilingual TED talk dataset (Qi et al., 2018) to

	Avg. #Coref. Clusters train/valid/test	#Samples train/valid/test
En-Ru	3.1/3.0/2.9	1.5M/10k/10k
En-De	4.4/4.4/4.4	206k/8k/2k

Table 1: Statistics of En-De and En-Ru datasets.

	Family	WO	PP	GP	GA
English	IE	SVO	♢	3SG	SEM
Russian	IE	SVO	♠	3SG	S-F
German	IE	SOV/SVO	♡	3SG	S-F
Spanish	IE	SVO	♡	1/2/3P	SEM
French	IE	SVO	♡	3SG	S-F
Japanese	JAP	SOV	♣	3P	♢
Romanian	IE	SVO	♠	3SG	S-F
Mandarin	ST	SVO	♡	3SG	♢
Vietnamese	AA	SVO	♠	♢	♢

Table 2: Properties of Languages in Our Experiments: WO (Word Order), PP (Pronouns Politeness), GP (Gendered Pronouns), and GA (Gender Assignment) denote language structural properties. IE (Indo-European), JAP (Japonic), ST (Sino-Tibetan), and AA (Austroasiatic) represent language families. Symbols ♢, ♡, ♣, and ♠ correspond to 'None', 'Binary', 'Avoided', and 'Multiple', respectively. The terms 3SG (Third Person Singular), 1/2/3P (First, Second, and Third Person), and 3P (Third Person) are used for pronoun references. SEM and S-F stand for Semantic and Semantic-Formal, respectively, in Gender Assignment.

assess the efficacy of our proposed method across a variety of language types, including different characteristics in pronouns, word order and gender assignment with specifics delineated in Table 2.

The En-Ru dataset comes from OpenSubtitle 2018 (Lison et al., 2018) by sampling training instances with three context sentences after tokenization and, thus, no document boundary information is preserved. In the En-De and multilingual datasets, document boundaries are provided. To maintain consistency in our translation settings during experiments, we tokenize all texts by using MeCab[1] for Japanese, Jieba[2] for Chinese, VnCoreNLP (Vu et al., 2018) for Vietnamese, and the SpaCy framework[3] for all other languages. We also apply a sliding window with a size of m sentences ($m = 4$) to each document to create a similar format to that of the En-Ru dataset. For the first $m - 1$ sentences, which do not have enough $m - 1$ context sentences in the m-to-m translation settings, we pad the beginning of these sentences with empty sentences, ensuring $m - 1$ context sentences for all samples in the dataset. For preprocessing, we apply the BPE (Byte-Pair Encoding) technique from Sennrich et al. (2016) with 32K merging operations to all datasets. To identify coreference clusters for the source language, i.e., English, we leveraged the AllenNLP framework[4] and employed the SpanBERT large model (Lee et al., 2018). After generating sub-word units, we adjust the word-wise indices of all members in coreference clusters using the offsets for sub-word units.

[1] https://taku910.github.io/mecab/.
[2] https://github.com/fxsjy/jieba.
[3] https://spacy.io.

4.2 Experiment Settings

Translation Setting In our experiments, we adopt the context-aware translation settings (m-to-m with $m = 4$) utilized in previous work (Zhang et al., 2020). For the context-agnostic setting, we translate each sentence individually.

Baselines Systems We adopt the Transformer model (Vaswani et al., 2017) as our two baselines: **Base Sent**, which was trained on source and target sentence pairs without context, and **Base Doc**, which was trained with contexts in the m-to-m setting as described in §2.2. To make a fair comparison with previous work that uses similar context-aware translation settings and enhance MT system at the encoder side, we employ the **G-Transformer** (Bao et al., 2021), **Hybrid Context** (Zheng et al., 2020), and **MultiResolution** (Sun et al., 2022). We also compare our approach with the **CoDoNMT** (Lei et al., 2022) model, which also integrates coreference resolution information to improve translation quality. Note that all aforementioned baselines utilize provided open-source code. Additionally, we trained a simple variant of a context-aware Transformer model similar to Base Doc, but differ in that it incorporated a coreference embedding, alongside the existing positional embedding, directly in to the encoder side of the model (**Trans+C-Embedding**). This coreference embedding is derived from the

[4] https://allenai.org.

original positional embedding in the encoder with the modification that all tokens within a coreference cluster share the same value as the left-most token in the same cluster. Note that it is intended as a simple baseline for a direct model as discussed in §3.

Our Systems We evaluate our proposed inference methods, including the original inference method in Transformer without reranking (**Trans+COREF**) or with reranking with the score from our sub-model (**Trans+COREF+RR**) using the coreference resolution task as denoted in Equation 13.

Hardware All models in our experiments were trained on a machine with the following specifications: An AMD EPYC 7313P CPU, 256GB RAM, a single NVIDIA RTX A6000 with 48GB VRAM, and CUDA version 11.3. For multilingual experiments, we used a single NVIDIA RTX 3090 GPU, Intel i9-10940X, 48GB VRAM, and CUDA version 12.1.

