ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200358

Learning Contextualized Sentence Representations for Document-Level Neural Machine Translation

Pei Zhang [1], **Xu Zhang** [2], **Wei Chen**[2], **Jian Yu**[2] , **Yanfeng Wang**[2] and **Deyi Xiong** [1] *

Abstract. Document-level machine translation incorporates inter-sentential dependencies into the translation of a source sentence. In this paper, we propose a new framework to model cross-sentence dependencies by training neural machine translation (NMT) to predict both the target translation and surrounding sentences of a source sentence. By enforcing the NMT model to predict source context, we want the model to learn "contextualized" source sentence representations that capture document-level dependencies on the source side. We further propose two different methods to learn and integrate such contextualized sentence embeddings into NMT: a joint training method that jointly trains an NMT model with the source context prediction model and a pre-training & fine-tuning method that pretrains the source context prediction model on a large-scale monolingual document corpus and then fine-tunes it with the NMT model. Experiments on Chinese-English and English-German translation show that both methods can substantially improve the translation quality over a strong document-level Transformer baseline.

1 Introduction

Neural machine translation (NMT) has achieved remarkable progress in many languages due to the availability of the large-scale parallel corpora and powerful learning ability of neural networks [1, 23, 26, 5]. However, most NMT systems translate a sentence without taking document-level context into account. Due to the neglect of inter-sentential dependencies, even the state-of-the-art NMT models lag far behind human translators on document-level translation [12].

Document-level machine translation has been therefore attracting more and more attention in recent years. A variety of approaches have been proposed, which can be roughly divided into two categories: leveraging discourse-level linguistic features [6, 29, 11] or encoding preceding/succeeding sentences into the model [33, 27, 15]. The former may need linguistic resources which are not easily available while the latter relies on parallel documents. Unfortunately, explicit document boundaries are often removed in parallel texts and it is difficult to automatically recover such boundaries.

In this paper, we propose a novel approach to document-level neural machine translation. Inspired by the success of contextualized word embeddings in various natural language tasks [2, 19, 20], we learn contextualized sentence embeddings for document-level NMT. Instead of encoding surrounding sentences into the NMT model, we try to learn a model to predict the previous or next sentence from the current sentence to be translated. This is similar to the skip-thought model [9] or the skip-gram model [16] at the word level. By training document-level NMT to predict surrounding source context, we hope the trained encoder to capture document-level dependencies in the sentence embedding of the current source sentence.

More specifically, we propose two methods to learn and integrate the contextualized sentence embeddings into document-level NMT. First, we explore a joint training method that simultaneously trains an encoder-decoder NMT model with a skip-thought model. As illustrated in Figure 1, we use a shared encoder to decode the target sentence, the previous and the next source sentence at the same time. Second, we jointly pretrain two encoder-decoder models to predict the previous and next source sentence respectively from the same current sentence on a monolingual document-level corpus and fine-tune the pretrained models with the document-level NMT model. In the pre-training & fine-tuning approach visualized in Figure 2 & 3, the model can use a very large-scale collection of monolingual corpus with document boundaries, which is easily available.

Our contributions can be summarized into three aspects:

- First, we propose a new framework to document-level NMT by learning contextualized sentence embeddings on the source side.
- Second, we further present two approaches to learning and incorporating contextualized sentence embeddings into document-level NMT: the joint training of NMT with a skip-thought model and the combination of pre-training with fine-tuning.
- Third, we validate the effectiveness of the proposed two methods based on the state-of-the-art NMT architecture Transformer [26] on both Chinese-English and English-German translation.

The paper is organized as follows. Section 2 overviews related work. Section 3 elaborates the two contextualized sentence embedding learning frameworks: the joint training and pre-training with fine-tuning. In section 4, we present the experimental settings and experiment results. In section 5, we conduct in-depth analyses and discuss our results, followed by our conclusion and future work in section 6.

2 Related Work

Document-level SMT Plenty of methods have been proposed for document-level statistical machine translation. Gong, Min, and Zhou [6] use a cache-based approach to model document-level machine translation. Meyer and Popescu-Belis [14] explore the discourse connectives to improve the quality of translation. Xiong et al. [29] propose to learn the topic structure of source document and then map the structure to the target translation. In addition to these approaches leveraging discourse-level linguistic features for document transla-

[1] Soochow University, China, email: pzhang.nmc@gmail.com, dyxiong@suda.edu.cn
[2] Sogou, China, email: {zhangxu216526, chenweibj8871, yujian216093, wangyanfeng}@sogou-inc.com
* Corresponding Author

