Breaking Corpus Bottleneck for Context-Aware Neural Machine Translation with Cross-Task Pre-training

Linqing Chen  Junhui Li  Zhengxian Gong  Boxing Chen
Weihua Luo  Min Zhang  Guodong Zhou

1School of Computer Science and Technology, Soochow University, Suzhou, China
2Alibaba DAMO Academy

Abstract

Context-aware neural machine translation (NMT) remains challenging due to the lack of large-scale document-level parallel dataset. To break the corpus bottleneck, in this paper we aim to improve context-aware NMT by taking the advantage of the availability of both large-scale sentence-level parallel dataset and source-side monolingual documents. To this end, we propose two pre-training tasks. One learns to translate a sentence from source language to target language on the sentence-level parallel dataset while the other learns to translate a document from deliberately noised to original on the monolingual documents. Importantly, the two pre-training tasks are jointly and simultaneously learned via the same model, thereafter fine-tuned on scale-limited parallel documents from both sentence-level and document-level perspectives. Experimental results on four translation tasks show that our approach significantly improves translation performance. One nice property of our approach is that the fine-tuned model can be used to translate both sentences and documents.

1 Introduction

Document-level context-aware neural machine translation (NMT) aims to translate sentences in a document under the guidance of document-level context. Recent years have witnessed great improvement in context-aware NMT with extensive attempts at effectively leveraging document-level context ([Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Maruf et al., 2019], to name a few). However, the performance of context-aware NMT still suffers from the size of parallel document dataset. On the one hand, unlike sentence-level translation models which could be well trained on large-scale sentence-level parallel datasets, the translation models of context-aware NMT may result in insufficient training. On the other hand, with only scale-limited source-side documents, the context encoders may fail to effectively extract useful context from the whole document. On the contrary, large-scale of parallel sentence corpora, and especially monolingual document corpora are much easier to find. In this paper, our goal is to break the corpus bottleneck for context-aware NMT by leveraging both large-scale sentence-level parallel dataset and monolingual documents. Specifically, we aim to use the former to boost the performance of translation models while employ the latter to enhance the context encoders’ capability of capturing useful context information.

There have been several attempts to boost context-aware NMT performance in the scenarios where the document-level parallel dataset is scale-limited, or even not available. On the one hand, sentence-level parallel dataset is a natural resource to use. For example, Zhang et al. (2018) propose a two-stage training strategy for context-aware NMT by pre-training the model on a sentence-level parallel dataset. On the other hand, Junczys-Dowmunt (2019) leverage large-scale source-side monolingual documents, in which they simply concatenate sentences within a document into a long sequence and explore multi-task training via the BERT-objective (Devlin et al., 2019) on the encoder. Due to that different models are usually required to model sentences and documents, however, it is challenging to effectively take them both in a single model.

In order to effectively and simultaneously model...
both sentence-level parallel dataset and monolingual documents, in this paper we propose a novel cross-task pre-training approach. As shown in Figure 1, we define two pre-training tasks. One learns to translate a sentence from source language to target language while the other learns to translate a document from deliberately noised to original. Importantly, the two pre-training tasks are jointly learned via the same model synchronously. Then we use document-level parallel dataset to fine-tune the properly pre-trained models. Similarly to the pre-training, we can fine-tune the models from both sentence-level and document-level perspectives. Experimental results on four document-level translation tasks show that our approach significantly improves translation performance, suggesting the effectiveness of our approach in modeling both sentence-level parallel dataset and monolingual documents. One nice property of our approach is that the fine-tuned models can be used to translate both sentences and documents.

2 Cross-Task Pre-training

In the following, we first describe our pre-training tasks defined upon sentence-level parallel dataset and large-scale monolingual documents (Section 2.1). Then we detail our model which caters such pre-training tasks (Section 2.2). Finally, we present our joint pre-training (Section 2.3).

2.1 Pre-training Tasks

We define two pre-training tasks in our pre-training. One is on sentence-level parallel dataset while the other is on monolingual documents.

Sentence-level Translation

Given large-scale sentence-level parallel dataset, our pre-training task is quite straight, i.e., sentence-level translation.

