Selecting Informative Context Sentence by Forced Back-Translation

Ryuichiro Kimura[†], **Shohei Iida**[†], **Hongyi Cui**[†], **Po-Hsuan Hung**[†],
Takehito Utsuro[†] **and Masaaki Nagata**[‡]
[†]Graduate School of Systems and Information Engineering, University of Tsukuba, Japan
[‡]NTT Communication Science Laboratories, NTT Corporation, Japan

Abstract

As one of the contributions of this paper, this paper first explores the upper bound of context-based neural machine translation and attempt to utilize previously unused context information. We found that, if we could appropriately select the most informative context sentence for a given input source sentence, we could boost translation accuracy as much as approximately 10 BLEU points. This paper next explores a criterion to select the most informative context sentences that give the highest BLEU score. Applying the proposed criterion, context sentences that yield the highest forced back-translation probability when back-translating into the source sentence are selected. Experimental results with Japanese and English parallel sentences from the OpenSubtitles2018 corpus demonstrate that, when the context length of five preceding and five subsequent sentences are examined, the proposed approach achieved significant improvements of 0.74 (Japanese to English) and 1.14 (English to Japanese) BLEU scores compared to the baseline 2-to-2 model, where the oracle translation achieved upper bounds improvements of 5.88 (Japanese to English) and 9.10 (English to Japanese) BLEU scores.

1 Introduction

Recently, neural machine translation (NMT) models (Sutskever et al., 2014; Luong et al., 2015; Vaswani et al., 2017) have made remarkable progress. Most NMT models are designed to translate a single sentence and do not accept input greater than one sentence, i.e., input sentences that include additional context information. However, recently, several approaches that attempt to translate inputs with more than one sentence have been proposed (Tiedemann and Scherrer, 2017; Libovický and Helcl, 2017; Maruf and Haffari, 2018; Miculicich et al., 2018; Bawden et al., 2018; Voita et al., 2018; Tu et al., 2018). These approaches to context-based NMT models can be roughly categorized according to the width of the context considered in those models. A typical approach is to consider the sentence immediately preceding the source sentence to be translated as the context (Tiedemann and Scherrer, 2017; Libovický and Helcl, 2017; Bawden et al., 2018; Voita et al., 2018). Context-based NMT models can be further categorized according to whether the source and context sentences are encoded using a single (Tiedemann and Scherrer, 2017) or multiple encoders (Libovický and Helcl, 2017; Bawden et al., 2018; Voita et al., 2018). Another approach considers a much wider context than the immediately preceding sentence, e.g., three preceding sentences (Miculicich et al., 2018), preceding sentences within the document (Tu et al., 2018), and all preceding and subsequent sentences within the document (Maruf and Haffari, 2018).

Such approaches to context-based NMT models possibly outperform existing models that only accept a single sentence to be translated. Note that we refer to the model that only accepts a single sentence as a "1-to-1" model. Among these existing models, the 2+2 or 2-to-2 model (Tiedemann and Scherrer, 2017) uses the sentence immediately preceding the source sentence to be translated as

	BLEU		Oracle BLEU	
	Ja-En	En-Ja	Ja-En	En-Ja
1-to-1 (baseline)	15.52	11.48	—	—
2-to-2 (baseline)	16.52	12.36	—	—
selection from 20-best of 2-to-2 (baseline) by 2-to-2 back-translation	16.69 / —	12.61 / —	—	—
1-to-1 + 2-to-2 (immediately preceding sent.)	17.04** / 16.51	13.24** / 12.47	18.15	15.61
1-to-1 + 2-to-2 (1st $\sim$ 5th preceding sents.)	17.25** / 16.67	13.50** / 13.14**	21.09	19.55
1-to-1 + 2-to-2 (1st $\sim$ 5th subsequent sents.)	17.04** / 16.52	13.46** / 13.13**	20.84	19.51
1-to-1 + 2-to-2 (1st $\sim$ 5th preceding + subsequent sents.)	17.26** / 16.68	13.02** / 12.81**	22.40	21.46

Table 1: Evaluation results (maximizing forced back-translation probability / maximizing back-translation sentence-BLEU) ($**$ represents significant difference ($p < 0.01$) against baseline 2-to-2 model)

