Enhancing Lexical Translation Consistency for Document-Level Neural Machine Translation

XIAOMIAN KANG, YANG ZHAO, and JIAJUN ZHANG, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and School of Artificial Intelligence, University of Chinese Academy of Sciences
CHENGQING ZONG, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, CAS Center for Excellence in Brain Science and Intelligence Technology, and School of Artificial Intelligence, University of Chinese Academy of Sciences

Document-level neural machine translation (DocNMT) has yielded attractive improvements. In this article, we systematically analyze the discourse phenomena in Chinese-to-English translation, and focus on the most obvious ones, namely lexical translation consistency. To alleviate the lexical inconsistency, we propose an effective approach that is aware of the words which need to be translated consistently and constrains the model to produce more consistent translations. Specifically, we first introduce a global context extractor to extract the document context and consistency context, respectively. Then, the two types of global context are integrated into a encoder enhancer and a decoder enhancer to improve the lexical translation consistency. We create a test set to evaluate the lexical consistency automatically. Experiments demonstrate that our approach can significantly alleviate the lexical translation inconsistency. In addition, our approach can also substantially improve the translation quality compared to sentence-level Transformer.

CCS Concepts: • **Computing methodologies** → **Machine translation;**

Additional Key Words and Phrases: Document-level translation, neural machine translation, lexical consistency, discourse phenomena

ACM Reference format:
Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2021. Enhancing Lexical Translation Consistency for Document-Level Neural Machine Translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 21, 3, Article 59 (December 2021), 21 pages.
https://doi.org/10.1145/3485469

1 INTRODUCTION

During the last few years, **neural machine translation** (**NMT**) has achieved remarkable progress and become the de facto standard paradigm of machine translation. A variety of effective NMT

The research work has been supported by the Natural Science Foundation of China under Grant No. 62006224.

Authors' addresses: X. Kang, Y. Zhao, and J. Zhang, Intelligence Building, No. 95, Zhongguancun East Road, Haidian District, 100190 Beijing, P. R. China; emails: {xiaomian.kang, yang.zhao, jiajun.zhang}@nlpr.ia.ac.cn; C. Zong, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, CAS Center for Excellence in Brain Science and Intelligence Technology, and School of Artificial Intelligence, University of Chinese Academy of Sciences, P.R. China; email: cqzong@nlpr.ia.ac.cn.

2375-4699/2021/12-ART59 $15.00
https://doi.org/10.1145/3485469

Table 1. An Example of Inconsistent Translations Indicated with the Underlines

Source	*<S1>* 证监会 指出 , 修订 后 的 准则 内容 涵盖 公司 治理 的 基本 原则 。 *<S2>* 证监会 指出 , 将 根据 新 准则 完善 相关 规章 。
Reference	*<S1>* The CSRC points out that the revised Code covers the basic principles of corporate governance. *<S2>* The CSRC also points out that it will comply with the new Code to improve relevant regulations.
SentNmt	*<S1>* The commission pointed out that the revised guidelines cover the basic principles of corporate governance. *<S2>* The CSRC said it would improve the relevant rules and regulations under the new standards.

methods underlying the encoder-decoder framework have been proposed to improve sentence-level translation quality due to the powerful end-to-end modeling [1, 4, 20, 36, 52, 54]. However, when fed an entire document, standard NMT systems have to translate sentences in isolation without considering the cross-sentence dependencies. Consequently, **document-level neural machine translation** (**DocNMT**) methods are explored to utilize inter-sentence contextual information to improve performance over sentences in a document [11, 12, 14, 41].

When translating a document, NMT systems have to handle discourse phenomena between sentences to generate more coherent translations. It is a new challenge that sentence-level translation does not need to face. It is widely recognized that different languages and genres present different discourse phenomena. Therefore, we carefully analyze the performance of NMT on discourse phenomena in three genres of Chinese-to-English translation tasks in Section 2, which, to the best of our knowledge, has not been systematically studied. We find that the most obvious phenomenon in Chinese-to-English translation is **lexical translation consistency**, which means that the repeated source words prefer to share the same target translations in the document [44]. Table 1 shows an inconsistent example containing two sentences in Chinese-to-English translation. A **sentence-level NMT** (**SentNmt** [36]) system translates the same named entity "证监会" in both sentences into "CSRC" and "commission", respectively. And the repeated notional word "准则 (Code)" is translated into two different words "guidelines" and "standards", which harms the discourse cohesion seriously.

In **statistical machine translation** (**SMT**), various approaches are proposed to encourage lexical translation consistency [7, 22, 28, 44]. However, it is seldom studied in NMT. Although existing DocNMT methods can alleviate the translation inconsistency through introducing cross-sentence contextual information, the problem is still serious. Majority of DocNMT methods mainly focus on the design of novel neural networks so that cross-sentence contextual information from different sources can be leveraged effectively [19, 21, 24–26, 31, 34, 37, 39, 48, 49, 51]. However, all contextual words are utilized in the same pattern. More specifically, existing DocNMT methods do not distinguish between repeated words and other words. As a result, they are not sensitive enough to the translation of repeated words that are usually regarded as the triggers for consistent translation [37, 44]. Recently, some researchers begin to focus on the evaluation and solution of discourse phenomena [2, 13, 38, 43]. Voita et al. [38] analyze the English-to-Russian subtitles dataset and propose a model that utilizes two-pass decoding to modify the results of sentence-level translation. The method improves the performance on four types of discourse phenomena including the translation consistency of entities. However, contextual words are still treated indiscriminately, and there are no explicit constraints on translation consistency. Meanwhile, the cases requiring consistent translations in Chinese-to-English tasks are more obvious and complex, not just limited to the translation of entities.

In this article, we propose an effective approach to enhance lexical translation consistency for the document-level translation (Section 3). We aim to explicitly provide consistent information from the global context so that the model can pay attention to repeated words and generate more consistent translations. Specifically, the approach consists of two modules that can be independent of the translation framework. First, a *Global Context Extractor* extracts two types of global context: document context and consistency context from all repeated source words. Then, *Consistency Enhancers* incorporate the global context into the output of encoder and decoder layers to enhance the lexical translation consistency.

We make the following contributions:

- We analyze the discourse phenomena of Chinese-to-English document-level translation in different genres. And statistics show that lexical translation consistency is the most obvious phenomenon.
- We explicitly model the lexical translation consistency for NMT. Repeated words and other words are treated differently, and two types of global context are utilized to provide a global and consistent constraint for each sentence. Independent of the translation network, our approach is easily adaptable to existing DocNMT models.
- We create a test set to evaluate the lexical translation consistency automatically. Experiments show that our approach can effectively alleviate the lexical translation inconsistency and perform much better than existing DocNMT models. It can also significantly improve translation quality over the Transformer.

2 OBSERVATION

We conduct a human study on the discourse phenomena to clarify the problems in Chinese-to-English DocNMT. We compare the distribution of discourse phenomena in three different genres: news, talks, and subtitles. The language in news is formal, while the style of subtitles is considered to be colloquial. The style of talks is somewhere in between.

Specifically, we train a sentence-level Transformer NMT model [36].[1] To analyze the performance of SentNmt, we randomly select 100 paragraphs for each genre.[2] Annotators are simultaneously shown the source-side paragraphs and their translations generated by NMT. Then we ask them to read the translations sentence-by-sentence. If a sentence is not appropriate, they are asked to refer to the standard answer and determine whether the errors can be resolved within the current sentence. We only focus on errors that have to be corrected with the aid of some contextual sentences.

