Modeling Consistency Preference via Lexical Chains for Document-level Neural Machine Translation

Xinglin Lyu[♣], **Junhui Li**[♣*], **Shimin Tao**[♠], **Hao Yang**[♠], **Ying Qin**[♠], **Min Zhang**[♣]
[♣]School of Computer Science and Technology, Soochow University, Suzhou, China
[♠]Huawei Translation Services Center, Beijing, China
xllv2020@stu.suda.edu.cn
{lijunhui,minzhang}@suda.edu.cn
{taoshimin,yanghao30,qinying}@huawei.com

Abstract

In this paper we aim to relieve the issue of lexical translation inconsistency for document-level neural machine translation (NMT) by modeling consistency preference for lexical chains which consist of repeated words in a source-side document and provide a representation of the lexical consistency structure of the document. Specifically, we first propose lexical-consistency attention to capture consistency context among words in the same lexical chains. Then for each lexical chain we define and learn a consistency-tailored latent variable, which will guide the translation of corresponding sentences to enhance lexical translation consistency. Experimental results on Chinese→English and French→English document-level translation tasks show that our approach not only significantly improves translation performance in BLEU, but also substantially alleviates the problem of the lexical translation inconsistency.

1 Introduction

Generally, the translations of source-side words repeated within a document tend to be consistent (Merkel, 1996; Carpuat, 2009; Türe et al., 2012; Guillou, 2013; Al Khotaba and Al Tarawneh, 2015; Lyu et al., 2021; Kang et al., 2021) while sentence-level neural machine translation (NMT) suffers from the serious problem of lexical translation inconsistency due to the lack of inter-sentence context. Although most recent studies in document-level NMT propose various context-aware models to better capture inter-sentence context, they do not handle specific discourse phenomena, e.g., lexical consistency. In this paper, we therefore study lexical consistency for document-level NMT by modeling consistency preference for lexical chains which represent the lexical consistency structure of a text.

Lexical translation consistency is a common discourse phenomenon. Many studies in statistical machine translation (SMT) (Merkel, 1996; Carpuat, 2009; Türe et al., 2012; Guillou, 2013; Al Khotaba and Al Tarawneh, 2015) discuss and apply the one translation per discourse hypothesis. More recently, Lyu et al. (2021) give a detailed analysis about the lexical translation consistency on Chinese→English translation task. In their analysis, the proportion of words related with this phenomenon reaches about 20 percent against all of words in the whole corpus. Furthermore, they find out that the translations of the words repeated within a document indeed tend to be consistent.

Intuitively, enhancing lexical translation consistency should consider inter-sentence context. However, existing researches in document-level NMT mainly focus on capturing inter-sentence context in general and do not explicitly handle specific discourse phenomena. As a result, these models have limited effect on enhancing lexical translation consistency. Different from them, both Lyu et al. (2021) and Kang et al. (2021) recently introduce auxiliary consistency losses to encourage the translations of the repeated source-side words being same.[1]

Alternatively, in this paper we propose a softer approach to model consistency preference for lexical chains in document-level NMT, where the lexical chains consist of repeated words in a source-side document. Specifically, we enhance the translation consistency of lexical chains from two aspects: 1) we propose a lexical-consistency attention to capture consistency context among words in the same lexical chains while we also use a general-context attention to capture general inter-sentence context; and 2) we propose a consistency-tailored latent variable for each lexical chain to model its

*Corresponding author: Junhui Li.

[1]Kang et al. (2021) additionally use a classifier to predict whether the translations of a pair of repeated words should be same or not.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6312 - 6326
December 7-11, 2022

consistency preference, which will guide the translation of the lexical chain in decoding. These latent variables are properly learned via a conditional variational autoencoder (CVAE) module, which does not explicitly constrain the translations of repeated source words to be same, and thus could ease over-correction. Experimental results on Chinese-to-English and French-to-English document-level translation tasks show that our approach not only significantly improves the translation performance in BLEU, but also greatly alleviates the problem of lexical translation inconsistency.

2 Related Work

There have been substantial studies in document-level NMT that focus on effective utilization of general inter-sentence context. These studies can be roughly categorized into three groups, including those who only consider the source-side inter-sentence context (Jean et al., 2017; Wang et al., 2017; Voita et al., 2018; Zhang et al., 2018; Tan et al., 2019; Yang et al., 2019; Mace and Servan, 2019; Kang et al., 2020; Xu et al., 2021; Fernandes et al., 2021), those who only consider the target-side inter-sentence context (or translation history) (Kuang et al., 2018; Tu et al., 2018; Xiong et al., 2019), and those who consider both the source and target inter-sentence context (Bawden et al., 2018; Maruf and Haffari, 2018; Maruf et al., 2019; Zheng et al., 2020; Bao et al., 2021; Sun et al., 2022). Different from ours, these studies leverage the inter-sentence context in a indistinguishable way and do not handle specific discourse phenomena. Therefore, although they achieve impressive improvement of translation accuracy with the expectation of alleviating the discourse issues in general, they usually have limited effect on enhancing lexical translation consistency.

There also exist many researches in MT that aim to enforce or encourage lexical translation consistency. In SMT, Carpuat (2009), Xiao et al. (2011) and Garcia et al. (2014, 2017) propose post-editing approaches to enforce lexical translation consistency by re-translating those repeated source words which have been translated differently. Tiedemann (2010a,b) and Gong et al. (2011) propose cache-based approaches to encourage translation consistency. Ma et al. (2011) and He et al. (2011) propose discriminative approaches to improve the consistency of translations. Türe et al. (2012) add three super-sentential "consistency features" to the translation model. Beside, Pu et al. (2017) propose to improve the translation consistency of repeated nouns in post-editing or/and reranking. In NMT, Lyu et al. (2021) and Kang et al. (2021) propose consistency losses which encourage the translations of repeated words being same. Different from theirs, our approach automatically models the translation consistency preference of repeated words via a CVAE module in a softer way without explicitly constraining their translations to be same.

3 Proposed Approach

As our goal is to model consistency preference for lexical chains, we first construct lexical chains of repeated words in the source-side document. Each lexical chain represents a repeated word that appear two or more times in the document. Then we encode source-side documents with prepared lexical chains by a consistency-aware encoder (Section 3.1). Meanwhile, we propose to learn a consistency-tailored latent variable for each lexical chain (Section 3.2) by a CVAE-based module. These learned latent variables are dynamically integrated into each decoding step to further enhance translation consistency (Section3.3). Finally, we define a joint training objective to optimize the model with the CVAE-based module (Section 3.4).

We define some notations before describing our approach. Given a parallel document pair $(\mathcal{X}, \mathcal{Y}) = \{X_i, Y_i\}|_{i=1}^{N}$ with N sentence pairs, we assume that each source sentence X_i consists of n words $(x_1^i, \cdots, x_n^i)$ while its target sentence Y_i consists of m words $(y_1^i, \cdots, y_m^i)$. From source document $\mathcal{X}$, we extract a set $\mathcal{S} = \{S_j\}|_{j=1}^{M}$ with M lexical chains. Specifically, each lexical chain $S_j = (a_k^j, b_k^j)\,|_{k=1}^{K}$ records all positions of a word repeated K times in document $\mathcal{X}$, where a and b indicate the sentence index and word index of a position, respectively. As shown in the bottom of Figure 1, lexical chain S = {(1,4), (2,3), $\cdots$, (N-1,4), (N,3)} indicates that x_4^1, x_3^2, $\cdots$, x_4^{N-1}, and x_3^N are a repeated word. See Appendix A for more about the construction and statistics of the lexical chains. We use d as the model size of embeddings and hidden states, and d_z as the size of consistency-tailored latent variable.