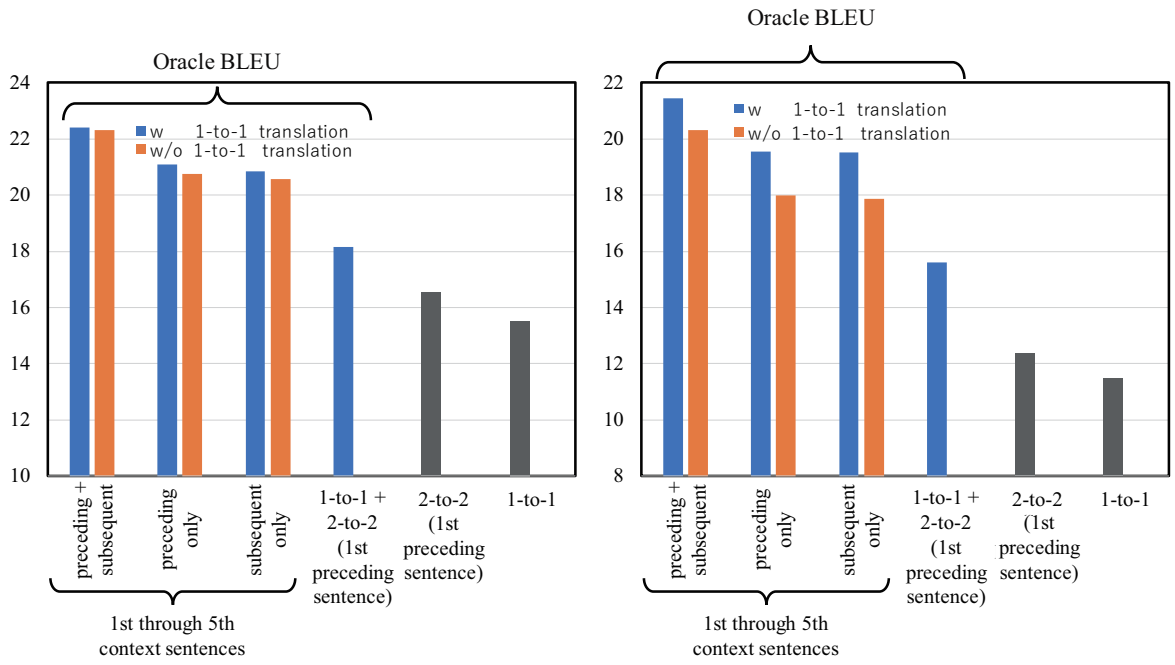
an extended context. Here the context sentence is concatenated to the source sentence using the ⟨CONCAT⟩ token. The 2-to-2 model is easy to implement into existing 1-to-1 models: however, it only considers the immediately preceding sentence as context. Thus, it is necessary to consider much wider contexts such as the second through fifth preceding sentences and the first through fifth subsequent sentences. We conducted an empirical study that revealed that, in some cases, among the first through fifth sentences preceding and subsequent to the source sentence, the most informative sentence, i.e, the sentence that returns the highest BLEU score, may not be the sentence immediately preceding the source sentence. We measured oracle BLEU scores by selecting context sentences that give the maximum sentence-BLEU scores among the five preceding and subsequent sentences, as shown in Table 1 and Figure 1. Then, we found that, if we could select the most informative context sentence for a given input source sentence, we can improve translation accuracy by as much as approximately 10 BLEU points, as indicated by the oracle BLEU scores in Table 1 and Figure 1. More specifically, compared to the baseline 2-to-2 model, the oracle translation achieved upper bound improvements of 5.88 (Japanese to English) and 9.10 (English to Japanese) BLEU scores.

Considering this result, within the framework of the 2-to-2 context based NMT model, this study explored how to select the most informative context sentences that give the highest BLEU score among the first five preceding and subsequent sentences[1]. Here, we used the Transformer model (Vaswani et al., 2017) as the base 1-to-1 model. To select the translation with the highest BLEU score among the 11 translations (i.e., those translated by the 1-to-1 and 10 2-to-2 models), we propose an approach that selects the translation that yields the highest forced back-translation probability when back-translating into the source sentence. The evaluation results shown in Table 1 demonstrate that the proposed approach achieves significant BLEU score improvements over the baseline 2-to-2 and 1-to-1 models. More specifically, over the baseline 2-to-2 model, the proposed approach achieved significant improvements of 0.74 (Japanese to English) and 1.14 (English to Japanese) BLEU scores.

2 Selective Extended Context Decoding

Tiedemann and Scherrer (2017) proposed the 2-to-2 model, which uses the sentence immediately preceding the source sentence to be translated as the extended context. We extend the 2-to-2 model by considering the first five preceding and first five subsequent sentences. In our extended 2-to-2 context-based NMT model, the immediately preceding sentence, the second through fifth preceding sentences, and the first through fifth subse-

[1] An obvious alternative to this approach is to simply employ 3-to-3 (or more) models using an approach similar to the 2-to-2 model that concatenates context sentences using the ⟨CONCAT⟩ token. However, due to the upper bound restriction of GPU memory, it is impractical to employ such 3-to-3 (or more) models. Furthermore, our preliminary evaluation result also indicates that the 3-to-3 model underperforms compared to the proposed approach.



(a) Ja-En (b) En-Ja

Figure 1: Oracle BLEU and BLEU scores of baseline 2-to-2 ($\boldsymbol{y_0^{22}(x_{-1}, x_0)}$) and 1-to-1 ($\boldsymbol{y_0^{11}(x_0)}$) models

quent sentences are considered candidates for concatenation to the source sentence. Then, among the first through fifth preceding and first through fifth subsequent sentences, we select the most informative context sentences in the 2-to-2 context-based NMT model.

