Data Augmentation using Back-translation for Context-aware Neural Machine Translation

Amane Sugiyama
The University of Tokyo
sugi@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science,
the University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Abstract

A single sentence does not always convey information required to translate it into other languages; we sometimes need to add or specialize words that are omitted or ambiguous in the source languages (*e.g.*, zero pronouns in translating Japanese to English or epicene pronouns in translating English to French). To translate such ambiguous sentences, we exploit contexts around the source sentence, and have so far explored context-aware neural machine translation (NMT). However, a large amount of parallel corpora is not easily available to train accurate context-aware NMT models. In this study, we first obtain large-scale pseudo parallel corpora by back-translating target-side monolingual corpora, and then investigate its impact on the translation performance of context-aware NMT models. We evaluate NMT models trained with small parallel corpora and the large-scale pseudo parallel corpora on IWSLT2017 English-Japanese and English-French datasets, and demonstrate the large impact of the data augmentation for context-aware NMT models in terms of BLEU score and specialized test sets on ja→en[1] and fr→en.

1 Introduction

Following the success of neural machine translation (NMT) models in sentence-level translation, context-aware NMT models have been studied to further boost the quality of translation (Jean et al., 2017; Tiedemann and Scherrer, 2017; Wang et al., 2017; Bawden et al., 2018; Voita et al., 2018; Maruf and Haffari, 2018; Miculicich et al., 2018; Maruf et al., 2019; Voita et al., 2019). These context-aware models take auxiliary inputs (contexts) to translate the source sentence which lacks information needed for translating into the target language (§ 2). Typically, contexts considered by context-aware NMT are surrounding sentences in the same document (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Maruf and Haffari, 2018; Voita et al., 2019), which provide beneficial information in translating zero pronouns, anaphoric pronouns, lexically ambiguous words, and so on.

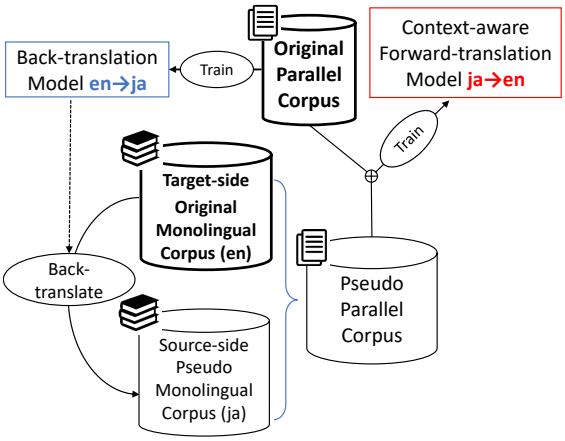

Figure 1: Overview of the data augmentation for context-aware NMT (Japanese to English in this case).

Although the context-aware NMT models outperform the baseline sentence-level NMT models in terms of BLEU score and some specialized test sets (Bawden et al., 2018; Voita et al., 2019; Müller et al., 2018), the reported gains, especially in BLEU score, are often marginal. We can think of several reasons for this; 1) the ratio of sentences (or linguistic phenomena) that require contexts for translation is small in the evaluation datasets, 2) the current context-aware models do not fully utilize the given contexts, 3) (narrow) contexts considered in context-aware NMT models do not include information required for translation, 4) the size of training data is not enough to effectively train context-aware NMT models. Although there are some studies that investigate the first to third

[1] http://www.tkl.iis.u-tokyo.ac.jp/~sugi/DiscoMT2019/

Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 35–44
Hong Kong, China, November 3, 2019.

aspects (Bawden et al., 2018; Voita et al., 2018; Imamura and Sumita, 2019), few studies have investigated the last possibility (§ 6), since there are few parallel corpora for context-aware translation; existing large-scale and high-quality parallel corpora are usually obtained by extracting reliable sentence alignments from translations by humans (Nakazawa et al., 2016; Pryzant et al., 2018). Considering that context-aware NMT models have larger input spaces than sentence-level models, they will demand larger training data to fully exert the models' performance.