3.1 Consistency-aware Encoding

We propose a consistency-aware encoder to encode source documents. As shown in Figure 1, different from standard Transformer encoder, the

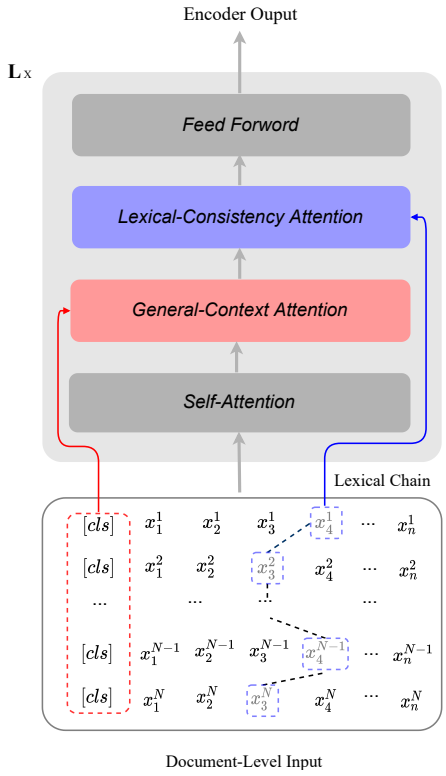

Figure 1: Consistency-aware encoder with a general-context attention sublayer and a lexical-consistency attention sublayer.

consistency-aware encoder equips with two additional attention sublayers, *Lexical-Consistency* attention and *General-Context* attention between the self-attention sublayer and the feed-forward sublayer. The two sublayers aim to capture consistency context and general inter-sentence context from source document, respectively.

Modeling General Inter-Sentence Context. As related studies show that modeling general inter-sentence context is helpful to improve translation performance, we propose *General-Context Attention* sublayer to properly modeling the general inter-sentence context.

Specifically, we follow BERT (Devlin et al., 2019) and add a special token $[cls]$ at the beginning of each sentence, as shown in the bottom of Figure 1. In the l-th encoder layer, we encode sentences $X_i|_{i=1}^{N}$ within document $\mathcal{X}$ with a multi-head attention function (Self-Attention in Figure 1) synchronously:

$$B^{(l)} = \text{LN}\left(\text{MultiHead}\left(A^{(l)}, A^{(l)}, A^{(l)}\right) + A^{(l)}\right), \quad (1)$$

where $\text{LN}(\cdot)$ is the layer normalization function (Ba et al., 2016), $A^{(l)} \in \mathbb{R}^{N\times n\times d}$ is the input sequence of the encoder layer, and $B^{(l)} \in \mathbb{R}^{N\times n\times d}$ is the output sequence of the self-attention sublayer. Then we feed $B^{(l)}$ to the General-Context attention sublayer to exchange information among sentences with document $\mathcal{X}$ via those $[cls]$ tokens:

$$C_{cls}^{(l)} = \text{MultiHead}\left(B_{cls}^{(l)}, B_{cls}^{(l)}, B_{cls}^{(l)}\right), \quad (2)$$

where $B_{cls}^{(l)} \in \mathbb{R}^{N\times d}$ is indexed from $B^{(l)}$ for $[cls]$ tokens. $C_{cls}^{(l)} \in \mathbb{R}^{N\times d}$ is further broadcast into shape $\mathbb{R}^{N\times n\times d}$ and added to $B^{(l)}$:

$$D^{(l)} = \text{LN}\left(B^{(l)} + C_{cls}^{(l)}\right), \quad (3)$$

where $D^{(l)} \in \mathbb{R}^{N\times n\times d}$ is the output of the General-Context attention sublayer.

Modeling Consistency Context via Lexical Chains. Given M lexical chains $\{S_j\}|_{j=1}^{M}$ in $\mathcal{X}$, we first index their states into $D_c^{(l)} \in \mathbb{R}^{M\times K\times d}$ from $D^{(l)}$. Inspired by Lyu et al. (2021), then we employ another multi-head attention (Lexical-Consistency Attention in Figure 1) to exchange information among tokens within a lexical chain:

$$E_c^{(l)} = \text{LN}\left(\text{MultiHead}\left(D_c^{(l)}, D_c^{(l)}, D_c^{(l)}\right) + D_c^{(l)}\right), \quad (4)$$

where $E_c^{(l)} \in \mathbb{R}^{M\times K\times d}$ is further used to replace their corresponding states in $D^{(l)}$ (i.e., $D_c^{(l)}$). We refer it as $F^{(l)}$ after the replacement. $F^{(l)} \in \mathbb{R}^{N\times n\times d}$ is fed into the feed-forward sublayer to get the final output of l-th encoder layer $G^{(l)} \in \mathbb{R}^{N\times n\times d}$:

$$G^{(l)} = \text{LN}\left(\text{FFN}\left(F^{(l)}\right) + F^{(l)}\right). \quad (5)$$

The output of the final encoder layer, i.e, $G^{(L)}$ (hereafter G for simplicity), will be used as the encoder output of the document. Specifically, we use g_j^i to indicates the hidden state for x_j^i, i.e., the j-th word in the i-th sentence.

3.2 Modeling Consistency Preference via Latent Variational Module

As shown in Figure 2, we propose a latent variational module to learn the distribution of consistency preference for every lexical chain. Given a lexical chain $S = (a_k, b_k)|_{k=1}^{K}$, next we describe how to learn its consistency-tailored latent vairable in training and inference, respectively.

Learning Consistency-Tailored Latent Variable in Training. For lexical chain $S = (a_k, b_k)|_{k=1}^{K}$, we first encode the chain and extract its potential translation. Then we produce a consistency-tailored latent variable $z \in \mathbb{R}^{d_z}$.

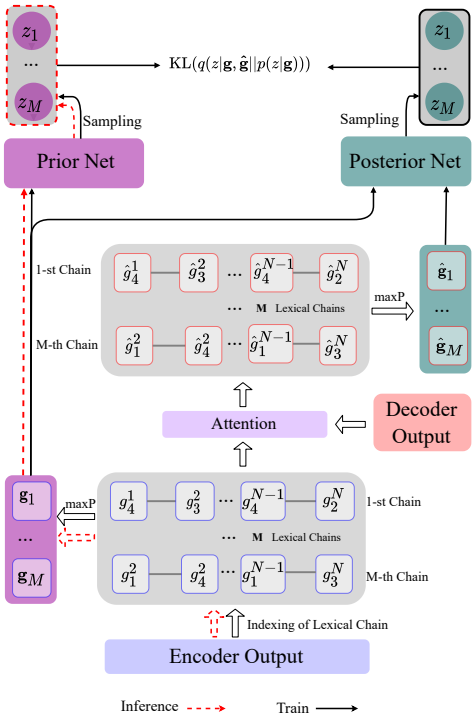

Figure 2: Illustration of the proposed latent variational module for modeling consistency preference.

