Encouraging Lexical Translation Consistency for Document-Level Neural Machine Translation

Xinglin Lyu, Junhui Li*, Zhengxian Gong, Min Zhang
School of Computer Science and Technology, Soochow University, Suzhou, China
xllv2020@stu.suda.edu.cn
{lijunhui,zhxgong,minzhang}@suda.edu.cn

Abstract

Recently a number of approaches have been proposed to improve translation performance for document-level neural machine translation (NMT). However, few are focusing on the subject of lexical translation consistency. In this paper we apply "one translation per discourse" in NMT, and aim to encourage lexical translation consistency for document-level NMT. This is done by first obtaining a word link for each source word in a document, which tells the positions where the source word appears at. Then we encourage the translations of those words within a link to be consistent in two ways. On the one hand, when encoding sentences within a document we properly exchange context information of those words. On the other hand, we propose an auxiliary loss function to better constrain that their translations should be consistent. Experimental results on Chinese↔English and English→French translation tasks show that our approach not only achieves state-of-the-art performance in BLEU scores, but also greatly improves lexical translation consistency.

1 Introduction

Unlike sentence-level neural machine translation (NMT), document-level NMT needs to not only model intra-sentence dependencies, but also consider a wide variety of inter-sentence discourse phenomena, such as coreference, lexical cohesion, semantic coherence, discourse relations. Motivated by the success of "one translation per discourse" in statistical machine translation (SMT) (Merkel, 1996; Carpuat, 2009; Türe et al., 2012; Guillou, 2013; Al Khotaba and Al Tarawneh, 2015), in this paper our goal is to encourage lexical translation consistency for document-level NMT.

Figure 1 shows an example of an input document and its output translated by a state-of-the-art sentence-level NMT system. The technical term

Source
#2 : … 国家 对 房地产业/fang_di_chan_ye 的 宏观 调控 政策 …
#3: … 去年 上海 房地产业/fang_di_chan_ye 各项 指标 …
#4: … 房地产业/fang_di_chan_ye 增加值 …
#7: 全年 房地产业/fang_di_chan_ye 增加值 为 670.23 亿 元 …
Sentence-Level NMT
#2 : … in 2005 , the state 's macroeconomic regulation of the real estate industry …
#3: … the various indicators for the shanghai real estate sector last year …
#4: … over the previous year , while real estate added value …
#7: the annual real estate market added amounted to 67.023 billion yuan …
Reference
#2 : … the state 's macro-control policies for the real estate sector …
#3: … various real estate sector indexes in shanghai … last year
#4: … while value added from the real estate sector …
#7: for the full year , the real estate sector registered value added of 67.023 billion yuan , …

Figure 1: An example of document-level Chinese-English translation from our development set NIST 2006, where the translations of source word 房地产业/*fang_di_chan_ye* tend to be consistent in reference.

房地产业/*fang_di_chan_ye*, occurring four times within a document, surprisingly obtains different translations while in its reference (human translation) it is translated consistently. Such inconsistent translations, however, tend to confuse readers in some cases.

Recent years have witnessed an increasing interest in document-level NMT, but most previous studies explore various context-aware models for better incorporating document-level context to improve translation performance without handling a specific discourse phenomenon (Maruf and Haffari 2018; Miculicich et al. 2018; Maruf et al. 2019, to name a few). As a way to encourage lexical translation consistency, Kuang et al. (2017) and Tu et al. (2018) cache recently translated words and/or their translations for translating future sentences. However, cache-based approaches may potentially guide the translation of future sentences in a wrong way since the cached translation could be incorrect. Rather than explicitly presenting lexical translations used

*Corresponding author: Junhui Li.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3265–3277
November 7–11, 2021.

in previous sentences as in cache-based approaches, in this paper we aim at improving lexical translation consistency in a softer way: we encourage translations of the same word in different positions of a document to be consistent. Specifically, we first obtain a word link for each source word in a document if it has, which tells the positions the source word appears at. To encourage translation consistency for words within a link, we exchange their context information when encoding sentences in a document. Moreover, we properly propose an auxiliary loss function to better constrain that the translations of these words should be consistent.

Overall, we make the following contributions.

- We propose a metric to properly measure lexical translation consistency, and provide a detailed study on lexical translation consistency in both Chinese↔English translation.
- We propose a novel approach to improve lexical translation consistency for document-level NMT. One nice property of our approach is that our models could synchronously translate sentences in a document, rather than translating them one by one as in cached-based approaches.
- Experimental results show that our approach outperforms various context-aware NMT models in BLEU. More importantly, our approach greatly improves lexical translation consistency.

2 Motivation

Given a parallel document pair $(\mathcal{S}, \mathcal{T})$, a source-side word w (stemmed to eliminate morphological differences if necessary) is one of **words of our interest** if it is a non-stop word and occurs two or more times in $\mathcal{S}$. For w, we conjecture that the translations (stemmed too if necessary) of w in $\mathcal{T}$ tend to be same. As shown in Figure 1, source word 房地产业*/fang_di_chan_ye* is consistently translated into *(the) real estate sector* in reference translation.