In this framework, this paper employs the notation described below. $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ $(i = 0, \pm1, \ldots, \pm5)$ denote the source and target sentences, respectively. $\boldsymbol{x_0}$ denotes the source sentence to be translated, while $\boldsymbol{y_0}$ denotes its translation in the target language. $\boldsymbol{x_{-1}}$ denotes the context sentence in the source language immediately preceding $\boldsymbol{x_0}$, $\boldsymbol{x_i}$ $(i = -2, \ldots, -5)$ the second through fifth preceding sentences, and $\boldsymbol{x_i}$ $(i = +1, \ldots, +5)$ the first through fifth subsequent sentences. In order to represent the direction of translation as the target language l ($\boldsymbol{x}$ or $\boldsymbol{y}$ in this paper) of the translation, the model m (1-to-1 or 2-to-2) used for the translation, the index i $(i = 0, \pm1, \ldots, \pm5)$ of the translated sentence, and the source sentence s and the context sentence c in the source language, this paper employs a general notation to represent the translated sentence in the target language as below:

$$l_i^m(c, s).$$

Translation by 1-to-1 Model

For example, the target sentence translated from $\boldsymbol{x_0}$ by the base 1-to-1 Transformer model is denoted as

$$\boldsymbol{y_0^{11}(x_0)}.$$

In this case, the source sentence s is $\boldsymbol{x_0}$, and is translated without a context sentence. Table 2 shows a typical Japanese subject zero pronoun case improved by the proposed informative context sentence selection approach by forced back-translation, where the bottom line represents the translation by the base 1-to-1 model. In Table 2, $\boldsymbol{y_0^{11}(x_0)}$, i.e., the translation of $\boldsymbol{x_0}$ by the base 1-to-1 model is:

If we leave now , we'll never get back .

Here, the base 1-to-1 model fails in the translation of the Japanese zero pronoun subject in $\boldsymbol{x_0}$, i.e., it is not translated as "you", but translated as "we".

Translation by Baseline 2-to-2 Model

The target sentence translated from $\boldsymbol{x_0}$ by the baseline 2-to-2 Transformer model which uses the sentence $\boldsymbol{x_{-1}}$ immediately preceding $\boldsymbol{x_0}$ as the extended context is denoted as

$$\boldsymbol{y_0^{22}(x_{-1}, x_0)}.$$

Source sentences		Target sentences		Forced back-translation probability / sentence-BLEU
—		Reference translation:	Walk out now and you may never return .	—
4th preceding sentence $\boldsymbol{x_{-4}}$:	私に逆らうなら お前 は何もなくなるぞ。(If you defy me , you will have nothing .)	Translation $\boldsymbol{y_0^{22}(x_{-4}, x_0)}$ by 2-to-2 model:	If you leave , you'll never get back .	$\mathbf{5.8\times10^{-8}}$ / **14.99**
Immediately preceding sentence $\boldsymbol{x_{-1}}$:	それが望みなのか？ (Is that what you want ?)	Translation $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$ by baseline 2-to-2 model:	If we leave now , we'll never get back .	2.8×10^{-10} / 13.55
Source sentence $\boldsymbol{x_0}$:	出て行けば、戻れなくなるぞ。	Translation $\boldsymbol{y_0^{11}(x_0)}$ by baseline 1-to-1 model:	If we leave now , we'll never get back .	1.3×10^{-8} / 10.57

Table 2: Example improvements over baseline 2-to-2 $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$ (Ja-En) (a) pronoun translation

In this case, $\boldsymbol{x_0}$ is concatenated with the immediately preceding sentence $\boldsymbol{x_{-1}}$ as " $\boldsymbol{x_{-1}}$ ⟨CONCAT⟩ $\boldsymbol{x_0}$", and the concatenated sentences are translated by the baseline 2-to-2 Transformer model. We denote the translated (concatenated) sentences as follows:

$$\boldsymbol{y_{-1}^{22}(x_{-1}, x_0)} \ \langle\text{CONCAT}\rangle \ \boldsymbol{y_0^{22}(x_{-1}, x_0)}$$

where $\boldsymbol{y_{-1}^{22}(x_{-1}, x_0)}$ and $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$ are the translations of $\boldsymbol{x_{-1}}$ and $\boldsymbol{x_0}$, respectively. In the case of Table 2, the immediately preceding sentence $\boldsymbol{x_{-1}}$ and the source sentence $\boldsymbol{x_0}$ are:

$\boldsymbol{x_{-1}}$: それが望みなのか？
(Is that what you want ?)
$\boldsymbol{x_0}$: 出て行けば、戻れなくなるぞ。
(Walk out now and you may never return .)

Then, $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$, i.e., the translation of $\boldsymbol{x_0}$ is:

If we leave now , we'll never get back .

Again, the baseline 2-to-2 model fails in the translation of the Japanese zero pronoun subject in $\boldsymbol{x_0}$, i.e., it is not translated as "you", but translated as "we".

Translation by 2-to-2 Model with a Context Sentence $\boldsymbol{x_{-4}}$

Similarly, the first line of Table 2 also shows the target sentence translated from $\boldsymbol{x_0}$ by the 2-to-2 Transformer model which uses the fourth sentence $\boldsymbol{x_{-4}}$ preceding $\boldsymbol{x_0}$ as the extended context. In this case, the translated sentence is denoted as

$$\boldsymbol{y_0^{22}(x_{-4}, x_0)}.$$

As shown in Table 2, the fourth preceding sentence $\boldsymbol{x_{-4}}$ and the source sentence $\boldsymbol{x_0}$ are:

$\boldsymbol{x_{-4}}$: 私に逆らうならお前は何もなくなるぞ。
(If you defy me , you will have nothing .)
$\boldsymbol{x_0}$: 出て行けば、戻れなくなるぞ。
(Walk out now and you may never return .)