The hidden states of the chain S, $\mathbf{g}_S \in \mathbb{R}^{K\times d}$ can be extracted from the encoder output G as:

$$\mathbf{g}_S = \left(g_{b_1}^{a_1}, \cdots, g_{b_k}^{a_K}\right). \tag{6}$$

To obtain the potential translation of the chain S from the target-side document $\mathcal{Y}$, we first use the decoder to synchronously get the hidden states H of all target-side sentences:

$$H = \text{Decoder}\left(G, \mathcal{Y}\right), \tag{7}$$

where $H \in \mathbb{R}^{N\times m\times d}$ is the output of the last decoder layer. Specifically, we denote the target-side hidden states of the i-th sentence Y_i as $H[i]$. As there is no explicit word-level alignment between the source-side and target-side documents, we could not directly obtain the chain's translation. Alternatively, we then employ an attention mechanism to implicitly learn its translation. For the k-th word in the chain S with its source-side hidden states $g_{b_k}^{a_k}$, we obtain its target-side counterpart $\hat{g}_{b_k}^{a_k} \in \mathbb{R}^d$ as weighted sum of the target-side hidden states $H[a_k]$:

$$\hat{g}_{b_k}^{a_k} = \text{Softmax}\left(g_{b_k}^{a_k} W_s \left(H[a_k]\right)^T\right) H[a_k], \tag{8}$$

where $W_s \in \mathbb{R}^{d\times d}$ is a trainable parameter matrix. Consequently, we obtain $\hat{\mathbf{g}}_S = \left(\hat{g}_{b_1}^{a_1}, \cdots \hat{g}_{b_K}^{a_K}\right)$, as the target-side hidden states of the chain S.

Once we obtain both the hidden states $\mathbf{g}_S$ of the chain S and its target-side counterpart $\hat{\mathbf{g}}_S$, we follow Wang and Wan (2019) and use eisotropic Gaussian distribution as the posterior distribution to sample the latent variable $z \in \mathbb{R}^{d_z}$:

$$z : q_\phi\left(z|\mathbf{g}_S, \hat{\mathbf{g}}_S\right) \sim \mathcal{N}\left(\mu, \sigma^2\mathbf{I}\right), \tag{9}$$

where $\mathbf{I}$ denotes the identify matrix, μ and σ are learned via neural networks:

$$\mu = \text{MLP}_\phi\left(\left[\text{maxP}\left(\mathbf{g}_S\right); \text{maxP}\left(\hat{\mathbf{g}}_S\right)\right]\right), \tag{10}$$

$$\log(\sigma^2) = \text{Softplus}\left(\text{MLP}_\phi\left(\left[\text{maxP}\left(\mathbf{g}_S\right); \text{maxP}\left(\hat{\mathbf{g}}_S\right)\right]\right)\right), \tag{11}$$

where $\text{MLP}\left(\cdot\right)$ and $\text{Softplus}\left(\cdot\right)$ are multi-layer perceptron and approximation of ReLU function, respectively. $\text{maxP}\left(\cdot\right)$ is MaxPooling function that converts chain-level hidden states $\mathbf{g}_S$ (or $\hat{\mathbf{g}}_S$) into a d-sized vector. $[\cdot;\cdot]$ is concatenation operation.

Learning Consistency-Tailored Latent Variable in Inference. In inference, the target-side hidden states of the chain S, i.e., $\hat{\mathbf{g}}_S$, is unavailable due to the unobservability of the ground-truth translation. Therefore, we sample the consistency-tailored latent variable $z \in \mathbb{R}^{d_z}$ for the chain S from a prior distribution which is conditioned only on the source-side hidden states of the chain S, i.e., $\mathbf{g}_S$:

$$z : p_\theta\left(z|\mathbf{g}_S\right) \sim \mathcal{N}\left(\mu', \sigma'^2\mathbf{I}\right). \tag{12}$$

Similarly, we employ another neural network to learn the prior distribution:

$$\mu' = \text{MLP}_\theta\left(\text{maxP}\left(\mathbf{g}_S\right)\right), \tag{13}$$

$$\log\left(\sigma'^2\right) = \text{Softplus}\left(\text{MLP}_\theta\left(\text{maxP}\left(\mathbf{g}_S\right)\right)\right). \tag{14}$$

The prior distribution is properly trained by approaching the posterior distribution via a KL divergence loss during the training stage:

$$J_{\text{con}}^S\left(\theta\right) = \text{KL}\left(q_\phi\left(z|\mathbf{g}_S, \hat{\mathbf{g}}_S\right) || \, p_\theta\left(z|\mathbf{g}_S\right)\right), \tag{15}$$

which enables the prior network learn the reliable consistency-tailored distribution even though the $\hat{\mathbf{g}}_S$ is unobservable.

3.3 Consistency-aware Decoding

After obtaining a consistency variables set $Z = \{z_j\}|_{j=1}^M$ either from the posterior distribution (at training stage) or the prior distribution (at inference stage) for the lexical chains set $\mathcal{S}$, we use the these

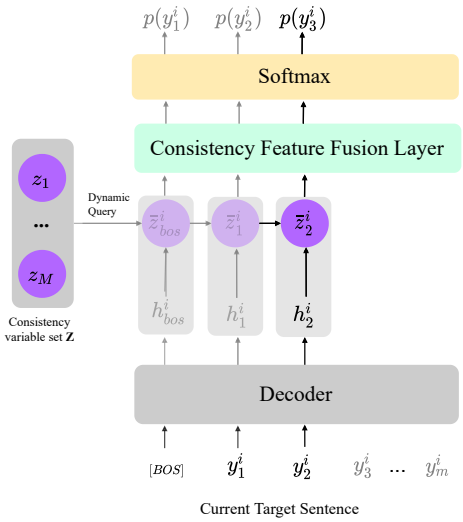

Figure 3: Illustration of consistency-aware decoding when generating the i-th sentence in a document.

consistency variables to enhance the translation consistency at each decoding step.

Dynamic Query over Consistency Variables. Figure 3 illustrates how to use these consistency-tailored latent variables in decoding for the i-th sentence. On the one hand, we note that the j-th variable z_j is relevant to the i-th sentence only if the j-th lexical chain contains words that are from sentence X_i. So we define function $f(z_j, X_i)$ which returns 1 if the j-th chain contains at least one word from X_i, otherwise 0. On the other hand, even z_j is relevant, it concerns to very few decoding steps. So we define a similarity function to compare the decoder hidden states against the latent variable. Overall, given the decoder output h_t^i at the t-th decoding step of sentence X_i, we perform dynamic query over latent variables $\{z_j\}|_{j=1}^M$ and obtain consistency feature $\bar{z}_t^i \in \mathbb{R}^{d_z}$:

$$\hat{z}_j = f(z_j, X_i) \cdot z_j, \tag{16}$$

$$\bar{z}_t^i = \sum_{j=1}^{M} \text{sim}(W_t h_t^i, \hat{z}_j) \cdot \hat{z}_j, \tag{17}$$

where $W_t \in \mathbb{R}^{d_z \times d}$ is a trainable parameter matrix, and $\text{sim}(\cdot,\cdot)$ is the cosine similarity function.[2]

Fusion of Consistency Feature. We fuse the consistency feature $\bar{z}_t^i$ of the t-th decoding step with the decoder output h_t^i in the output layer as o_t^i:

$$o_t^i = \text{Tanh}(W_z[\bar{z}_t^i; h_t^i] + b_z), \tag{18}$$

where $W_z \in \mathbb{R}^{d \times (d+d_z)}$ and $b_z \in \mathbb{R}^d$ are trainable parameters. Finally, o_t^i is fed to a linear transformation and a Softmax layer to get the probability distribution of y_t^i, i.e.,:

$$p(y_t^i) = \text{Softmax}(W_o o_t^i), \tag{19}$$

where we use target-side word embedding parameters as $W_o \in \mathbb{R}^{|V| \times d}$, and $|V|$ is vocabulary size.