Lexical Translation Consistency Metric. To properly evaluate lexical translation consistency, we propose *lexical translation consistency ratio* (LTCR), which is based on word-alignment. Let us assume that source word w appears k times in $\mathcal{S}$. Based on word alignment between $\mathcal{S}$ and $\mathcal{T}$, we obtain its k translations,[1] i.e., $(t_1, \cdots, t_k)$, where t_i may consist of zero, one or more words. Then we define the metric for word w as:

$$\text{LTCR}\,(w) = \frac{\sum_{i=1}^{k}\sum_{j=i+1}^{k} \mathbb{1}\,(t_i = t_j)}{C_k^2} \times 100\% \quad (1)$$

where the denominator C_k^2 denotes the size of the combination of translation set $(t_1, \cdots, t_k)$, and function $\mathbb{1}\,(t_i = t_j)$ returns 1 if t_i is same as t_j, otherwise 0. The metric illustrates how frequent translation pairs of w is same within a document. The higher the metric value is, the more likely w is translated consistently. Taking source word 房地产业*/fang_di_chan_ye* in Figure 1 as an example, its *LTCR* is 100% for reference translation and 0% for sentence-level NMT.

In above we calculate *LTCR* for a single word in a document. Likewise, we could apply the metric to all source words that are of our interest in a parallel document pair, or a document-level parallel dataset by summing up all these words' corresponding numerators and denominators, respectively.

Statistics on Reference Translation and Automatic Translation. To better understand lexical consistency in translation, we take a concrete Chinese-English (ZH-EN) manually word-aligned document-level parallel corpus (LDC2015T06) as representative to study how consistent the lexical translation is in ZH→EN and EN→ZH translation. The corpus consists of 268 documents with 6741 sentences in total from domains including broadcast, newswire, and web data.

Moreover, for sentence-level NMT translation we perform word alignment to obtain word-level translation.[2] Table 1 compares the lexical translation consistency in ZH→EN and EN→ZH translation of LDC2015T06. From it, we observe that although translation diversity is usually encouraged, *LTCR* still reaches 74.24% and 63.11% in ZH→EN and EN→ZH reference translation, respectively. This confirms our conjecture that the translations of same source words tend to be consistent. We also note that the consistency is different among different types of words. For example, the consistency for nouns is much higher than those of other word types in both translation directions. Unfortunately, the consistency in automatic translation is much lower than that in reference translation, indicating there exists much room to improve lexical consistency in document-level machine translation. Finally, it also shows that the percentages of words of

[1] To obtain translation, we filter out determiners.

[2] The word aligners are trained on our machine translation datasets by fast_align (Dyer et al., 2013).

ZH→EN				EN→ZH			
POS	**Gold**	**Auto**	**Percentage(%)**	**POS**	**Gold**	**Auto**	**Percentage(%)**
Noun	80.98	50.74	13.57	Noun	75.79	55.49	8.83
Verb	57.86	35.96	2.48	Adj	66.83	51.94	1.77
Adv	61.23	30.77	1.43	Verb	46.93	29.97	1.34
Adj	81.77	52.83	0.72	Adv	49.72	30.58	0.99
Others	75.96	30.92	3.16	Others	64.97	33.17	1.62
All	74.24	43.13	20.92	All	63.11	36.02	15.06

Table 1: *LTCR* values in ZH→EN and EN→ZH translation of LDC2015T06. The columns of **Gold** and **Auto** indicate that *LTCR* is computed against reference (sentence-level NMT) translation with gold (auto) word-alignment. The column of **Percentage** indicates the proportion of the interest words against all source words.

our interest are quite high, i.e., 20.92% and 15.06% in ZH and EN documents, respectively.

3 Encouraging Lexical Translation Consistency via Word Links

As our goal is to encourage lexical consistency in document-level translation, we first obtain word links, each of which tells the positions that a word appears in a document (Section 3.1). To encourage translation consistency among words in the same link, on the one hand we exchange their information when encoding sentences within a document (Section 3.2). On the other hand, we properly propose an auxiliary loss function to better constrain the translations of these words being consistent (Section 3.3).

3.1 Obtaining Word Links

We define some notations before describing our approach. Given a document-level parallel pair $(\mathcal{S}, \mathcal{T}) = (S_i, T_i)|_{i=1}^{N}$ with N sentence pairs, we assume that each source sentence $S_i = (s_{i,j})|_{j=1}^{n}$ consists of n words. Given document $\mathcal{S}$, we use $\mathcal{V}$ to denote the collection of words of our interest in $\mathcal{S}$, which are non-stop words and appear two or more times.

For word $s_{i,j}$ if it exists in $\mathcal{V}$, we maintain a link list $L_{i,j} = (a_{i,j,k}, b_{i,j,k}, m_{i,j,k})|_{k=1}^{K}$ with K triples, which tells the other K positions where $s_{i,j}$ appears.[3] Specifically, in a triple (a, b, m), a and b indicate the sentence index and word index of a position respectively while $m \in \{0, 1\}$ is a padding mask and indicates (a, b) is a real position pair or a fake one.

Specially, for cases where $s_{i,j}$ appears more than K times in $\mathcal{S}$, we choose the top K closest ones to construct its word link.[4]

[3] We do not include $s_{i,j}$ itself in $L_{i,j}$.

[4] According to our preliminary experimentation, the effect of different ways of choosing K positions is negligible.

3.2 Encoding Documents with Word Links

Now each word of our interest in a document is equipped with a word link. In encoding, we take documents as input units by synchronously encoding sentences within a document. Figure 2 shows our encoder layer which encodes documents with word links.

3.2.1 Sentence Position Embedding

Since words in a link list may appear in different sentences, a Transformer encoder can not distinguish the sentence positions of the linked words and the current word. Therefore, we introduce sentence position embedding to distinguish the positions of these words.

Formally, given the i-th sentence S_i in $\mathcal{S}$, we project each word $s_{i,j}$ into a word embedding $e_{i,j} \in \mathbb{R}^d$, a (intra-sentence) position embedding $pe_j \in \mathbb{R}^d$, and a sentence position embedding $spe_i \in \mathbb{R}^d$, where d is the size of embedding and hidden state throughout the entire model. Then, we perform an addition operation to unify them into a single input, i.e., $e_{i,j} + pe_j + spe_i$. Note that both the word embeddings and the sentence position embeddings are trainable parameters while the (intra-sentence) position embeddings are sinusoidal (Vaswani et al., 2017).