Then, the concatenated sentences "$\boldsymbol{x_{-4}}$ ⟨CONCAT⟩ $\boldsymbol{x_0}$" are translated into:

$$\boldsymbol{y_{-4}^{22}(x_{-4}, x_0)} \ \langle\text{CONCAT}\rangle \ \boldsymbol{y_0^{22}(x_{-4}, x_0)}.$$

Here, $\boldsymbol{y_0^{22}(x_{-4}, x_0)}$, i.e., the translation of $\boldsymbol{x_0}$ is:

If you leave , you'll never get back .

This time, the fourth preceding source sentence $\boldsymbol{x_{-4}}$ includes the Japanese pronoun "お前"

(mostly translated as "you" in English in the training corpus): thus, the translation $\boldsymbol{y_0^{22}(x_{-4}, x_0)}$ by the 2-to-2 model successfully includes the translation of the Japanese zero pronoun subject in $\boldsymbol{x_0}$ as "you". This then contributes to having the highest forced back-translation probability and sentence-BLEU score with the reference translation compared to $\boldsymbol{y_0^{11}(x_0)}$ (translated by the base 1-to-1 model) and $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$ (translated by the baseline 2-to-2 model), in which the Japanese zero pronoun subject is translated as "we" in both cases. This analysis clearly indicates that the baseline 2-to-2 model is insufficient relative to correctly translating Japanese zero pronouns into English.

Translation by 2-to-2 Model with a Context Sentence $\boldsymbol{x_i}$ ($i = \pm1, \ldots, \pm5$)

More generally, in addition to translation $\boldsymbol{y_0^{11}(x_0)}$ obtained by the base 1-to-1 Transformer model, we prepare 10 translated sentences $\boldsymbol{y_0^{22}(x_{-1}, x_0), \ldots, y_0^{22}(x_{-5}, x_0)}$ and $\boldsymbol{y_0^{22}(x_{+1}, x_0), \ldots, y_0^{22}(x_{+5}, x_0)}$ as candidate translations, each of which is generated using the 2-to-2 model based on the standard Transformer model. Each $\boldsymbol{y_0^{22}(x_i, x_0)}$ ($i = \pm1, \ldots, \pm5$) of these 10 translated sentences is generated by the 2-to-2 model, where one of the first through fifth preceding and subsequent sentences $\boldsymbol{x_i}$ ($i = \pm1, \ldots, \pm5$) is used as the context sentence of the 2-to-2 model[2]. In the 2-to-2 model, only one of the five preceding and subsequent sentences $\boldsymbol{x_i}$ ($i = \pm1, \ldots, \pm5$) is concatenated to the source sentence $\boldsymbol{x_0}$ using the ⟨CONCAT⟩ token as:

$$\boldsymbol{x_i} \ \langle\text{CONCAT}\rangle \ \boldsymbol{x_0}.$$

Then, the concatenated sentences are translated by the 2-to-2 Transformer model. We denote the translated (concatenated) sentences as follows:

$$\boldsymbol{y_i^{22}(x_i, x_0)} \ \langle\text{CONCAT}\rangle \ \boldsymbol{y_0^{22}(x_i, x_0)}$$

where $\boldsymbol{y_i^{22}(x_i, x_0)}$ and $\boldsymbol{y_0^{22}(x_i, x_0)}$ are the translations of $\boldsymbol{x_i}$ and $\boldsymbol{x_0}$, respectively.

3 Selecting Informative Context Sentences with Maximum Forced Back-translation Probability

In the proposed method of selecting a translation among the 11 candidate translations $\boldsymbol{y_0^{11}(x_0)}$, $\boldsymbol{y_0^{22}(x_{\pm1}, x_0), \ldots, y_0^{22}(x_{\pm5}, x_0)}$, we select the translation that yields the highest forced back-translation probability when back-translating into the source sentence. In this context, forced back-translation is defined as forced decoding from a translated target sentence to its source sentence.

Here, assume the source sentence $\boldsymbol{x_0}$ of word length n with a context sentence $\boldsymbol{x_i}$ is given. For the back-translation translation model, we used the 2-to-1 Transformer model with the setup described in Section 5, rather than the 1-to-1 Transformer model. This is simply because, in forced back-translation into $\boldsymbol{x_0}$, the 2-to-1 model considers both $\boldsymbol{y_i^{22}(x_i, x_0)}$ and $\boldsymbol{y_0^{22}(x_i, x_0)}$, while the 1-to-1 model considers $\boldsymbol{y_0^{22}(x_i, x_0)}$ (translation of the source sentence $\boldsymbol{x_0}$) only, but not $\boldsymbol{y_i^{22}(x_i, x_0)}$ (translation of the context sentence $\boldsymbol{x_i}$). We assume that considering both $\boldsymbol{y_i^{22}(x_i, x_0)}$ and $\boldsymbol{y_0^{22}(x_i, x_0)}$ in forced back-translation will yield forced back-translation probabilities that are significantly informative[3].