3.4 Joint Training Objective

We have introduced a KL-divergence loss to learn the consistency variable in Section 3.2. The joint objective function of our model $J(\theta)$ over document pair $(\mathcal{X}, \mathcal{Y})$ is defined as:

$$J(\theta) = J_{\text{NMT}}(\theta) + \alpha J_{\text{con}}(\theta), \tag{20}$$

$$J_{\text{NMT}}(\theta) = -\sum_{i,t} \log p\left(y_t^i | y_{<t}^i, \bar{z}_t^i, \mathcal{X}\right), \tag{21}$$

$$J_{\text{con}}(\theta) = \sum_j J_{\text{con}}^{S_j}(\theta), \tag{22}$$

where α determines the contributions of KL-divergence loss $J_{\text{con}}(\theta)$, and $J_{\text{NMT}}(\theta)$ is the cross entropy loss function.

4 Experimentation

As inspired by the conclusion in Guillou (2013) and Lyu et al. (2021) that lexical translation consistency is encouraged in Chinese (ZH)→English (EN) and French (FR)→English (EN) human translation, we evaluate our approach on {ZH, FR}→EN document-level translation tasks.

4.1 Experimental Settings

Datasets. For ZH→EN (News), we follow Zhang et al. (2018) and use document parallel corpora from LDC as the training set, NIST2006 dataset as the development set and combination of NIST2002, 2003, 2004, 2005 and 2008 as the test set. For ZH→EN (TED), the training set is from the IWSLT 2014 and 2015 evaluation (Cettolo et al., 2012, 2015). We use dev2010 as the development and combine tst2010-2013 as the test set. For FR→EN (TED), the training set is from the 2015 evaluation (Cettolo et al., 2015). We use dev2010 as the development and combine tst2010-2013 as the test set. More statistics and preprocessing of the experimental datasets are in Appendix A.

Model details. We use *OpenNMT* (Klein et al., 2017) as the sentence-level Transformer and extend it. For the dimension of latent variables, we set d_z = 96. See Appendix B for more details of implementation of model.

[2] We map similarity γ from [-1, 1] to [0, 1] via $0.5(\gamma + 1)$.

Type	Model	ZH→EN (News)			ZH→EN (News, *2M Pretraining*)		
		s-BLEU	d-BLEU	LTCR	s-BLEU	d-BLEU	LTCR
Sent2Sent	Transformer	40.55	43.11	56.87	47.61	49.67	63.26
Doc2Sent	+LC-Attn	‡42.39	‡44.71	61.31	‡48.23	‡50.60	64.16
	+LC-Attn + ConVar	‡42.54	‡44.81	64.09	‡48.40	‡50.95	**64.86**
	+LC-Attn + ConVar + GC-Attn	‡**43.27**	‡**45.39**	**64.56**	‡**48.89**	‡**51.07**	64.72
Previous context-aware NMT models							
Doc2Sent	HAN (Miculicich et al., 2018)	41.58	43.61	57.01	48.01	50.37	63.44
Doc2Doc	G-Trans (Bao et al., 2021)	42.91	44.97	60.77	48.60	50.95	64.12
Doc2Sent	W-Link (Lyu et al., 2021)	42.69	44.83	63.88	48.31	50.64	64.55

Table 1: Performance (BLEU and LTCR scores) on the test set of ZH→EN (News). *2M Pretraining* indicates we use 2M sentence pairs at the first training stage. ‡ indicates that the improvement in BLEU is significant over Transformer at 0.01 (Koehn, 2004). Scores with **bold**/underline indicate the top/second best performance.

Type	Model	ZH→EN (TED)			FR→EN (TED)		
		s-BLEU	d-BLEU	LTCR	s-BLEU	d-BLEU	LTCR
Sent2Sent	Transformer	18.17	24.59	59.35	39.34	43.67	82.98
Doc2Sent	+LC-Attn	‡19.76	‡25.83	66.94	‡41.05	‡44.97	85.41
	+LC-Attn +ConVar	‡20.04	‡26.00	70.22	‡41.35	‡45.36	**88.21**
	+LC-Attn +ConVar +GC-Attn	‡20.42	‡**26.38**	**70.40**	‡**41.62**	‡45.60	87.94
Previous context-aware NMT models							
Doc2Sent	HAN (Miculicich et al., 2018)	18.97	25.01	61.33	40.89	44.77	82.76
Doc2Doc	G-Trans (Bao et al., 2021)	19.61	26.12	62.59	41.52	**45.87**	85.37
Doc2Sent	W-Link (Lyu et al., 2021)	**20.47**	25.92	68.33	41.21	45.44	86.57

Table 2: Performance (BLEU and LTCR scores) on the test sets of ZH/FR→EN (TED) translation tasks. ‡ indicates that the improvement in BLEU is significant over Transformer at 0.01 (Koehn, 2004). Scores with **bold**/underline indicate the top/second best performance.

Training and inferring strategy. To train the models more effectively, we follow Lyu et al. (2021) and adapt a two-stage training strategy. In the first training stage, we use the sentence pairs to *pretrain* the sentence-level modules with the training objective $J_{\text{NMT}}(\theta)$ while in the second training stage we train all modules with the joint training objective $J(\theta)$. To alleviate the degeneration problem of the variational framework, we follow Liang et al. (2021) and apply KL annealing. Other training settings are in Appendix B. In inferring, we set the beam size to 5 and the length penalty to 0.6.

Evaluation metrics. We report both sentence-level metric (s-BLEU) and document-level metric (d-BLEU) to evaluate the quality of the translation. For all translation tasks, we report case-insensitive BLEU score calculated by *multi-bleu.perl* script. To evaluate the ability of enhancing the lexical translation consistency, we follow Lyu et al. (2021) and report the LTCR (Lexical Translation Consistency Ratio) score.[3] We also report the Herfindahl Hirschman Index (HHI) score in Appendix D.

[3]Different from Lyu et al. (2021) that uses `fast-align` (Dyer et al., 2013), we use `awesome-align` (Dou and Neubig, 2021) with higher alignment accuracy to do word alignment.

4.2 Experimental Results

Besides sentence-to-sentence (Sent2Sent) Transformer baseline, we compare our performance to three representative context-aware models: `HAN` (Miculicich et al., 2018) which models document-level context for better translating source sentences,[4] `G-Transformer` (Bao et al., 2021) which directly views source documents as long sequences and perform seq2seq translation with a long sequence-tailored Transformer,[5] and `W-Link` (Lyu et al., 2021) which encourages lexical translation consistency via word links. Among them, `HAN` and `W-Link` are document-to-sentence (Doc2Sent) models while `G-Transformer` is a document-to-document (Doc2Doc) model. For fair comparison, we run their source code or our re-implementation with our experimental settings.

Results on ZH→EN (News). Table 1 shows the performance on the test set of ZH→EN (News). From it, we first observe that using lexical-consistency attention alone (+LC-Attn) to capture consistency-context via lexical chain significantly improves translation performance in both BLEU (from 40.55 to 42.39 in s-BLEU) and LTCR (from

[4]https://github.com/idiap/HAN_NMT

[5]https://github.com/baoguangsheng/g-transformer

56.87 to 64.56). It is also not surprised to observe that although incorporating the latent variational module (+LC-Attn +ConVar) slightly improves BLEU score (e.g., from 42.39 to 42.54 in s-BLEU), it significantly boosts the LTCR score from 61.31 to 64.09. This suggests that the proposed consistency-tailored latent variable is effective in modeling consistency preference for document translation. In contrast, furthering modeling general document-level context via the General-Context Attention (+LC-Attn + ConVar +GC-Attn) has limited effect in LTCR performance while it further improves translation performance from 42.59 to 43.27 in s-BLEU. This is reasonable since modeling general inter-sentence context does not aim to resolve a particular discourse phenomenon. Similar performance trend is also observed when using more sentence pairs (e.g., 2M) for pretraining. Comparing the performance when using 0.8M and 2M sentence pairs for pretraining, we observe that although there exists a big performance gap in BLEU, the gap of LTCR is quite small (e.g., 64.56 v.s. 64.72).