Document-level Restoration

Given monolingual documents, our pre-training task is to restore a document from a noised version. To this end, we deliberately corrupt documents by following the two pre-training objectives, which are inspired by both gap sentence objective (Zhang et al., 2020) and masked language model objective (Devlin et al., 2019).

- Context-Aware Gap Sentence Restoration (CA-GSR). Given a document $S$ with $N$ sentences, we randomly select $M$ sentences as gap sentences and replace them with a mask token [MASK1] to inform the model. The gap sentence ratio is, therefore $M/N$. For each selected gap sentence, we use its left and right neighbours as input while the gap sentence serves as output. To mimic document-level translation task, in the selection the first and the last sentences are always not selected while any two consequent sentences are not both selected.

- Context-Aware Masked Sentence Restoration (CA-MSR). Given a sentence $X$, we follow BERT and randomly select 15% tokens in it. The selected tokens are (1) 80% of time replaced by a mask token [MASK2], or (2) 10% of time replaced by a random token, or (3) 10% of time unchanged. For a sentence, we use its masked $\hat{X}$ as input while the original $X$ serves as output.

Both CA-GSR and CA-MSR are applied simultaneously with the noised document as context. For convenience of presentation, we use a concrete example to illustrate the input and output of our document-level restoration task. As shown in Figure 2, let assume that a document $X$ contains 6 sentences and the third and fifth sentences (i.e., $X_3$ and $X_5$) are selected as gap sentences while the others are not. On the one hand, for a sentence which is not selected as gap sentence, e.g., $X_1$, we use its masked version (e.g., $\hat{X}_1$) as input while try to predict its original sentence (e.g., $X_1$). On the other hand, for a gap sentence, e.g., $X_3$, we concatenate its left and right neighbouring sentences with separator [MASK1] and try to predict the gap sentence (e.g., $X_3$). As shown in Figure 2, sentences from $S_1$ to $S_6$ constitute document-level input $S$ while sentences from $T_1$ to $T_6$ make up output $T$. Note that we do not include either gap sentences themselves or their masked version in $S$, in case the document
This paper was retracted. For more information, see https://aclanthology.org/2021.acl-long.222.

The self-attention sub-layer takes \( A_i^{(k)} \in \mathbb{R}^{n \times d_m} \) as input and computes a new sequence \( B_i^{(k)} \) with the same length via multi-head attention function:

\[
B_i^{(k)} = \text{MultiHead}\left(q = A_i^{(k)}, k = A_i^{(k)}, v = A_i^{(k)}\right),
\]

where the output \( B_i^{(k)} \) is in the shape of \( \mathbb{R}^{n \times d_m} \), and \( q, k, v \) represent the query and key-value pairs in attention mechanism respectively. For the first encoder layer, \( A_1^{(1)} \) is the addition of \( S_i \)'s word embedding and its position embedding while for other layers, \( A_i^{(k)} \) is the output of the preceding encoder layer.

In the \( k \)-th encoder layer, the sentence representation sub-layer takes \( B_i^{(k)} \) as input and computes a vector to represent the sentence through a linear combination with a vector of weights as:

\[
\alpha_i^{(k)} = \text{softmax}\left(W^2 \tanh \left(W^1 \left(B_i^{(k)}\right)^T\right)\right)
\]

where \( W^1 \in \mathbb{R}^{d_m \times d_m} \) and \( W^2 \in \mathbb{R}^{d_m} \) are model parameters. The output \( \alpha_i^{(k)} \) is a \( n \)-sized vector. Then the representation vector of sentence \( S_i \) is the weighted sum of its hidden states:

\[
C_i^{(k)} = \alpha_i^{(k)} B_i^{(k)},
\]

where \( C_i^{(k)} \) is a \( d_m \)-sized vector. We then stack vectors of all sentences in \( S \) into \( c^{(k)} \), i.e., \( C^k = \left[C_1^{(k)}, \ldots, C_N^{(k)}\right] \).