The forced back-translation probability score of the source sentence word x_j ($1 \leq j \leq n$) of $\boldsymbol{x_0}$ is expressed as follows.

$$b_j \ = \ -\log p\Big(\ x_j | x_{<j}, \ \boldsymbol{y_i^{22}(x_i, x_0)}, \ \boldsymbol{y_0^{22}(x_i, x_0)} \ \Big)$$

From $\boldsymbol{y_i^{22}(x_i, x_0)}$ and $\boldsymbol{y_0^{22}(x_i, x_0)}$, the forced back-translation probability score of the entire source sentence $\boldsymbol{x_0}$ is obtained as the sum of each b_j.

$$\mathrm{B}\Big(\ \boldsymbol{x_0}, \ \boldsymbol{y_i^{22}(x_i, x_0)}, \ \boldsymbol{y_0^{22}(x_i, x_0)} \ \Big) \ = \ \sum_j b_j$$

Similarly, the forced back-translation probability score of the entire source sentence $\boldsymbol{x_0}$ for the base 1-to-1 model is obtained as below:

$$b_j \ = \ -\log p\Big(\ x_j | x_{<j}, \ \boldsymbol{y_0^{11}(x_0)} \ \Big)$$

$$\mathrm{B}\Big(\ \boldsymbol{x_0}, \ \boldsymbol{y_0^{11}(x_0)} \ \Big) \ = \ \sum_j b_j$$

Finally, among the 11 candidate translations $\boldsymbol{y_0^{11}(x_0)}$, $\boldsymbol{y_0^{22}(x_{\pm1}, x_0), \ldots, y_0^{22}(x_{\pm5}, x_0)}$, we select the translation that yields the highest

[2] We examined how many of the 10 translations $\boldsymbol{y_0^{22}(x_{\pm1}, x_0), \ldots, y_0^{22}(x_{\pm5}, x_0)}$ are exactly the same as $\boldsymbol{y_0^{11}(x_0)}$. The rates of cases where none of the 10 translations was exactly the same as $\boldsymbol{y_0^{11}(x_0)}$. were 54% for Japanese to English and 63% for English to Japanese.

[3] In the evaluation discussed in Section 7.1, forced back-translation using the 1-to-1 model achieved merely the same BLEU scores as that of the 2-to-1 model.

forced back-translation probability B when back-translating into the source sentence $\boldsymbol{x_0}$ as below:

$$\underset{i=0,\pm1,...,\pm5}{\operatorname{argmax}} \begin{cases} \mathrm{B}\Big(\boldsymbol{x_0},\ \boldsymbol{y_0^{11}(x_0)} \Big) & (i=0) \\ \mathrm{B}\Big(\boldsymbol{x_0},\ \boldsymbol{y_i^{22}(x_i, x_0)}, & (i \neq 0) \\ \qquad \boldsymbol{y_0^{22}(x_i, x_0)} \Big) \end{cases}$$

Employing the forced back-translation probability differs from existing approaches (Rapp, 2009; Li and Jurafsky, 2016; Goto and Tanaka, 2017; Kimura et al., 2017) that incorporate back-translation from the translated target sentence to the source sentence. Rapp (2009) employed the BLEU score between the source sentence and source language sentence back-translated from the target translated sentence in an automatic MT evaluation context. Li and Jurafsky (Li and Jurafsky, 2016) proposed to re-rank decoded translations based on mutual information between source and target sentences x and y i.e., the probabilities $p(y \mid x)$ and $p(x \mid y)$. Goto and Tanaka (2017) and Kimura et al. (2017) also employed the ratio of forced back-translation probabilities in the context of detecting untranslated content in NMT. These approaches differ from the proposed use of the forced back-translation probability[4].

4 Selecting Informative Context Sentences with Maximum Back-translation Sentence-BLEU

Rapp (2009) proposed an approach of using BLEU score between the source sentence and source language sentence back-translated from the target translated sentence in an automatic MT evaluation context. Based on Rapp (2009), we employ another approach to selecting informative context sentences, where back-translation sentence-BLEU is maximized. As in the case of selecting informative context sentences with maximum forced back-translation probability presented in the previous section, candidate translations are the same as those 11 candidates $\boldsymbol{y_0^{11}(x_0)}, \boldsymbol{y_0^{22}(x_{\pm1}, x_0)}, \ldots, \boldsymbol{y_0^{22}(x_{\pm5}, x_0)}$. For each of those 11 candidate translations, its back-translation back-tran$\big(i\big)$[5] into the source language is given as below:

$$\text{back-tran}\big(i\big) = \begin{cases} \text{back-tran}^{11}\big(\boldsymbol{y_0^{11}(x_0)} \big) & (i=0) \\ \text{back-tran}^{21}\big(\boldsymbol{y_i^{22}(x_i, x_0)}, & (i \neq 0) \\ \qquad \boldsymbol{y_0^{22}(x_i, x_0)} \big) \end{cases}$$

Then, we measure the sentence-BLEU score between the source sentence $\boldsymbol{x_0}$ and each back-translation. We then select the one that gives the highest sentence-BLEU score.

$$\underset{i=0,\pm1,...,\pm5}{\operatorname{argmax}} \ \text{sent-BLEU}\big(\boldsymbol{x_0},\ \text{back-tran}\big(i\big) \big)$$

5 Dataset and Experimental Setup

The dataset used for the oracle translation statistics and the BLEU evaluation comprised 2,083,576 English and Japanese parallel sentence pairs from Opensubtitles 2018 (Lison et al., 2018). Note that we followed Tiedemann and Scherrer (2017) to create the extended context dataset. Here, 90% of the dataset (1,876,624 sentence pairs) was used for training, 5% (104,379 sentence pairs) for development, and 5% (102,573 sentence pairs) for oracle statistics and evaluation. Here, of these 102,573 sentence pairs, only 10,000 pairs were actually used for oracle statistics and evaluation[6] [7]. Throughout the paper, we approximate that all the 2-to-2 models are trained with the immediately preceding sentence as the context.