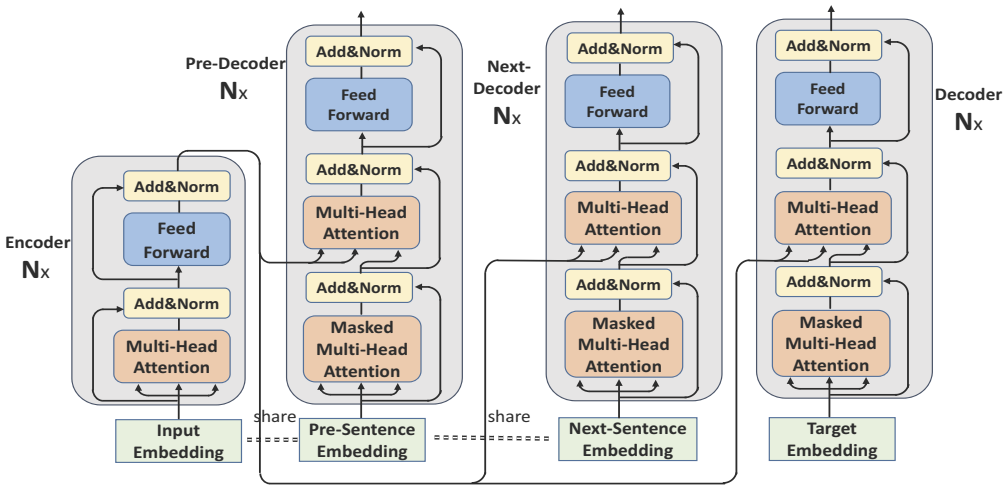

Figure 1. Architecture of the joint training model with an encoder learning the representation of current source sentence and three decoders (Pre-Decoder, Next-Decoder and Decoder) for predicting the previous source sentence, next source sentence and target sentences.

tion, Garcia et al. [4] incorporate new word embedding features into decoder to improve the lexical consistency of translations.

Document-level NMT In the context of neural machine translation, previous studies first incorporate contextual information into NMT models built on RNN networks. Tiede-mann and Scherrer [24] use extended source language context to improve the robustness of translation. Tu et al. [25] propose a cache to record the hidden state of each steps of the encoder and decoder as contextual information for word generation. Wang et al. [28] use a cross-sentence context-aware approach to resolve ambiguities and inconsistencies of translation. Maruf and Haffari [13] propose to use memory networks [3] for document-level NMT. Kuang et al. [11] propose a cache-based model for document-level NMT, where a static cache is used to store topical words while a dynamic cache [7] is for words generated in previous translations.

For document-level NMT based on the Transformer, Zhang et al. [33] propose to explore previous sentences of the current source sentence as the document information, which is further exploited by the encoder and decoder via attention networks. Xiong et al. [30] use multiple passes of decoding with Deliberation Network [31] to improve the translation quality. When translating the current sentence, translation results of other sentences in the first-pass decoding are used as the document information. In order to improve translating anaphoric pronouns, Voita et al. [27] propose context-aware NMT. Different from these methods, we train document-level NMT to predict surrounding sentences rather than encoding them into NMT or integrating translations of surrounding sentences into NMT. Miculicich et al. [15] present a hierarchical attention model to capture context information and integrate the model into the original NMT.

Pretrained Language Models Our work is also related to the recent pretrained language models [21, 2, 19, 20, 34] that learn contextualized word embeddings. By learning dynamic context-related embeddings, the pretrained language models significantly improve the representation learning in many downstream natural language processing tasks. In our models, we also learn "contextualized" sentence embeddings by putting a source sentence in its context. We believe that learning the representation of a sentence in its surrounding context is helpful for document-level NMT as the learned representation connection to the surrounding sentences. The way that we learn the "contextualized" sentence embeddings is similar to the skip-thought model [9] in that we also predict surrounding source sentences from the current source sentence. But in addition to learning contexualized sentence representations, we integrate the source context prediction model into document-level NMT via two methods: the joint training method and the pre-training & fine-tuning method.

3 Learning Contextualized Sentence Embeddings for Document-Level NMT

In this section, we introduce the proposed two methods on the Transformer for document-level NMT.

3.1 Joint Training

When NMT translates a source sentence, we want the sentence representation obtained by the encoder to contain information that can predict both the target sentence and surrounding source sentences. For this, we introduce a joint training method over the Transformer to jointly train an NMT model and a Transformer-based skip-thought model, both of which share the same encoder. Meanwhile, we select the previous and next sentences of the current source sentence as the surrounding sentences to be predicted .