In this study, we hypothesize that context-aware NMT models can benefit from an increase of the training data more than sentence-level models, and confirm this by performing data augmentation using back-translation (Sennrich et al., 2016b) (§ 6) for context-aware NMT models. We propose to assist the training of context-aware NMT models using pseudo parallel data which is automatically generated by back-translating a large monolingual data (§ 3, Figure 1). The back-translation model here is trained on an existing parallel corpus. Since context-aware models are designed to recover information that is absent from the source sentence but should be present in the target sentence, back-translation can produce effective training data if it could naturally drop the information to be recovered in translating sentences in the target language into the source language.

We evaluate our method on IWSLT2017 data sets (Cettolo et al., 2012), which are collections of subtitles of TED Talks, on two language pairs: English-Japanese (en-ja) and English-French (en-fr) (§ 4). We exploit BookCorpus (Zhu et al., 2015), Europarl v7 (Koehn, 2005), and the record of the National Diet of Japan as monolingual corpora for back-translation (§ 5). Experimental results revealed that the data augmentation improved the translation in terms of BLEU score (Papineni et al., 2002) and the accuracy on specialized test sets for context-aware NMT.

The contribution of this paper is as follows:

- We **first evaluated data augmentation on context-aware NMT**, and confirmed BLEU improvement on en↔fr and ja→en datasets,
- **developed a new specialized test set for evaluating ja→en** context-aware translation, and
- **confirmed that the data augmentation improves context-aware translation** through the existing en→fr (Bawden et al., 2018) and our specialized test set for ja→en translation.

2 Context-aware NMT Models

To incorporate contexts to translate sentences, recent studies on NMT have explored context-aware models which take sentences around the source sentence as auxiliary inputs. Typical contexts considered in those models are a few sentences that precede the source sentence.

The context-aware NMT models are grouped into two types: single-encoder models that apply a sentence-level NMT model to the source sentence concatenated after their contexts (preceding sentence(s)) (Tiedemann and Scherrer, 2017; Bawden et al., 2018) and multi-encoder models that design an additional context encoder to process the contexts (Jean et al., 2017; Wang et al., 2017; Bawden et al., 2018; Voita et al., 2018; Maruf and Haffari, 2018; Miculicich et al., 2018; Tu et al., 2018; Maruf et al., 2019). In what follows, we briefly review these models.

Single-encoder models take the preceding sentence(s) as the contexts in addition to the source sentence and concatenate them with a special symbol <CONC> (Tiedemann and Scherrer, 2017). The concatenated sentences are then translated using an existing sentence-level NMT model.

There are two subtypes of the single-encoder models that differ in handling contexts in the target language. The first model, which we refer to as 2-to-1, only considers contexts in the source language, and is trained on pairs of the source sentence with the preceding sentence(s) and the target sentence. It learns a mapping from the source sentence with its context to the target sentence. The second one, which we refer to as 2-to-2, considers contexts in both the source and target languages. 2-to-2 models are trained on pairs of the source sentence with the preceding sentence(s) and the target sentence with the preceding sentence(s). At test time of a 2-to-2 model, the decoder receives the encoder hidden states and the translation of the previous sentence, which has been generated in the previous translation step. We analogically refer to the standard sentence-level NMT models as 1-to-1 to highlight the difference in input and output.

Multi-encoder models take the preceding sentence(s) as the contexts, and use additional neural networks to encode the contexts. Several net-

work architectures have been explored for this additional encoder (Jean et al., 2017; Wang et al., 2017; Bawden et al., 2018; Voita et al., 2018; Maruf and Haffari, 2018; Miculicich et al., 2018; Tu et al., 2018).

In this study, we adopt the standard single-encoder model (Tiedemann and Scherrer, 2017) in our experiments (§ 4), since both single-encoder and multi-encoder models are reported to outperform the sentence-level models and the performance gap between the two context-aware models are marginal. We then focus on investigating the impact of additional pseudo parallel training data generated by back-translation. Note that the single-encoder models are simpler, and we can employ the well-established NMT architectures such as Transformer (Vaswani et al., 2017) without any modifications for sequence-to-sequence transformation.

