Effective Use of Target-side Context for Neural Machine Translation

Hideya Mino [1,2] **Hitoshi Ito** [1] **Isao Goto** [1] **Ichiro Yamada** [1] **Takenobu Tokunaga** [2]
[1] NHK Science & Technology Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo 157-8510, Japan
{mino.h-gq,itou.h-ce,goto.i-es,yamada.i-hy}@nhk.or.jp
[2] Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
take@c.titech.ac.jp

Abstract

Through the progress made in a sentence-level neural machine translation (NMT), a context-aware NMT has been rapidly developed to exploit previous sentences as context. Recent work in the context-aware NMT incorporates source- or target-side contexts. In contrast to the source-side context, the target-side context causes a gap between training that utilizes a ground truth sentence and inference using a machine-translated sentence as context. This gap leads to translation quality deteriorating because the translation model is trained with only the ground truth data that cannot be used in the inference. In this paper, we propose sampling both the ground truth and the machine-translated previous sentences of the target-side for the context-aware NMT. The proposed method can make the translation model robust against mistakes and biases made at the inference. Models using our proposed approach show improvements over models using the previous approaches in English ↔ Japanese and English ↔ German translation tasks.

1 Introduction

Neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) has made great progress in translating sentences in isolation. In contrast to isolated sentences, sentences in documents cannot be correctly translated without context outside the current sentence (Läubli et al., 2018). To address this problem, various context-aware NMT models have been developed to exploit preceding and/or succeeding sentences in source- and/or target-side languages as context (Wang et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Voita et al., 2019; Maruf and Haffari, 2018; Maruf et al., 2019; Agrawal et al., 2018; Kuang et al., 2018; Miculicich et al., 2018). Bawden et al. (2018) used NMT models with source- and target-side contexts in the previous sentence, and reported that the source-side context was effective for improving the translation quality. However, the target-side context led to lower quality even though they highlighted the importance of the target-side context. Agrawal et al. (2018) also pointed out that the use of the target-side context increases the risk of error propagation. We considered that the reason for the low quality was a gap between a training and an inference when using target-side context in the model. At the training phase, the ground truth sentences (references) were used as the target-side context, and at the inference phase, the sentences predicted by a translation model were used. The predicted sentence includes translation errors such as mistranslation, under-translation, and over-translation not included in the ground truth sentence. Even if there are no translation errors in the predicted sentence, there is a bias of translationese that is outputs of machine translation tend to be simpler and more standardized than human-translated texts (Toral, 2019). Therefore, the predicted target-side sentence tends to have lower diversity than the ground truth sentence. This different use, called exposure bias (Bengio et al., 2015; Ranzato et al., 2016), led to lower translation quality. The following example, extracted from the IWSLT2017 Japanese-English dataset (Cettolo et al., 2012), presents the previous target-side ground truth and predicted sentences (ground truth context and predicted context) in addition to the source sentence and target sentences (reference and vanilla NMT output).

Proceedings of the 28th International Conference on Computational Linguistics, pages 4483–4494
Barcelona, Spain (Online), December 8-13, 2020

Ground truth context: She lays eggs, she feeds the larvae – so an ant starts as an egg, then it's a larva.
Predicted context: And when they get their eggs, they get their eggs, and the queen is there.
Source sentence: 脂肪を吐き出して幼虫を育てます
Reference: She feeds the larvae by regurgitating from her fat reserves.
Vanilla NMT output: They take fat and they raise the larvae of their larvae.

This example indicates two things: (a) the context information, which is a previous target-side sentence in this example, is effective to translate the source sentence correctly, and (b) the ground truth context and predicted context are different. First, we can find that both the ground truth context and the predicted context include significant information (that is "she" for the ground truth context and "queen" for the predicted context) for a correct translation, where the subject of the source is "she." The vanilla NMT trained without the contexts fails to translate the subject of the sentence shown in the vanilla NMT output of the above example. Second, the predicted context includes some translation errors of mistranslation, under-translation, and over-translation not included in the ground truth context. This means that there is the exposure bias.