Compared to the three previous context-aware NMT models, our approach achieves the best performance in both BLEU and LTCR scores. Specifically, we note that `G-Transformer` consistently achieves higher LTCR scores than `HAN` and baseline as it views a source document as a long sequence and obtains its translation sentence by sentence. Therefore, to some extent it implicitly encourages lexical translation consistence among target-side sentences. However, different from Doc2Sent models, Doc2Doc models could not translate sentences within a document in a synchronous way.

Results on other translation tasks. Table 2 shows translation performance on the test sets of the ZH/FR→EN (TED) translation tasks. From it we have similar conclusions as on ZH→EN (News) that *lexical-consistency attention* contributes improvement on both BLEU and LTCR while *consistency-tailored latent variables* and *general-context attention* mainly contribute on improvement on LTCR and BLEU, respectively.

5 Discussion

Next, we take ZH→EN (News) translation as a representative to discuss how our proposed approach improves translation performance. We also provide more discussion in Appendix D∼G.

Model	Coherence	Cohesion
Transformer	0.6830	0.8627
HAN	0.6894	0.8878
G-Trans	0.6953	0.9003
W-Link	0.6973	0.8998
Ours	0.7038	0.9087
Reference	**0.7140**	**0.9257**

Table 3: Results of discourse coherence and cohesion on Zh→EN (News) test set.

5.1 Discourse Coherence

Lapata and Barzilay (2005) propose to measure discourse coherence as the degree of similarity between sentences in a document. They view the representation of a sentence as the mean (centroid) of the distributed vectors of its words, and the similarity between two sentences Y_i and Y_j as the cosine of their means. Similar to Liang et al. (2021), we use Word2Vec[6] to learn the 100-sized word vectors on the English Gigaword.[7] Given target-side documents $\{\mathcal{Y}\}$, we use the averaged cosine similarity between the current sentence Y_i and its adjacent Y_{i+1} as the discourse coherence:

$$\text{Coherence} = \frac{1}{\sum_{\mathcal{Y}} (|\mathcal{Y}| - 1)} \sum_{\mathcal{Y}} \sum_{i=1}^{|\mathcal{Y}|-1} \text{sim}(Y_i, Y_{i+1}). \quad (23)$$

Table 3 (left) shows the discourse coherence scores on ZH→EN (News) test set. It reveals that all context-aware models, including `HAN`, `G-Trans`, `W-Link`, and ours have better coherence in document translation than the baseline system while our model achieves the best coherence.

5.2 Discourse Cohesion

Next Sentence Prediction (NSP) is proposed as a pre-training task in BERT (Devlin et al., 2019). Given two sentences (Y_i, Y_j), NSP will compute the probability of Y_j being the next sentence of Y_i. We propose to measure the discourse cohesion by the mean of NSP probabilities of all adjacent sentence pairs in the test set.

Specifically, the discourse cohesion on the translation $\{\mathcal{Y}\}$ of test set is calculated as following:

$$\text{Cohesion} = \frac{1}{\sum_{\mathcal{Y}} |\mathcal{Y}| - 1} \sum_{\mathcal{Y}} \sum_{i=1}^{|\mathcal{Y}|-1} \text{NSP}([Y_i; Y_{i+1}]). \quad (24)$$

We use the `bert-base-uncased` model from *Huggingface* (Wolf et al., 2020) to compute the probability of NSP.[8] Table 3 (right) shows the discourse

[6] http://word2vec.googlecode.com/svn/trunk/
[7] https://catalog.ldc.upenn.edu/LDC2009T13
[8] https://huggingface.co/bert-base-uncased

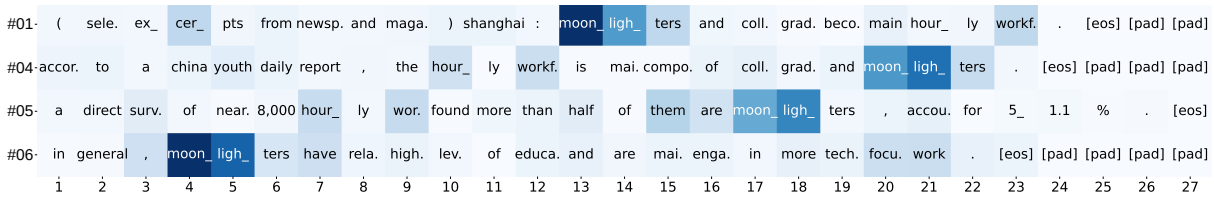

Figure 4: Visualization of the attention score from source token 兼职/*jian_zhi* against the target-side tokens in the corresponding sentences. The token 兼职/*jian_zhi* appears in the 1st, 4th, 5th and 6th source-side sentences. Each pixel illustrates the attention score. The darker it is, the higher the attention score is.

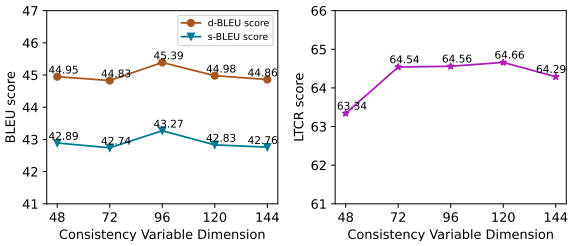

Figure 5: Effect of dimension of consistency-tailored latent variables on BLEU score and LTCR score of ZH→EN(News) test set.

Annotator	Equal	Better	Worse
1	39%	40%	21%
2	43%	37%	20%
Average	41%	39%	20%

Table 4: Human evaluation results on 500 sentences from our test set. We compare our approach with sentence-level Transformer.

cohesion scores of various systems on the test set. It suggests that the utilization of the inter-sentence context improves discourse cohesion while our model achieves the best cohesion.

5.3 Attention in Target-side Information

In Section 3.2, we employ an attention mechanism (Equation 8) to extract the target-side counterpart for each source token in lexical chains. To verify whether the attention module could effectively map the source token in lexical chain to its most-related translation, we visualize the attention score of a lexical chain in development set.

As shown in Figure 4, source token 兼职/*jian_zhi* appears in the 1st, 4th, 5th and 6th sentences of the source document, and we observe the attention module consistently assigns higher attention scores to *moon_* and *ligh_* tokens in their corresponding target-side translation. The averaged and summed attention weights of *moon_* and *ligh_* is 0.63. This confirms our conjecture that we can rely on the attention module to implicitly learn the translation of a source token.

5.4 Dimension of Consistency Variables

As shown in Section 5, we model the consistency preference by introducing a consistency-tailored latent variable $z \in \mathbb{R}^{d_z}$. Figure 5 presents the translation performance of *+LC-Attn + ConVar + GC-Attn* system when d_z ranges from 48 to 144. As shown, it improves performance when increasing d_z from 48 to 96 while it no longer improves or even hurts performance on BLEU score and LTCR score when d_z increases from 96 to higher.