3 Data Augmentation for Context-aware NMT using Back-translation

We hypothesize that context-aware NMT models can benefit from an increase of the training data more than sentence-level NMT models, and experimentally confirm this by training and evaluating context-aware NMT models with additional training data. We propose to use data augmentation based on back-translation (Sennrich et al., 2016a) to obtain the additional training data for context-aware NMT models. We hereafter refer to (final) source-to-target translation as *forward-translation* to distinguish it with (target-to-source) back-translation for data augmentation.

The pseudo parallel data is automatically generated by back-translating large target-side monolingual corpora (target→source). Since monolingual corpora can be obtained more easily than bilingual parallel corpora which are aligned at sentence level, the back-translation allows us to train a context-aware NMT model with larger data. We can expect the resulting pseudo parallel corpora to contain more cases from which the model can learn to use contexts in translation.

Back-translation for data augmentation The data augmentation in this study follows the existing back-translation strategies for NMT (Sennrich et al., 2016a; Imamura et al., 2018; Edunov et al., 2018) except that we assume a context-aware model for the forward-translation; the monolingual data for back-translation must be a set of documents each of which consists of contiguous sentences. This data-augmentation approach would naturally benefit context-aware models more than sentence-level models because the former are to handle a larger input/output space, which makes them more complex as a mapping task.

Here, we describe our training process to obtain pseudo training data for translation from the source (L_S) to the target (L_T) language.

Train a back-translation model ($L_T \rightarrow L_S$)
Given a (small) parallel data for source language L_S and target language L_T, we first train a back-translation model $L_T \rightarrow L_S$ on the parallel data.

Back-translate L_T monolingual data into L_S
We next back-translate a large L_T (target-side) monolingual data to generate pseudo L_S (source-side) monolingual data, which forms pseudo parallel data together with the original target-side monolingual data. Note that sentential alignments are naturally obtained through the translation.

Train a forward-translation model ($L_S \rightarrow L_T$)
We then train the forward-translation model from the original parallel data augmented with the obtained pseudo parallel data.

The pseudo parallel data has merits and demerits against human-translated parallel data which is automatically aligned. The pseudo parallel data is inferior to the human-translated parallel data in that it is generated automatically by a possibly inaccurate machine translation system. However, it does not contain mismatches of sentence boundaries between the target and the obtained (back-translated) source monolingual data, in contrast to the human-translated data where, for example, a source sentence can correspond to multiple target sentences.

On back-translation model We can use either a sentence-level or context-aware NMT model for back-translation. In the following experiments, we first adopt 2-to-1 NMT model as a back-translator for data augmentation, and evaluate the impact of the data augmentation on the translation performance of context-aware NMT models. We then compare those results with results obtained by the data augmentation using 1-to-1 and 2-to-2 models instead of 2-to-1 model for back-translation.

	# sentence pairs	avg. source length	avg. target length
en→ja	223k / 0.87k / 1.54k	24.7 / 28.0 / 24.6	25.4 / 27.9 / 24.5
ja→en	212k / 0.87k / 1.54k	22.3 / 28.0 / 24.6	22.8 / 27.9 / 24.5
en→fr	222k / 0.89k / 1.56k	22.1 / 27.2 / 24.3	23.5 / 28.0 / 25.8
fr→en	222k / 0.89k / 1.56k	23.5 / 28.0 / 25.8	22.1 / 27.2 / 24.3

Table 1: Statistics of IWSLT2017 corpora: the number of sentence pairs and the average length (number of tokens per sentence) for the train / dev / test portions.

We can expect context-aware NMT models to moderately omit redundant information as humans do and to yield more natural translations when back-translating, especially if the source language L_S prefers to omit redundant expressions (*e.g.*, zero pronouns in Japanese). It would produce a better training data from which the forward-translation model can learn to restore the omitted information referring to context.