In this paper, we propose a controlled sampling with both the ground truth sentence and the predicted sentence of the target-side context to alleviate the gap between the training and the inference, inspired by the scheduled-sampling method (Bengio et al., 2015). At the first epoch, we train the translation model with only the ground truth contexts, without the predicted contexts. Then we change the training data for each epoch in order to gradually force the translation model to deal with noise. The predicted contexts are generated by the vanilla NMT model. We implement the proposed method on a concatenation-based context-aware NMT and a multi-encoder context-aware NMT. Experimental results on English $\leftrightarrow$ Japanese translation tasks in the News corpus and IWSLT2017 dataset (Cettolo et al., 2012) and English $\leftrightarrow$ German translation tasks in the IWSLT2017 dataset show that our proposed method can improve on the previous model in terms of bilingual evaluation understudy (BLEU) scores.

2 Context-aware NMT

In this section, we introduce two existing techniques for context-aware NMT that have been actively developed in recent work.

2.1 Concatenation-based Context-aware NMT

This model concatenates the current sentence and the previous sentence with a special token (such as $\_BREAK\_$) as described in (Tiedemann and Scherrer, 2017). They proposed two methods in the inference phase that translate both the previous and the current source sentence (2-TO-2) and only the current sentence (2-TO-1). The 2-TO-2 model outputs the special token between the previous and current translated sentence that shows the concatenated position. Hence, the translation result of the current sentence can be obtained by extracting tokens following the special token and discarding preceding tokens.

2.2 Multi-encoder Context-aware NMT

A multi-encoder context-aware NMT models encode the previous sentence and the current sentence separately, and the two generated vectors are combined to form a single vector for the input to the decoder. In this model, there are two types of options: where to integrate and how to integrate. First, there are two choices for "where to integrate": outside or inside the decoder. We used the approach of the integration inside the decoder with a parallel attention because it has the best reported results as described in (Kim et al., 2019). This method relates the context to the target history independently of the current source-side sentence and makes the decoding speed faster. Second, there are three choices for "how to integrate" as shown in (Bawden et al., 2018): an attention concatenation, an attention gate or a hierarchical attention. We used the attention gate because it achieved the best results in exploratory experiments. The attention gate makes a single vector c with the previous sentence vector c_r and the current sentence vector c_s as the below equations:

$$g = \tanh(W_r c_r + W_s c_s) + b_g, \tag{1}$$

$$c = g \odot (W_t c_r) + (1 - g) \odot (W_u c_s), \tag{2}$$

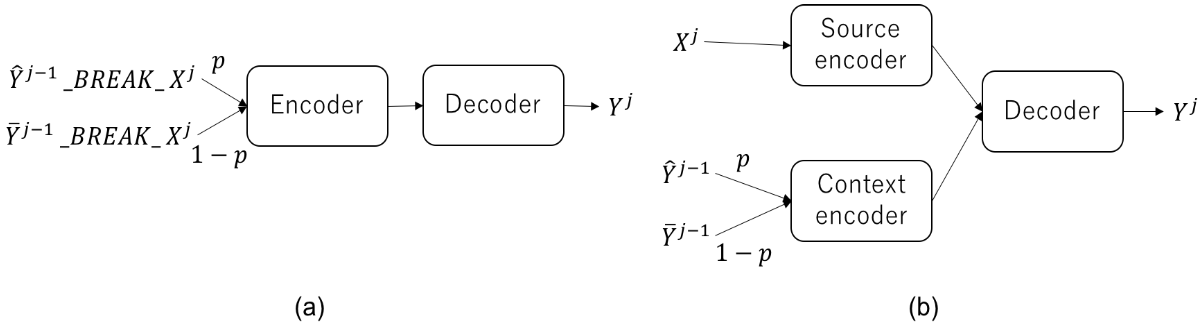

Figure 1: The architecture of our method with a single encoder. (a) Concatenate the previous sentence and the current sentence with a special token (such as $\_BREAK\_$) and multi-source encoder. (b) Combine the outputs from the encoders with a simple projection and sum.

where b_g is a bias vector, W_r, W_s, W_t, W_u are weight matrices, and $\odot$ is an element-wise multiplication. A gate g is learnt from two sentence vectors in order to give differing importance to the elements of each vector.