6 Oracle Translation of Context-based NMT

When measuring the oracle sentence-BLEU score, for each source sentence $\boldsymbol{x_0}$, we select the sentence

[4]The proposed approach is included among those that consider a much wider context than the immediately preceding sentence, e.g., the approaches proposed by Miculicich et al. (2018), Tu et al. (2018), and Maruf and Haffari (2018).

[5]For the back-translation translation model, we used the 1-to-1 Transformer model (denoted as back-tran^{11}) when back-translating $\boldsymbol{y_0^{11}(x_0)}$ ($i = 0$), while we used the 2-to-1 Transformer model (denoted as back-tran^{21}) with the setup described in section 5 when back-translating $\boldsymbol{y_0^{22}(x_i, x_0)}$ ($i \neq 0$, i.e., translated from $\boldsymbol{x_0}$ with a context sentence by the 2-to-2 model).

[6]In training and development, the encoder rejects input sentences (source sentence concatenated with the context sentence for the 2-to-2 models) with greater than 50 tokens. Average token length of the 10,000 pairs for oracle statistics and evaluation is 7.9 (English) and 6.9 (Japanese).

[7]Experimental setup is as follows: Tokenizers are Moses tokenizer (Koehn et al., 2007) for English and MeCab (`http://taku910.github.io/mecab/`) for Japanese tokenization. OpenNMT-py (Klein et al., 2017) is used for training and testing NMT models. 50,000 vocabulary sizes are employed for both English and Japanese. Embedding sizes are 512. Encoder and decoder are with six layers with batch size as 4,096 and dropout rate as 0.3 and 100,000 steps for training. Adam optimizer (Kingma and Ba, 2015) is used. One NVIDIA Tesla P100 16GB GPU is used. MTEval Toolkit (`https://github.com/odashi/mteval`) is used to measure BLEU, and Moses decoder's sentence-bleu.cpp is used to measure sentence-BLEU.

with the maximum sentence-BLEU score among the candidate translations after obtaining 11 candidates ($\boldsymbol{y_0^{11}(x_0)}$ translated by the 1-to-1 model and $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i{=}\pm1, \ldots, \pm5)$ translated by the 2-to-2 models). Figure 1 shows the oracle BLEU scores for the following seven cases:

(i) among $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = \pm1, \ldots, \pm5)$ with and without $\boldsymbol{y_0^{11}(x_0)}$
(ii) among $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = -1, \ldots, -5)$ with and without $\boldsymbol{y_0^{11}(x_0)}$
(iii) among $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = +1, \ldots, +5)$ with and without $\boldsymbol{y_0^{11}(x_0)}$
(iv) between $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$

and the BLEU scores of the baseline 1-to-1 ($\boldsymbol{y_0^{11}(x_0)}$) and 2-to-2 models with the immediately preceding sentence as the context ($\boldsymbol{y_0^{22}(x_{-1}, x_0)}$). For all three 2-to-2 model cases with the candidate translation obtained by the 1-to-1 model, the oracle BLEU increased by including $\boldsymbol{y_0^{11}(x_0)}$. Furthermore, the oracle BLEU score increases as more candidates are considered. Table 1 shows that, by considering $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = -5, \ldots, -2, +1, \ldots, +5)$ in addition to $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$, the oracle BLEU score improves by approximately four points for Japanese to English and six points for English to Japanese. These results indicate that longer contexts yield obvious benefit for the 2-to-2 context-based NMT model, which is the primary motivation for selecting informative context sentences in that model.

7 Evaluation

7.1 Evaluation Results

For both English to Japanese and Japanese to English directions, Table 1 shows the BLEU scores obtained by selecting the translation candidate that maximizes the forced back-translation and the back-translation sentence-BLEU score. For the proposed method, we compare the following translation candidate cases:[8] (i) between $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$, (ii) among $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = -1, \ldots, -5)$, (iii) among $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = +1, \ldots, +5)$, (iv) among $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = \pm1, \ldots, \pm5)$. Compared to the BLEU scores of $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$, all BLEU scores obtained by the proposed method demonstrate significant improvement ($p < 0.01$), except for the Japanese to English translation obtained by maximizing the back-translation sentence-BLEU score.

By comparing the BLEU scores of $\boldsymbol{y_0^{11}(x_0)}$, $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$, the oracle among them, and the selection between them by maximizing the forced back-translation, the selection between $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$ by maximizing forced back-translation achieves BLEU scores that are comparable to the oracle BLEU scores. Thus, we conclude that the proposed method contributes to selecting better translation between those candidates. However, the proposed method cannot select informative context sentences among $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = -5, \ldots, -2, +1, \ldots, +5)$, because the results obtained by adding $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = -5, \ldots, -2, +1, \ldots, +5)$, to translation candidates $\boldsymbol{y_0^{11}(x_0)}$ and $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$ yields little or no gain in BLEU score. Note that this does not coincide with improving the oracle BLEU score by approximately four points for Japanese to English and six points for English to Japanese with the overall 11 translation candidates. Thus, it can be concluded that further study is required to appropriately select the informative context sentences among the 11 candidates such that the BLEU score becomes much closer to the oracle BLEU score.