3.2.2 Encoder

As shown in Figure 2, the encoder consists of M identical encoder layer, which consists of three sub-layers, i.e., a self-attention sub-layer, a word-link-attention sub-layer, and a feed-forward sub-layer. Next we use sentence $S_i = (s_{i,j})|_{j=1}^{n}$ to illustrate the encoding process.

Self-Attention Sub-Layer. In the m-th encoder layer, it takes $A_i^{(m)} \in \mathbb{R}^{n \times d_m}$ as input and computes a new sequence $B_i^{(m)}$ with the same length

via multi-head attention function:

$$B_i^{(m)} = \text{LayerNorm}\left(\text{MultiHead}\left(A_i^{(m)}, A_i^{(m)}, A_i^{(m)}\right)\right) + A_i^{(m)}, \tag{2}$$

where LayerNorm is the layer normalization function (Ba et al., 2016), and the output $B_i^{(m)}$ is of shape $\mathbb{R}^{n\times d}$. For the first encoder layer, $A_i^{(1)}$ is the input of the encoder while for other layers, $A_i^{(m)}$ is the output of the $(m-1)$-th encoder layer.

Word-Link-Attention Sub-Layer. Since we encode sentences within document $S_i|_{i=1}^N$ synchronously, we obtain $B_i^{(m)}\Big|_{i=1}^N$ from the self-attention sub-layer of the m-th layer. Let us assume that word $s_{i,j}$ in sentence S_i is of our interest and has a word link list $L_{i,j}$. Then we use the list to index the states of its K linked words from $B_i^{(m)}\Big|_{i=1}^N$. We use $C_{i,j}^{(m)} \in \mathbb{R}^{K\times d}$ to denote the indexed states. Consequently, this sub-layer uses another multi-head attention function to exchange information among linked words:

$$D_{i,j}^{(m)} = \text{LayerNorm}\left(\text{MultiHead}\left(B_{i,j}^{(m)}, C_{i,j}^{(m)}, C_{i,j}^{(m)}\right)\right) + B_{i,j}^{(m)}. \tag{3}$$

Specifically, if $s_{i,j}$ is out of our interest and does not have a word link list, we set $D_{i,j}^{(m)} = B_{i,j}^{(m)}$.

Feed-Forward Sub-Layer. In the m-th encoder layer, this sub-layer is applied to each position separately and identically by two linear transformations with a ReLU activation in between.

$$E_i^{(m)} = \text{LayerNorm}\Big(\max\left(0, D_i^{(m)}W^{F1} + b^{F1}\right)W^{F2} + b^{F2}\Big) + D_i^{(m)}, \tag{4}$$

where $W^{F1}, W^{F2} \in \mathbb{R}^{d\times d}$, and $b^{F1}, b^{F2} \in \mathbb{R}^d$ are model parameters. The output of the final layer, i.e., $E_i^{(M)}$ will be used as the output of the encoder.

3.3 Consistency Constraint Loss

After encoding sentences within a document, we properly extract useful information from document-level context via deliberately obtained word links. We expect the extracted information from document-level context can enhance the translations of the same words being more consistent, i.e., the states of the same words within a document being closer. Let us assume that word $s_{x,y}$, i.e., the y-th word in the x-th sentence is in the word-link list of word $s_{i,j}$. We use $E_{i,j}^{(M)}$ and $E_{x,y}^{(M)}$ to denote their hidden states of our encoder with word-link attention sub-layer. Meanwhile we use $\widetilde{E}_{i,j}^{(M)}$ and $\widetilde{E}_{x,y}^{(M)}$ to denote their hidden states of a vanilla Transformer encoder, i.e., the encoder without the word-link attention sub-layer. Since our encoder has exchanged context information between $s_{i,j}$ and $s_{x,y}$ while the vanilla encoder has not, we expect that the two states $E_{i,j}^{(M)}$ and $E_{x,y}^{(M)}$ are closer than $\widetilde{E}_{i,j}^{(M)}$ and $\widetilde{E}_{x,y}^{(M)}$.[5]

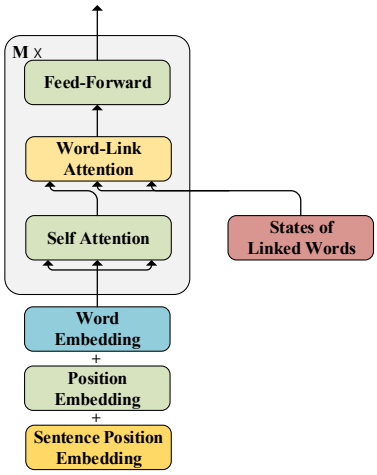


Figure 2: Our proposed encoder with word-link attention sub-layer. Note that states of linked words are indexed from the whole document.

According to Section 3.2, our encoder returns $E_i^{(M)}\Big|_{i=1}^N$ for document $S_i|_{i=1}^N$. We use $\widetilde{E}_i^{(M)}\Big|_{i=1}^N$ to denote the outputs of its corresponding vanilla encoder. [6] To encourage that our encoder would generate closer hidden states for a pair of linked words than the vanilla encoder, we follow previous work on visual semantic embedding (Kiros et al., 2014) and define a consistency constraint loss.