3 Proposed Method

We propose a sampling method that can be applied to the context-aware NMTs in the previous section to alleviate the gap between the training and inference phases in context-aware translation tasks with target-side context.

3.1 Model

The approach of the method is to train the target-side context by using both ground truth and predicted sentences in encoders of an NMT architecture. Figure 1 shows two types of architectures to implement our model: the concatenation-based context-aware NMT and the multi-encoder context-aware NMT in the previous section. Let $X^j = (x_1^j, ..., x_L^j)$ be the j-th source sentence in a document, $\hat{Y}^{j-1} = (\hat{y}_1^{j-1}, ..., \hat{y}_M^{j-1})$ be the $(j-1)$-th ground truth sentence, and $\bar{Y}^{j-1} = (\bar{y}_1^{j-1}, ..., \bar{y}_N^{j-1})$ be the $(j-1)$-th predicted sentence. L, M, and N show the number of tokens of each sentence. The predicted sentences are generated with a non-context-aware NMT beforehand. Let p be a sampling rate to use either the ground truth sentence $\hat{Y}^{j-1}$ or the predicted sentence $\bar{Y}^{j-1}$. When $p = 1$, the model is trained with only the ground truth sentences as the target-side context, while when $p = 0$, the model is trained with only the predicted sentences. During the inference, the previous predicted sentence is always inputted, i.e., $p = 0$. Then, for the architecture in Figure 1 (a), the vector outputted by the source encoder is inputted to the decoder. The decoder outputs a translation result Y^j of only the current source sentence. For the architecture of Figure 1 (b), the two resulting vectors outputted by the source and context encoders are combined to form a single vector, and a target-side sentence Y^j is predicted through the decoder.

3.2 Controlled Sampling for the Context-aware NMT

The appropriate sampling rate p of the training phase, whether we use the predicted sentence $\bar{Y}^{j-1}$ or the ground truth sentence $\hat{Y}^{j-1}$, depends on a task. In the case of training with only the ground truth sentence ($p = 1$), a translation model can be trained effectively because the model can concentrate on training to acquire the information for predicting the next sentence. Intuitively, it is appropriate to use either the ground truth sentence $\hat{Y}^{j-1}$ or the predicted sentence $\bar{Y}^{j-1}$ as the target-side context during the training when the predicted sentences are high enough quality to acquire the necessary information for predicting the next sentence. However, since the predicted sentences contain noise like mistranslation, under-translation, and over-translation, a translation model has to deal with not only the acquisition of the information but also its own mistakes (noise) in the target-side context during the inference phase. Furthermore, even if there is no noise in the predicted sentence, there is a bias of translationese. Outputs of

Corpus name	Language pair	Train	Development	Test
News	English-Japanese	220,180	2,000	2,000
TED Talk	English-Japanese	194,170	879	1,285
	English-German	203,998	888	1,080

Table 1: Number of sentences in each dataset.

machine translation tend to be simpler and more standardized than original human-translated texts (Toral, 2019). Therefore, predicted sentences tend to have lower diversity than ground truth sentences. To train both target-side sentences with different features effectively, we propose a controlled sampling method inspired by curriculum learning approach (Bengio et al., 2009) and scheduled sampling (Bengio et al., 2015). The idea is to change the training data for each epoch in order to gradually force the translation model to deal with noise. Our method involves the following steps:

1. At the beginning of training, the initial sampling rate $p_1 = 1$ at which the translation model is totally trained with the ground truth sentence is used.