2.1 Types of Discourse Phenomena

The results of manual analysis are shown in Table 2. For news and talks, lexical and tense consistency account for a major proportion. News tends to report objective events, where repeated entities and facts are frequently mentioned and usually require consistent translations. Tense inconsistency mainly occurs in the translation of declarative sentences, which are difficult to distinguish past and present tenses without tenses of contextual sentences. As a contrast, for colloquial

[1] In order to obtain good sentence-level translations, we mixed the training data of News, TED Talks, and Subtitles together to alleviate the lack of data. Experiments show that the translation quality is far superior to models trained on individual genre-specific datasets. The details of training and data processing are described in Section 5.

[2] We treat each paragraph as a document in this article. Analyzed paragraphs come from the public test set *newstest2019* in WMT2019 for news, and *tst-2014~2015* in IWSLT2017 for TED talks. For subtitles, we extract the analysis set from the original training dataset provided by Wang et al. [40]. The selected sentences will not be used in training.

Table 2. The Proportion of Different Types of Discourse Phenomena in Chinese-to-English Translations

Type	lexical consistency	tense consistency	pronoun translation	connective	ellipsis	ambiguity	other
News	43.9%	24.5%	9.2%	4.6%	6.9%	5.4%	5.5%
TED	35.4%	27.0%	8.7%	10.6%	7.3%	6.1%	4.9%
Subtitles	21.9%	17.8%	19.6%	3.1%	16.5%	11.5%	9.6%

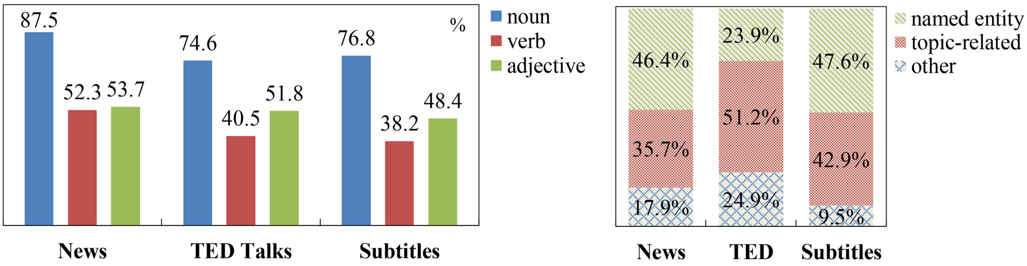

Fig. 1. The types of lexical translation consistency. Left: the proportion of consistent translations in repeated source words with different part-of-speech tags. Right: the proportion of different types in repeated nouns translated consistently.

subtitles, the pronoun translation, ellipsis, and ambiguity are significantly more obvious. In Chinese, a pro-drop language, zero anaphora exists widely and confuses the choice of pronouns.

Although different genres have different distribution of discourse phenomena, it can be found that lexical inconsistency is always one of the most serious issues in Chinese-to-English translations.

2.2 Lexical Translation Consistency

We further analyze the cases of lexical translation consistency in the real human references, which will guide us to design methods to alleviate serious lexical inconsistency. We analyze the references of above selected paragraphs for each genre.

2.2.1 Trigger of Translation Consistency. Intuitively, the repeated source words are more likely to be translated into the same. However, one possible issue is that some non-repeated source words (usually pronoun or coreference) may also be forced to translate into the same. Fortunately, we find that the cases are rare, which are less than 15.2% in all cases of target-side consistency. Therefore, we focus on most cases where the repeated target words are translated from repeated source words. We regard the repeated source words as the triggers of translation consistency, which can be obtained easily by character matching.

In addition, for repeated phrases consisting of multiple words, considering the word-by-word translation paradigm of NMT, we regard the multiple words in one phrase as different words. And as long as one of the words is translated differently, the target phrase is inconsistent.

2.2.2 Types of Consistency. Inspired by Guillou [9], we discuss the translation consistency of repeated source words with three different part-of-speech tags: nouns, verbs, and adjectives. Figure 1 Left shows the proportion of consistent translations (e.g., proportion = #repeated at both source- and target-side / #repeated at source-side). It can be found that repeated nouns are more like to be translated consistently than verbs and adjectives. The repeated verbs and adjectives translated consistently are just over or less than half. Their translations are so flexible that it is difficult to

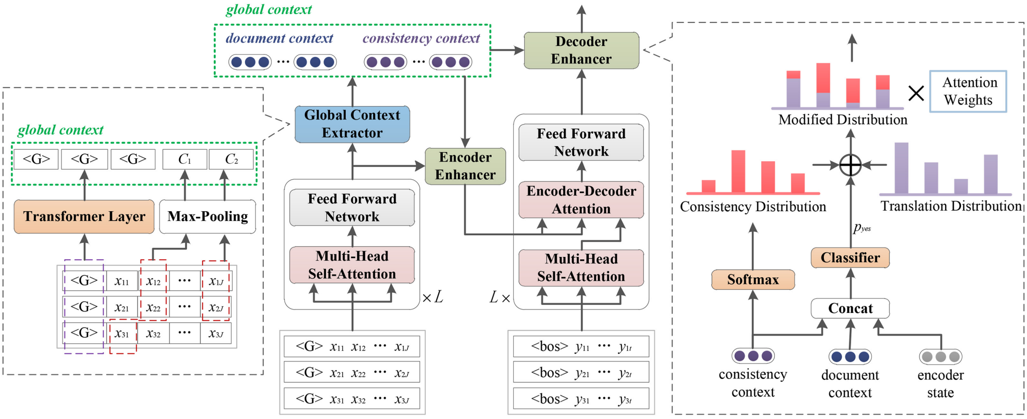

Fig. 2. Overview of our approach. Sentences in a document are translated in parallel. The global document context vectors and consistency context vectors are obtained by Global Context Extractor (whose details are shown in the left dotted box) on the top of encoder blocks. Then, extracted context is utilized to modify the encoder states by Encoder Enhancer, and the prediction probability distribution by Decoder Enhancer ((whose details are shown in the right dotted box), respectively.

determine the consistency. In contrast, the repeated nouns translated consistently are more than 75%. In news, a formal genre, the proportion is more obvious (about 87.5%). Inconsistent instances are usually due to the fact that some repeated source-side nouns are translated into coreferential words with different forms.

In the nouns translated consistently, the named entities are common. As shown in Figure 1 Right, consistent named entities is about 46.4% in news. One reason for the large proportion of entities in subtitles is that the names of characters in movies are often called. Besides, there are many consistent general nouns related to the topic (which may be important to construct lexical chains). The cases are more than 50% in TED Talks.

In conclusion, the lexical translation consistency is the most significant phenomenon in Chinese-to-English document-level translations. Most of the repeated nouns are translated into the same, while verbs and adjectives are not. It inspires us to focus on the consistent translation of nouns.

3 APPROACH

To alleviate the lexical translation inconsistency, our approach consists of two modules. *Global Context Extractor* extracts two types of global context (Section 3.1). The document context vectors are obtained by special tokens, and the consistency context vectors are generated from individual encoder states of repeated words[3] in the document. *Consistency Enhancers* utilize the extracted global vectors to enhance the translation consistency at the outputs of encoder and decoder blocks (Section 3.2). In the decoder enhancer, we learn a consistency classifier to determine whether a repeated source word should be translated consistently. Our approach translates sentences in parallel. The overall architecture is shown in Figure 2.

3.1 Global Context Extractor

We extract two types of global context: document context and consistency context.