4 Experimental settings

This section describes experimental settings to evaluate the impact of the data augmentation on context-aware NMT models. We conduct translation experiments on two language pairs for both directions: Japanese→English (hereafter, ja→en), English→Japanese (en→ja), French→English (fr→en), and English→French (en→fr) using publicly available corpora of spoken language that are used in the previous studies.

Datasets (parallel corpora) For all the language pairs, we use IWSLT2017 corpus[2] (Cettolo et al., 2012) as the original (human-translated) parallel data. This corpus is made from subtitles of TED Talks. The English subtitles are transcription of the talks and the subtitles in the other languages are translations of the English subtitles. We consider each talk as a document. We use `dev2010` for development and `tst2010` for evaluation in each language pair. The statistics of IWSLT2017 corpus used in our experiments are listed in Table 1.

Datasets (monolingual corpora) For ja→en and fr→en translations, we exploit BookCorpus (Zhu et al., 2015) as the monolingual data. BookCorpus is a collection of English e-books available on the Web.[3] We extract paragraphs from BookCorpus that consist of more than 9 sentences and treat them as single documents. For en→ja and en→fr translation, we adopt the record of the National Diet of Japan[4] (hereafter, DietCorpus) and Europarl corpus v7[5] (Koehn, 2005) as the monolingual data, respectively. We use the French part of Europarl as a monolingual corpus in our experiments considering its domain being close to that of IWSLT2017 (most documents in Europarl corpus consist of conversation of multiple persons but each block of contiguous utterances given by a single person tends to be long so it can be assumed to be locally monologue like IWSLT2017) and it consists of contiguous sentences, which meets our demand.

Preprocessing We normalize punctuation of the English and French datasets and perform tokenization and truecasing using Moses toolkit verion 4.0.[6] We tokenize the Japanese datasets using MeCab verion 0.996 with ipadic dictionary verion 2.7.0.[7] For each language pair, we finally split datasets into subword units using SentencePiece (verion 0.1.81)[8] with unigram language model. The SentencePiece model is trained using the original parallel corpus (IWSLT2017 corpus) following (Sennrich et al., 2016a; Imamura et al., 2018). The vocabulary size is 16k shared by the source and target languages.

Prior to training, all 1-to-1 back-translation models and 1-to-1 forward-translation models for ja→en and en↔fr, we remove from the training datasets sentence pairs in which the source or target sentence contains more than 64 tokens. We set a larger limit of 128 in training the 1-to-1 forward-translation model of en→ja since the Japanese monolingual corpus DietCorpus has longer sentences on average and the limit of 64 is too small to cover an adequate proportion of sentence pairs

[2] https://wit3.fbk.eu/mt.php

[3] We used a crawler available at https://github.com/soskek/bookcorpus

[4] https://www2.ninjal.ac.jp/lrc/index.php

[5] https://www.statmt.org/europarl/

[6] http://www.statmt.org/moses/

[7] https://taku910.github.io/mecab/

[8] https://github.com/google/sentencepiece

	# sentences	avg. source length	avg. target length
en→ja	1030k	31.9	39.7
ja→en	6493k	16.4	14.9
en→fr	2223k	26.8	30.0
fr→en	6493k	16.0	14.9

Table 2: Statistics of the target-side monolingual corpora and their source-side counterparts obtained by back-translation: the number of sentences in the original corpora and the average length in the pseudo parallel data used to train 1-to-1 models.

in the monolingual corpus. Prior to training 2-to-X forward-translation models, we removed pairs of concatenated sentences where the source or target contains more than 128 tokens except en→ja forward-translation with the length limit of 200 for the same reason as above. The statistics of the datasets we used to train 1-to-1 models are shown in Table 1 and 2.[9]

NMT models For all NMT models, we adopted Transformer (Vaswani et al., 2017) as the core neural model architecture. We implemented it using Tensorflow[10] verion 1.12.0. Both encoder and decoder comprise 6 blocks, the dimension of the embedding layers is 512 and the dimension of the FFN layers is 2048. The source and target embedding weights and the decoder pre-softmax weights are all shared. Training is performed using Adam optimizer (Kingma and Ba, 2015) with a learning rate conditioned on the training steps following the original Transformer. Each batch contains about 16384(= 128^2) tokens, and hence the number of sentences in a batch varies.