Our joint training model is shown in Figure 1. The model is composed of one encoder and three decoders, where the encoder is to encode the source sentence in question and the three decoders are to predict the previous source sentence, next source sentence and target sentence, respectively. The network stuctures of the encoder and three decoders are the same as the Transformer encoder / decoder [26]. In order to train the joint model, we collect document-level parallel training instances $(s_i, s_{i-1}, s_{i+1}, y_i)$, where s_i is the current source sentence. s_{i-1} is the previous sentence of s_i, s_{i+1} is the next sentence of s_i and y_i is the target sentence. The task of the joint model is to predict s_{i-1}, s_{i+1} and y_i at the same time given s_i. For this, the multi-head attention network between each decoder and the encoder is constructed and automatically learned during training.

The loss function for the joint model is computed as follows:

$$Loss = Loss_{tgt} + \mu * Loss_{pre} + \lambda * Loss_{next} \quad (1)$$

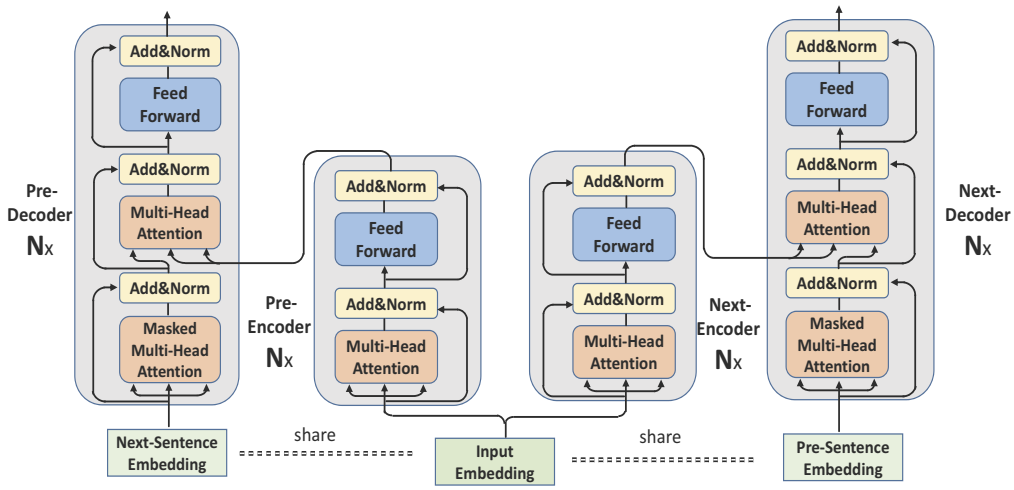

Figure 2. Architecture of the pre-training model where two encoder-decoder models are jointly trained to predict the previous and next source sentence from the current source sentence.

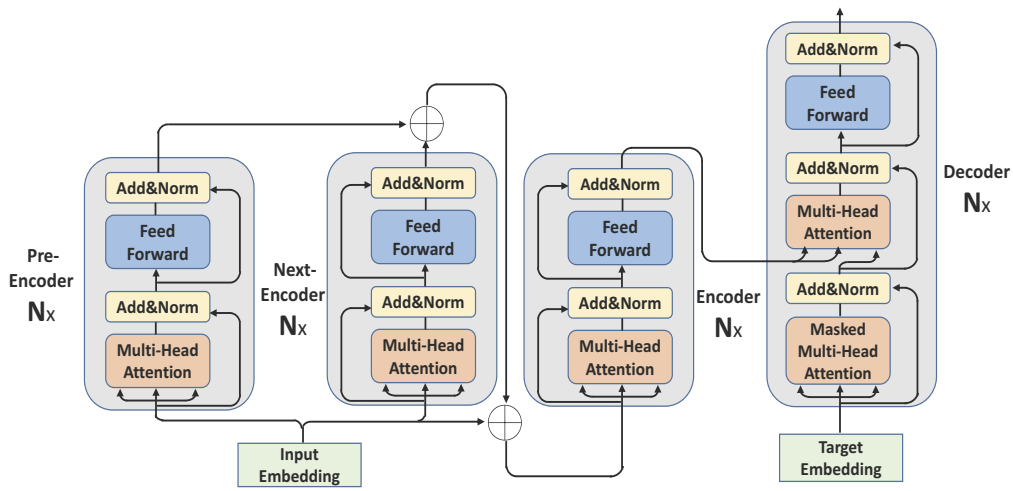

Figure 3. Architecture of the fine-tuning model where the sums of the outputs of the pre-trained pre-encoder and next-encoder are input to the NMT encoder.

where $Loss_{tgt}$ is the loss from the target decoder. $Loss_{pre}$ is the loss of predicting the previous source sentence and $Loss_{next}$ the next source sentence. We update model parameters according to the gradient of the joint loss. We use two hyperparameters μ and λ for $Loss_{pre}$ and $Loss_{next}$, respectively. The s_i, s_{i-1} and s_{i+1} share the same source embeddings.