[3] Actually, the proposed approach itself does not distinguish the part-of-speech in repeated words, so we use “repeated words” when introducing the method in the following. But in fact, as suggested in the conclusion of Section 2.2, we enhance and evaluate the translation consistency of “repeated nouns” in our experiments.

Suppose that there are I sentences and N different repeated words in a document. The document context is a set of context-aware vectors $\mathcal{V} = \{v_1, \ldots, v_I\}$, which provides the document-level contextual information for each sentence to improve translation quality. The generation of each document context vector v_i does not distinguish whether words are repeated or not. Inspired by the popular pre-training language models such as GPT [29] and BERT [5], we add a special symbol "$\langle G \rangle$" at the beginning of each sentences. We believe the symbol can encode its sentence-level information well by the self-attention mechanism. After encoding, each hidden state of "$\langle G \rangle$" is extracted as the input to a Transformer layer (containing a multi-head self-attention sub-layer and feed forward network sub-layer) to model the dependencies among sentences in the document. Therefore, for each sentence, we can obtain a corresponding document context $v_i \in \mathbb{R}^d$, where d indicates the hidden size.

The consistency context is a set of N global consistency-aware vectors $\mathcal{U} = \{u_1, \ldots, u_N\}$. For each repeated word, the extractor collects all individual encoder states belonging to the word from all sentences to generate a global consistency context vector. Specifically, a token $x_{i,j}$ in sentence $X_i = \{x_{i,1}, \ldots, x_{i,J}\}$ is encoded into an individual state $h_{i,j} \in \mathbb{R}^d$. If $x_{i,j}$ belongs to the nth repeated word, the extractor extracts the state $h_{i,j}$ of $x_{i,j}$ into the corresponding set $\mathcal{H}_n$ that stores all the states of words belonging to the nth repeated word in the entire document. Then, we generate a unique global consistency context vector $u_n \in \mathbb{R}^d$ for words belonging to the nth repeated word as follows:

$$u_n = \text{Maxpooling}\left(\mathcal{H}_n\right), \tag{1}$$

where the element-wise max-pooling operation takes all states of words belonging to the nth repeated word as inputs and outputs a vector. The input size is variable.

3.2 Consistency Enhancer

We integrate the extracted global context vectors into the standard NMT model for two purposes. First, the source repeated words should know the information of each other in the encoding process. Second, the decoding probability distribution should be encouraged to be similar when translating the same repeated word. Therefore, we design two types of consistency enhancers to integrate the global context vectors to modify the encoder states at encoder-side and the prediction probability distribution at decoder-side, respectively.

3.2.1 Encoder Enhancer. The encoder enhancer integrates the document context and the consistency context into the encoder states of source words. We arrange our encoder enhancer on the top of standard sentence-level encoder blocks underlying Transformer framework. For a word $x_{i,j}$, whose encoder state is denoted by $h_{i,j}$, there are two cases:

(1) If $x_{i,j}$ is a repeated word, we define the group number of repeated words to which $x_{i,j}$ belongs as $g_{i,j}$, where $1 \leq g_{i,j} \leq N$. Then, $h_{i,j}$ is integrated with the corresponding consistency context vector $u_{g_{i,j}}$ via a gated sum operation as follows:

$$r_{i,j} = \sigma(\text{FNN}([u_{g_{i,j}}; h_{i,j}])), \tag{2}$$

$$\tilde{h}_{i,j} = r_{i,j} \odot u_{g_{i,j}} + (1 - r_{i,j}) \odot h_{i,j}, \tag{3}$$

where FNN $(\cdot)$ denotes a feed forward network. σ stands for the sigmoid function. $[\cdot;\cdot]$ concatenates elements into a vector. The gate weight $r_{i,j}$ balances the sentence-level individual encoder state and the document-level shared consistency context.

(2) If $x_{i,j}$ is a non-repeated word, its encoder state is integrated with the corresponding document context v_i instead of the consistency context $u_{g_{i,j}}$ under the same network defined by Equation (2) and Equation (3).

The special symbol "$\langle G \rangle$" directly copies v_i as its final encoder state.

3.2.2 Decoder Enhancer. The decoder enhancer aims to make the probability distribution of the translations more similar when translating the words that should be translated consistently. However, different from the encoder enhancer that has been informed of the repeated words in source sentences, the decoder enhancer faces two issues. (1) Not all repeated source words need to be translated consistently. (2) When decoding a target word $y_{i,t}$ at time t, it is unknown whether $y_{i,t}$ is translated from a repeated source word, or which repeated word should be translated.

Therefore, we leverage a consistency classifier and encoder-decoder attention weights to alleviate the two issues, respectively. Specifically, the final prediction probability distribution is computed in three steps. It is noted that the calculation of the first two steps is independent of the current decoding state, so they can be immediately performed after the global document context and consistency context are extracted.

Step1. For the nth repeated word, we generate a consistency probability distribution $\mathcal{P}_n^{CON}$ using the consistency context vector u_n as follows:

$$\mathcal{P}_n^{CON} = \text{softmax}\left(W\left(W_1\ u_n\right) + b\right), \tag{4}$$

where W and b are learnable parameters and we share them with the original NMT model. $W_1 \in \mathbb{R}^{d \times d}$ is a transfer matrix. The $\mathcal{P}_n^{CON}$ is supposed to constrain the original translation probability distribution $\mathcal{P}_{i,t}^{NMT}$ at decoding time t.

Step2. We use a **consistency classifier** to estimate the consistent probability as the confidence of a repeated source word being translated consistently. We define it as a binary classification task. Therefore, for a word $x_{i,j}$, there are two cases:

(1) If $x_{i,j}$ is a repeated word whose group number is $g_{i,j}$, its confidence to be translated consistently is calculated by a two-layers perceptron as follows:

$$s_{i,j} = \sigma(W_3(W_2[u_{g_{i,j}}; v_i; h_{i,j}] + b_2) + b_3), \tag{5}$$

where $W_2 \in \mathbb{R}^{3d \times d}$, $b_2 \in \mathbb{R}^d$, $W_3 \in \mathbb{R}^{d \times 1}$, and $b_3 \in \mathbb{R}^1$ are model parameters.

(2) If $x_{i,j}$ is a non-repeated word, the confidence is defined as $s_{i,j} = 0$.

Step3. We calculate the final prediction probability distribution with the aid of the encoder-decoder attention weights that bridge the current target word $y_{i,t}$ and the source words in $X_i = \{x_{i,1}, \ldots, x_{i,J}\}$. We average the encoder-decoder multi-head attention over heads and layers. Then the averaged attention weights are fed into a softmax function to output the normalized attention weight vector $A_{i,t} = \{a_{i,t,1}, \ldots, a_{i,t,J}\}$, $A_{i,t} \in \mathbb{R}^J$. We denote the jth element as $a_{i,t,j}$, which measures the contribution of the source word $x_{i,j}$ to the generation of target word $y_{i,t}$.

The final probability distribution at time t is calculated by:

$$\mathcal{P}_{i,t} = \sum_{j=1}^{J} a_{i,t,j} * \left[s_{i,j} * \mathcal{P}_{g_{i,j}}^{CON} + (1 - s_{i,j}) * \mathcal{P}_{i,t}^{NMT}\right]. \tag{6}$$

The following example explains the calculation of the final probability distribution. And the figure 3 shows the steps of the example.