Hyperparameters We use the same parameters, including the number of training epochs, learning rate, batch size, etc., for all models in our experiments. Specifically, we train all models for 40 epochs when both losses of coreference and translation in the valid set show unchanging or no improvements.

For translation tasks, we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 1e-9$, along with an inverse square root learning rate scheduler. All *dropout* values are set to 0.1, and the learning rate is set to $7e-5$. We use a batch size of 128 and 32 for experiments on the English-Russian and English-German datasets, respectively. Other parameters follow those in Vaswani et al. (2017).

For coreference tasks, we adopt parameters from Kirstain et al. (2021), with some modifications to accommodate our GPU memory. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 1e-9$, with a learning rate of $7e-5$. *Dropout* value is set to 0.3, *top lambda* (the percentage of all spans to keep after filtering) is set to 0.4, *hidden size* is set to 512, and the *maximum span length* is set to 10. The *maximum cluster value*s are set to 8 and 20 for the English-Russian and English-German datasets, respectively. To rerank the N-best translations, we use Equation 13 and perform a grid search on the validation set with a step size of 0.0001 to select the optimal value for β from -2 to 2.

4.3 Metrics

BLEU We employ SacreBLEU (Post, 2018) as an automated evaluation metric to assess the quality of translations in our experiments.

BARTScore We follow Yuan et al. (2021) and use the *mbart-large-50* model (*mBART*)[5] to compute the average BARTScore of all translation to measure semantic equivalence and coherence between references and translations. In this metric, the higher value, the better semantic equivalence and coherence.

COMET We also utilize the COMET[6] metric (Rei et al., 2020), a neural network-based measure, since it is highly correlated to human judgment in prior work by Freitag et al. (2022).

4.4 Results

The main results of our experiments are presented in Table 3. Our results indicate that training the baseline Transformer model with both context and target sentences (Base Doc) results in better performance than training with only target sentences (Base Sent) in the En-Ru dataset. This finding is consistent with those reported by Voita et al. (2019), in which more contextual information is helpful to achieve better translation. However, in the En-De dataset, the Base Doc system performs worse compared to the Base Sent system. This discrepancy can be explained by the different methodologies used in constructing the En-De and En-Ru datasets. For the En-De datasets, both context-aware and context-agnostic datasets are compiled from the same pool of non-duplicate sentences. However, for the En-Ru datasets, the context-agnostic dataset is created by removing context sentences from the context-aware dataset (Voita et al., 2019), which results in varying numbers of non-duplicate sentences between these context-agnostic and context-aware datasets.

When comparing our systems with the Transformer model (Base Doc), our approaches, both Trans+COREF and Trans+COREF+RR, have proven effective in enhancing translation quality by

[5] https://huggingface.co/facebook/mbart-large-50.

[6] COMET-20 model (*wmt20-COMET-da*).

	En - Ru			En - De		
	BL ↑	BS ↑	CM↑	BL ↑	BS ↑	CM↑
Base Sent	29.46	−9.695	82.87	22.76	−6.178	68.06
Base Doc	29.91	−9.551	83.40	21.54	−6.200	66.91
Hybrid Context (Zheng et al., 2020)	29.96	−9.590	83.45	22.05	−6.236	66.97
G-Transformer (Bao et al., 2021)	30.15	−9.691	83.13	22.61	−6.090	68.36
MultiResolution (Sun et al., 2022)	29.85	−9.763	81.76	22.09	−6.099	67.99
DoCoNMT (Lei et al., 2022)	29.92	−9.552	83.03	22.55	−6.197	67.93
Trans+C-Embedding	30.13	−9.522	83.43	22.54	−6.092	68.80
Trans+COREF	30.39*	−9.501†	**83.56$^{\bullet}$**	23.57**	−6.088†	69.17$^{\diamond}$
Trans+COREF+RR	**30.43***	**−9.500**†	**83.56$^{\bullet}$**	**23.60****	**−6.086**†	**69.21**$^{\diamond}$

(*) and (*) indicate statistical significance (Koehn, 2004) at $p < 0.02$ and $p < 0.01$, respectively, compared to the *Base Doc* system and all other baseline systems. ($^{\diamond}$), (†), and ($^{\bullet}$) signify statistical significance at $p < 0.05$ compared to all baselines, all except Trans+C-Embedding and G-Transformer, and all except Trans+C-Embedding, Hybrid Context, and G-Transformer, respectively.

Table 3: The results of all main experiments. BL, BS and CM are abbreviations for BLEU, BARTScore and COMET, respectively. The best performance per metric are in bold text.

explaining the decision of translation through predicting coreference information. This is demonstrated by the superior BLEU scores ($+0.52$ in En-Ru and $+2.06$ in En-De for the Trans+COREF+RR), BARTScore, and COMET observed when comparing across different settings and language pairs.