In order to train joint model, we take the following two steps:

- First, we train the joint model to minimize the joint loss function on the collected training set and obtain the best joint model after it converges on the training data.
- Second, we remove the pre-decoder and next-decoder from the joint model and continue to train the reserved NMT encoder and decoder on the parallel training set $\{(s_i, y_i)\}$ to optimize $Loss_{tgt}$.

The trained NMT model in this way will be used to examine the effectiveness of the proposed joint training method on the test set.

3.2 Pre-training and Fine-Tuning

The joint training model requires parallel documents as training data. However, large-scale parallel corpora with document boundaries are not easily available. On the contrary, there are plenty of monolingual documents. Therefore, we further propose a pre-training strategy to train an encoder on monolingual documents. We want the pretrained encoder to capture inter-sentential dependencies by learning to predict surrounding sentences from a current source sentence.

As shown in Figure 2, we jointly pre-train two encoder-decoder models. From a large-scale set of monolingual documents, we can collect a huge amount of triples (s_{i-1}, s_i, s_{i+1}) as training instances for the pre-training model. The pre-training task trains one encoder-decoder model to predict the previous sentence s_{i-1} from s_i and the other s_{i+1} from s_i. Source word embeddings are shared by the two encoder-decoder models and they are jointly trained to optimize the following loss:

Table 1. BLEU scores of the two methods on Chinese-English translation (trained on the 900K-sentence corpus). "‡": statistically significantly better than the baseline ($p < 0.01$).

Methods	NIST06	NIST02	NIST03	NIST04	NIST05	NIST08	AVG
Baseline	38.14	41.42	42.09	42.7	40.59	30.65	39.49
Joint Training							
Pre	38.63	42.65	42.77	43.05	40.84	30.81	39.96
Next	38.94	42.69	42.36	43.08	41.14	31.24	40.1
Pre+Next	39.11 ‡	42.72 ‡	43.28 ‡	42.67	40.99	31.97 ‡	40.33
Pre-training and Fine-tuning							
Pre	39.39	43.04	42.62	42.83	40.78	31.9	40.23
Next	39.26	42.13	42.97	42.78	41.08	32	40.2
Pre+Next	39.11 ‡	**43.05** ‡	42.58	43.29	41.44 ‡	31.73 ‡	40.42
Pre-training for Joint Training							
Pre+Next	**39.46** ‡	42.19 ‡	**43.39** ‡	**43.42**	**41.74** ‡	**32.43** ‡	**40.63**

Table 2. BLEU scores for the pre-training & fine-tuning method on Chinese-English translation (trained on the 2.8M-sentence corpus). "‡": statistically significantly better than the baseline ($p < 0.01$).

Methods	NIST06	NIST02	NIST03	NIST04	NIST05	NIST08	AVG
Baseline	43.05	43.85	44.84	46.03	43.43	36.26	42.89
Pre	43.48	45.2	45.76	46.34	44.45	**37.16**	43.78
Next	43.15	45.01	46.3	46.3	44.91	37	43.9
Pre+Next	**43.62**	**45.78** ‡	**46.32** ‡	**46.42** ‡	**45.23** ‡	37.12 ‡	**44.17**

$$Loss = Loss_{pre} + Loss_{next} \tag{2}$$

Once we have the pre-trained model, we can continue to fine tune the pretrained two encoders (pre-encoder and next-encoder) with the Transformer model. The details are shown in Figure 3. The fine-tuning model contains three encoders and one decoder. The two encoders, namely the pre-encoder and next-encoder that encode the previous and next source sentences during the fine tuning, are from the pretrained model. In order to fine tune the pretrained encoders with the NMT encoder-decoder model, we use three strategies. First, the source word embeddings of the pre-encoder, next-encoder and NMT encoder can be initialized by the word embeddings from the pre-trained model. Second, the input to the NMT encoder is the sum of the outputs of the pre-/next-encoder and word embeddings of the current source sentence, formulated as follows:

$$\begin{aligned} encoder\_input = {} & input\_embedding \\ & + pre\_encoder\_output \\ & + next\_encoder\_output \end{aligned} \tag{3}$$

Third, during the fine-tuning stage, the shared source word embeddings and parameters of the pre-encoder and next-encoder continue to be optimized.

4 Experiments

We conducted experiments on Chinese-English and English-German translations to evaluate the effectiveness of the proposed methods.

4.1 Experimental Setting

For Chinese-English translation, we selected corpora LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and a portion of data from the corpus LDC2004T08 (Hong Kong Hansards/Laws/News) as our bilingual training data, which contain 2.8M sentences. We then selected from the 2.8M-sentence training data 94K parallel documents with explicit document boundaries, containing 900K parallel sentences. Each selected parallel document consists of 11 sentences on average. We used NIST06 dataset as our development set and NIST02, NIST03, NIST04, NIST05, NIST08 as our test sets. The development and test datasets contain 588 documents and 5,833 sentences in total. Each document has 10 sentences averagely. We also collected a large-scale monolingual (Chinese) document corpus from CWMT[3] and Sogou Labs[4], with 25M sentences and 700K documents. On avegrage, each documents contains 35 sentences.