Example. Suppose a document has two different repeated words. Therefore, we obtain two consistency distribution $\mathcal{P}_1^{CON}$ and $\mathcal{P}_2^{CON}$ at *step1*. A source sentence with five words is $\{w_1^1, w_2, w_3^2, w_4^1, w_5\}$, where w_1^1 and w_4^1 belong to the first repeated word, w_3^2 belongs to the second, and w_2 and w_5 are non-repeated words. At *step2*, we can obtain consistent confidence of each word, assuming that they are $[0.45, 0, 0.65, 0.78, 0]$, respectively. For time t, the averaged encoder-decoder attention weight $A_t = [0.15, 0.3, 0.25, 0.1, 0.2]$, and the original probability distribution is

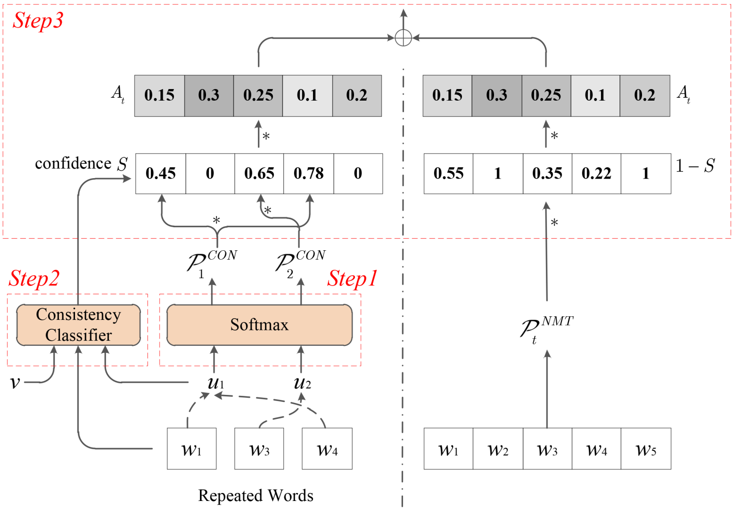

Fig. 3. The calculation steps of the example used to explain the decoder enhancer.

$\mathcal{P}_t^{NMT}$. Therefore, at *step3*, the final distribution $\mathcal{P}_t = (0.15 * 0.45 + 0.1 * 0.78) * \mathcal{P}_1^{CON} + (0.25 * 0.65) * \mathcal{P}_2^{CON} + (0.15 * 0.55 + 0.3 + 0.25 * 0.35 + 0.1 * 0.22 + 0.2) * \mathcal{P}_t^{NMT}$.

3.3 Training

Our approach translates sentences in a document in parallel, and the consistency context is extracted over the document. Therefore, when training, we shuffle the data over document. Our model is trained in two stages, which has proved effective [25, 26, 34]. First, a standard sentence-level Transformer is pre-trained to ensure a fine initialization of encoder-decoder attention weights. Then, we add our global context extractor and consistency enhancers to optimize parameters. The newly introduced consistency classifier is trained together with the original translation components. Suppose a document has I sentences and N different repeated words, our optimization goal is to minimize the negative following log-likelihood loss:

$$\mathcal{L} = -\sum_{i=1}^{I}\sum_{t=1}^{T} \log P\left(y_{i,t} \mid y_{i,<t}, X_i\right) - \sum_{n=1}^{N}\sum_{m=1}^{|\mathcal{H}_n|}\left[k_{n,m}\log s_{n,m} + (1-k_{n,m})\log(1-s_{n,m})\right], \quad (7)$$

where $k_{n,m}$ is the golden label of the mth word that belongs to the nth group of repeated words whose size is $|\mathcal{H}_n|$. If its translation is consistent, $k_{n,m} = 1$. Otherwise, $k_{n,m} = 0$. Section 5.1.2 describes the automatic annotation process of $k_{n,m}$.

4 TEST SET

It is generally acknowledged that standard machine translation metrics (e.g., BLEU) are not sensitive enough to discourse phenomena [42]. Recently, some works create contrastive test sets to evaluate specific phenomena [2, 13, 38]. Each test instance consists of a positive and several negative translations with incorrect phenomena. Models are evaluated by the proportion of instances whose generation probability of positive translation is higher than negative ones. Voita et al. [38] construct contrastive test sets for English-to-Russian subtitles to evaluate four types of discourse phenomena including the translation consistency of named entities (what they call lexical cohesion in their article). However, the hand-crafted test sets may not carry over to practical scenarios [17].

Table 3. An Instance of Test Set Evaluating the Lexical Translation Consistency

Source	*<S1>* 这 是 一部 奇妙 的 多轨道 [电影]^1。 *<S2>*...... *<S3>* 这 部 [电影]^1 是 你 的 [意识]^2 流 ... *<S4>*...... *<S5>*...... *<S6>* 从此, 关于 [意识]^2 方面 的 科学 [研究]^3 ... *<S7>* [意识]^2 [研究]^3 的 核心 是 寻找 ... 特定 的 [意识]^2 状态 之 间 的 相关性。 *<S8>*......
Reference	*<S1>* It's an amazing multi-track [movie]^1. *<S2>*...... *<S3>* This [movie]^1 is your stream of [consciousness]^2 ... *<S4>*...... *<S5>*...... *<S6>* Since then, scientific work on [consciousness]^2 ... *<S7>* the centerpiece of [consciousness]^2 study has been the search ... correlations between certain states of [consciousness]^2. *<S8>*......
Consistent Instances	(1) *Trigger*: 电影; *Type*: general; *Candidate*: [movie, film, cine]; *Index*: [S1, S3] (2) *Trigger*: 意识; *Type*: general; *Candidate*: [consciousness, mentality]; *Index*: [S3, S6, S7, S7]

[Repeated words] and inconsistent ones are marked.

Table 4. Statistics of Consistent Instances in the Test Set for Lexical Translation Consistency

Genre	**#Para.**	**# Consistent Instance (CI)**				**CI Avg. len**
		Total	Entity	General	Para Avg. CI	
News	150	537	257	280	3.28	2.67
TED	150	317	112	205	2.11	2.84
Subtitles	150	188	95	93	1.25	2.13

Their target-side context has been already assumed, so the test sets cannot evaluate the quality of generated context sentences. Meanwhile, they cannot evaluate the real translation results. In practice, the generation of subsequent sequence is affected by previous generated words.

As a result, we pick a test set from the real data to evaluate the lexical consistency of practical translations. Table 3 shows a test instance. Each test instance is a paragraph pair contains several consistent instances annotated manually. The annotation process contains four steps: (1) We collect triggers of translation consistency, i.e., repeated source-side nouns. (2) We extract triggers translated consistently, and record corresponding sentence indexes. For a repeated source word, we check its target-side words. If lemmas of corresponding target words are the same, it is a consistent instance.[4] (3) We annotate the triggers' types: named entity or general words. (4) We expand the possible translations of the repeated source word. We extract the top-20 candidate translations of the repeated source word from the lexical table of Moses [18], and keep the lemmas of candidates having correct meaning. We create a test set that contains 150 paragraphs for each of the domains: News, TED Talks, and Subtitles. Table 4 shows the statistics of consistent instances.

When evaluating a result, we conduct a simple matching automatically. For each consistent instance in the tested paragraph, we use lemmas of generated sentences by NLTK toolkit. We check the generated sentences in the index list one-by-one, and extract the lemmas that belong to the candidate list. The number of lemmas extracted from each sentence must be equal to the number of times the sentence is indexed. For example, for the consistent instance triggered by "意识" in Table 3, "consciousness/mentality" has to appear twice in the translation of "*<S7>*". After that, if extracted lemmas are exactly the same, we believe this instance is translated consistency. The

[4] We do not consider the complex case where target words are partially repeated in this article.