The actual output of this sub-layer is \( \text{LayerNorm}(B_i^{(k)} + A_i^{(k)}) \), where \( \text{LayerNorm} \) is the layer normalization function. For simplicity, we do not include the residual addition and layer normalization functions in our sub-layers. Note that the sentence representation sub-layer is the only exception which does not have residual addition and layer normalization.
2.2.2 Adapting Context-Aware Model to Sentence-Level Translation

In the first pre-training task, sentence-level translation is context-agnostic and does not require the global context encoder. Therefore, it only uses the sentence encoder and decoder, as shown in Figure 3 (b). Moreover, we turn off the gate mechanism by setting \( H'_i = H_i \). Since we share the two sub-layers of self-attention and feed forward between the sentence encoder and the global context encoder, updating the model by sentence-level translation will have direct impact on the global context encoder too.

2.3 Joint Pre-training Process

As shown in our experimentation, we share the same vocabulary for pre-training tasks. To train the above two pre-training tasks with a single model, we follow the strategy used in Johnson et al. (2017) and add a preceding language tag to each source and target sentence.

Our joint pre-training on two tasks falls into the paradigm of multi-task learning (MTL). In training...
stage, we take turns to load the training data of these pre-training tasks. For example, we update model parameters on a batch of training instances from the first task, and then update parameters on a batch of training instances of the other, and the process repeats.

3 Fine-tuning on Document-Level Parallel Dataset

3.1 Fine-tuning Tasks
Similar to pre-training tasks, we define the following two different fine-tuning tasks from both sentence-level and document-level.

Sentence-level Translation We first extract sentence-level parallel sentence pairs from the document-level parallel dataset for fine-tuning. This fine-tuning task enables the fine-tuned model to translate sentences. In fine-tuning, this task is processed as same as the sentence-level translation task in pre-training.

Document-level Translation Given a parallel document \((x, y)\) with \(N\) sentence pairs \((x_i, y_i)\) \(\forall i\). This fine-tune task is to translate source document \(x\) into target document \(y\). In fine-tuning, this task takes parallel documents as input units and is processed as same as the document-level restoration task in pre-training.

3.2 Fine-tuning Process
The fine-tuning process is quite similar as the pre-training process in Section 2.3. Specifically, we add a preceding language tag to each sentence. Meanwhile in fine-tuning, we alternatively load batches of the two fine-tuning tasks.

4 Experimentation
To test the effect of our approach in leveraging sentence-level parallel dataset and monolingual documents, we carry out experiments on Chinese-to-English (ZH-EN) and English-to-German (EN-DE) translation.

4.1 Experimental Settings

Pre-training data settings. The ZH-EN sentence-level parallel dataset contains 2.0M sentence pairs with 54.8M Chinese words and 60.8M English words.\(^4\) We use WMT14 EN-DE translation dataset as the EN-DE sentence-level parallel dataset which consists of 4.4M sentence pairs.\(^5\)

We use Chinese Gigaword (LDC2009T27) and English Gigaword (LDC2012T21) as monolingual document dataset for ZH-EN and En-DE translation, respectively. For efficient training, we split long documents into sub-documents with at most 30 sentences. We have 2.6M (7.3M) sub-documents with 24M (102M) sentences in total for Chinese (English). Upon the monolingual documents, we prepare training instances for the document-level restoration task and set gap sentence ratio to 20%.

All Chinese sentences are segmented by Jieba\(^6\) while all English and German sentences are tokenized by Moses scripts (Koehn et al., 2007).\(^7\) For ZH-EN (EN-DE) translation, we merge the source and target sentences of the parallel dataset and the monolingual document and segment words into sub-words by a BPE model with 30K (25K) operations (Sennrich et al., 2016).

Fine-tuning data settings. For ZH-EN, we have one translation task on news domain. The document-level parallel corpus of training set include 41K documents with 780K sentence pairs.


For EN-DE, we test three translation tasks in domains of TED talks, News-Commentary and Europarl.

- TED, which is from IWSLT 2017 MT track (Cettolo et al., 2012). We combine test2016 and test2017 as our test set while the rest as the development set.
- News, which is from News Commentary v11 corpus.\(^9\) We use news-test2015 and news-test2016 as the development set and test set, respectively.
- LDC2009T15, LDC2010T03.