5.5 Interpretability of Consistency Variables

In Section 3.2, we adapt a dynamic way to obtain the consistency feature $\bar{z}$ at each decoding step (Equation 17) over all consistency variables Z. To examine what learned by latent variational module (i.e., interpretability of consistency variables) and whether the module could effectively extract the most-related consistency feature at decoding steps, we use an example from the development to visualize the similarity scores against its relevant consistency variables at each decoding step.

As shown in Figure 6, the sentence has 6 relevant consistency variables. It shows that at certain decoding steps, the query module assigns higher similarity score to their corresponding most-related consistency variable. For example, at decoding steps for predicting *heilongjiang* and *ecological*, it assigns 0.62 and 0.69 similarity scores to z(黑龙江) and z(生态), respectively, while it assigns very low scores to other consistency variables. This suggests that the information inside consistency variables is closely tied to their corresponding most-related words, which confirms our conjecture that we can rely on the query module to implicitly extract the most relevant consistency feature at each decoding step.

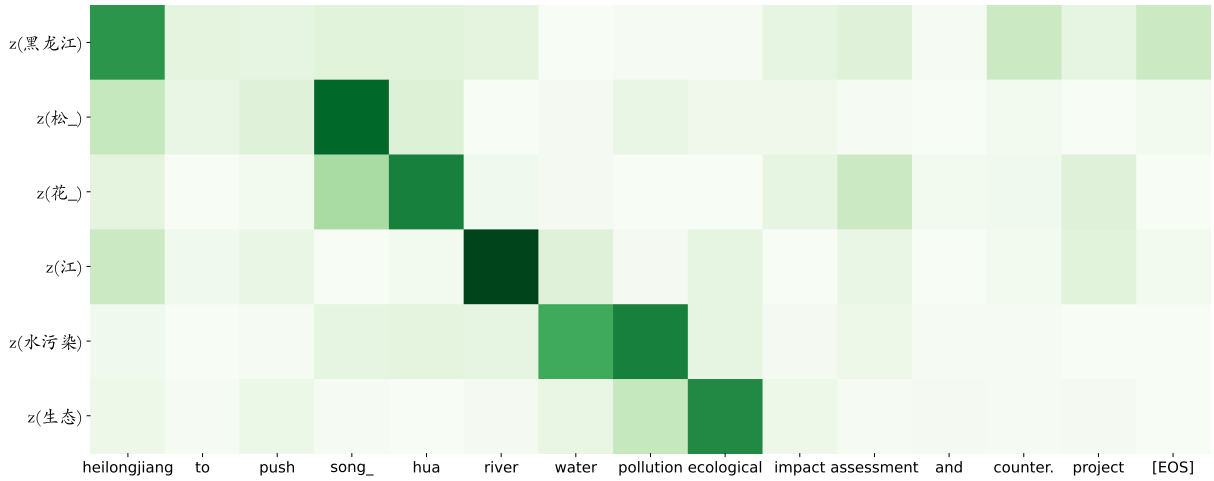

Figure 6: Visualization of similarity scores in Equation 17 from the tokens in translation (*horizontal*) against relevant consistency variables (*vertical*). $z(x)$ denotes the consistency variable of the lexical chain of x in the source-side document. The darker each pixel is, the higher the similarity score is.

5.6 Human Evaluation

Similar to Lyu et al. (2021), we randomly select 500 sentences from the test set and conduct a human evaluation on them. For each selected source-side sentence, we assign its translation background and corresponding two generated translations from the sentence-level Transformer and our approach to two human annotators without order. The translation background consists of its two preceding and two future source-side sentences and their target-side references. Following Voita et al. (2019) and Lyu et al. (2021), the annotators are asked to pick one of the tree options: (1) the first translation is better, (2) the second translation is better, and (3) the two translations are equal quality. Both annotators are postgraduate students and not involved in other parts of the paper.

Table 4 shows the results of human evaluation. It shows that on average 41% cases have equal quality. Among the other cases, the annotators have an obvious preference for our approach since it outperforms Transformer in 66% cases. We also provide a case study in Appendix G.

6 Conclusion

In this paper, we have proposed a lexical-chain based approach to alleviate the issue of translation inconsistency for document-level NMT. We first use lexical-consistency attention to capture consistency context among words in the same lexical chains. Then we learn a consistency-tailored latent variable for each lexical chain to model consistency preference in translation. Experimental results on Chinese→English and French→English document-level translation tasks show that our approach could both improve translation performance in BLEU and enhance lexical translation consistency.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive feedback. This work was supported by the National Natural Science Foundation of China (Grant No. 62036004).

Limitations

The lexical chains in this work consist of repeated words in the source-side document. However, some source words with similar semantics but different morphology also potentially have same translations. For example, both 名誉/*ming_yu* and 声望/*sheng_wang* can be translated into *reputation*. Therefore, these lexical chains in this work are limited in diversity. Introducing more diversity into lexical chains, e.g., synonym-based lexical chains proposed in Xiong et al. (2013), will be explored in our future work. We also note that the computation of both the LTCR and HHI scores is based on whether translations of a repeated word are consistent or same. However, it does not take the reference into account and ignores the correctness of these translations. Therefore, introducing a more appropriate metric to evaluate the translations of repeated words from both consistency and correctness aspects will also be explored in future work.

References

Eissa Al Khotaba and Khaled Al Tarawneh. 2015. Lexical discourse analysis in translation. *Education and Practice*, 6(3):106–112.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *Computing Research Repository*, arXiv:1607.06450.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of ACL*, pages 3442–3455.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1304–1313.

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

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*, pages 261–268.

Mauro Cettolo, Niehues Jan, Stüker Sebastian, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign. In *Proceedings of IWSLT*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of EACL*, pages 2112–2128.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of ACL*, pages 644–648.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André FT Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of ACL*, pages 6467–6478.

Eva Martínez Garcia, Carles Creus, Cristina Espana-Bonet, and Lluís Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *Prague Bulletin of Mathematical Linguistics*, 108:85–96.

Eva Martínez Garcia, Cristina Espana-Bonet, and Lluís Màrquez. 2014. Document-level machine translation as a re-translation process. *Procesamiento del Lenguaje Natural*, 53:103–110.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of EMNLP*, pages 909–919.

Liane Guillou. 2013. Analysing lexical consistency in translation. In *Proceedings of DiscoMT*, pages 10–18.

Yifan He, Yanjun Ma, Andy Way, and Josef van Genabith. 2011. Rich linguistic features for translation memory-inspired consistent translation. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 456–463.

Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic validation of terminology translation consistenscy with statistical method. In *Proceedings of EAMT*, pages 269–274.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machinetranslation benefit from larger context? *Computing Research Repository*, arXiv:1704.05135.

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 EMNLP*, pages 2242–2254.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2021. Enhancing lexical translation consistency for document-level neural machine translation. *Transactions on Asian and Low-Resource Language Information Processing*, 21:1–21.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceeding of ICLR*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL, System Demonstrations*, pages 67–72.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

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 ACL, System Demonstrations*, pages 177–180.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of COLING*, pages 596–606.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of IJCAI*, pages 1085–1090.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Modeling bilingual conversational characteristics for neural chat translation. In *Proceedings of ACL*, pages 5711–5724.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of EMNLP*, pages 3265–3277.

Yanjun Ma, Yifan He, Andy Way, and Josef van Genabith. 2011. Consistent translation using discriminative learning-a translation memory-inspired approach. In *Proceedings of ACL*, pages 1239–1248.