In practice, similar to Chen et al. (2020), we introduce a small neural network *projection head* that maps representations, i.e. $E_i^{(M)}\Big|_{i=1}^N, \widetilde{E}_i^{(M)}\Big|_{i=1}^N$, to a space where a consistency constraint loss is applied during training. We use MLP with one hidden layer to obtain Z and $\widetilde{Z}$ (i.e. $Z_i^{(M)}\Big|_{i=1}^N, \widetilde{Z}_i^{(M)}\Big|_{i=1}^N$) by $Z = g(E) = W^{(1)}\sigma(W^{(2)}E))$ and $\widetilde{Z} = g(\widetilde{E})$, where

[5] Although $\widetilde{E}_{i,j}^{(M)}$ and $\widetilde{E}_{x,y}^{(M)}$ are not directly used to train the model, there are in the semantic space as $E_{i,j}^{(M)}$ and $E_{x,y}^{(M)}$. See Appendix E for performance comparison by using E and $\widetilde{E}$.

[6] For simplicity, rather than training an independent vanilla encoder, we use our proposed encoder without the word-link attention sub-layers.

σ is a ReLU non-linearity, and $W^{(1)}, W^{(2)} \in \mathcal{R}^{d \times d}$ are model parameters. As shown in Appendix C, we find it beneficial to define the consistency constraint loss on $\left(Z, \widetilde{Z}\right)$'s rather than $\left(E, \widetilde{E}\right)$'s.

After that, the consistency constraint loss is defined as follow:

$$J_{\text{CC}}(\theta) = \sum_{\mathcal{S}} \sum_{i,j,k} \max\Big\{0, \gamma - D\left(Z_{i,j}^{(M)}, Z_{a_{i,j,k},b_{i,j,k}}^{(M)}\right) + D\left(\widetilde{Z}_{i,j}^{(M)}, \widetilde{Z}_{a_{i,j,k},b_{i,j,k}}^{(M)}\right)\Big\} \tag{5}$$

where θ are the parameters in our model, D is a distance function , i.e., cosine distance between two vectors, and γ is a margin, $a_{i,j,k}$ and $b_{i,j,k}$ denote the sentence and word indexes of word $s_{i,j}$'s k-th linked word, respectively.[7]

Finally, the joint objective function of our model $J(\theta)$ is define as:

$$J(\theta) = J_{\text{NMT}}(\theta) + \alpha J_{\text{CC}}(\theta) \tag{6}$$

where α determines the contribution of consistency constraint loss, and $J_{NMT}(\theta)$ is the cross entropy loss function, i.e.,

$$J_{\text{NMT}}(\theta) = -\sum_{(\mathcal{S},\mathcal{T})} \sum_{i,j} \log p\left(t_{i,j} | t_{i,<j}, \mathcal{S}\right) \tag{7}$$

4 Experimentation

To verify the effectiveness of our proposed approach, we carry out experiments on ZH↔EN translation tasks of two different domains: news and TED talks. As inspired by the conclusion in Guillou (2013) that lexical consistency is encouraged in English-French human translation, we also validate our approach on EN→FR translation.

4.1 Experimental Setup

Datasets. For ZH↔EN (News), the training data is composed from LDC. We use the NIST2006 dataset as the development set and combine NIST2002, 2003, 2004, 2005 and 2008 as the test set. Note that in the development and test sets every Chinese document has four aligned English documents, thus for ZH→EN translation one Chinese sentence has four references. In turn for EN→ZH translation each English sentence has one reference, and the numbers of sentences in development and test sets are four times those of ZH→EN translation, e.g., 4 × 879 and 4 × 5473, respectively.

For ZH↔EN (TED), the dataset is from the IWSLT 2014 and 2015 (Cettolo et al., 2012, 2015) evaluation. We use dev2010 as the development set and combine tst2010-2013 as the test set. For both ZH↔EN translations, every source sentence has one translation reference.

For EN→FR, we use IWSLT 2015 (Cettolo et al., 2015) evaluation as training data. For development and testing, we use dev2010 as the development set and combine tst2010-2013 as test set and every source sentence has one translation reference.

See Appendix A for more statistics and preprocessing of the experimental datasets.

Training Strategy. To compute the consistency constraint loss $J_{\text{CC}}(\theta)$, sentences are required to be encoded twice, i.e., one for encoding with the word-link attention sub-layer and the other for encoding without it. Therefore, including this loss function from the beginning may break the balance between optimizing the encoder and the decoder, and make it hard for the training to properly converge. To alleviate this problem, we divide the whole training process into two stages. In the first stage, we train the models to convergence with the cross entropy loss $J_{\text{NMT}}(\theta)$ only while in the second stage, we combine the consistency constraint loss $J_{\text{CC}}(\theta)$ and train the models with the joint loss. Actually, the second training stage acts like a fine-tuning, in which we use a smaller learning rate and fewer training steps.

Model Setting. We use *OpenNMT* (Klein et al., 2017) as the implementation of the Transformer and extend it. For the number of linked words with the current word, we set K = 6. The margin size γ in the consistency constraint loss is set to 0.2 while the weight α in joint objective function is set to 0.01. Other model settings are in Appendix B.

Evaluation. For all translation tasks, we report case-insensitive BLEU score as calculated by the *multi-bleu.perl* script.

4.2 Experimental Result

Besides sentence-level Transformer, we also compare our approach to three previous Transformer-based context-aware NMT models: HAN (Miculicich et al., 2018),[8] SAN (Maruf et al., 2019),[9] and

[7] In implementation, we need to use $m_{i,j,k}$ to distinguish those padding words in link lists.