\(^4\)It consists of LDC2002E18, LDC2003E07, LDC2003E14, news part of LDC2004T08, LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, LDC2010T03. Note that they are also included in ZH-EN parallel dataset.

\(^5\)http://www.statmt.org/wmt14/translation-task.html

\(^6\)https://github.com/messense/jieba-rs

\(^7\)As related studies, we lowercase English sentences in ZH-EN while truecase English and German sentences in EN-DE.

\(^8\)It consists of LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, LDC2010T03.

\(^9\)http://www.casmacat.eu/corpus/news-commentary.html
Europarl, which is extracted from the Europarl v7. The training, development and test sets are obtained through randomly splitting the corpus.

All above EN-DE document-level parallel datasets are downloaded from Maruf et al. (2019). Similar to fine-tuning datasets, the pre-processing steps consist of word segmentation, tokenization, long document split. Then we segment the words into subwords using the BPE models trained on pre-training datasets. See Appendix A for more statistics of the fine-tuning datasets.

Model settings. We use OpenNMT (Klein et al., 2017) as the implementation of Transformer and implement our models based on it. For all translation models, the numbers of layers in the context encoder, sentence encoder and decoder (i.e., $N_c$, $N_e$, and $N_d$ in Fig. 3) are set to 6. The hidden size and the filter size are set to 512 and 2048, respectively. The number of heads in multi-head attention is 8 and the dropout rate is 0.1. In pre-training, we train the models for 500K steps on four V100 GPUs. The number of heads in multi-head attention is 8 and the dropout rate is 0.1. In pre-training, we train the models for 500K steps on four V100 GPUs.

Evaluation. For evaluation, we use two metrics: BLEU (Papineni et al., 2002) and Meteor (Lavie and Agarwal, 2007) to evaluate translation quality.

4.2 Experimental Results

Main results. Table 1 shows the performance of our approach, where Ours-sent and Ours-doc indicate the performance achieved by our approach when we use sentences or documents as input units, respectively. In the scenario where both sentence-level parallel dataset and monolingual documents are not used, we directly train our models from scratch with the two fine-tuning tasks on the fine-tuning datasets. #2 and #3 in the table show that our model is capable of translating both sentences and documents. Interestingly, when we use sentences as translation units, our models (i.e., #2 Ours-sent) outperform sentence-level Transformer baseline (i.e., #1 who uses sentences as input units in both training and inferring) over all translation tasks with improvement of averaged 1.36 BLEU and 1.72 Meteor. Moreover, when we use documents as translation units, our models (i.e., #3 Ours-doc) achieve further improvement by modeling document-level context. Compared to previous studies, it also shows that our approach surpasses all context-aware baselines on ZH-EN and EN-DE (TED) tasks and achieves the state-of-the-art on average.

In the scenario where both sentence-level parallel dataset and monolingual documents are used, similar performance trends also hold. For example, #5 Ours-sent significantly exceeds Transformer
Table 2: Ablation studies on ZH-EN and EN-DE (News) translation tasks. Hereafter, we use Ours for Ours-doc, i.e., using documents as input units.

<table>
<thead>
<tr>
<th>Model</th>
<th>Bi-sent</th>
<th>Mo-doc</th>
<th>ZH-EN (BLEU)</th>
<th>EN-DE (News) (BLEU)</th>
<th>Meteor</th>
<th>Meteor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans.</td>
<td>X</td>
<td>X</td>
<td>39.64</td>
<td>27.56</td>
<td>22.03</td>
<td>41.37</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>X</td>
<td>42.77</td>
<td>28.46</td>
<td>24.70</td>
<td>44.38</td>
</tr>
<tr>
<td>Trans.</td>
<td>✓</td>
<td>✓</td>
<td>46.99</td>
<td>33.46</td>
<td>26.89</td>
<td>47.01</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>48.03</td>
<td>34.27</td>
<td>28.32</td>
<td>48.16</td>
</tr>
</tbody>
</table>

Table 3: Performance on ZH-EN translation with respect to different fine-tuning strategies and different input units in inferring.