Valentin Mace and Christophe Servan. 2019. Using whole document context in neural machine translation. In *Proceedings of IWSLT*.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of ACL*, pages 1275–1284.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of NAACL*, pages 3092–3102.

I Dan Melamed. 1997. Measuring semantic entropy. In *Proceeding of SIGLEX Workshop on Tagging Text with Lexical Semantics*, pages 41–46.

Magnus Merkel. 1996. Consistency and variation in technical translation: a study of translators ' attitudes. In *Proceedings of Unity in Diversity, Translation Studies Conference*, pages 137–149.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of EMNLP*, pages 2947–2954.

Xiao Pu, Laura Mascarell, and Andrei Popescu-Belis. 2017. Consistent translation of repeated nouns using syntactic and semantic cues. In *Proceedings of EACL*, pages 948–957.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725.

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 ACL*, pages 3537–3548.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of EMNLP-IJCNLP*, pages 1576–1585.

Jörg Tiedemann. 2010a. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15.

Jörg Tiedemann. 2010b. To cache or not to cache? experiments with adaptive models in statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 189–194.

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.

Ferhan Türe, Douglas W Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of NAACL*, pages 417–426.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of EMNLP-IJCNLP*, pages 877–886.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of ACL*, pages 1264–1274.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of EMNLP*, pages 2826–2831.

Tianming Wang and Xiaojun Wan. 2019. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In *Proceedings of IJCAI*, pages 5233–5239.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R é mi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*, pages 38–45.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. *Machine Translation Summit XIII*, 13:131–138.

Deyi Xiong, Yang Ding, Min Zhang, and Chew Lim Tan. 2013. Lexical chain based cohesion models for document-level statistical machine translation. In *Proceedings of EMNLP*, pages 1563–1573.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of AAAI*, pages 7338–7345.

Hongfei Xu, Deyi Xiong, Josef Van Genabith, and Qiuhui Liu. 2021. Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *Proceedings of IJCAI*, pages 3933–3940.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proceedings of EMNLP-IJCNLP*, pages 1527–1537.

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 EMNLP*, pages 533–542.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Toward making the most of context in neural machine translation. In *Proceedings of IJCAI*, pages 3983–3989.

A Experimental Datasets and Preprocessing

For ZH→EN(News), the training set consists of 41,341 documents with 0.8M sentence pairs. In addition, we use a larger sentence-level training set with 2M sentence pairs (including the 0.8M from the above document parallel training set) for pre-training to build a strong baseline. The sentence-level training set consists of LDC2002E18, LDC2003E07, LDC2003E14, news part of LDC2004T08 and the document-level training set from LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, LDC2010T03. The development (/test) set contains 79 (/509) documents with 1,649 (/5,146) sentence pairs.

For ZH→EN(TED), the training set consists of 3124 documents and 0.3M sentence pairs. The development (test) set contains 8 (/56) documents with 887 (/5,232) sentence pairs.

For FR→EN(TED), the training set consists of 3124 documents and 0.2M sentence pairs. The development (test) set contains 8 (/46) documents with 887 (/4,632) sentence pairs.

For all translation tasks, the English and French sentences are tokenized and lowercased by Mose toolkit (Koehn et al., 2007) [9] while the Chinese sentences are segmented by Jieba.[10] For ZH→EN (News) and ZH/FR→EN (TED), we segment the source and target sentences into sub-words by a BPE model with 32K and 21K merged operations (Sennrich et al., 2016), respectively.

We split long documents in training datasets into sub-documents with at most 20 sentences for efficient training. When constructing lexical chains, we use the most-frequency 1,000 source words in corresponding training set as the stop-word list and only consider words that are not in the stop-word list and appear two or more times in a document. Besides, the construction of the lexical chains is done before applying BPE operation. When applying BPE operation, a lexical chain could be split into multiple chains if the corresponding word is split into multiple sub-words. For example, the lexical chains of 科学家 and 科学家们 will be split into five different lexical chains, i.e., the chains of 科学_, 家, 科学_, 家_ and 们, since the source words 科学家 and 科学家们 are segmented into [科学_, 家] and [科学_, 家_, 们], respectively. That is to say, the two lexical chains of 科学_ are *not* merged since they are from different words 科学家 and 科学家们. Table 5 presents statistics about the lexical chains. It shows that the proportion of source words, the average length and the average number of the lexical chains vary across different translation tasks.

[9] https://github.com/moses-smt/mosesdecoder

[10] https://github.com/fxsjy/jieba

B Model Setting and Training

For all translation models, the hidden size and the filter size are set to 512 and 2048, respectively. The number of heads in multi-head attention is set to 8. For models on ZH→EN (News and TED), the numbers of layers in the encoder and the decoder are set 6, while for FR→EN (TED), we change the numbers to 4. For models on ZH→ EN (News) under + *2M Pretrain* setting, we set the dropout to 0.1. For other models, we set the dropout to 0.3.

In the first training stage, we train the sentence-level modules for 200K steps, warm-up steps as 8K, learning rate as 1.0 while in the second training stage, we continue to train all modules for 50K steps, learning rate 0.5. The weight of lexcial-consistency loss α gradually increases from 0 to 1.0 over the first 20K steps. We train all models on eight V100 GPUs with batch-size 4096 and use Adam with $\beta 1 = 0.9$, $\beta 2 = 0.98$ for optimization (Kingma and Ba, 2015).

C Model Parameters and Training Speed

As shown in Section 3, we use *General-Context Attention, Lexical-Consistency Attention* and *Latent Variational Module* to model consistency preference. Next we analyze the number of parameters and training speed of these various models.

Table 7 shows the number of parameters and training speed of these models. In total our ap-

Set	ZH→EN(News)			ZH→EN(TED)			FR→EN (TED)		
	Perc.	Avg. K	Avg. M	Perc.	Avg. K	Avg. M	Perc.	Avg. K	Avg. M
Training	19.25	3.04	22.32	15.57	2.79	9.81	14.23	2.47	9.99
Development	20.21	3.33	34.01	16.61	3.22	202.75	14.12	2.84	121.51
Test	21.12	3.87	22.79	17.11	3.16	107.91	15.27	3.01	87.97

Table 5: Statistics of the lexical chains on the training, development and test sets. *Perc.* indicates the percentage of words in lexical chains against all source-side words. *Avg. K* and *Avg. M* denote the average length of chains and the average number of chains in a document, respectively. Note *Perc.*, *Avg. K* and *Avg. M* on training sets are counted over sub-documents.

Model	ZH→ EN(News)	ZH→EN(News, *2M Pre.*)	ZH→EN (TED)	FR→EN(TED)
Transformer	70.67	74.97	71.30	87.87
+LC-Attn	72.90	75.72	76.81	89.35
+LC-Attn +ConVar	75.83	**76.29**	78.60	**90.76**
+LC-Attn +ConVar +GC-Attn	**76.15**	76.17	**79.04**	90.35
Previous context-aware NMT models				
HAN(Miculicich et al., 2018)	71.02	74.76	71.77	87.71
G-Trans(Bao et al., 2021)	72.77	75.74	73.77	88.90
W-Link(Lyu et al., 2021)	75.01	76.11	77.35	89.82

Table 6: HHI scores on test sets of ZH→EN (News), ZH→EN(TED) and FR→EN(TED) translation tasks.

Model	#Param.	Speed
Transformer	69.71M	11.31K
+LC-Attn	76.13M	9.59K
+LC-Attn + ConVar	79.23M	6.73K
+LC-Attn + ConVar + GC-Attn	85.56M	5.65K

Table 7: Comparison of the number of parameters and training speed among different models on ZH→EN News. **#Param.** denotes the number of parameters in millions. **Speed** denotes the training speed measured in source words per second.