Another important comparison with a baseline is also shown as "selection from 20-best of 2-to-2 (baseline) by 2-to-2 back-translation" in Table 1. With this baseline, it is intended to examine whether the five preceding and subsequent sentences introduced in the proposed method are sufficiently informative compared to other well studied translation candidates such as n-best translations. Specifically, the baseline 2-to-2 model with the immediately preceding sentence as the context is employed to generate 20-best translations, and then, out of those generated 20-best translations, the one with the maximum forced back-translation into the source sentence is selected[9]. As shown in Table 1, this baseline performed worse than the proposed approach. From this result, it is obvious that the proposed approach of introducing five preceding and subsequent sentences as the context is

[8] Throughout the evaluation results of this paper, when obtaining the forced back-translation probability for $\boldsymbol{y_{11}}$, we used the 1-to-1 Transformer model as the back-translation translation model.

[9] We compare the 2-to-2 and 2-to-1 models in the step of forced back-translation here, where the 2-to-2 model outperformed the 2-to-1 model. In Table 1, we show the results obtained by the 2-to-2 model.

category of phenomena	Ja-En		En-Ja	
	succeed	fail	succeed	fail
synonymous expression	14	12	17	10
pronoun	**7**	2	2	0
untranslated by baseline	**5**	0	**10**	0
article	0	1	0	0
other	11	2	7	3
manually judged				
(comparable)	10	18	13	27
(baseline wins)	3	0	1	0
(baseline loses)	0	15	0	10
total	50	50	50	50

Table 3: Distribution of oracle translation phenomena (through manual analysis of 50 examples) (proposed method succeeds / fails in identifying those oracle translations)

category of phenomena	Ja-En		En-Ja	
	2-to-2 wins	1-to-1 wins	2-to-2 wins	1-to-1 wins
synonymous expression	16	20	14	22
pronoun	2	2	1	2
untranslated by 1-to-1	**8**	0	**12**	2
article	1	0	0	0
other	7	8	3	4
manually judged				
(comparable)	15	15	18	16
(1-to-1 wins)	1	0	2	0
(2-to-2 wins)	0	5	0	4
total	50	50	50	50

Table 4: Distribution of phenomena where baseline 2-to-2 $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$ wins v.s. 1-to-1 $\boldsymbol{y_0^{11}(x_0)}$ wins (through manual analysis of 50 examples)

much more informative than 20-best translations with just the preceding sentence as the context.

7.2 Analysis of Improvements and Errors

To analyze typical cases relative to the improvements and errors of the proposed approach, we randomly select 50 success cases and 50 failure cases when identifying oracle translations using the proposed method. Specifically, we first collect cases where oracle translation is selected from $\boldsymbol{y_0^{11}(x_0)}$ or $\boldsymbol{y_0^{22}(x_i, x_0)}$ $(i = -5, \ldots, -2, +1, \ldots, +5)$, rather than from $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$. Then, from these cases, we randomly select 50 examples for each of the following cases.

(a) Proposed approach (maximizing forced back-translation) successfully identifies collected oracle translations.
(b) Proposed approach (maximizing forced back-translation) fails to identify collected oracle translations.

Then, we manually categorize the 50 examples (for each case) according to the phenomena in Table 3.

For both Japanese to English and English to Japanese, nearly 30∼40% are categorized as "synonymous expression", where the proposed approach of maximizing forced back-translation successfully selects the oracle translation that includes the synonymous expression rather than exactly the same expression (as in the reference translation). Due to this synonymous expression, the sentence has the highest sentence-BLEU score and is selected as the oracle translation. Although this phenomenon is top ranked among others, it is also top ranked among the failure cases. Thus, it is necessary to incorporate other criteria to reduce the failure cases.

For comparison, we also categorize the phenomena of randomly selected 50 cases when the baseline 2-to-2 model outperforms the 1-to-1 model, i.e., translation $\boldsymbol{y_0^{22}(x_{-1}, x_0)}$ by the baseline 2-to-2 model achieves a sentence-BLEU score that is greater than that of translation $\boldsymbol{y_0^{11}(x_0)}$ by the 1-to-1 model. We also categorize the phenomena of randomly selected 50 cases of its opposite, i.e., when the 1-to-1 model outperforms the baseline 2-to-2 model. These results are shown in Table 4. As can be seen, even in the comparison of the baseline 2-to-2 and 1-to-1 models, the "synonymous expression" category is top ranked.