Compared to the G-Transformer system described in Bao et al. (2021), our system shows an improvement in both inference approaches (Trans+COREF and Trans+COREF+RR). In the En-Ru dataset, our system achieves a higher BLEU score by $+0.24$, while in the En-De dataset, it demonstrates a larger improvement of $+1.14$ in the same metric (Trans+COREF). Additionally, our method outperforms the G-Transformer in terms of the BARTScore and COMET for both the En-Ru and En-De datasets. One possible explanation for these results is that the G-Transformer is specifically designed to map each sentence in the source language to only a single sentence in the target language during both training and inference steps. This design choice helps mitigating issues related to generating very long sequences. However, when the dataset size is small, as in the case of the En-De dataset, the G-Transformer encounters difficulties in selecting useful information. In contrast, our approach effectively selects useful information indirectly through the coreference explanation sub-model, especially when dealing with small-sized datasets, which allows our system to outperform under the scenarios with limited dataset size. Our method also surpasses the Transformer model with additional position embedding (Trans+C-Embedding), which relied on coreference information using a direct modeling approach.

In the results of the multilingual TED talk dataset in Table 4, where we compare our proposed method to Transformer models and the best baselines in Table 3, our method also surpasses other baselines within $+1.0$ to $+2.3$ BLEU scores. These findings provide further evidence that our approach is effective in improving translation quality and can be applied to diverse language types.

We provide an example of translations from our systems as well as other baseline systems in Table 5. In this example, the correct translation of the phrase in the last sentence, *моей командой* (*my team*), is determined by identifying which word refers to *''my''*, in this case, *i* and *me*. Both the G-Transformer and Trans+C-Embedding systems fail to capture these mentions and consequently produce an incorrect translation, *мою команду*. Despite correctly translating *моей*, the Base Doc system's phrase *встретимся в моей команде* is grammatically

	Es ↑	Fr ↑	Ja↑	Ro ↑	Zh ↑	Vi ↑
Base Sent	37.23	37.75	12.11	24.35	12.38	31.74
Base Doc	36.22	36.89	10.13	23.27	11.66	31.22
G-Transformer	36.46	37.88	12.27	24.63	12.07	32.69
Trans+COREF	**38.13***	**39.01***	**12.93***	**25.56***	**13.18***	**33.51***

*With statistically significance (Koehn, 2004) at $p < 0.01$ compared to other systems.

Table 4: The results of multilingual dataset in the BLEU metric. The highest results are in bold text.

Input	*but i 'm different . _eos do me just one favor . _eos before you make any decision ...* **_eos meet my team .**
Base Doc	*но я другой . _eos сделай мне одолжение . _eos прежде чем ты примешь решение ...* **_eos встретимся в моей команде .**
G-Transformer	*но я другой . _eos сделай мне одолжение . _eos прежде чем ты примешь решение ...* **_eos встреть мою команду .**
Trans+C-Embedding	*но я другой . _eos сделай мне одолжение . _eos прежде чем принять решение ...* **_eos встретить мою команду .**
Trans+DEC+RR	*но я другой . _eos сделай мне одолжение . _eos прежде чем принять решение ...* **_eos познакомься с моей командой .**
Reference	*но я изменился . _eos выполни только одну мою просьбу . _eos прежде чем ты решишь что-то ...* **_eos познакомься с моей командой .**

Table 5: Example of translations. The context and target sentences are highlighted in *italics* and **bold**, respectively. Translations of the Trans+DEC+RR and Trans+DEC are identical. Underlined words indicate the same mention entity.

incorrect and deviates from the original English *''meet my team''*. Conversely, our systems capture this reference accurately, yielding a translation consistent with the reference.

5 Analysis

Contribution of Coreference Explanation We conducted experiments by adjusting the value of α in Equation 12 during the training of the Trans+COREF without reranking. The experimental results in Table 6 indicate that for medium-sized corpora, selecting a value of α that is either too small or too large negatively impacts translation quality. The optimal range for α is $0.8 \leq \alpha \leq 2$.

Conditioning on Source and Target Language We conducted a study on coreference explanation on the En-De dataset (Maruf et al., 2019) with coreference cluster information as in §4.1 by ablating the information from the translation so that it conditions only on the input information (Trans+ENC). This setting can be regarded as a conventional multi-task setting in which coreference on the input-side is predicted together with its translation. Specifically, we replace the input representation for coreference resolution sub-model from $\textsc{Dec}(\cdot)$ in Equation 11 to $\textsc{Enc}(\cdot)$ in Equation 14 as follows

$$\mathbf{H}''_{coref} = \textsc{Enc}\boldsymbol{H}^{l_e}_{enc}. \quad (14)$$

	α	BLEU ↑
Base Sent	–	29.46
Base Doc	–	29.91
Trans+COREF	0.8	30.36
	1.0	30.31
	2.0	**30.39**
	3.0	30.27
	4.0	30.15
	10.0	30.00

Table 6: Ablation results on the En-Ru dataset with different weight α. The highest result is in bold text.