2. The value of p_e is updated every epoch and strictly decreased via the inverse sigmoid decay of the following equation:

$$p_e = \frac{k}{k + \exp(e - 2/k)}, \tag{3}$$

where $e = 2, 3, ...$ is the number of training epochs, which start from 2. k is a hyper-parameter to control the speed of convergence where $k \geq 1$ depends on the baseline translation quality. The higher the translation quality, the smaller the value of k (set near to one). Under the above procedure, the model learns to handle the noise and be more robust against the noise in the inference.

4 Experiments

To evaluate our method, we experimented on two document-level parallel datasets in two language pairs for machine translation.

4.1 Data

We used two parallel corpora as follows.

- **News (Japanese $\leftrightarrow$ English).** We used a content-equivalent translated news corpus (Mino et al., 2020). The corpus was low-noise parallel data made by manually translating Japanese news articles into English in a content-equivalent manner. We removed news titles from development and test sets.

- **TED Talk (Japanese $\leftrightarrow$ English, English $\leftrightarrow$ German).** We used the IWSLT 2017 (Cettolo et al., 2012) datasets based on the TED Talks where each talk is considered a document. We used the "train" set for training and the "dev2010" set for validation for all the tasks. The "tst2014" was used for testing English $\leftrightarrow$ Japanese tasks and the "tst2015" was used for testing English $\leftrightarrow$ German tasks.

Table 1 shows the statistics of each corpus.

4.2 Systems

Our method can be implemented to various types of context-aware NMT systems. In this paper, we applied our method to the concatenation-based context-aware NMT model (Figure 1 (a)) and the multi-encoder context-aware NMT model (Figure 1 (b)). Though various numbers of sentences can be utilized as context in each model, we used only one previous sentence for each model to reduce memory consumption.

4.3 Settings

We used the Moses toolkit[1] to clean and tokenize the English and German data and KyTea (Neubig et al., 2011) to tokenize the Japanese data. Then, we used a vocabulary of 48K units based on a joint byte-pair encoding (BPE) (Sennrich et al., 2016) for the source and target. The frequency threshold for the vocabulary filter was set to 35.

For the concatenation-based context-aware NMT, we used the encoder and decoder of the transformer model (Vaswani et al., 2017), which was a state-of-the-art NMT model. The transformer model contains a multi-headed attention mechanism, applied as self-attention, and a position-wise fully connected feed-forward network. The encoder converted the received source-language sentence into a sequence of continuous representations, and the decoder generated the target-language sentence. We implemented this system with the Sockeye toolkit (Hieber et al., 2018). For the multi-encoder context-aware NMT, we also used the transformer model and implemented it by modifying the architecture as shown in (Littell et al., 2019). All models were trained on an Nvidia P100 Tesla GPU. In training each model, we applied stochastic gradient descent (SGD) with Adam (Kingma and Ba, 2015) as the optimizer, using a learning rate of 0.0002, multiplied by 0.7 after every eight checkpoints. We set the batch size to 5000 tokens and the maximum sentence length to 200 BPE units for the concatenation-based context-aware NMT and to 100 BPE units for the multi-encoder context-aware NMT. For the other hyperparameters of the models, we used the default Sockeye parameter values. We applied early stopping with a patience of 32. Decoding was performed through beam search with a beam size of 5, and we did not apply ensemble decoding with multiple models, although this could have improved the translation quality. The hyperparameter k of our proposed method in Equation (3) was set to 2.

To evaluate the translation quality, we trained five models with different seeds and used the median BLEU score of the five translation results. We calculated case-insensitive BLEU (Papineni et al., 2002) scores by using multi-bleu.perl[2].

4.4 Baseline Methods

To measure the effectiveness of our proposed approach, we consider the following baselines.

- **Sentence-level translation (Single-Sentence).** To compare translations with and without context information beyond the current sentence, we used a single sentence translation as the baseline: non-context-aware translation. The Single-Sentence method was implemented using the transformer model.