<table>
<thead>
<tr>
<th>Fine-Tuning</th>
<th>Inferring-Input</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/ sentence-level</td>
<td>document sentence</td>
<td>50.03</td>
</tr>
<tr>
<td>w/o sentence-level</td>
<td>sentence</td>
<td>49.58</td>
</tr>
</tbody>
</table>

Table 4: Accuracy (%) of discourse phenomena.

<table>
<thead>
<tr>
<th>Model</th>
<th>Bi-sent</th>
<th>Mo-doc</th>
<th>deixis</th>
<th>lex.c</th>
<th>ell.inf.</th>
<th>ell.VP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans.</td>
<td>X</td>
<td>X</td>
<td>50.0</td>
<td>45.3</td>
<td>52.0</td>
<td>27.3</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>X</td>
<td>62.3</td>
<td>47.9</td>
<td>64.9</td>
<td>36.0</td>
</tr>
<tr>
<td>Trans.</td>
<td>✓</td>
<td>✓</td>
<td>50.9</td>
<td>46.4</td>
<td>67.2</td>
<td>75.6</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>81.9</td>
<td>61.7</td>
<td>70.6</td>
<td>80.5</td>
</tr>
</tbody>
</table>

5 Discussion

Next we use ZH-EN translation to analyze more on how our approach affects translation performance. See Appendix B for parameter analysis and statistics of the pre-trained models.

5.1 Effect of Joint Fine-tuning

In Section 3 we alternate sentence-level translation and document-level translation in fine-tuning. We investigate the effect of including sentence-level translation as a fine-tuning task. Table 3 compares the performance with respect to different fine-tuning strategies and different input units in inferring. When we use documents as input units in inferring, the joint fine-tuning strategy provides no advantage. However, when the input units are sentences, the joint fine-tuning strategy outperforms the one not including sentence-level translation in fine-tuning.

5.2 Analysis of Discourse Phenomena

We also want to examine whether the proposed approach actually learns to utilize document context to resolve discourse inconsistencies. Following Voita et al. (2019b) and Zheng et al. (2020), we use the same datasets to train model and contrastive test set for the evaluation of discourse phenomena for English-Russian by Voita et al. (2019b). There are four test sets in the suite regarding deixis, lexicon consistency, ellipsis (inflection and verb phrase). Each testset contains groups of contrastive examples consisting of a positive translation with correct discourse phenomenon and negative translations with incorrect phenomena. The goal is to figure out if a model is more likely to generate a cor-

This paper was retracted. For more information, see https://aclanthology.org/2021.acl-long.222.
5.3 Pronoun Translation


Table 5 lists the performance of pronoun translation. From it we observe that our proposed approach can well improve the performance of pronoun translations.

5.4 Effect of Gap Sentence Ratio

A significant hyper-parameter in the pre-training task of document-level restoration is the gap sentence ratio. A low ratio makes the document-level restoration less challenging while choosing gap sentences at a high ratio makes the global context have more overlapped. Table 6 shows that we achieve the best performance when the ratio is set as 20%.

5.5 Effect of Pre-training Objectives

As shown in Figure 2, we include two pre-training objectives in document-level restoration, i.e., CA-GSR and CA-MSR. To investigate the effect of CA-GSR, we use CA-MSR as the only objective in this pre-training task. In this way, the $s_3$ and $s_5$ in Figure 2 (a), for example, will be $\hat{x}_3$ and $\hat{x}_5$, respectively. Table 7 compares the performance when the pre-training task is of CA-MSR objective or combination of CA-GSR and CA-MSR. It shows the combining objective achieves better performance than using CA-MSR alone.

6 Related Work

We describe related studies in the following two perspectives.

6.1 Context-Aware NMT

Cache/Memory-based approaches (Tu et al., 2018; Kuang et al., 2018; Maruf and Haffari, 2018; Wang et al., 2017) store word/sentence translation in previous sentences for future sentence translation. Various approaches with an extra context encoders are proposed to model 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; Huo et al., 2020), 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).

Besides, there have been several attempts to improve context-aware NMT with monolingual document data. To make translations more coherent within a document, 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 translations. Finally, Junczys-Dowmunt (2019) use source-side monolingual documents to explore multi-task training via the BERT-objective on the encoder. They simply concatenate sentences within a document into a long sequence, which is different from our approach.