For English-German translation, we used the WMT19 bilingual document-level training data[5], which contains 39k documents with 855K sentence pairs as training set. We collected a large-scale English monolingual document corpus[6] from WMT19 with 10M sentences and 410k documents. We used the newstest2019 development set as our development set and newstest2017, newstest2018 as test sets, which contain 123 documents with 2,998 sentence pairs and 252 documents with 6,002 sentence pairs respectively.

We used the byte pair encoding [22] to decompose words into small sub-word units for both languages. We used the case-insensitive 4-gram NIST BLEU score as our evaluation metric [18] and the script "mteval-v11b.pl" to compute BLEU scores. All the out-of-vocabulary words were replaced with a token "UNK".

As mentioned before, we used the Transformer to construct our models and implemented our models on an open source toolkit THUMT [32]. We set hidden size to 512 and filter-size to 2,048. The number of encoder and decoder layers was 6 and the number of attention heads was 8. We used Adam [8] for optimization. The learning rate was set to 1.0 and the number of warm-steps was 4000. We set batch size as 4,096 words for iterations. We used four TITAN

[3] http://nlp.nju.edu.cn/cwmt-wmt.
[4] https://www.sogou.com/labs/resource/list_news.php.
[5] https://s3-eu-west-1.amazonaws.com/tilde-model/rapid2019.de-en.zip.
[6] http://data.statmt.org/news-crawl/en-doc/news-docs.2015.en.filtered.gz.

Table 3. BLEU scores (average results on the 5 test sets) for the two-step training strategy for the "joint training" and "Pre-Training + Joint Training" methods abbreviated into "PT + JT" for limited space on the 900K-sentence corpus.

Methods	Joint Training	PT + JT
Pre+Next (step 1)	40.12	40.35
Pre+Next (step 2)	40.33	40.63

Table 4. Comparison between the joint training method and the explicit contextual integration method on Chinese-English translation.

Methods	AVG
Baseline	39.49
Joint Training	40.33
Explicit Contextual Integration	40.59

Xp GPUs for training and two TITAN Xp GPUs for decoding. Additionally, during decoding, we used the beam search algorithm and set the beam size to 4. We used bootstrap resampling method [10] to conduct statistical significance test on results.

We compared our models against the Transformer [26] as our baseline and some previous document-level NMT models [33, 24].

4.2 The Effect of the Joint Training

The experimental results of the joint training method are shown in Table 1. "Pre" / "Next" indicate that we use only the pre-decoder / next-decoder in the joint training model illustrated in Figure 1. We set $\mu = 0.5$ and $\lambda = 0.5$ for the "Pre" and "Next" methods according to the experiment results on the development set. The BLEU scores of these two methods are close to each other, indicating that the previous and next sentences have almost the same influence on the translation of the current sentence. When we use "Pre+Next" method to predict the previous and next sentence at the same time ($\mu = 0.5, \lambda = 0.3$ set according to results on the development set), the performance is better than "Pre" and "Next" alone achieving an improvement of +0.84 BLEU points over the baseline.

4.3 The Effect of the Pre-training & Fine-Tuning

We trained the pre-training model on the 25M-sentence monolingual document corpus and then conducted the fine-tuning as shown in Figure 3 on the two different parallel datasets: the 900K-sentence parallel corpus and the 2.8M-sentence parallel corpus. Sentences from the same document in the former corpus exhibit strong contextual relevance to each other. However, in the latter corpus, not all documents have clear document boundaries. When we train the pre-training model on the monolingual document corpus, we did not shuffle sentences in documents and kept the original order of sentences in each documents. The results are shown in Table 1 and Table 2. Similar to the joint training, we can train the pre-training model with a single encoder, either the pre-encoder or the next-encoder, to obtain the results for "Pre" or "Next". Of course, we can train the two encoders together to have the results for "Pre+Next". As shown in Table 1 & 2, the "Pre" and "Next" in the pre-training & fine-tuning model obtain comparable improvements over the two baselines. The improvements over the Transformer baseline without using any contextual information are +0.93 and +1.28 BLEU points on the two corpora, respectively.

Table 5. Comparison to other document-level NMT methods on Chinese-English translation.