[8] HAN: https://github.com/idiap/HAN_NMT

[9] SAN: https://github.com/sameenmaruf/selective-attn

Model	News			TED		
	#Param (M)	BLEU	LTCR	#Param (M)	BLEU	LTCR
Transformer	**68.97**	40.34	56.14	**59.88**	18.39	52.51
+word-link	75.42	41.50‡	59.77	66.31	19.03‡	56.74
+word-link +CC-loss	76.01	**42.57**‡§	**63.88**	66.91	**20.44**‡§	**62.79**
HAN (Miculicich et al., 2018)	75.73	41.38‡	56.01	66.67	18.93‡	53.10
SAN (Maruf et al., 2019)	74.86	41.80‡	57.13	65.75	19.33‡	54.33
MCN (Zheng et al., 2020)	75.23	41.58‡	55.81	66.14	19.90‡§	52.21

Table 2: Performance (BLEU and LTCR scores) on the test sets of ZH→EN translation. #Param denotes the number of parameters in millions. ‡ and § indicate that the improvement in BLEU is significant over Transformer/+*word-link* at 0.01, tested by bootstrap resampling (Koehn, 2004).

Model	News			TED		
	#Param (M)	BLEU	LTCR	#Param (M)	BLEU	LTCR
Transformer	**68.97**	16.36	51.33	**59.88**	11.77	43.97
+word-link	75.42	17.13‡	54.89	66.31	12.44‡	47.86
+word-link +CC-loss	76.01	**18.23**‡§	**59.01**	66.91	**13.11**‡§	**52.87**
HAN (Miculicich et al., 2018)	75.73	17.26‡	51.28	66.67	12.26‡	44.45
SAN (Maruf et al., 2019)	74.86	18.00‡§	53.66	65.75	12.99‡§	45.27
MCN (Zheng et al., 2020)	75.23	17.90‡§	52.11	66.14	12.71‡	44.39

Table 3: Performance (BLEU and LTCR scores) on test sets of EN→ZH translation.

Model	TED		
	#Param (M)	BLEU	LTCR
Transformer	**45.71**	40.76	47.26
+WL	50.65	41.57‡	49.33
+WL +CC	51.25	**42.94**‡§	**54.22**
HAN	52.49	41.75‡	48.55
SAN	51.62	41.67‡	49.32
MCN	51.64	42.05‡	49.21

Table 4: Performance (BLEU and LTCR scores) on test sets of EN→FR translation. Here +WL is for +Word-link, and +CC for +CC-loss.

MCN (Zheng et al., 2020).[10] For fair comparison, we run their source code with our model settings. Note that the above context-aware NMT models aim to improve the translation accuracy (i.e., BLEU) without focusing on resolving a particular discourse phenomenon.

Chinese-English Translation. Table 2 lists the performance of ZH→EN translation on both News and TED talk domains. From the table, we have the following observations.

- Exchanging information via words within word links (i.e., + *word-link*) achieves significant improvement in BLEU over (sentence-level) Transformer, suggesting that extracting information from document-level context via our deliberately designed word links is effective. Upon the setting of + *word-link*, constraining the translations of words within a link (i.e., +*CC-loss*) to be consistent with our proposed loss function achieves further significant improvement in BLEU. Comparing to Transformer, our approach gains +2.23 and +2.05 BLEU on the two domains, respectively.

- In terms of LTCR, both +*word-link* and +*CC-loss* greatly improve lexical translation consistency. For example, with +*word-link* +*CC-loss* our approach achieves +7.74% and +10.28% LTCR on the two domains, respectively.

- Though the three previous context-aware NMT models significantly outperform Transformer in terms of BLEU, their performance of LTCR is very close to that of Transformer, suggesting that these models have very limited effect in encouraging lexical translation consistency. Compared to these models, our approach achieves better performance in BLEU while more importantly, it greatly improves the performance in LTCR.

- With the word-link attention sub-layer, our approach introduces additional 10.87% parameters and have similar number of parameters as the previous context-aware NMT models.

English-Chinese Translation. Table 3 shows the performance results of EN→ZH translation on the two domains. From it, We observe a similar performance trend as ZH→EN translation. For example, our approach gains +1.87 BLEU and

[10]MCN: https://github.com/Blickwinkel1107/making-the-most-of-context-nmt

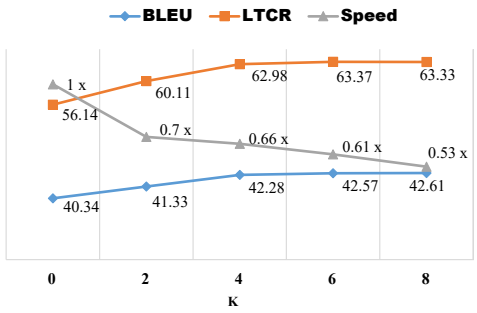


Figure 3: Performance on the test set of ZH→ EN (News) with different number of linked words.

+1.34 on the two domains over Transformer, respectively. Meanwhile, we achieve +7.68% LTCR and +8.90%, respectively.

English-French Translation. Table 4 shows the performance results of EN→FR translation on the TED domain. From it, We also observe a similar performance trend as ZH→EN translation. Our approach gains +2.18 BLEU and +6.96% LTCR over Transformer, respectively.

5 Discussion

Next, we take ZH→EN translation on news domain as a representative to discuss how our proposed approach improves translation performance. See Appendix for more discussion.

5.1 Effect of Hyper-parameter K

Among the words of our interests, the valid lengths of their word links differ greatly. As shown in Table 5, about 79.68% of our interested words have a word link whose valid length is 6 or less.

Length	Count	Per.	Length	Count	Per.
1	3444	36.61	2	1797	19.10
3	896	9.52	4	695	7.38
5	390	4.14	6	273	2.90
>6	1912	20.32	All	9407	100.00

Table 5: Statistics of valid lengths of word links on the development set NIST2006. Per. is for percentage (%).