- **Context-aware translation with a source-side previous sentence (Src-Context), a target-side previous ground truth sentence (Trg-Context GT), and a target-side predicted sentence (Trg-Context Pred.)** In the case of the context-aware translation, two types of sentences can be used as context: source-side sentences and target-side sentences. To investigate which type of context is effective, we separately used source-side sentences and target-side sentences as context. For the target-side sentences, since either the ground truth sentences or the predicted sentences can be used, we used both for the baseline methods. As a result, we compared six types of baseline methods for the context-aware translation: three concatenation-based context-aware NMT systems with Src-Context (Bawden et al., 2018; Agrawal et al., 2018), Trg-Context GT, and Trg-Context Pred and three multi-encoder context-aware NMT systems with Src-Context and Trg-Context GT (Bawden et al., 2018), and Trg-Context Pred respectively. All baseline methods of the context-aware translation were implemented to the concatenation-based context-aware NMT and the multi-encoder context-aware NMT the same as the proposed method.

4.5 Results

We list the experimental results of English↔Japanese tasks with News and TED Talk corpora and English↔German tasks with TED Talk corpus in Tables 2 and 4. First, we show the experimental results

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

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

		Context in training							
		Trg	Trg	News		TED Talk			
Method	NMT	GT	MT	EN-JA	JA-EN	EN-JA	JA-EN	EN-DE	DE-EN
Single-Sentence	vanilla			41.99	24.23	15.69	8.48	24.04	28.46
Trg-Context GT	concat.	✓		41.45	24.40	14.72	7.55	25.09	29.11
Trg-Context Pred.	concat.		✓	42.07	23.83	15.32	7.37	23.99	29.05
Proposed method (Fig. 1 (a))	concat.	✓	✓	42.89	24.62	15.33	8.36	25.09	29.74
Trg-Context GT	multi-enc.	✓		42.40	24.80	16.15	8.98	25.96	30.25
Trg-Context Pred.	multi-enc.		✓	42.45	24.31	16.27	9.15	25.98	30.23
Proposed method (Fig. 1 (b))	multi-enc.	✓	✓	42.79	24.86	16.37	9.66	26.01	30.60

Table 2: English↔Japanese (News and TED Talk) and English↔German (TED Talk) translation results of the comparison between the target-side context-aware NMTs: vanilla is the non-context-aware transformer NMT, concat. is the concatenation-based context-aware NMT, and multi-enc. is the multi-encoder context-aware NMT.

using our proposed method and the other target-side context-aware NMTs that are Trg-Context GT and Trg-Context Pred. Then we show the experimental results using source-side context-aware NMTs that is Src-Context to compare the effectiveness between the source-side context and the target-side context.

Effectiveness of our proposed method in News and TED Talk corpora. Table 2 shows the results of the context-aware NMTs using the target-side context. In the experimental results with the News corpus, the context-aware NMTs using the target-side context improved the translation performance compared with the vanilla NMT. There are no clear differences between the concatenation-based context-aware NMT and the multi-encoder context-aware NMT. For the TED corpus, the multi-encoder context-aware NMT achieved better results than the concatenation-based context-aware NMT. Both our proposed methods improved the translation quality compared with the baseline methods: Single-Sentence, Trg-Context GT, and Trg-Context Pred. To evaluate the effectiveness of our proposed method in a different domain, we experimented with the English↔Japanese dataset of another corpus: TED Talk corpus. The experimental results of the multi-encoder context-aware NMTs show they are superior at exploiting the context, whereas the concatenation-based context-aware NMT did not improve. Thus, our proposed method of the multi-encoder context-aware NMT improved the translation quality compared with the baseline methods in the English↔Japanese task in different domains.

Effectiveness in a different language pair. We experimented using English↔Japanese and English↔German datasets of TED Talk corpus to determine the tendencies of the translation quality on different language pairs. For the both language pairs, our proposed method of the multi-encoder context-aware NMTs achieved better results than all the baseline systems. Thus, our proposed method was found to be effective in English-Japanese and English-German language-pairs.