Methods	900k	2.8M
Baseline	39.49	42.89
(Tiedemann and Scherrer [24])	38.83	-
(Zhang et al. [33])	39.91	43.52
Our work	40.63	44.17

4.4 The Effect of the Pre-training for the Joint Training

Both the joint training and pre-training are able to improve the performance in our experiments. We further conducted experiments to test the combination of them. Particularly, in the combination of two methods, we used the source embeddings learned from the pre-training model to initialize the source embeddings of the joint training model. It can be seen from Table 1 that such a combination of the two methods achieves the highest BLEU scores with +1.14 BLEU points higher than the Transformer baseline and +0.3 BLEU points higher than the single joint training model.

4.5 The Effect of the Two-Step training for the Joint Training

As mentioned in Section 3.1, we take two steps to train the joint training model. Here we carried out experiments to examine this two-step training method. Results are shown in Table 3, from which we can clearly see that the two-step training method is able to improve both the single joint training model and the combination of the joint training with the pre-training model.

4.6 Comparison to the Explicit Integration of Preceding/Succeeding Source Sentences into NMT

In our joint training method, during the testing phase, we do not explicitly use any contextual information for the current source sentence as the jointly trained NMT model implicitly learns the potential context information by learning to predict the preceding / succeeding source sentence during the training phase. In order to study how much potential context information can be captured by this implicit method, we conducted a comparison experiment. We obtain the representations of the preceding / succeeding source sentences via the pre-encoder and next-encoder and integrate them into the encoder of the current sentence by using self-attention that treats the encoder outputs of the current sentence as q and the encoder outputs of the previous / next sentence as k and v. We refer to this method as "Explicit Contextual Integration". The results are shown in Table 4. The explicit contextual integration method is better than our joint training method by only 0.26 BLEU points. Comparing this with the improvement of 0.84 BLEU points achieved by our joint training method over the baseline, we find that the joint training method is able to learn sufficient contextual information for translation in an implicit way.

4.7 Comparison with Other Document-Level NMT Methods

We further conducted experiments to compare our methods with the following previous document-level methods on Chinese-English translation:

Table 6. BLEU scores on English-German translation.

Methods	BLEU
Baseline	17.08
Joint Training	17.89
Pre-training & Fine-tuning	18

Table 7. BLEU scores (average results on the 5 test sets) of the two encoders vs. the single shared encoder for the pre-training model on the 900K-sentence corpus.

Methods	BLEU
Baseline	39.49
Single Shared Encoder	39.95
Two Encoders	40.42

- (Tiedemann and Scherrer [24]): concatenating previous sentence with the current sentence and inputting them into the encoder.
- (Zhang et al. [33]): using previous sentences of the current source sentence as document information, which is further exploited by the encoder and decoder via attention networks.

Table 5 shows the comparison results in terms of BLEU scores. It is clear that our method outperforms these two methods. Although the concatenation method [24] can improve document-level NMT over the RNNSearch model [17], it fails to improve the Transformer model. Compared with the method by Zhang et al. [33] that is also based on Transformer, our work outperforms their model by 0.72 and 0.65 BLEU points on the two datasets.

4.8 Results on English-German Translation

The results are shown in Table 6, which shows that our two methods also outperforms the baseline by 0.81 and 0.92 BLEU points on English-German translation, respectively.

5 Analysis

Experiments on the two methods show that using the context of previous and next sentence for document translation can improve NMT performance. In this section, we provide further analyses and discussions on the two methods.

5.1 Analysis on the Two Encoders for the Pre-training Model

As we have described in Section 3.2, we use two separate encoders in the pre-training model to encode the current source sentence for predicting the previous and next sentence. Only the word embeddings are shared in these two encoders. An alternative to these two encoders is to use a single encoder that is shared by the pre-decoder and next-decoder.

We conducted experiments to compare the two encoders vs. the single shared encoder for the pre-training model. Results are shown in Table 7. Obviously, the two-encoder network is better than the single shared encoder. This indicates that the two separate encoders are better at capturing the discourse dependencies between the previous and the current sentence and between the next and the current sentence than the single encoder. We conjecture that this is because the dependencies on the previous sentence are different from those on the next sentence.

Table 8. BLEU scores for the ablation study of the fine-tuning model on the two corpora. "Pre+Next+Input Embedding" is the method that inputs the sum of the outputs from the pre- and next-encoder and input embeddings into the fine-tuning model. "Input Embedding" only uses input embeddings of the pre-training model for the input to the fine-tuning model.

Methods	900k	2.8M
Baseline	39.49	42.89
Input Embedding	39.83	43.55
Pre+Next+Input Embedding	40.33	44.17

Table 9. The number of parameters used in the baseline and our method and performance comparison.