	BLEU ↑	P↑	R↑	F1↑
Trans+ENC	22.98	**85.02**	75.69	80.08
Trans+COREF	**23.57**	82.63	**78.31**	**80.41**

*The MUC metric counts the changes required to align the system's entity groupings with the gold-standard, focusing on adjustments to individual references.

Table 7: Evaluation of Trans+ENC and Trans+COREF systems using BLEU and MUC* metrics on the validating set of the En-De dataset. The highest results are in bold text.

As shown in Table 7, the conventional multi-task learning setting of Trans+ENC performed lower than Trans+COREF, which indicates the benefits of fusing information from the predicted translation. We further examine the entity heat maps derived from self-attention weights of both the translation and coreference sub-models in Base Doc, Trans+ENC, and Trans+COREF systems for the input *"I_0 hate that you_0 're leaving . Well, Grandma 's not doing well . So you_1 have to drop everything to go take care of her ? Yes, William, I_1 do ."* from the human annotated test set from Voita et al. (2019). In this particular example, the coreference clusters are defined as*[I_0, William], [you_0, you_1, I_1]*, and *[Grandma, her]*. To provide visual representations, we depict the average self-attention values from the last encoder layer of these three systems. This choice is based on their tendency to focus more on semantic or abstract information (Clark et al., 2019).

Figure 1 displays the entity heat maps, which illustrate the behavior of self-attention in different systems in the translation sub-model. In the Base Doc system, self-attention primarily concentrates on local sentences while disregarding information between sentences. In contrast, the Trans+ENC system exhibits the ability to focus on inter-sentences. However, when it comes to tokens within coreference clusters, the focused values are incorrect for certain clusters, such as *[I_0, William]*. On the other hand, the Trans+COREF system not only exhibits inter-sentential focus in its self-attention heat map but also accurately depicts the focused values for tokens within coreference clusters.

Figure 2 demonstrates the entity heat maps in the coreference sub-model. In the Trans+ENC system, self-attention mainly concentrates on entities within the local scope and immediate adjacent sentences. However, when comparing these high attention values with links in the coreference clusters, a significant proportion is found to be incorrect, i.e., *[Grandma; I_1]*. On the other hand, the self-attention in Trans+COREF exhibits a more balanced distribution of focus across all entities within the input. This balanced distribution results in considerably fewer errors when compared to self-attention in the Trans+ENC system. These findings align with the MUC metric (Vilain et al., 1995), which is based on the minimum number of missing links in the response entities compared to the key entities, with details, particularly the F1 score, provided in Table 7. Note that we use reference translations to form $\boldsymbol{H}^{l_d}_{dec}$ in Equation (11) for identifying coreference clusters. Additionally, we generate gold label coreference clusters using the AllenNLP framework, as discussed in Section 4.1.

Impact of the Context Size We conducted experiments with the COREF (Trans+COREF) and the Transformer (Base Doc) systems by exploring different context sizes in m-to-m settings ranging from 2 to 4. The experimental results in Figure 3 demonstrate that the Base Doc system significantly drops the translation quality when the context gets longer, while Trans+COREF consistently achieves gains as we incorporate more context. This result also indicates the use of the coreference sub-model is able to capture contextual information better than the baseline.

Impact of Coreference Explanation We conduct experiments by reranking all translation hypotheses with varying beam sizes during inference by the Equation (13) to assess the impact of coreference explanation sub-model on the En-Ru dataset (Voita et al., 2019). Figure 4 illustrates the results of our experiments measured by BLEU score. Our findings indicate that reranking with the sub-model COREF yields improved results, with differences ranging from 0.02 to 0.09. We also report oracle BLEU score in Figure 5, which is measured by selecting a hypothesis sentence that gives the maximum sentence-BLEU scores among potential hypotheses, to verify the potentially correct translations in an N-best list. The results of this experiment with differences ranging from 0.2 to 0.4 suggest that using the sub-model COREF has more potential to generate correct translations. Despite the relatively minor difference in