A significant hyper-parameter in our proposed model is K, i.e., the number of words in every word link (Section 3.1). A low value makes the information exchanging among sentences within a document not sufficient while a high value increases the cost of computation. We compare the performance and training consumed time for five different K values. Note that our model is equivalent to sentence-level Transformer when K is 0.

Linked Words	Model	BLEU	LTCR
-	Transformer	40.34	56.14
Of same stem	+WL	41.50	59.77
	+WL +CC	42.57	63.88
Random	+WL	40.83	55.68

Table 6: Performance comparison when linked list contains the positions of words with same stem, or random positions.

Figure 3 shows the performance over different values of K. It shows that when K increases from 0 to 6, we observe consistent improvement on both BLEU and LTCR. The performance tends to be stable at $K = 6$ since no further improvement is achieved by increasing K to 8. Meanwhile, increasing K slightly slows down the training speed. Compared to Transformer (i.e., $K = 0$, 12700 **toks/sec**), our approach with $K = 6$ (7800 **toks/sec**) spends 39% more training time, consumed by the word-link attention sub-layers and the computation of consistency constraint loss.

5.2 Effect of Random Linked Word Positions

As shown in Section 3.1, the word link of word $s_{i,j}$ contains the other positions where $s_{i,j}$ appears at. To validate that the improvement achieved indeed comes from exchanging information among words with same stem, we perform a contrastive experiment by replacing the positions in word links with random positions. Note that in this way it does not make sense to apply the consistency constraint loss (+CC-loss) since the linked words are random.

Table 6 compares the performance. On the one hand, replacing words in word lists with random words still achieves +0.49 BLEU over Transformer. This suggests that even randomly exchanging information cross sentences is helpful. On the other hand, using random linked words does not bring LTCR improvement over Transformer. This in turn may suggest that the BLEU improvement achieved by our approach is mainly contributed by improved lexical translation consistency.

5.3 Performance on LDC2015T06

In Section 2 we use word-aligned document-level parallel corpus LDC2015T06 to analyze lexical consistency in translation. Table 7 compares the LTCR performance of our approach to those of the gold and sentence-level NMT scenarios. It shows that our approach (e.g., *+word-link +CC-loss*) achieves higher LTCR than Transformer over

POS	Gold	Trans.	+word-link	+word-link +CC-loss
Noun	80.98	50.74	53.66	58.11
Verb	57.86	35.96	38.72	38.19
Adv	61.23	30.77	32.68	35.66
Adj	81.77	52.83	53.41	56.63
Others	75.96	30.92	32.91	34.11
All	74.24	43.13	45.01	48.34

Table 7: LTCR values in ZH→EN translation of LDC2015T06. **Trans.** indicates sentence-level Transformer which achieves 8.39 BLEU while **+word-link** and **+word-link +CC-loss** achieve 8.66 BLEU and 9.61, respectively.

Model	Test
Trans.	68.39
+word-link	68.97
+word-link +CC-loss	**69.23**

Table 8: Accuracy of pronoun translations on the test set of ZH→EN (News).

all POS tags, especially for nouns. Meanwhile, the performance gap behind that of reference translation suggests that there still exists room for further improvement.

5.4 Pronoun Translation

We follow Miculicich et al. (2018) and Tan et al. (2019) to evaluate coreference and anaphora using the reference-based metric: accuracy of pronoun translation (Werlen and Popescu-Belis, 2017).

Table 8 lists the performance of pronoun translation. From it we observe that our approach also improves the performance of pronoun translation while exchanging context information among linked words (i.e., +word-link) contributes more than the consistency constraint loss (i.e., +CC-loss).

5.5 Human Evaluation

We conduct a human evaluation on 500 sentences randomly selected from our test set. Let us assume that the i-th sentence S_i in a document-level parallel pair $(\mathcal{S}, \mathcal{T})$ is selected. Then we provide

Annotator	Equal	Better	Worse
1	42%	36%	22%
2	51%	32%	17%
Avg.	47%	34%	19%

Table 9: Human evaluation results on 500 sentences from our test set when compare our approach (*+word-link +CC-loss*) with sentence-level Transformer.

Model	BLEU	LTCR
Trans.	40.34	56.14
+word-link	41.50	59.77
w/o SPE	41.01	58.93

Table 10: Performance comparison when we introduce the SPE (Sentence Position Embedding) to indicate sentence position of words, or not.

Exchange Information Function	BLEU	LTCR
Muti-head Attention	41.50	59.77
Average Pooling	41.03	59.11

Table 11: Performance comparison when we use different functions to exchange information among the linked words.

two annotators with a group of source sentences and translations, i.e., $(S_{i-2}, S_{i-1}, S_i, S_{i+1}, S_{i+2})$ and $(T_{i-2}, T_{i-1}, ?, T_{i+1}, T_{i+2})$, where *?* is S_i's translation of either our approach or the sentence-level Transformer. Besides, translation *?* is provided in random order with no indication which model it is from. Following Voita et al. (2019a), the task is to pick one of the three options: (1) the first translation is better, (2) the second translation is better, and (3) the translations are equal quality. The two annotators are asked to avoid the third option if they could give preference to one of the translations.

Table 9 shows the human evaluation results. In average the annotators mark 47% cases as having equal quality. Among the others, our approach outperforms Transformer in 64% cases, suggesting that overall the annotators have a strong preference for our approach over Transformer.

5.6 Effect of Sentence Position Embedding

As shown in Section 3.2.1, we introduce sentence position embedding (SPE) to indicate the sentence position of words. To analyze that the effects of it on our proposed approach, we perform a contrastive experiment.