Results using the ground truth context for inference. Recent work of context-aware NMT exploiting the target-side context, including this paper, uses the predicted sentence in the inference phase. To investigate the upper bound scores using the target-side context, we used the ground truth sentence in both the training and the inference. Table 3 shows the translation results when the ground truth sentence is used in the inference phase of the Trg-Context GT method. Our proposed method achieved comparable results to the Trg-Context GT method with the ground truth sentence.

Comparison between the source-side context and target-side context. Table 4 shows the results of the source-side context-aware NMTs (Src-Context). Our proposed method achieved comparable results to the Src-Context when using the multi-encoder context-aware NMT. As a result, we concluded that the target-side context improves the document-level translation quality as effectively as the source-side context does. Since the source-side context-aware NMT and the target-side context-aware NMT require different resources (the source-side and target-side contexts, respectively), the target-side context-aware NMT does not compete against the source-side context-aware NMT. We consider that it is important to improve the target-side context-aware NMT without requiring the resources of source-side context.

	News		TED Talk			
Method	EN-JA	JA-EN	EN-JA	JA-EN	EN-DE	DE-EN
Single-Sentence	41.99	24.23	15.69	8.48	24.04	28.46
Trg-Context GT with the predicted sent.	42.40	24.80	16.15	8.98	25.96	30.25
Trg-Context GT with the ground truth sent.	42.50	25.09	16.39	9.51	25.98	30.33
Proposed method	42.79	24.86	16.37	9.66	26.01	30.60

Table 3: Translation results of the multi-encoder context-aware translation system trained with a target-side previous ground truth context when the ground truth context is used in the inference phase (Trg-Context GT with the ground truth sent.) and when the predicted sentence is used in the inference phase (Trg-Context GT with the predicted sent.).

		Context in training								
			Trg	Trg	News		TED Talk			
Method	NMT	Src	GT	MT	EN-JA	JA-EN	EN-JA	JA-EN	EN-DE	DE-EN
Single-Sentence	vanilla				41.99	24.23	15.69	8.48	24.04	28.46
Src-Context	concat.	✓			42.87	24.40	15.17	8.79	25.96	30.03
Proposed method (Fig. 1 (a))	concat.		✓	✓	42.89	24.62	15.33	8.36	25.09	29.74
Src-Context	multi-enc.	✓			42.06	24.37	15.58	9.33	25.97	29.55
Proposed method (Fig. 1 (b))	multi-enc.		✓	✓	42.79	24.86	16.37	9.66	26.01	30.60

Table 4: English↔Japanese (News and TED Talk) and English↔German (TED Talk) translation results of the comparison between the use of source and target contexts: vanilla is the non-context-aware transformer NMT, concat. is the concatenation-based context-aware NMT, and multi-enc. is the multi-encoder context-aware NMT.

Examples. We show the output examples of the Japanese-to-English task with the TED Talk corpus in Table 5. "The previous ground-truth target sentence" and "The previous predicted target sentence" show sentences used as context information. "The current sentence" shows a source sentence to be translated. The multi-encoder context-aware NMTs were used for the four context-aware methods. Japanese sentences of the examples include zero anaphora, where pronouns can be omitted when they are pragmatically or grammatically inferable from intra- and inter-sentential context (Okumura and Tamura, 1996; Iida et al., 2016). The first example (#1) is the simplest case. The previous and the current sentences are same for the English-side ("We choose to go to the moon."). On the other hand, for the Japanese-side, the subject pronoun of the current sentence ("we") is omitted as zero anaphora. Due to zero anaphora, the Single-Sentence NMT failed to translate the current sentence. In contrast, the other four context-aware NMTs succeeded in translating the current sentence. From this example, it appears that the context-aware NMTs try to resolve the zero anaphora and are effective to translate documents correctly. The same as the first example, since the pronoun ("him") is omitted in the current sentence of the second example (#2), the sentence is not translated correctly with the Single-Sentence NMT. Among the context-aware NMTs, while the Src-Context and our proposed method NMTs translate the pronoun correctly, the Trg-Context GT and Trg-Context Pred. NMTs mistranslate even though a clue to predict the pronoun is contained in the previous predicted target contexts. It appears that Trg-Context GT and Trg-Context Pred. NMTs could not encode the target-side context appropriately whereas our proposed method could exploit the target-side context. The third example (#3) is the case where only the proposed method can predict the pronoun ("his") correctly.