Back-translation For each language pair, back-translation models are trained on IWSLT2017 corpora. Monolingual data are back-translated by using 2-to-1 models with beam size of 5.

Forward-translation For each language pair, we train 1-to-1, 2-to-1 and 2-to-2 models while varying the size of pseudo parallel data used to augment the original parallel data. We train ja→en and fr→en models on 0k (none), 500k, 1000k, 2000k and 4000k pseudo data, en→ja models on 0k (none), 500k and 1000k pseudo data, and en→fr models on 0k (none), 500k, 1000k and 2000k pseudo data. At test time, we perform translation with beam size of 8.

Evaluation using BLEU We evaluate the translation quality of the forward-translation with BLEU scores (Papineni et al., 2002), computed by `multi-bleu.perl` in the Moses toolkit, after decoding the subwords by SentencePiece.

Evaluation using specialized test sets Also, we perform evaluation on en→fr and ja→en translation using an existing (Bawden et al., 2018) and a newly-created specialized test sets for evaluating context-aware NMT. These datasets are designed to assess whether NMT models capture intersentential contexts.

Both test sets consist of questions to be asked to the model. In each question, given a source sentence, source-side context, target-side context and two translation candidates, models must determine which one of the two candidates is correct as a translation for the source sentence on the basis of the translation scores (in our experiments, we compute translation scores from log-likelihood of the sequences with length-normalization (Johnson et al., 2017)). Both test sets are designed so that sentence-level models always achieve 50% accuracy.

For en→fr 2-to-2 models,[11] we exploit the existing discourse test sets tailored by Bawden et al. (2018). The test set include coreference test set and coherence/cohesion test set. The coreference test set contains 200 questions, which require NMT models to implicitly resolve anaphora to translate anaphoric pronouns. The coherence/cohesion test set contains 200 questions to test how well NMT models maintain discourse-level consistency. Note that this dataset was made on the basis of OpenSubtitles2016 corpus (Lison and Tiedemann, 2016), and, in some questions, the context and the main sentence form a dialogue; the domain does not fully match that of our parallel corpus (TED talks, monologue).

For ja→en models, following (Bawden et al., 2018), we newly created a specialized test set. Referring to (Nagata and Morishita, 2019), we tailored test cases focusing on zero pronouns in Japanese for the specialized test set as follows.[12] First, we choose two contiguous sentences, as the

[9] Training of 2-to-X models is done using different subsets of the whole pseudo parallel data (due to the different cleaning standards stated in this paragraph). Since the statistics of the pseudo parallel data are almost identical, we provide here the statistics of 1-to-1 as representative.

[10] https://www.tensorflow.org/

[11] We only evaluate 2-to-2 models because some questions in the datasets require target-side contexts to answer.

[12] We developed a new ja→en test set since Nagata and Morishita (2019) does not release their test set.

# pseudo train (# sent. pairs)	en→ja / # train: 212k			ja→en / # train: 211k			en→fr / # train: 222k			fr→en / # train: 222k		
	1-to-1	2-to-1	2-to-2	1-to-1	2-to-1	2-to-2	1-to-1	2-to-1	2-to-2	1-to-1	2-to-1	2-to-2
0k	**12.47**	**12.88**	12.42	11.07	11.32	11.76	36.77	36.83	37.03	35.73	36.16	36.29
500k	12.32	12.79	**12.54**	11.92	12.68	13.04	38.08	38.05	38.16	37.22	37.07	37.37
1000k	11.98	11.99	12.28	12.03	12.80	13.20	**38.11**	37.63	38.55	37.11	37.20	37.90
2000k	n/a	n/a	n/a	11.84	12.91	**13.57**	37.98	**38.30**	**38.79**	37.36	**37.86**	37.86
4000k	n/a	n/a	n/a	**12.14**	**13.06**	13.51	n/a	n/a	n/a	**37.47**	37.44	**38.01**

Table 3: BLEU scores of the sentence-level and context-aware models with data augmentation: All the models are trained on the original parallel corpora and the pseudo parallel data generated by back-translation, while varying the size of pseudo training data from 0 (no pseudo training data) to 4000k.