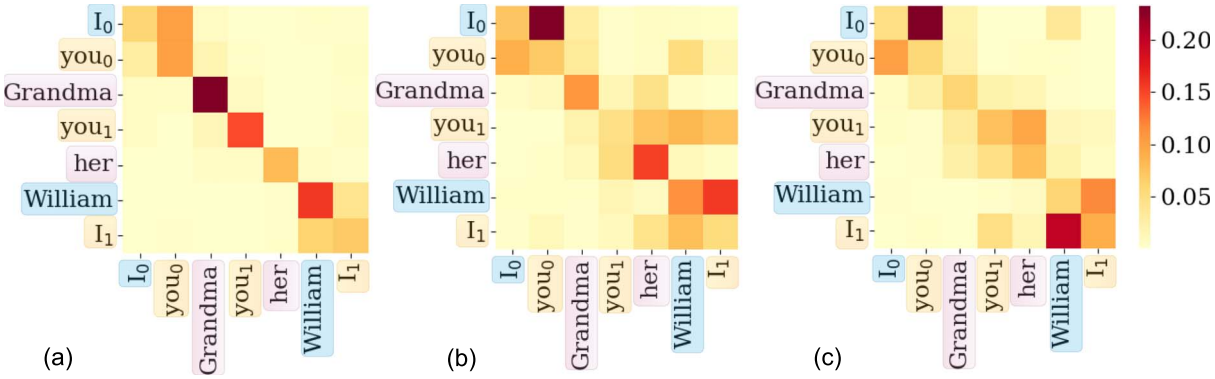

Figure 1: Entity heat maps of self-attentions: (a) Base Doc, (b) Trans+ENC and (c) Trans+COREF.

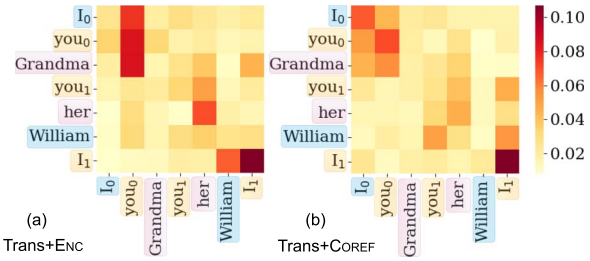

Figure 2: Entity heat maps of self-attentions in the coreference resolution sub-model.

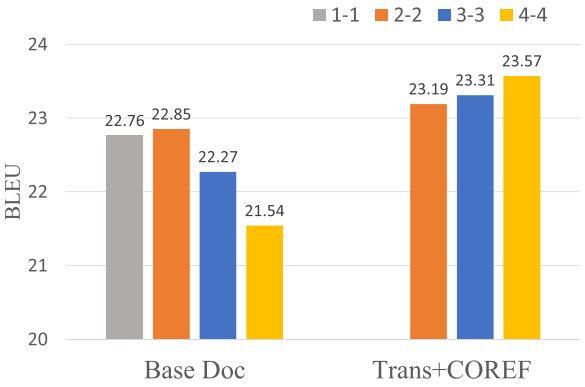

Figure 3: Translation results on En-De datasets with different m-to-m translation settings from $m = 2$ to $m = 4$. The result in the $m = 1$ setting serves as the Base Sent reference. The α in Equation 12 is set to 4.0.

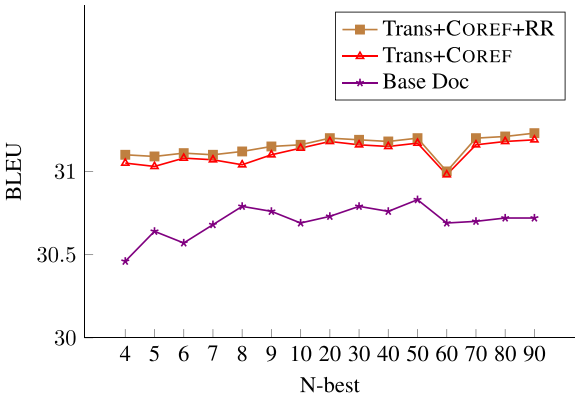

Figure 4: The results with N-best variants on the En-Ru dataset (Voita et al., 2019).

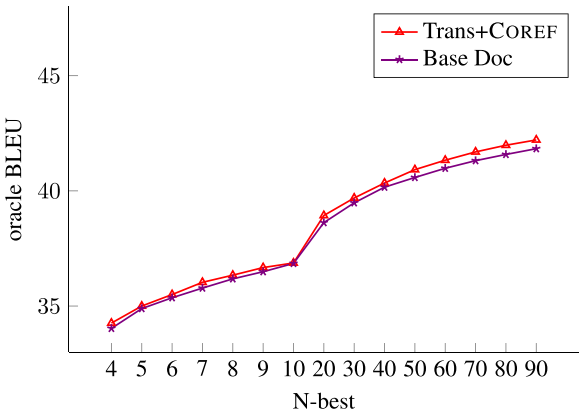

Figure 5: The results with N-best variants using the oracle BLEU metric on the En-Ru dataset (Voita et al., 2019).

the oracle BLEU score between the Trans+COREF and the Base Doc systems, indicating a similarity in their candidate space, the beam search process yields better results with the Trans+COREF when compared with the Base Doc system. This reflects the differences in BLEU scores between Trans+COREF and Base Doc. The performance gap in the BLEU score between the Trans+COREF and Trans+COREF+RR could potentially be further maximized by incorporating the coreference resolution during the beam search at the expense of more computational costs. We intend to explore this possibility in future research.