5 Related Work

NMT models exploiting context beyond the current sentence are an active research area. Bawden et al. (2018), Läubli et al. (2018), Müller et al. (2018), and Voita et al. (2018) showed that the document-level

	Input and ref. (Upper side) and method (Lower side)	Sentences
#1	The previous ground truth target sentence	**We** choose to go to the moon.
	The previous predicted target sentence	**We** decided to go to the moon.
	The current sentence	月に行くことにしたんです
	Reference	**We** choose to go to the moon.
	Single-Sentence	**I** went to the moon.
	Src-Context	**We** decided to go to the moon.
	Trg-Context GT	**We** decided to go to the moon.
	Trg-Context Pred.	**We** decided to go to the moon.
	Proposed method	**We** decided to go to the moon.
#2	The previous ground truth target sentence	On the day after September 11 in 2001, I heard the growl of a sanitation truck on the street, and I grabbed my infant son and I ran downstairs and there was a man doing his paper recycling route like **he** did every Wednesday.
	The previous predicted target sentence	On September 11, 2001, I heard a young man flying in a city, and he went down, and a man came down the stairs, he was going down the stairs, he was going down a piece of paper, and **he** was doing it every Wednesday.
	The current sentence	あの日に一日中仕事をしてくれていることに感謝しようとしました でも涙が出てきたのです
	Reference	And I tried to thank **him** for doing his work on that day of all days, but I started to cry.
	Single-Sentence	And I was trying to appreciate the fact that I was doing my work every day, but tears were coming out.
	Src-Context	And on that day, I decided to thank **him** for doing my work, but I came up with tears.
	Trg-Context GT	And I was going to thank you for that day, and I was going to thank you for doing my work every day, but I got to tears.
	Trg-Context Pred.	And on that day, I thank you very much for doing everything in that day, but tears come up and tears come up.
	Proposed method	And I was grateful for that day, and I tried to thank **him** for doing work all the day, but I had tears in my tears.
#3	The previous ground truth target sentence	I would like to talk to you about a story about a small town **kid**.
	The previous predicted target sentence	I'm going to talk about **a boy** in a village.
	The current sentence	名前は知りませんが彼の話はできます
	Reference	I don't know **his name**, but I do know his story.
	Single-Sentence	I don't know **the name**, but I can tell you his story.
	Src-Context	You don't know **the name**, but you can talk to him.
	Trg-Context GT	I don't know **the name**, but I can tell him.
	Trg-Context Pred.	He doesn't know **the names**, but he can tell the story.
	Proposed method	I don't know **his name**, but I can tell you his story.

Table 5: The output examples using the multi-encoder context-aware NMT in the Japanese→English task of the TED Talk corpus: the upper-side shows the input and reference sentences and the lower-side shows outputs of each method.

context helps to maintain lexical, tense, deixis, and ellipsis consistencies in the current sentence. Furthermore, the context can resolve anaphoric pronouns and other discourse characteristics. Most context-aware NMT systems are modified to take additional source- and/or target-side context as their input. Recent work in the context-aware NMT can be classified into multi-input models and multi-encoder models. Tiedemann and Scherrer (2017) proposed concatenating the previous and the current source sentences with a special token (such as _*BREAK*_) as multi-input to a translation model. Bawden et al. (2018) extended multi-encoder NMT systems (Zoph and Knight, 2016; Libovický and Helcl, 2017; Wang et al., 2017) to exploit the previous source sentence as the context. These systems combine the information from the current sentence and the context of the source-side with either concatenation, gating or hierarchical attention. Kim et al. (2019) investigated the difference in a multi-input model and two types of multi-encoder NMT models, where the first method is an integration outside the decoder, which combines encoder representations of all input sentences before being fed to the decoder, and another method is to integrate each encoder representation inside the decoder.