Source
context: 父親 は 何 か 呟い て い た 。
sentece: どう も ドア の ほう に 向き なおっ て いる らしい 。
Target
context: My **father** murmured something.
correct: He seems to be turning towards the door.
incorrect: She seems to be turning towards the door.

Source
context: 母親 は 何 か 呟い て い た 。
sentence: どう も ドア の ほう に 向き なおっ て いる らしい 。
Target
context: My **mother** murmured something.
correct: She seems to be turning towards the door.
incorrect: He seems to be turning towards the door.

Figure 2: An example pair of questions in our ja→en test set; the underlined pronouns refer to the boldfaced nouns, and do not appear in the source Japanese sentences (zero pronouns).

source sentence and its context, denoted by S and C_{s_1} respectively, from a Japanese corpus, Keyaki Treebank (Butler et al., 2018), and translate them into English, which result in a correct translation and the target-side context, denoted by T_1 and C_{t_1}, respectively. Next, we write an incorrect translation T_2 and source/target contexts C_{s_2}, C_{t_2} with which the incorrect translation could be correct. Then, using these sentences, we make two questions:

Q_1: given $S, C_{s_1}, C_{t_1}, T_1, T_2$, choose T_1 or T_2

Q_2: given $S, C_{s_2}, C_{t_2}, T_1, T_2$, choose T_1 or T_2

For Q_1 and Q_2, the correct answer is T_1 and T_2, respectively. By iterating this process, we made 100 questions. Note that sentence-level models achieve exactly 50% accuracy on this test set. Unlike the en→fr test set, all the questions are answerable without seeing the target-side context. Some of the created questions are shown in Figure 2.

5 Results and Analysis

In this section, we first report the impact of the data augmentation on sentence-level and context-aware NMTs (§ 5.1). We next investigate whether the translation performance with the data augmentation is affected by the type of translation system used for back-translation: single-sentence NMT or context-aware NMT (§ 5.2). We then confirm that the data augmentation improves ja→en and en→fr translation that requires contexts by using the two discourse-oriented test sets (§ 5.3). We finally show some translation examples (§ 5.4).

5.1 Impact of the size of pseudo training data

Table 3 lists the BLEU scores of sentence-level and context-aware NMT models while varying the size of pseudo parallel data. In what follows, we interpret results in detail.

ja→en and en↔fr models A comparison among 1-to-1, 2-to-1, and 2-to2 models provides a certain trend; context-aware models (2-to-X) are better than the sentence-level model (1-to-1), and the target-side contexts contribute to the translation quality (2-to-1 vs. 2-to-2). The impact of the pseudo parallel data is clear: adding pseudo parallel data to a certain extent results in higher BLEU scores; 2-to-X models achieve the best performance with more pseudo data than 1-to-1 models. In other words, context-aware models with auxiliary inputs benefit from more pseudo parallel data, as we have expected; 2-to-2 models benefit from the largest pseudo training data.

We additively obtained the gain in BLEU by using the pseudo parallel data in addition to using contexts. This results in a large improvement in BLEU scores: +2.50 (11.07 → 13.57) in ja→en, +2.02 (36.77 → 38.79) in en→fr, and then +2.28 (35.73 → 38.01) in fr→en.

	pseudo train	train	dev	test
en	14 / 27 / 45	13 / 20 / 32	14 / 23 / 35	13 / 20 / 31
ja	18 / 34 / 56	13 / 21 / 33	13 / 22 / 37	12 / 20 / 32

Table 4: The quartile of the number of tokens per sentence in each dataset: train, dev and test indicate the train, dev, and test sets of IWSLT2017 corpus. The English portion of the pseudo train dataset is the translation of the Japanese monolingual corpus, DietCorpus.