Model	s-BLEU	d-BLEU	LTCR
Our model	43.27	45.39	64.56
w/o GC-Attn	42.54	44.81	64.09
w/o LC-Attn	42.46	44.79	62.78
w/o ConVar	43.17	45.33	61.99
Transformer	40.55	43.11	56.87

Table 8: Ablation study on ZH→EN (News) translation task.

proach (+LC-Attn +ConVar +GC-Attn) introduces 22.5% additional parameters and encumbers the training speed down 50.1% compared to sentence-level Transformer. It also shows that though the latent variational module (+ConVar) slightly increases the number of parameters, it significantly lowers the training speed. This is reasonable since the computation of post distribution is time-consuming in training (as shown in Figure 2).

D Performance on HHI score

Beside LTCR, we also evaluate the lexical translation consistency by Herfindahl-Hirschman Index (HHI) score (Melamed, 1997; Itagaki et al., 2007; Guillou, 2013) which is commonly accepted as measurement of market concentration. Specifically, for a lexical chain S with size K, we assume that the chain has n various translations. Then we compute HHI score of the chain as following:

$$\text{HHI}(S) = \sum_{i}^{n} R_i^2, \tag{25}$$

where R_i is the ratio of the i-th translation against the total number of translations (i.e., K). Finally, given a corpus with N lexical chains, we compute corpus-level HHI score as:

$$\text{HHI} = \sum_{j}^{N} \frac{|S_j|}{\sum_i^N |S_i|} \text{HHI}(S_j), \tag{26}$$

where $|S|$ returns the size of chain S.

Table 6 lists HHI scores on test sets of different translation tasks. We observe the similar trend as that of LTCR score while our systems (+LC-Attn +ConVar or +LC-Attn +ConVar +GC-Attn) achieve the best performance in HHI on all translation tasks.

E Ablation Study

Table 8 shows the performance of ablation study on ZH→EN (News) translation task. This further confirms our conclusions in Section 4.2. First, the general-context attention is effective to improve translation performance in BLEU while it has limited effect in LTCR. In contrast, incorporating the

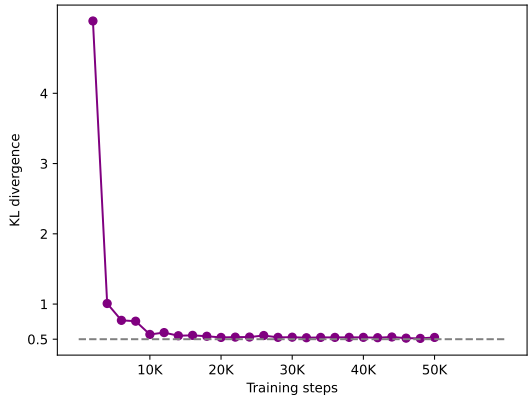

Figure 7: KL divergence (per chain) of consistency latent variable on development set.

latent variables contributes more to the improvement of LTCR while it has negligible effect in BLEU. Finally, lexical-consistency attention plays important role in improving both BLEU and LTCR.

F KL divergence

Following Liang et al. (2021), we analyze whether our CVAE module works well for modeling consistency preference. Figure 7 shows that the KL divergence of consistency latent variables maintains around 0.5 during the second training stage, which indicates that the degeneration problem of variational framework does not appear in our model. The consistency-tailored latent variable learned by our CVAE module plays its corresponding role.

G Case Study

We use two examples to illustrate how our proposed approach helps translation. As shown in Figure 8, we observe that in the first example, the sentence-level model may confuse readers by translating source word 书画/*shu_hua* into three different translations, i.e., *painting and calligraphy*, *painting*, and *writing and painting*. In contrast, our approach consistently and correctly translate it into *painting and calligraphy*. In the second example, the source word 佩里斯/*pei_li_si* is consistently translated into *belize* by our model while it is translated into two different translations, i.e., *petris* and *belize*, by the sentence-level model.

We note that over-corrected cases would be caused by our model. As shown in Figure 9, the source word 信任投票/xin_ren_tou_piao is repeatedly translated into *the trust voting* by our model while it is translate into *the trust vote* and *the vote of confidence* in both the reference and the sentence-level translation, respectively.

Source	**<#1>** 庆 香港 回归 5 周年 公务员 书画/*shu_hua* 大赛 将 举行 **<#2>** ... 公务员 —— 人民 的 公仆 " 为 创作 主题 的 公务员 书画/*shu_hua* 大赛 ， 将 在 港 、 澳 和 内地 公务员 中 推开 。 **<#5>** 大赛 将 突出 香港 回归 5 周年 的 喜庆 气氛 ， 充分 展示 中国 书画/*shu_hua* 的 艺术 魅力 。
Reference	**<#1>** painting and calligraphy competition of government employees ... . **<#2>** ... painting and calligraphy competition of government employees on the theme " ... . **<#5>** ... china and fully demonstrate the artistic charm of the chinese painting and calligraphy .
SentNMT	**<#1>** civil servants ' painting and calligraphy competition ... . **<#2>** ... a civil service painting competition with the theme of " state civil servants ... **<#5>** ... region (hksar) and demonstrate the artistic beauty of chinese writing and painting .
Ours	**<#1>** civil service painting and calligraphy competition to celebrate ... . **<#2>** ... a civil service painting and calligraphy competition with the theme of " national ... . **<#5>** ... reunification and fully demonstrate the artistic charm of chinese painting and calligraphy .

Source	**<#3>** 斯里兰卡 司法 、 宪法 、 种族 事务 和 民族 一体化部 部长 佩里斯/*pei_li_si* 在 当天 举行 的 记者 招待会 上 说 ， 斯政府 和 猛虎 ... 。 **<#5>** 佩里斯/*pei_li_si* 说 ， 政府 打算 在 亚洲 某个 国家 与 猛虎 ... 。
Reference	**<#3>** peiris , minister of sri lankan legislature , constitution and ethnic integration , said at a press conference ... . **<#5>** peiris said the government intends to hold talks ... . 。
SentNMT	**<#3>** ... racial affairs and ethnic onslaught minister petris said at a press conference on the same day that ... . **<#5>** belize said the government planned to hold talks ... .
Ours	**<#3>** ... racial affairs and national anatomy minister belize said at a press conference on the same day that ... . **<#5>** belize said the government planned to hold talks ... .

Figure 8: Examples of document-level ZH→EN (News) translation from the test set.

Source	**<#1>** 巴基斯坦 总统 穆夏拉夫 赢得 参众两院/*can_zhong_liang_yuan* 信任投票/*xin_ren_tou_piao* **<#2>** ... 及 反对党 抵制 的 情况 下 ， 赢得 国会 参众两院/*can_zhong_liang_yuan* 的 信任投票/*xin_ren_tou_piao* 。
Reference	**<#1>** pakistani president musharraf won the trust vote in senate and lower house . **<#2>** ... won the vote of confidence in both senate and lower house that legitimizes his ruling until 2007 . ".
SentNMT	**<#1>** pakistani president musharraf wins the trust voting by the senate and houses **<#2>** ... won the vote of confidence in the house and houses of congress, amid the abandonment ... .
Ours	**<#1>** pakistani president musharraf won the trust voting of the senate and house **<#2>** ... won the trust voting of the senate and house of congress under the abandoned ... .

Figure 9: An example of over-corrected consistency on document-level ZH→EN (News) translation from the test set.