Table 5. Statistics of the #sentence in Different Datasets

Dataset	Zh→En News	Zh→En TED	Zh→En Subtitles	En→De News	En→De TED	En→De Europarl
Training	0.31M	0.23M	2.14M	0.24M	0.21M	1.67M
Development	2.00K	0.88K	1.09K	2.17K	8.97K	3.59K
Test	3.98K	6.05K	1.15K	3.00K	2.27K	5.13K

M: million. K: thousand.

accuracy of consistent instances is utilized to evaluate the ability of models resolving inconsistent translation.

5 EXPERIMENTS

5.1 Data Preparation

5.1.1 Datasets. For Chinese-to-English (Zh→En) translation, we evaluated our approaches on three different genres. For the news genre, we used News-Commentary v14 provided by WMT19[5] for training. *newstest2017* and *newstest2018* were used for development and testing, respectively. For talks, we used TED Talks in IWSLT17.[6] We used *dev-2010* as the development set and *tst-2010~2013* as the test set, as Miculicich et al. [26] do. For subtitles, the dataset was collected by Wang et al. [40] from the subtitles of television episodes.[7] We removed the sentences used to analyze discourse phenomena from the original training set.

For English-to-German (En→De) translation, we conducted experiments on the same datasets Maruf et al. [25].[8] Specifically, there were three datasets. TED Talks was also from the IWSLT17. We took *tst-2016~2017* as the test set, and others as the development set. For the news genre, News-Commentary v11 corpus was used for training. The *newstest2015* and *newstest2016* in WMT were used as the development set and test set, respectively. Europarl was extracted from Europarl v7 and split by the SPEAKER tag.

The corpora statistics are listed in Table 5. All above datasets are provided document boundaries. Considering the memory limitation, the original documents with more than 16 sentences were forced to split into paragraphs, and we treated each paragraph as one document in our experiments.

Chinese sentences were segmented into words by our in-house toolkit. English and German datasets were tokenized by the Moses toolkit.[9] Focusing on the consistency of nouns, we run Standford part-of-speech tagger [23] for source sentences and removed stop words to extract the repeated nouns. Words were segmented by byte-pair encoding with 30K merge operations [30]. It is noted that only the sub-words segmented from the repeated words are regarded as repeated words, and the same sub-words of different words are not shared. For example, the sub-word “No” in “Nosair” and the sub-word “No” in “Nordlund” belong to different repeated groups.

5.1.2 Annotation of Lexical Consistency. To train the consistency classifier, we automatically annotated the repeated source words translated consistently by the alignment tool [6]. Specifically, we removed stop words and run the alignment tool for the source-reference pairs. A repeated source word is assumed to be consistent if the frequency of repetition in its aligned target words is

[5] http://data.statmt.org/news-commentary/v14.
[6] https://wit3.fbk.eu/mt.php?release=2017-01-trnted.
[7] https://github.com/longyuewangdcu/tvsub.
[8] https://github.com/sameenmaruf/selective-attn/tree/master/data.
[9] https://github.com/moses-smt/mosesdecoder/tree/master/scripts.

equal to the frequency of repetition in source.[10] In this way, we can annotate the repeated words that are translated consistently.

5.2 Baselines and Details

We compared our approach with following methods.

- **SentNmt** [36] is a standard Transformer model with the "base" version parameters.
- **Cache** [34] is a model utilizing the translation history. The model read the states of fixed-size generated words stored in a cache. Then, the weighted state is used to modify the decoder state in RNN framework. We re-implemented the method on Transformer. The cache size was set to 25 words suggested in their article.
- **DocT** [51] encodes previous context sentences through an extra encoder, and introduces contextual information into each encoder and decoder Transformer layers.
- **HAN** [26] adds a hierarchical attention network on the top of the last encoder and decoder layer to model sentence-level and word-level information in previous sentences. We adopted the "HAN encoder + HAN decoder" strategy that achieved the best performance.
- **SAN** [25] calculates the weights of sentence-level and word-level context hierarchically. When calculating the attention weights, it utilizes the sparsemax function instead of softmax to focus on relevant sentences. We choose the "offline" model that use the context of entire document to integrate into the encoder with the "sparse-soft H-Attention". The layers of encoder and decoder in their article were set to 4 but we set 6 in our experiments for a fair comparison with other models.
- **MmcNmt** [53] encodes each source sentences independently and integrates the source-side context at the top of encoder. It translates a document sentence-by-sentence with Transformer-XL net to integrate the target-side history context.
- **X +** ***Our*** stands for the models with our approach as attachments. Our global context extractor and consistency enhancers are independent of the translation model. Therefore, it can be added into existing DocNMT models.

We implemented all our models based on the toolkit THUMT [50].[11] The parameters were the "base" version of the Transformer. Specifically, we used 6 layers of encoder and decoder with 8 attention heads. The hidden size h and feed-forward layer size were 512 and 2,048, respectively.

When training, we shuffled the data over paragraphs to ensure all sentences in a document were processed in parallel. The batch size was 3,000 tokens. We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We employed label smoothing with a value of 0.1 and dropout with a rate of 0.1. During inference, we used multi-bleu.perl[12] to compute the BLEU [27] score. The beam size was set to 4.

6 RESULTS

6.1 Translation Quality

We first study the impact of our approach on translation quality. Table 6 shows the average BLEU scores on test sets.[13] Towards improving lexical consistency that is not sensitive to BLEU, our approach should at least ensure that it will not negatively affect the translation quality.

[10] It is noted that we consider different target words with the same stem to be repetitive. And repeated source words with partially repeated translations are considered inconsistent.

[11] https://github.com/thumt/THUMT.

[12] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl.

[13] The significance test is conducted by the script "bootstrap-hypothesis-difference-significance.pl" in Moses.

Table 6. Performance of Our Approach and Baselines on BLEU (%)

Model	Zh→En			En→De		
	News	TED	Subtitles	News	TED	Europarl
SentNmt [36]	13.17	16.97	29.10	22.78	23.28	28.72
Cache [34]	13.45	17.32	29.34	23.39	23.71	29.25
DocT [51]	13.71	17.75	29.82	23.08	24.00	29.35
HAN [26]	13.89	17.90	30.02	25.03	24.58	29.58
SAN [25]	13.84	17.69	30.06	24.76	24.23	29.72
MmcNmt [53]	14.06	**18.64**	30.11	24.91	**25.10**	**30.40**
Our	13.93	17.72	29.90	$24.51_{\dagger}$	$24.53_{\dagger}\ddagger$	$29.63_{\dagger}$
HAN + *Our*	$14.16^{*}_{\dagger}\ddagger$	$18.15^{*}_{\dagger}$	$30.29^{*}_{\dagger}$	$\mathbf{25.11}_{\dagger}\ddagger$	$24.89^{*}_{\dagger}\ddagger$	$29.96^{*}_{\dagger}\ddagger$
SAN + *Our*	$\mathbf{14.19}^{*}_{\dagger}\ddagger$	$18.03^{*}_{\dagger}$	$\mathbf{30.38}^{*}_{\dagger}\ddagger$	$24.96_{\dagger}$	$25.03^{*}_{\dagger}\ddagger$	$30.07^{*}_{\dagger}\ddagger$

Our approach is always significantly better than SentNmt and Cache. †, ∗, ‡: statistically significantly (p-values < 0.05) better than DocT, HAN, and SAN.