Methods	Parameters	BLEU
Baseline	90.2M	39.49
Joint Training	90.2M	40.33
Pre-training & Fine-tuning	128M	40.42
No Pre-training	128M	38.8

5.2 Ablation Study for the Fine-Tuning Model

As we mentioned in Section 3.2, we use the sum of input word embeddings, the output of the pre-encoder and next-encoder loaded from the pre-training model as the input to the NMT encoder. Here we empirically investigate the impact of these three parts on the fine-tuning model via ablation study. From the results shown in Table 8, we can find that if we use only the pretrained word embeddings as the input word embeddings to the NMT encoder of the fine-tuning model, we can obtain improvements of +0.34 and +0.66 BLEU points over baseline on the two corpora. When we add the outputs from both the pre-encoder and next-encoder, we obtain further improvements of +0.5 and +0.62 BLEU points over the input word embeddings. This clearly suggests that the pretrained two encoders learn additional contextual information.

5.3 Analysis on the Additional Parameters

Table 9 shows the number of parameters used by the baseline and our method. The joint training method uses the same number of parameters as the baseline, achieving a higher BLEU score. The number of parameters in the pre-training & fine-tuning method is about 30% larger than that of the baseline due to the additional two encoders for the preceding and succeeding source sentence. If we do not pre-train the model, the performance significantly drops, even worse than the baseline. This suggests that the improvement of the pre-training & fine-tuning method over the baseline is achieved not because of using additional parameters but the learned contextual knowledge in the pre-training on the monolingual document data.

5.4 Analysis on Translations

We take a deep look into the translations generated by the baseline Transformer and our best model. We find that our methods can improve translation quality by disambiguating word senses with document context (as shown in the first example in Table 10) or by making the document-level translations more consistent (as illustrated in the second example in Table 10) and so on. In the first example, "cui ruo" is ambiguous with two senses of "weak" or "fragile", which can be translated correctly by our model since it has learned the exactly contextualized sentence embeddings. In the second example, our model translates "fa xian" in both sentences into the same translation "detected", because the meaning of the "an jian" (case) mentioned by

Table 10. Translation examples of the baseline Transformer and our best model.

SRC	dang ran , mu qian fei zhou di qu de ***wen ding ju mian*** yi ran bi jiao ***cui ruo*** .
REF	of course the ***stability*** in africa at present is still ***fragile*** .
Transformer	of course , the current ***situation*** in africa is still relatively ***weak*** .
Our best model	of course , the current ***stability*** in the african region is rather ***fragile*** .
SRC	(1) zai xiang gang jing wu chu du pin ke zhi chu , zhe shi xiang gang jing fang de shou ci ***fa xian*** . (2) xiang gang jing fang chen , zai guo qu de yi nian zhong , gong ***fa xian*** shu zong an jian .
REF	(1) the dangerous drug division of hong kong police department points out that this is the first ***discovery*** by the police . (2) according to hong kong police, they have ***discovered*** several cases last year .
Transformer	(1) hkp pointed out that this was the first time that the hong kong police had ***found*** . (2) the hong kong police said that in the past year , a number of cases were ***detected*** .
Our best model	(1) hkp pointed out that this was the first time that the hong kong police (hkp) had ***detected*** it . (2) according to the hong kong police , in the past year , a number of drug cases were ***detected*** .

the second sentence can be predicted from the meanings of "du pin" (drug) and "jing fang" (police) in the first sentence.

We made a further statistical analysis on these two kinds of translation improvements. We randomly selected five documents with 48 sentences in total from the test datasets for translation. Among these sentences, our model can successfully deal with translations of ambiguous words in 32 sentences. For the document-level consistent translation problem, we found that 4 sentences among 7 sentences with such consistent phenomena were correctly and consistently translated by our model.

6 Conclusion and Future Work

In this paper, we propose a new framework to document-level NMT by learning a contextualized representation for a source sentence to be translated. We have presented two methods for learning and integrating such representations into NMT. The joint training method learns contextualized sentence embeddings simultaneously with the prediction of target translations. The pre-training & fine-tuning method learns the contextualized representations on a large-scale monolingual document corpus and then fine-tunes them with NMT. Experiments and analyses validate the effectiveness of the proposed two methods.

The proposed two methods can be extended in several ways. First, we would like to train the pre-training model with more monolingual documents so as to learn better contextualized sentence representations. Second, we also want to adapt the current framework from the source side to the target side in the future.

ACKNOWLEDGEMENTS

The present research was supported by the National Natural Science Foundation of China (Grant No. 61861130364 and 61622209) and the Royal Society (London) (NAF\R1\180122). We would like to thank the anonymous reviewers for their insightful comments.

REFERENCES

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

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', *CoRR abs/1810.04805*, (2018).

[3] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning, 'Key-value retrieval networks for task-oriented dialogue', in *n Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 37–49, (2017).