To further understand the impact of the coreference explanation sub-model on translation results, we perform an experiment on the contrastive test in Voita et al. (2019), which contains

	D ↑	EI ↑	EV ↑	L ↑
Base Doc	83.32	70.20	62.20	46.0
Trans+COREF	**85.64**	**71.20**	**65.2**	**46.4**

Table 8: Experimental results on the contrastive test (Voita et al., 2019). D, EI, EV and L are abbreviations for Deixis, Ellipsis Infl, Ellipsis Vp and Lexical Cohesion, respectively. Note that we only utilized the text described in §4.1, while other studies may incorporate additional sentence-level bilingual and monolingual texts associated with Voita et al. (2019).

human-labeled sentences to evaluate discourse phenomena and relies on the source text only, to verify whether our method can solve phenomena at the document level. Table 8 presents the results this experiment, which indicate that our system outperforms the Base Doc system in all aspects. These results demonstrate the significant contribution of the coreference explanation sub-model to the MT system.

Impact of Coreference Accuracy We carried out experiments to assess the impact of varying accuracies within the external coreference framework, which was reported in 80.4% of the F1 score on the MUC metric for the English CoNLL-2012 shared task in Lee et al. (2018), on the overall translation quality. This was achieved by randomly omitting members from coreference clusters while ensuring that each valid cluster retained a minimum of two members, i.e., removing you_1 from the cluster $[you\_0, you\_1, I\_1]$ in Figure 1.

Table 9 presents the outcomes of these experiments, where a slight reduction in translation quality is observed as members of coreference clusters are randomly dropped. Remarkably, even with the omission of up to half of the cluster members, the results continue to exceed the performance of the Base Doc system. This implies that our method could be robust and effective, particularly for languages with limited accuracy in coreference resolution tasks.

Impact of the Corpus Size We randomly sampled training instances from the En-Ru dataset and varied the sample sizes to 200,000 (comparable size to the En-De dataset), 500,000, and

	Pruning (%)	BLEU ↑	
		− RR	+RR
Base Doc	–	21.54	–
Trans+COREF	0	23.57	23.60
	10	23.43	23.44
	20	23.40	23.41
	30	23.29	23.29
	50	22.86	22.86

Table 9: Experimental results on dropping coreference clusters on the En-De dataset. RR means reranking with the coreference sub-model using Equation 13.

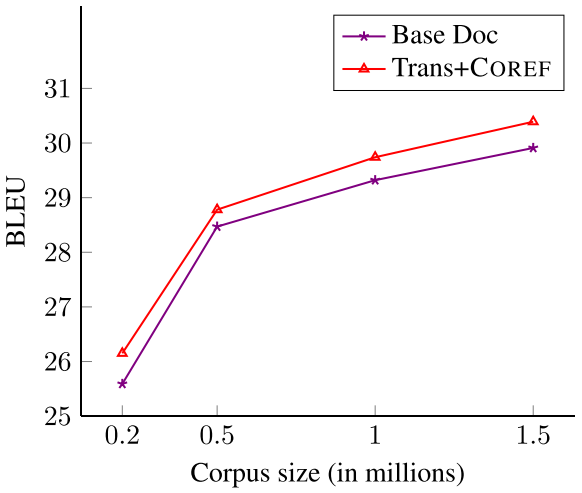

Figure 6: Translation results on the En-Ru dataset (Voita et al., 2019) with different sample sizes.

1,000,000. Subsequently, we evaluated the contribution of the COREF sub-model (Trans+COREF) and the Transformer (Base Doc) on these datasets of different sample sizes. Figure 6 illustrates the results of these experiments. Our proposed system outperforms the Transformer model (Base Doc) across all sample sizes in the test set. Notably, this improvement is not limited to the small dataset size setting but similar trends are observed for medium-sized datasets. These results indicate that our system consistently outperforms the transformer model and achieves improved translation qualities regardless of the dataset size.

Remaining Challenges and Unresolved Questions While our proposed method and existing works enhance translation accuracy for

	Accuracy (%) ↑
Base Doc	12.71
G-Transformer	14.45
Trans+COREF	18.50

Table 10: Accuracy of translating the word *we* into Vietnamese (173 samples).

certain linguistic phenomena, challenges persist, particularly in handling deixis. Unlike straightforward scenarios where additional context aids in accurately translating deictic terms (e.g., determining the speakers in a conversation to correctly translate the words *I* and *You*), some instances require a comprehensive understanding of the provided text's content to achieve correct pronoun translation. Consider the following example from the test data of the English-Vietnamese dataset (Qi et al., 2018): *"Oh my god! you're right! who can* $we_{[chúng\ ta]}$ *sue? Now Chris is a really brilliant lawyer, but he knew almost nothing about patent law and certainly nothing about genetics. I knew something about genetics, but I wasn't even a lawyer, let alone a patent lawyer. So clearly* $we_{[chúng\ tôi]}$ *had a lot to learn before* $we_{[chúng\ ta]}$ *could file a lawsuit."* In this context, the English word *we* is translated as either *chúng tôi* ($we_{[chúng\ tôi]}$) or *chúng ta* ($we_{[chúng\ ta]}$), reflecting the exclusion or inclusion of the listener. This example underscores the importance of contextual nuances in translating pronouns like *we* or *us* from English to Vietnamese, where the choice between *chúng tôi* and *chúng ta* is critical.