Table 7. Performance of Our Approach and Baselines on the Test Set of Lexical Translation Consistency

Model	BLEU	News Acc.		TED Acc.		Subtitles Acc.		Total Acc.
		Entity	General	Entity	General	Entity	General	
SentNmt	19.82	59.1	56.1	56.3	47.3	49.5	43.0	53.4
Cache [34]	20.24	62.3	60.4	60.7	52.7	52.6	49.5	57.7
DocT [51]	20.51	60.3	58.2	61.6	50.7	51.6	44.1	55.8
HAN [26]	20.69	58.0	58.6	59.8	51.2	52.6	48.4	55.8
SAN [25]	20.71	63.0	57.5	57.1	50.7	53.7	46.2	56.1
MmcNmt [53]	20.85	61.1	58.9	64.3	52.2	54.7	49.5	57.5
Our	**20.87**	**68.5**	**63.6**	**70.5**	**59.0**	**61.1**	**52.7**	**63.4**

Average BLEU scores (%) and accuracy (Acc.) of different genres of consistent instances (%) are reported.

Results show that our approach is superior to SentNmt significantly, with +0.76, +0.75, and +0.80 BLEU gains for Zh→En News, TED, and Subtitles, respectively. For En→De, our approach still improves the BLEU scores by +1.73, +1.25, and +0.91 on News, TED, and Europarl, respectively.

Compared with existing DocNMT models HAN and SAN, our model achieves better or comparable translation quality. Although it is slightly lower than HAN and SAN in some datasets, the difference is not significant. Despite the higher BLEU achieved by MmcNmt, our goal-oriented approach performs better translation consistency (shown in Table 7) and is easy to combine with some existing models. When attached to HAN or SAN, our model does not reduce or even further improves the BLEU of original models.

Different from existing methods that utilize complex networks to capture the attention relationship between words in a long context sequence, our approach only uses a simple symbol to encode sentence-level contextual information. We suppose the improvement of BLEU by our model with global context mainly benefits from two points. First, the encoder states are enhanced by the global document contextual information. Second, the words that need to be consistent are translated more correctly, which directly affects the generation of subsequent sequences.

6.2 Translation Consistency

We then investigate the effectiveness of our approach to improve translation consistency on our specific lexical consistency test set.

Table 8. Results of Different Enhancers on Average BLEU and Total Consistency Accuracy

#	Model	BLEU	Accuracy
1	SentNmt [36]	19.82	53.4
2	+ Encoder Enhancer (EE)	20.59	56.5
3	− consistency context	20.48	55.3
4	− document context	20.03	54.6
5	+ Decoder Enhancer (DE)	20.12	59.1
6	+ EE + DE	20.87	63.4

As shown in Table 7, our approach performs best on both the translation quality and consistency when compared with other methods. The BLEU scores are averaged on three genres, and our approach is statistically significantly (p-values < 0.05) better than SentNmt, Cache, and DocT. SentNmt model only achieves total 53.4% consistency accuracy. Although existing DocNMT models leverage the cross-sentence context to achieve better BLEU, the improvement of consistency accuracy is still limited. Among them, Cache [34] and MmcNmt [53] perform better than other baselines, which is due to the utilization of translation history.

In contrast, our approach with encoder and decoder enhancers explicitly models the translation consistency and is more sensitive to repeated words. Compared with SentNmt model, the accuracy on the lexical consistency is 63.4%, which is better than the sentence-level model by 10.0%. Compared with other DocNMT models, the improvements of our approach in lexical consistency are also obvious. Our approach achieves the highest BLEU score on the test set. Meanwhile, the accuracy on the lexical consistency is significantly higher than other DocNMT methods.

In addition, we compare the performance of methods in different types of consistent lexical. For each genre, the translation consistency of general national words is harder to achieve than named entities. And compared with general words, the accuracy of named entities can obtain higher improvement through our approach. (For News, TED, and Subtitles, that is +9.4% vs. +7.4%, +14.2% vs. +11.7%, and +11.6% vs. +9.7%, respectively).

6.3 Effect of Consistency Enhancer

We compare the effect of different enhancers, and discuss the contributions of two types of global context. Experiments are conducted on consistency test set. The BLEU is averaged on three genres.

As shown in Table 8, the encoder enhancer and decoder enhancer behave differently. And interestingly, they seem to present a complementary relationship (row 2 vs. row 5). The encoder enhancer plays a more important role in the BLEU improvement, while the decoder enhancer is more helpful to enhance the consistency accuracy. The combination of two enhancers (row 6) is superior to single one in both translation quality and consistency.

The encoder enhancer introduces both document context and consistency context. The consistency context enhances repeated source words that can construct lexical chains throughout the document. However, due to the relative sparseness of repeated words, the improvement of using consistency context alone (row 4) is limited. As a contrast, document context are integrated into most of the encoder states to enhance source-side representation. During decoding, the symbol "$\langle G \rangle$" whose encoder state is corresponding document context vector can offer contextual information directly. As row 3 shows, document context can utilize general global information effectively to improve translation quality. But it is not sensitive to consistency.

The decoder enhancer is specifically designed for consistency. Results show that it can well constrain the model to produce consistent translations (row 5 vs. row 1). However, without the

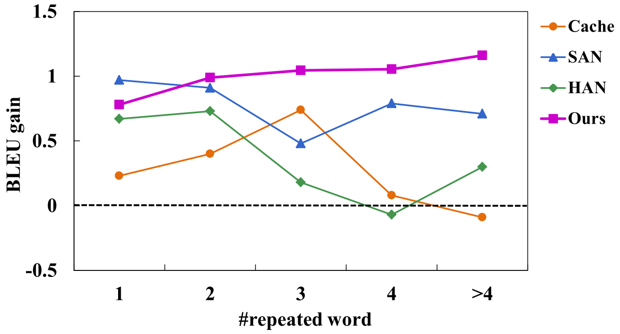

Fig. 4. BLEU gains over the SentNmt model (the black dotted lines) on the sentences with repeated words.

Table 9. Accuracy of Predicting Whether Repeated Words Need to be Translated Consistently

	Total	News			TED			Subtitles		
		E	*G*	Total	*E*	*G*	Total	*E*	*G*	Total
Majority	79.5	90.0	82.0	85.6	83.8	69.2	72.3	87.2	68.6	76.6
Our	82.9	91.2	84.9	87.8	85.0	75.6	77.7	88.5	73.3	79.9

"Majority" means that all repeated words are forced to be translated consistently. E: entity, G: general.

document context, decoder enhancer cannot obtain cross-sentence contextual information except for repeated words.

6.4 Sentences with Repeated Words

Figure 4 shows the BLEU gains over the SentNmt model on sentences with the different numbers of the repeated words in Chinese-to-English TED test sets. When a sentence contains repeated words, our approach always achieves the BLEU gains over SentNmt.

More importantly, with the increase of the number of repeated words in a sentence, the BLEU improvement of our approach ("*Ours*" in Figure 4) gets higher and higher. In contrast, the other three DocNMT models do not share this property. It proves that our approach can effectively utilize global and consistent information. The more repeated words, the closer the connection between current sentence and other sentences through the consistency context vectors, and the decoder enhancer can also impose stronger constraints on the generation of sequences.

6.5 Results of Consistency Classification

Considering that not all repeated words are consistent in target-side, our approach introduces a consistency classifier to explicitly determine whether a repeated word need to be translated consistently. Our test set has been annotated consistent instances, so it is used to measure the accuracy of the classification. Table 9 shows the results. Different from the accuracy of consistent instances in Table 7, the accuracy in Table 9 is calculated at the word level.