6.2 Pre-training for Document-Level NMT

While there are substantial studies on improving sentence-level NMT with pre-training, we limit ourselves here to pre-training for document-level (context-aware) NMT. BART (Lewis et al., 2020) is a denoising auto-encoder model which learns to reconstruct the original document from a noised version. Inspired by BART, mBART (Liu et al.,
2020) is a model trained on a mixed corpus containing monolingual documents of different languages. Both BART and mBART concatenate sentences in one document into a long sequence, and thus fall into a standard sequence-to-sequence (seq2seq) framework. This is very different from our cross-task pre-training, in which we combine both context-agnostic learning and context-aware learning in a single model.

7 Conclusion

In order to leverage both large-scale sentence-level parallel dataset and source-side monolingual documents for context-aware NMT, in this paper, we have proposed a novel cross-task pre-training approach, which simultaneously learns to translate a sentence from source language to target language while denoising a document from deliberately noised to original. Upon the pre-trained models, we fine-tune them with document-level parallel dataset from both sentence-level and document-level perspectives. Experimental results on multiple document-level translation tasks have demonstrate the effectiveness of our approach. Finally, we also provide insights on how context-aware NMT benefits from our approach.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62036004 and 61876120).

References


A Experimental Datasets

Table 8 summarizes statistics of the four translation tasks. Note that we split long documents into sub-documents with at most 30 sentences for efficient training.

<table>
<thead>
<tr>
<th>Set</th>
<th>ZH-EN</th>
<th>EN-DE (Europarl)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#SubDoc #Sent</td>
<td>#SubDoc #Sent</td>
</tr>
<tr>
<td>Training</td>
<td>47,758 781,524</td>
<td>132,721 1,666,904</td>
</tr>
<tr>
<td>Dev</td>
<td>82 1,664</td>
<td>273 3,587</td>
</tr>
<tr>
<td>Test</td>
<td>627 5,833</td>
<td>415 5,134</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set</th>
<th>EN-DE (TED)</th>
<th>EN-DE (News)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#SubDoc #Sent</td>
<td>#SubDoc #Sent</td>
</tr>
<tr>
<td>Training</td>
<td>7,491 206,126</td>
<td>10,352 236,287</td>
</tr>
<tr>
<td>Dev</td>
<td>326 8,967</td>
<td>112 2,169</td>
</tr>
<tr>
<td>Test</td>
<td>87 2,271</td>
<td>184 2,999</td>
</tr>
</tbody>
</table>

Table 8: Statistics of the training, development, and test sets of the four translation tasks.

B More Result Analysis

B.1 Model Parameters

Table 9 presents the numbers of parameters for ZH-EN and EN-DE translations. Note that for all EN-DE translation tasks, the numbers of parameters are same as the vocabulary for them are shared. The table shows that our models introduce very limited parameters to encode document-level context.

<table>
<thead>
<tr>
<th>Model</th>
<th>ZH-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer</td>
<td>80.6M</td>
<td>61.4M</td>
</tr>
<tr>
<td>Ours</td>
<td>86.2M</td>
<td>64.0M</td>
</tr>
</tbody>
</table>

Table 9: Model parameters for ZH-EN and EN-DE translations.

B.2 Statistics on Our Pre-trained models

Table 10 presents statistics on our two pre-trained models for ZH-EN and EN-DE translations. With 500K training steps and within 120 (130) hours we complete 3.0 (1.2) and 35 (20) passes over the sentence-level parallel dataset and monolingual document dataset for Chinese (English), respectively.

<table>
<thead>
<tr>
<th>Translation</th>
<th>#Epoch on Bi-sent</th>
<th>#Epoch on Mo-doc</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZH-EN</td>
<td>35</td>
<td>3.0</td>
<td>120h</td>
</tr>
<tr>
<td>EN-DE</td>
<td>20</td>
<td>1.2</td>
<td>130h</td>
</tr>
</tbody>
</table>

Table 10: Statistics on our two pre-trained models for ZH-EN and EN-DE translations.