Building on the insights from the described example, we extracted all samples that presented similar linguistic challenges, in which a correctly translated sample must ensure that every instance of the word *we* is accurately translated. Table 10 presents the accuracy of translating the word *we* into the correct Vietnamese. While our method surpasses other baseline models in performance, it still exhibits lower accuracy in comparison to the deixis-related outcomes of the contrastive test for Russian (Voita et al., 2019). This discrepancy highlights the phenomenon as a significant challenge that warrants further investigation.

Computational Cost We present a detailed comparison of the parameter count and training time per epoch for our proposed method alongside other baselines in Table 11. When compared to the G-Transformer, our method uses fewer parameters, takes less time to train, and yet achieves better performance. On the other hand, the Base Doc system uses the fewest parameters and trains the quickest, but its results are notably underperforming.

	No. of Params	Training Time	En-De ↑
Base Doc	92.03	407	21.54
MultiResolution	92.03	610	22.09
G-Transformer	101.48	566	22.61
Hybrid Context	65.78	1,776	22.05
CoDoNMT	92.03	638	22.55
Trans+COREF	98.59	503	23.57

Table 11: Number of parameters (in million), training time for one epoch (in seconds), and results of systems (in the BLEU metric) on the En-De dataset.

6 Related Work

Multi-task learning has primarily been utilized in MT tasks to integrate external knowledge into MT systems. Luong et al. (2016), Niehues and Cho (2017), and Eriguchi et al. (2017) have employed multi-task learning with different variations of shared weights of encoders, decoders, or attentions between tasks to effectively incorporate parsing knowledge into sequence-to-sequence MT systems.

For incorporating coreference cluster information, Ohtani et al. (2019), Xu et al. (2021), and Lei et al. (2022) incorporate coreference cluster information to improve their NMT models. Ohtani et al. (2019) integrates coreference cluster information into a graph-based NMT approach to enhance the information. Similarly, Xu et al. (2021) uses the information to connect words across different sentences and incorporates other parsing information to construct a graph at the document-level, resulting in an improvement in translation quality. Lei et al. (2022) employs coreference information to construct cohesion maskings and fine-tunes sentence MT systems to produce more cohesive outputs. On the other hand, Stojanovski

and Fraser (2018) and Hwang et al. (2021) leverage coreference cluster information through augmented steps. They either add noise to construct a coreference-augmented dataset or use coreference information to create a contrastive dataset and train their MT systems on these enhanced datasets to achieve better translation performance. For context-aware MT, Kuang et al. (2018) and Tu et al. (2018) focus on utilizing memory-augmented neural networks, which store and retrieve previously translated parts in NMT systems. These approaches help unify the translation of objects, names, and other elements across different sentences in a paragraph. In contrast, Xiong et al. (2019) and Voita et al. (2019) develop a multiple-pass decoding method inspired by the Deliberation Network (Xia et al., 2017) to address coherence issues, i.e., deixis and ellipsis in paragraphs. They first translate the source sentences in the first pass and then correct the translations to improve coherence in the second pass. Mansimov et al. (2020) introduce a self-training technique, similar to domain self-adaptation, to develop a document-level NMT system. Meanwhile, various methods aim to encapsulate contextual information, i.e., hierachical attention (Maruf et al., 2019), multiple-attention mechanism (Zhang et al., 2020; Bao et al., 2021), and recurrent memory unit (Feng et al., 2022).[7] In a data augmentation approach, Bao et al. (2023) diversify training data for the target side language, rather than only using a single human translation for each source document.

Recently, Wang et al. (2023) have shown that state-of-the-art Large Language Models (LLMs), i.e., GPT-4 (OpenAI et al., 2024), outperform traditional translation models in context-aware MT. In other approaches, Wu et al. (2024) and Li et al. (2024) have developed effective fine-tuning and translation methods for lightweight LLMs; however, the efficacy of NMT models can exceed that of lightweight LLMs, varying by language pair.

7 Conclusion

This study presents a context-aware MT model that explains the translation output by predicting coreference clusters in the source side. The model comprises two sub-models, a translation sub-model and a coreference resolution sub-model, with no modifications to the translation model. The coreference resolution sub-model predicts coreference clusters by fusing the representation from both the encoder and decoder to capture relations in the two languages explicitly. Under the same settings of the En-Ru, En-De, and the multilingual datasets, and following analyses on the coreference sub-model's contributions, the impacts of context and corpus size, as well as the type of information utilized in the sub-model, our proposed method has proven effective in enhancing translation quality.