"Majority" force all repeated words to be translated the same. Its accuracy indicates that most repeated words are translated consistently in real, and named entities tend to be more consistent in translations than general nouns. The accuracy of entities is much higher than general words. Our classifier performs better than "Majority" with total +3.4% accuracy. It mainly benefits from the general words in informal genres, i.e., TED and Subtitles.

Table 10. Results of Our Approach With and Without Pre-Training

#	**Pre-training**		*newstest2018*	*our News test set*	
	LDC	News	BLEU	BLEU	Accuracy
1	×	×	13.70 (+0.53)	13.82 (+0.61)	62.4 (+4.9)
2	×	✓	13.93 (+0.76)	14.11 (+0.90)	65.9 (+8.4)
3	✓	✓	20.05 (+6.88)	19.85 (+6.64)	68.3 (+10.8)

✓ indicates that the corpus is used for pre-training, while × means not.
Numbers in brackets are the gains compared with SentNmt.

Table 11. Statistics of Parameters, Training, and Decoding Speeds (tokens/sec. ↑)

Model	**#Params**	v_{train}	v_{test}
SentNmt [36]	75.2 M	4,809	353.8
HAN [26]	87.9 M	2,805	261.3
SAN [25]	82.6 M	3,327	294.7
Ours	80.5 M	3,765	341.5

Actually, the expression is flexible when translating a repeated word, which makes consistent prediction difficult. This is one reason why we adopt softer strategies that utilize the classification probability as a confidence to weighted the distribution in Equation (6), rather than the hard ones specifying consistent words in advance.

6.6 Effect of Pre-training

Our approach applies the two-step training strategy, which has been widely used in DocNMT methods [25, 26, 31, 34, 51]. We discuss the effect of the pre-training in the news genre, so that the experiment using extra dataset can be conducted in the same domain. Table 10 shows the BLEU on the standard test set *newstest2018*, and the performance on our lexical consistency test sub-set in the news genre.

Compared with SentNmt, our approach trained from scratch (row 1) achieves significant +0.53 BLEU gains on standard test set, which is slightly lower than other models (shown in Table 6) using pre-training strategy [25, 26, 51]. On the targeted test set, the improvement of BLEU is +0.61. The accuracy of consistent instances is 62.4%, which goes beyond other DocNMT methods (shown in Table 7). With the pre-training using internal 0.31M News data (row 2), the performance is further improved. The improvement of consistent accuracy is +8.4%. In particular, when we use the mixed data of both News data and 2.0M extra LDC data to pre-train our model (row 3), the results are improved by +6.88 and +6.64 BLEU on the two test set, respectively. The overall results prove the effectiveness of the two-step training strategy.

The results show that the pre-training can improve both the translation quality and lexical consistency. On the targeted test set, the model gains 6.03 BLEU and 5.9% lexical consistency accuracy improvement (row 3 vs. row 1), respectively. We think the improvement of lexical consistency is mainly due to the improvement of translation quality of repeated words, especially repeated named entities, which benefits from the large-scale sentence-level pre-training.

6.7 Parameters and Speeds

Table 11 shows the parameters and speeds on Zh→En TED task. Our model introduces 5.2M extra parameters (that mainly come from the Transformer layer of document context in the global

Table 12. Examples to Show that Our Approach Generates Consistent Translations

Source	*<S1>* 我 的 下一项 发明 是 , 我 想 做 一个 电力 栅栏 。 *<S2>* 我 知道 , 电力 栅栏 已经 有 了 ...
Reference	*<S1>* My next invention is, I want to make an electric fence. *<S2>* I know electric fence is already invented ...
SentNmt	*<S1>* My next invention is, I want to make a power fence. *<S2>* I know the electricity fence has got it ...
HAN	*<S1>* My next invention is, I want to make a power fence. *<S2>* I know that the electric fences are already there ...
SAN	*<S1>* My next invention is, I want to make a electric fence. *<S2>* I know the electricity fence has already existed ...
Our	*<S1>* My next invention is, I want to make an electric fence. *<S2>* I know that the electric fence is already there ...
Source	*<S1>* 诺塞尔 一开始 并 未 被 指认 参与 谋杀 ... *<S2>* 诺塞尔 最终 被 指控 参与 这 场 犯罪 谋划 。
Reference	*<S1>* Nosair was initially found not guilty of the murder ... *<S2>* Nosair would eventually be convicted for his involvement in the plot.
SentNmt	*<S1>* Norman attorney's not shown in the beginning ... *<S2>* Norézier ended up charges to participate in this crime.
HAN	*<S1>* Nosel was not identified at first for the murder ... *<S2>* Norl was ultimately charged with this crime.
SAN	*<S1>* Nosel didn't even get involved in murder ... *<S2>* Normatl was ultimately charged to participate in this crime.
Our	*<S1>* Nousl was not identified for murder at the beginning ... *<S2>* Nousl was ultimately charged with the crime.

context extractor) to the SentNmt model. It translates all sentences in a document simultaneously. Because of the relatively simple network to utilize context, the decoding speed of our approach is similar to the sentence-level system, and is 15.9% faster than SAN.

6.8 Case Study

Table 12 shows two examples to demonstrate that our approach can constrain the model to translate the repeated source words consistently. In the first example, the repeated word " 电力 " in two sentences is translated into different forms ("power", "electricity", and "electric") by DocNMT methods without consistent constraint. Compared with that, our approach generate the same translation "electric". In the second example, only our approach can translate the repeated named entity " 诺塞尔 " into the consistent word "Nousl".

6.9 Results of English-to-Russian Discourse Phenomena

Voita et al. [38] propose a two-pass CADec model to handle the scenarios where sentence-level parallel corpus is large-scale but document-level parallel corpus is rare. CADec utilizes both source-side context sentences and their target-side results translated by a SentNmt. Voita et al. construct English-to-Russian contrastive test sets to evaluate four types of discourse phenomena: deixis, lexical cohesion, inflection ellipses, and VP ellipses. Each test instance is assigned three given context sentences. Its evaluation method has been described in Section 4.

Table 13. Accuracy (%) of English-to-Russian Discourse Phenomena

Model	deixis	lexical cohesion	inflection ellipses	VP ellipses
SentNmt [36]	50.0	45.9	53.0	28.4
CADec [38]	81.6	58.1	72.2	80.0
CADec + *Our*	82.3	70.2	74.6	80.8

Their experiments show that CADec can improve the discourse phenomena well. However, the test set is not friendly to models using the source-side context. In their contrastive instances, the source-side context is the same, so results of models only using the source-side context cannot change with the target-side context. Therefore, to compare with their method, we add our approach to the original CADec. We use the same experiment settings and datasets. The accuracy of discourse phenomena is shown in Table 13.

Compared with original CADec, our approach can improve the performance of discourse phenomena. We mainly focus on the results of lexical cohesion, i.e., the translation consistency of named entities in our article. It can be found that our approach can improve the accuracy of lexical cohesion significantly. Our approach is lexical consistency oriented by explicitly modeling the consistency and distinguishing repeated and non-repeated words.