Table 10 compares the performance. The SPE slightly improves BLEU (+ 0.49) and LTCR (+ 0.84%) over word-link Transformer without SPE. This is suggest that SPE for document-level NMT is helpful. We will explore more about it in the future work.

5.7 Analysis of Exchanging Information among Linked Words

As shown in Section 3.2.2, we use the multi-head attention function to exchange information among

linked words. To valid the effectiveness of this method, we perform a contrastive experiment by replacing multi-head attention function in Eq. 3 with the average pooling function Eq. 8.

$$D_{i,j}^{(m)} = \text{LayerNorm}\left(\text{Avg}\left(C_{i,j}^{(m)}\right)\right) + B_{i,j}^{(m)}. \quad (8)$$

Table 11 lists the performance of translation when we use different functions to exchange information among linked words. From it we observe that the multi-head attention function performs better. This in turn may suggest that simply averaging hidden states of linked words to exchange information lead to the mediocrity of cross-sentence information.

6 Related Work

There has been substantial work in SMT that either encourages or enforces lexical translation consistency. For example, Xiao et al. (2011) and Garcia et al. (2014, 2017) propose post-editing approaches to re-translate those source words which have been translated differently in a document. Tiedemann (2010a,b) and Gong et al. (2011) propose cache-based approaches to remember translation history. Discriminative learning approaches (Ma et al., 2011; He et al., 2011) are also proposed to fix lexical translation non-consistency. Besides, Carpuat (2009) and Türe et al. (2012) demonstrate that applying "one translation per discourse" constraint in SMT leads to better translation quality.

Moving to NMT, most of document-level NMT studies have proposed various context-aware NMT models to leverage either local context, e.g., previous sentences (Jean et al., 2017; Wang et al., 2017; Zhang et al., 2018; Bawden et al., 2018; Voita et al., 2018, 2019b; Yang et al., 2019), or entire document (Maruf and Haffari, 2018; Mace and Servan, 2019; Maruf et al., 2019; Tan et al., 2019; Xiong et al., 2019; Zheng et al., 2020; Kang et al., 2020). However, different from ours, these studies aim to improve the translation accuracy without handling a specific discourse phenomena. Kuang et al. (2017) and Tu et al. (2018) cache recently translated words and/or their translations which could be used to increase lexical consistency when translate future sentences. However, cache-based approaches require to translate sentences in a document one by one and may potentially guide the translation of future sentences in a wrong way since the cached translations could be incorrect. Experimental results in related studies (Zhang et al., 2018; Miculicich et al., 2018) have shown that the improvement of cache-based approaches is limited in BLEU over (sentence-level) Transformer. Our approach is different from cached-based approach as we translate sentences within a document synchronously, and more importantly it does not explicitly suggest any translation.

There also exists many studies in NMT that aim to resolve discourse phenomena in post-process. For example, to make translation outputs of a document more coherent, Voita et al. (2019a) propose DocRepair trained on monolingual target language documents to correct the inconsistencies in sentence-level translation while Yu et al. (2020) train a context-aware language model to re-rank sentence-level translation candidates.

7 Conclusion

In this paper, we apply "one translation per discourse" in NMT, and have proposed an approach to encourage lexical translation consistency. This is done by first obtaining a word link for each source word in a document, which tells the positions the source word appears at. Then we encourage the translations of words within a link to be consistent by both exchanging their context information in encoding, and using an auxiliary loss to constrain their translation being consistent. Experimental results on Chinese↔English and English→French translation tasks show that our approach not only achieves higher BLEU scores than various context-aware NMT models, but also greatly improves 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 and 61876120).

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.

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*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607.

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

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.

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.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *Computing Research Repository*, arXiv:1411.2539.

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. 2017. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of COLING*, pages 596–606.

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.

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.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

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

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of ACL*, pages 1198–1212.

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.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *Proceedings of Workshop on Discourse in Machine Translation*, pages 17–25.

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

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

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.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

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

For ZH↔EN on news domain, the training data set consists of LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, and LDC2010T03.

Table 12 summarizes statistics of the translation tasks. Note that we split long documents in training datasets into sub-documents with at most 20 sentences for efficient training.

Table 13 presents the percentage of words of our interest against all source-side words in the five translation tasks. It shows that the percentage of words of our interest varies across different translation tasks.

For ZH↔EN, the English sentences are tokenized and lowercased by Moses toolkit (Koehn et al., 2007)[11] while the Chinese sentences are segmented by Jieba.[12] For News (TED), we segment the source and target sentences into sub-words by a BPE model with 32K (21K) merged operations (Sennrich et al., 2016).

For EN→FR, all English and French sentences are tokenized and lowercased by Moses toolkit, we use BPE with 32K merged operations to segment words into sub-word units.

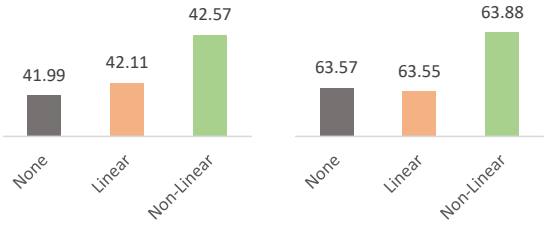
[11] https://github.com/moses-smt/mosesdecoder
[12] https://github.com/fxsjy/jieba

Set	ZH↔EN (News)		ZH↔EN (TED)		EN→FR (TED)	
	#Doc	#Sent	#Doc	#Sent	#Doc	#Sent
Training	41341	831485	3124	389421	1706	207323
Dev	79	1649	8	887	8	887
Test	509	5146	46	4632	46	4632

Table 12: Statistics of the training, development, and test sets of the translation tasks.