[4] Eva Martinez Garcia, Creus Carles, Cristina Espana-Bonet, and Lluis Marquez, 'Using word embeddings to enforce document-level lexical consistency in machine translation', in *The Prague Bulletin of Mathematical Linguistics*, pp. 85–96, (2017).

[5] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, 'Convolutional sequence to sequence learning', in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1243–1252. JMLR. org, (2017).

[6] Zhengxian Gong, Zhang Min, and Guodong Zhou, 'Cache-based document-level statistical machine translation', in *Conference on Empirical Methods in Natural Language Processing*, pp. 909–919, (2011).

[7] Edouard Grave, Armand Joulin, and Nicolas Usunier, 'Improving neural language models with a continuous cache', in *Proceedings of the International Conference on Learning Representations*, (2017).

[8] Diederik Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', in *CoRR abs/1412.6980*, (2015).

[9] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, 'Skip-thought vectors', in *International Conference on Neural Information Processing Systems*, pp. 3294–3302, (2015).

[10] Philipp Koehn, 'Statistical significance tests for machine translation evaluation', in *Proceedings of Empirical Methods in Natural Language Processing*, pp. 388–395, (2004).

[11] Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou, 'Cache-based document-level neural machine translation', in *International Conference on Computational Linguistics*, (2018).

[12] Samuel LÃaubli, Rico Sennrich, and Martin Volk, 'Has machine translation achieved human parity? a case for document-level evaluation', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4796, (2018).

[13] Sameen Maruf and Gholamreza Haffari, 'Document context neural machine translation with memory networks', in *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1275–1284, (2018).

[14] Thomas Meyer and Andrei Popescu-Belis, 'Using sense-labeled discourse connectives for statistical machine translation', in *Joint Workshop on Exploiting Synergies Between Information Retrieval & Machine Translation*, pp. 129–138, (2012).

[15] Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson, 'Document-level neural machine translation with hierarchical attention networks', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2947–2954, (2018).

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', in *Proceedings of the International Conference on Learning Representations*, (2013).

[17] Robert Ostling, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen, 'The helsinki neural machine translation system', in *Proceedings of the Second Conference on Machine Translation*, pp.

338–347, (2017).
[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 'BLEU: a method for automatic evaluation of machine translation', in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, (2002).
[19] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contxtualized word representations', *CoRR abs/1802.05365*, (2018).
[20] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, 'Improving language understanding with unsupervised learning.', *Technical report, OpenAI*, (2018).
[21] Prajit Ramachandran, Peter Liu, and Quoc Le, 'Unsupervised pretraining for sequence to sequence learning', *CoRR abs/1611.02683*, (2016).
[22] Rico Sennrich, Barry Haddow, and Alexandra Birch, 'Neural machine translation of rare words with subword units', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, (2015).
[23] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, 'Sequence to sequence learning with neural networks', in *Advances in Neural Information Processing Systems*, pp. 3104–3112, (2014).
[24] Jörg Tiedemann and Yves Scherrer, 'Neural machine translation with extended context', in *Proceedings of the Third Workshop on Discourse in Machine Translation,*, pp. 82–92. Association for Computational Linguistics, (2017).
[25] Zhaopeng Tu, Liu Yang, Shuming Shi, and Zhang Tong, 'Learning to remember translation history with a continuous cache', in *Transactions of the Association for Computational Linguistics*, pp. 407–420, (2017).
[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in Neural Information Processing Systems*, pp. 5998–6008, (2017).
[27] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov, 'Context-aware neural machine translation learns anaphora resolution', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1264–1274, (2018).
[28] Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu, 'Exploiting cross-sentence context for neural machine translation', in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 1–9, (2018).
[29] D. Xiong, G. Ben, M. Zhang, Y. Liu, and Q. Liu, 'Modeling lexical cohesion for document-level machine translation', in *International Joint Conference on Artificial Intelligence*, pp. 2183–2189, (2013).
[30] Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang, 'Modeling coherence for discourse neural machine translation', in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7338–7345, (2018).
[31] Xia Yingce, Tian Fei, Wu Lijun, Lin Jianxin, Qin Tao, Yu Nenghai, , and Liu Tie-Yan, 'Deliberation networks: Sequence generation beyond one-pass decoding', in *In Advances in Neural Information Processing Systems*, pp. 1784–1794, (2017).
[32] Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Cheng Yong, Maosong Sun, Huanbo Luan, and Liu Yang, 'Thumt: An open source toolkit for neural machine translation', *CoRR, abs/1706.06415*, (2017).
[33] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Fei Fei Zhai, and Yang Liu, 'Improving the transformer translation model with document-level context', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 533–542, (2018).
[34] Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng, 'Document embedding enhanced event detection with hierarchical and supervised attention', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 414–419, (2018).