It is interesting to compare the second and third ranked categories, i.e., "pronoun translation" and "untranslated by baseline," among Japanese to English and English to Japanese in Tables 3 and 4. The "pronoun translation" category is ranked high only in the Japanese to English case with the proposed approach (Table 3). Table 2 shows a typical Japanese subject zero pronoun case and its detail is described in section 2. With the "untranslated by baseline / 1-to-1" categories, it is obvious from Table 3 and Table 4 that the proposed approach outperforms the baseline 2-to-2 model for

Source sentences		Target sentences		Forced back-translation probability / sentence-BLEU
—		Reference translation:	Every pain you suffered was punishment for your sins .	—
Immediately preceding sentence x^{-1}:	お前の一挙一動がここにお前を導いた。(Every step you took led you to here .)	Translation $y_0^{22}(x_{-1}, x_0)$ by baseline 2-to-2 model:	Every suffering you suffered was your punishment .	3.4×10^{-19} / 27.64
Source sentence x_0:	お前の受けた全ての苦しみはお前の 罪 に対する罰だった。	Translation $y_0^{11}(x_0)$ by baseline 1-to-1 model:	All the suffering you've had was your punishment .	5.3×10^{-19} / 16.62
2nd subsequent sentence x^{+2}:	お前の命を奪うために悪魔 が送ったものを見よ! (See what the devil has sent to claim you .)	Translation $y_0^{22}(x_{+2}, x_0)$ by 2-to-2 model:	All your suffering was punishment for your sins .	$\mathbf{5.9\times10^{-14}}$ / **57.18**

Table 5: Example improvements over baseline 2-to-2 $y_0^{22}(x_{-1}, x_0)$ (Ja-En) (b) untranslated by baseline

both Japanese to English and English to Japanese directions. In addition, the baseline 2-to-2 model outperforms the 1-to-1 model. Thus, it can be concluded that the matter of untranslated content in context-based NMT can be handled consistently by appropriately extending the range of the context considered within a certain framework of context-based NMT models such as the 2-to-2 model. For example, as shown in Table 5, the baseline 2-to-2 model fails to produce the word "sins" in its translation. In contrast, the translation $y_0^{22}(x_{+2}, x_0)$ obtained by the 2-to-2 model with the second subsequent source sentence x_{+2} as the context sentence successfully includes the word "sins," probably because the second subsequent source sentence x_{+2} includes "悪魔" ("devil").

By examining the cases of improvements over the baseline 2-to-2 model, we observe that the essential advantage of the proposed approach is that the measure of forced back-translation probability can distinguish translation errors from relatively acceptable translations, with which the sentence-BLEU score with the reference translation is typically higher than that of the baseline 2-to-2 model. As a result, we conclude that it is unnecessary to consider a context with a much greater number of sentences, such as 3-to-3 (or higher) models.

8 Conclusion

Within the framework of the 2-to-2 context-based NMT model, this paper has explored how to select the most informative context sentences that provide the highest BLEU score among the five preceding and five subsequent sentences. In future, we plan to compare the proposed method to an existing approach (Li and Jurafsky, 2016) that incorporates back-translation into the MT framework. In addition, we plan to incorporate monolingual techniques such as BERT (Devlin et al., 2018) and neural coreference resolution (Lee et al., 2017), to evaluate whether context sentences (i.e., the second through fifth sentences preceding the source sentence and the first through fifth sentences subsequent to the source sentence) are in fact informative. Also, in the context of translation quality estimation techniques (Specia et al., 2015), the proposed approach of estimating the quality of translation by maximizing forced back-translation is novel and has never been studied so far in the task of translation quality estimation.

References

Bawden, R., R. Sennrich, A. Birch, and B. Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proc. NAACL-HLT*, pages 1304–1313.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *CoRR*, volume abs/1810.04805.

Goto, I. and H. Tanaka. 2017. Detecting untranslated content for neural machine translation. In *Proc. 1st NMT*, pages 47–55.

Kimura, R., Z. Long, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. 2017. Effect on reducing untranslated content by neural machine translation with a large vocabulary of technical terms. In *Proc. 7th PSLT*, pages 13–24.

Kingma, D. P. and J. Ba. 2015. Adam: A method for stochastic optimization. *Proc. ICLR*.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. 55th ACL*, pages 67–72.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL*, pages 177–180.

Lee, K., L. He, M. Lewis, and L. Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proc. EMNLP*, pages 188–197.

Li, J. and D. Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. In *CoRR*, volume abs/1601.00372.

Libovický, J. and J. Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proc. 55th ACL*, pages 196–202.

Lison, P., J. Tiedemann, and M. Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proc. 11th LREC*, pages 1742–1748, May 7-12, 2018.

Luong, T., H. Pham, and C. D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. 2015 EMNLP*, pages 1412–1421.

Maruf, S. and G. Haffari. 2018. Document context neural machine translation with memory networks. In *Proc. 56th ACL*, pages 1275–1284.

Miculicich, L., D. Ram, N. Pappas, and J. Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proc. EMNLP*, pages 2947–2954.

Rapp, R. 2009. The back-translation score: Automatic mt evaluation at the sentence level without reference translations. In *Proc. 47th ACL and 4th IJCNLP*, pages 133–136.

Specia, L., G. Paetzold, and C. Scarton. 2015. Multi-level translation quality prediction with QuEst++. In *Proc. 53rd ACL and 7th IJCNLP System Demonstrations*, pages 115–120.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural machine translation. In *Proc. 27th NIPS*, pages 3104–3112.

Tiedemann, J. and Y. Scherrer. 2017. Neural machine translation with extended context. In *Proc. 3rd DiscoMT*, pages 82–92.

Tu, Z., Y. Liu, S. Shi, and T. Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of ACL*, 6:407–420.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proc. 30th NIPS*, pages 5998–6008.

Voita, E., P. Serdyukov, R. Sennrich, and I. Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proc. 56th ACL*, pages 1264–1274.