Set	ZH→EN (News)	ZH→EN (TED)	EN→ZH (News)	EN→ZH (TED)	EN→FR (TED)
Training	21.75	18.04	15.43	14.44	19.89
Dev	22.21	18.30	18.88	17.96	17.96
Test	24.12	19.41	17.27	19.25	19.25

Table 13: Percentages (%) of words of our interest.



(a) BLEU score (b) LTCR score

Figure 4: BLEU and LTCR scores on the test set of ZH→EN (News) with different projection heads $g(.)$.

B Model Settings

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. The dropout rate is 0.1. For models on ZH↔EN, the numbers of layers in the encoder and the decoder are set to 6, while for models on EN→FR, we change the numbers to 4. We train the models on two V100 GPUs with batch-size 4096 and use Adam with $\beta 1 = 0.9, \beta 2 = 0.98$ for optimization (Kingma and Ba, 2015). In the first training stage, we train the models for 150K steps, warm-up steps as 8K, learning rate as 1.0 while in the second training stage, we continue to train the models for 50K steps, warm-up steps as 4K, learning rate as 0.5. In inferring, we set the beam size to 5.

C Effect of Non-linear Projection Head

We take ZH→EN translation on news domain as example to study the importance of including a projection head, i.e. $g(.)$. Figure 4 shows LTCR and BLEU scores using three different architecture for the head: (1) identity mapping; (2) linear projection and (3) the default non-linear projection with one additional hidden layer (and ReLU activation).We observe that a non-linear projection is better than a linear projection (+0.46 BLEU and +0.32% LTCR), and much better than no projection (+0.58 BLEU and +0.31% LTCR).

D More Words of Our Interest, More Improvement?

Model	<=20%	20 ∼ 40%	>40%
Transformer	38.82	41.01	28.92
+word-link +CC-loss	39.81	43.82	33.23
Δ	+ 0.99	+ 2.81	+ 4.31

Table 14: BLEU Performance comparison over different subsets with different percentages of words of our interest.

We study if our approach performs better, i.e, more BLEU improvement over Transformer when there are more words of our interest in a document. To this end, we divide all documents in the test set into three subsets with different percentages of words of our interest:

- <=20%, which includes 137 documents with 1,449 sentences;
- 20 ∼ 40%, which includes 362 documents with 3,606 sentences;
- >40%, which includes 10 documents with 91 sentences.

As shown in Table 14, we observe that our approach indeed achieves more improvement over documents with higher percentages of words of our interest. For example, when the percentage is bigger than 40%, we achieve +4.31 BLEU gain.

Source

#1: （ 国际 ） 中智 签订 关于 实施 动植物 卫生 措施 备忘录 备忘录/bei_wang_lu
#2: ... 17日 在 这里 签订 关于 《 实施 卫生 和 植物 卫生 措施 协议 》 的 备忘录/bei_wang_lu。
#3: 根据 备忘录/bei_wang_lu ， 中智 双方 将 按照 该 协议 的 规则 以及 世界 动物 卫生 ...
#4: 备忘录/bei_wang_lu 规定 ， 双方 应 严格 按照 两 国 签署 的 议定书 或 商定 的 检验 ...
#5: ... 总局 副局长 葛志荣 和 智利 农业部 代部长 巴雷拉 在 备忘录/bei_wang_lu 上 签字 。

Sentence-Level NMT

#1: (international) zhongji signed memorandum on measures to implement animal and plants
#2: ... inspection general of china and the chilean ministry of agriculture signed a memorandum here on ...
#3: under the mou , both sides will , in accordance with the rules of the agreement and the standards developed by...
#4: the mou stipulates that the two sides should strictly adhere to the protocol or the agreed inspection and ...
#5: the memorandum was signed by the deputy director-general of the state of quality inspection and inspection of ...

Word-Link NMT

#1: (international) ciq sign memorandum on implementation of animal and plant health measures
#2: ... and the state ministry of agriculture of chile signed a memorandum on the implementation of health ...
#3: under the memorandum , the two sides will , in accordance with the rules of the agreement and the standards ...
#4: the memorandum stipulates that both sides should strictly implement inspection and quarantine of animals ...
#5: the memorandum was signed by the deputy director of the state administration of quality and inspection of ...

Reference

#1: (international) china and chile sign memorandum on application of animal and plant sanitary measures
#2: ... national quality inspection bureau and the ministry of agriculture of chile signed here a memorandum ...
#3: according to the memorandum , china and chile will formulate inspection and quarantine requirements for the ...
#4: the memorandum also stipulates that both sides should conduct inspection and quarantine of the imported and ...
#5: ... quality inspection bureau , and barrera , acting minister of agriculture of chile , signed the memorandum .

Figure 5: An example of document-level Chinese-English translation from our test set.

Encoder output	BLEU	LTCR
$\widetilde{E}$	39.83	56.22
E	41.50	59.77
Trans.	40.34	56.14

Table 15: Performance comparison when Using E and $\widetilde{E}$ as encoder output

E Performance Comparison When using E or $\widetilde{E}$ as Encoder Output

Table 15 lists the performance. It is not surprising that the performance of using $\widetilde{E}$ as encoder output is lower than that of using E since the former does not use any contextual information. This suggests that although $\widetilde{E}$ is not directly used to train the model, it is in the semantic space as E.

F Qualitative Analysis

We use an example to illustrate how word-link method helps translation (Figure 5). From it we observe that our proposed approach (Word-Link NMT) can effectively alleviate the translation inconsistency issue in document-level NMT, source word 备忘录/*bei_wang_lu* is consistently translated into *memorandum* by our model.