7 RELATED WORK

Document-level translation is an important branch of machine translation [8, 16, 32, 33, 47]. In this article, our goal is to enhance the lexical translation consistency in DocNMT. Actually, this issue has been widely studied in the age of SMT. Xiao et al. [44] define the ambiguous words that need to be translated consistently, and re-translate them using words in candidate sets by two ways. Ture et al. [35] introduce three features to encourage consistency. Garcia et al. [7] design a feature that scores lexical consistency using word embeddings, and a change operation affecting how the translation search space is explored. Some researchers also systematically analyze the system behaviors on consistency [3, 9, 10, 28]. Other related works in SMT propose methods to improve the lexical cohesion, which mainly takes into account the repetition and hyponymy of words [45, 46]. With the rise of deep learning, NMT has surpassed SMT in many translation tasks. Some issues and their solutions applicable to the SMT framework should be reconsidered in the encoder-decoder framework. For DocNMT, our analysis has shown that lexical translation inconsistency is serious. However, there are few studies on the problem.

Existing DocNMT methods mainly focus on how to encode and use cross-sentence contextual information. Many works utilize fixed size previous source-side sentences as context [2, 12, 41]. Voita et al. [39] explore the context-aware model on the Transformer to show that cross-sentence attention can learn anaphora resolution. Zhang et al. [51] encode the context sentences with an extra encoder, and integrate them into each encoder and decoder layers. Miculicich et al. [26] propose a hierarchical attention network to model sentence-level attention. Yang et al. [49] use the capsule network to model relations between context. Researches also take advantage of previous target-side context [26, 38]. Kuang et al. [19] design dynamic and topic caches to store the word embedding of translation history and global topics, respectively. Tu et al. [34] utilize the hidden states of translation history. Voita et al. [38] add a second-pass decoder to leverage the source and the translations of SentNmt to improve the performance on discourse phenomena. They also propose a repair model to utilize target-side monolingual data to learn a document-level language model [37]. On the other hand, some other works explore a larger context in the entire document [15, 24, 31]. Maruf et al. [25] propose a hierarchical selective attention network. The attention weights are sharpened to attend sentences and words. Xiong et al. [48] model the discourse coherence of the entire document with a two-pass decoder and a reward teacher. Considering that the improvement of standard MT metrics cannot fully reflect the resolution of discourse phenomena, some works focus on the evaluation of phenomena [1, 38]. However, all existing DocNMT models are insensitive to repeated words in context. They fail to carefully analyze the translation consistency and explicitly model it.

8 CONCLUSION AND FUTURE WORK

In this article, we analyze the discourse phenomena of Chinese-to-English translation in different genres. The analysis shows that lexical translation inconsistency is the most frequent errors in DocNMT. We also summarize the types of translation consistency, and create a test set to evaluate the lexical consistency automatically.

To alleviate the lexical inconsistency, we propose an explicit approach to enhance the lexical translation consistency. Specifically, we extract two types of global contextual information. Repeated source words in the document are extracted to generate consistent contextual vectors, which are integrated into encoder-side and decoder-side, to modify the encoder representation and constrain the generation of consistent translations, respectively. Compared with existing DocNMT models, experiments on different datasets show that our approach can substantially improve lexical translation consistency. Meanwhile, it can also improve translation quality of SentNmt significantly. In the future, we will explore the translation consistency at phrase-level and larger granularity.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their insightful comments and suggestions.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representation 2015*.

[2] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Vol. 1. Association for Computational Linguistics, 1304–1313.

[3] Marine Carpuat and Michel Simard. 2012. The trouble with SMT consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 442–449.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1724–1734.

[5] 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,* Vol. 1. Association for Computational Linguistics, 4171–4186.

[6] Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 644–648.

[7] Eva Martínez Garcia, Carles Creus, Cristina España-Bonet, and Lluís Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics* 108, 1 (2017), 85–96.

[8] Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 909–919.

[9] Liane Guillou. 2013. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, 10–18.

[10] Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic validation of terminology translation consistency with statistical method. In *Proceedings of the MT Summit XI*. 269–274.

[11] Sebastien Jean and Kyunghyun Cho. 2019. Context-Aware learning for neural machine translation. CoRR, abs/1903.04715.

[12] Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? CoRR, abs/1704.05135.

[13] Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2964–2975.

[14] Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2242–2254.

[15] Xiaomian Kang and Chengqing Zong. 2020. Fusion of discourse structural position encoding for neural machine translation. *Chinese Journal of Intelligent Science and Technologie* 2, 2 (2020), 144–152.

[16] Xiaomian Kang, Chengqing Zong, and Nianwen Xue. 2019. A survey of discourse representations for chinese discourse annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing* 18, 3 (2019), 1–25.

[17] Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the 4th Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, 24–34.

[18] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, 177–180.

[19] 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*. Association for Computational Linguistics, 596–606.

[20] 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*. Association for Computational Linguistics, 1412–1421.

[21] Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3505–3511.

[22] Yanjun Ma, Yifan He, Andy Way, and Josef van Genabith. 2011. Consistent translation using discriminative learning: A translation memory-inspired approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1239–1248.

[23] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 55–60.

[24] Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* Vol. 1. Association for Computational Linguistics, 1275–1284.

[25] 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,* Vol. 1. Association for Computational Linguistics, 3092–3102.

[26] Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2947–2954.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

[28] Xiao Pu, Laura Mascarell, and Andrei Popescu-Belis. 2017. Consistent translation of repeated nouns using syntactic and semantic cues. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* Vol. 1. Association for Computational Linguistics, 948–957.

[29] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving Language Understanding by Generative Pre-Training*. Technical report. OpenAI.

[30] 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,* Vol. 1. Association for Computational Linguistics, 1715–1725.

[31] Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 1576–1585.

[32] Mei Tu, Yu Zhou, and Chengqing Zong. 2013. A novel translation framework based on rhetorical structure theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* Vol. 2. Association for Computational Linguistics, 370–374.

[33] Mei Tu, Yu Zhou, and Chengqing Zong. 2014. Enhancing grammatical cohesion: Generating transitional expressions for SMT. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,* Vol. 1. Association for Computational Linguistics, 850–860.

[34] 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 (2018), 407–420. DOI: https://doi.org/10.1162/tacl_a_00029

[35] Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, 417–426.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł. ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems 30.* Curran Associates, Inc., 5998–6008.

[37] Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-Aware monolingual repair 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*. Association for Computational Linguistics, 877–886.

[38] 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.* Association for Computational Linguistics, 1198–1212.

[39] 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,* Vol. 1. Association for Computational Linguistics, 1264–1274.

[40] Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence.* AAAI Press, 1–9.

[41] 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*. Association for Computational Linguistics, 2826–2831.

[42] Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 1060–1068.

[43] KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 5971–5978.

[44] Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Proceedings of the Machine Translation Summit*, Vol. 13. 131–138.

[45] Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lv, and Qun Liu. 2013. Modeling lexical cohesion for document-level machine translation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence.* AAAI Press, 21832189.

[46] 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 the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1563–1573.

[47] Deyi Xiong, Min Zhang, and Xing Wang. 2015. Topic-based coherence modeling for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 3 (2015), 483–493.

[48] Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7338–7345.

[49] 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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 1527–1537.

[50] Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017. Thumt: An open source toolkit for neural machine translation. CoRR, abs/1706.06415.

[51] Jiacheng Zhang, Huanbo Luan, Maosong Sun, FeiFei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 533–542.

[52] Yang Zhao, Jiajun Zhang, Chengqing Zong, Zhongjun He, and Hua Wu. 2019. Addressing the under-translation problem from the entropy perspective. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. AAAI Press, 451–458.

[53] 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 29th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 3983–3989.

[54] Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics* 7, 5 (2019), 91–105. DOI: https://doi.org/10.1162/tacl_a_00256

Received August 2020; revised April 2021; accepted September 2021