In addition to the use of the source-side previous sentences as context, the use of the target-side previous sentences was also studied. Agrawal et al. (2018) proposed the concatenating translation model with up to two previous sentences for the target-side in addition to up to three previous sentences and one next source sentence for the source-side. They reported that a large number of previous target-side sentences deteriorates performance because of error propagation. Though using various types of the context such as both source and target previous sentences may improve the translation quality, it tends to require a huge memory and a long computational time. Bawden et al. (2018) applied the use of the target-side context to the multi-encoder model, though they reported deteriorated BLEU scores. They used the previous reference during training, and the previous output of a baseline translation model during inference as target-side context. In this paper, we focus mainly on the effective use of the target-side context. Recent work on the use of target-side sentences as context shows the importance of using target-side context. However, the improvement in translation quality is not stable.

The main idea of our work is borrowed from a scheduled sampling (Bengio et al., 2015). Bengio et al. (2015) proposed the schedule sampling method to solve an exposure bias problem, which refers to a gap between maximum likelihood estimation training and inference for an auto-regressive language model. Zhang et al. (2019) proposed a sampling method with decay to alleviate the exposure bias in a recurrent neural network (RNN) based NMT model. They showed that the method to feed either the ground truth or the previous predicted words as context effectively reduced the gap between training and inference. We extended this scheduled-sampling idea to the context-aware NMT where the model was trained with the ground truth previous sentence of the target-side and the model generates translation results with the predicted previous sentence of the target-side.

6 Conclusion

In this work, we have presented a controlled sampling method for a context-aware neural machine translation (NMT) using target-side previous ground-truth and predicted sentences as context. Our aim is to tackle the exposure bias of the target-side context between the training and the inference on the basis of the scheduled-sampling method. We implemented and evaluated the proposed method on two types of context-aware NMT systems: concatenation-based context-aware and a multi-encoder context-aware NMTs. The experimental results verified the effectiveness of our proposed method in terms of BLEU.

Acknowledgements

The research results have been achieved by “Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation,” the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 11–20, Alacant, Spain. European Association for Machine Translation.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June. Association for Computational Linguistics.

Yoshua Bengio, Jrme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In Andrea Pohoreckyj Danyluk, Lon Bottou, and Michael L. Littman, editors, *ICML*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1171–1179. Curran Associates, Inc.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 200–207, Boston, MA, March. Association for Machine Translation in the Americas.

Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. 2016. Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1244–1254, Austin, Texas, November. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 24–34.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. 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*, pages 4791–4796, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada, July. Association for Computational Linguistics.

Patrick Littell, Chi-kiu Lo, Samuel Larkin, and Darlene Stewart. 2019. Multi-source transformer for Kazakh-Russian-English neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 267–274, Florence, Italy, August. Association for Computational Linguistics.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia, July. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Hideya Mino, Hideki Tanaka, Hitoshi Ito, Isao Goto, Ichiro Yamada, and Takenobu Tokunaga. 2020. Content-equivalent translated parallel news corpus and extension of domain adaptation for nmt. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3616–3622, Marseille, France, May. European Language Resources Association.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium, October. Association for Computational Linguistics.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA, June. Association for Computational Linguistics.

Manabu Okumura and Kouji Tamura. 1996. Zero pronoun resolution in Japanese discourse based on centering theory. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark, September. Association for Computational Linguistics.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 273–281, Dublin, Ireland, August. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia, July. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China, November. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark, September. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy, July. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California, June. Association for Computational Linguistics.