[7] In Feng et al. (2022), they provided source code without instructions. We tried to reuse and reimplement their method; however, we cannot reproduce their results in any efforts. They did not reply our emails for asking training details. We therefore decided not to include their results in Table 3.

Limitations

In this study, the hidden dimension size in the coreference resolution sub-model is smaller than typical state-of-the-art systems, i.e., 512 vs. 2048, potentially limiting its accuracy and negatively impacting the quality of translation. Additionally, this study requires fine-tuning for a certain hyperparameter that combines the coreference resolution sub-model and the translation model to achieve satisfactory results.

Acknowledgments

The authors are grateful to the anonymous reviewers and the action editor who provided many insightful comments that improve the paper. This work was supported by JSPS KAKENHI grant number JP21H05054.

References

Chinatsu Aone and Scott William. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129, Cambridge, Massachusetts, USA. Association for Computational Linguistics. https://doi.org/10.3115/981658.981675

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Guangsheng Bao, Zhiyang Teng, and Yue Zhang. 2023. Target-side augmentation for document-level machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10725–10742, Toronto, Canada. Association for Computational Linguistics.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1–6, 2021*, pages 3442–3455. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Emanuele Bugliarello and Naoaki Okazaki. 2020. Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1618–1627, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.147

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4828

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 72–78, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2012

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. 2022. Learn to remember: Transformer with recurrent memory for document-level machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.105

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Yongkeun Hwang, Hyeongu Yun, and Kyomin Jung. 2021. Contrastive learning for context-aware neural machine translation using coreference information. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10–11, 2021*, pages 1135–1144. Association for Computational Linguistics.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine

translation benefit from larger context? *CoRR*, abs/1704.05135v1. Version 1.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.3

Dan Klein and Christopher D. Manning. 2002. Conditional structure versus conditional estimation in NLP models. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, July 6–7, 2002*, pages 9–16. https://doi.org/10.3115/1118693.1118695

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics. https://doi.org/10.3115/1073445.1073462

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 – June 1, 2003*. 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.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1018

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Yikun Lei, Yuqi Ren, and Deyi Xiong. 2022. CoDoNMT: Modeling cohesion devices for document-level neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5205–5216, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2024. Enhancing document-level translation of large language model via translation mixed-instructions. *arXiv preprint arXiv:2401.08088v1*. Version 1.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1166

Elman Mansimov, Gábor Melis, and Lei Yu. 2020. Capturing document context inside sentence-vel neural machine translation models with self-training. *CoRR*, abs/2003.05259v1. Version 1. https://doi.org/10.18653/v1/2021.codi-main.14

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.

Joseph F. McCarthy and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20–25 1995, 2 Volumes*, pages 1050–1055. Morgan Kaufmann.

Jan Niehues and Eunah Cho. 2017. Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the Second Conference on Machine Translation*, pages 80–89, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4708

Takumi Ohtani, Hidetaka Kamigaito, Masaaki Nagata, and Manabu Okumura. 2019. Context-aware neural machine translation with coreference information. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-6505

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo,

Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe, Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774v6*. Version 6.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2084

Maria Refinetti, Alessandro Ingrosso, and Sebastian Goldt. 2023. Neural networks trained with SGD learn distributions of increasing complexity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28843–28863. PMLR.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2209

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. https://doi.org/10.18653/v1/P16-1162

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9573–9585. Curran Associates, Inc.

Dario Stojanovski and Alexander Fraser. 2018. Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference*

on Machine Translation: Research Papers, pages 49–60, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6306

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548. Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.279

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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. https://doi.org/10.18653/v1/W17-4811

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420. https://doi.org/10.1162/tacl_a_00029

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Marc B. Vilain, John D. Burger, John S. Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding, MUC 1995, Columbia, Maryland, USA, November 6–8, 1995*, pages 45–52. ACL. https://doi.org/10.3115/1072399.1072405

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. https://doi.org/10.18653/v1/P19-1116

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. https://doi.org/10.18653/v1/P18-1117

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1301

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for

document-level machine translation. *arXiv preprint arXiv:2401.06468v2*. Version 2.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 – February 1, 2019*, pages 7338–7345. AAAI Press. https://doi.org/10.1609/aaai.v33i01.33017338

Mingzhou Xu, Liangyou Li, Derek F. Wong, Qun Liu, and Lidia S. Chao. 2021. Document graph for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8435–8448, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.663

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for 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 5696–5701, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1571

Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomás Kociský. 2017. The neural noisy channel. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. OpenReview.net.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*, pages 27263–27277.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, pages 1081–1087. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.81

Boyuan Zheng, Patrick Xia, Mahsa Yarmohammadi, and Benjamin Van Durme. 2023. Multilingual coreference resolution in multiparty dialogue. *Transactions of the Association for Computational Linguistics*, 11:922–940. https://doi.org/10.1162/tacl_a_00581

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, IJCAI 2020*, pages 3983–3989. ijcai.org. https://doi.org/10.24963/ijcai.2020/551