# pseudo train	1-to-1 back-trans.			2-to-1 back-trans.			2-to-2 back-trans.		
	1-to-1	2-to-1	2-to-2	1-to-1	2-to-1	2-to-2	1-to-1	2-to-1	2-to-2
0k	11.07	11.32	11.76	(same to the left)					
500k	12.02	12.90	13.02	11.92	12.68	13.04	12.15	12.65	13.35
1000k	12.41	12.99	13.22	12.03	12.80	13.20	12.43	13.19	13.49
2000k	**12.49**	**13.35**	**13.57**	11.84	12.91	**13.57**	12.59	13.40	**13.79**
4000k	12.23	13.02	13.34	**12.14**	**13.06**	13.51	**12.78**	**13.34**	13.58

Table 5: The BLEU scores of ja→en context-aware models trained with pseudo parallel data generated by 1-to-1 and 2-to-2 back-translation: The scores of the models trained on pseudo data generated by 2-to-1 back-translation are excerpted from Table 3.

en→ja models The additional data did not contribute to the translation quality, which indicates that the data augmentation using back-translation was not effective. This is partly due to difficult ja→en (back-)translation, and partly due to the difference between the original and pseudo parallel corpora. As shown in Table 4, there is clearly a gap between the original and pseudo parallel corpora in terms of the number of tokens per sentence. In IWSLT2017 datasets, the average number of tokens per sentence is almost equivalent between English and Japanese while in the pseudo parallel data English sentences are significantly shorter than the Japanese counterparts. This implies that some information has been lost in back-translating the Japanese monolingual corpus into English, and thus mismatches of the contents of the sentences in the two languages are likely to occur.

5.2 1-to-1 vs. 2-to-1 back-translation

To confirm the effect of using context-aware models instead of sentence-level models for back-translation, we additionally train ja→en models using pseudo parallel data generated by 1-to-1 and 2-to-2 back-translation. We train 1-to-1, 2-to-1, and 2-to-2 models on 500k, 1000k, 2000 and 4000k pseudo data. We conduct an evaluation using BLEU and the specialized test set we created (reported later in § 5.3), and compare the results with those trained on pseudo data generated by 2-to-1 back-translation.

Table 5 shows the evaluation results in BLEU. We observe comparable effect of the two back-translation methods, 1-to-1 and 2-to-1, on the forward-translation, whereas the 2-to-2 back-translation results in slightly higher scores of the forward-translation over the other two methods.

	Coref.	Coherence /cohesion
Bawden et al. (2018) / # train: 29M		
2-TO-2 (single-encoder best)	63.5	52.0
S-HIER-TO-2 (multi-encoder best)	72.5	**57.0**
2-to-2 (this paper) / # train: 222k		
(# pseudo train) 0k	70.0	51.0
500k	76.5	51.5
1000k	78.0	52.5
2000k	**78.5**	52.5

Table 6: Results of 2-to-2 models on the en→fr specialized test sets (accuracy in %).

# pseudo train	1-to-1 back-t.		2-to-1 back-t.		2-to-2 back-t.	
	2-to-1	2-to-2	2-to-1	2-to-2	2-to-1	2-to-2
0k	78	79	(same to the left)			
500k	87	84	85	89	83	89
1000k	**91**	89	81	89	**88**	88
2000k	86	90	88	**93**	87	**90**
4000k	85	**93**	**91**	**93**	86	89

Table 7: Results of 2-to-X models on the ja→en specialized test sets (accuracy in %).

5.3 Evaluation of context-aware translation using specialized test sets

Table 6 and 7 show results on the en→fr and ja→en specialized test sets, respectively. In what follows, we interpret results in detail.

Source	
context	彼女$_1$ の 20 代 も 困難 な もの でし た が Φ_2 それ 以前 の 人生 は もっと 困難 に 溢れ て い まし た
sentence	Φ_3 診察 中 何 度 も 涙 を 流し まし た が 「 家族 は 選べ ない けど 友達 は 選べ る 」 と その たび に 言っ て 気持ち を 落ち着か せ て い まし た
Target	
context	and as hard as **her**$_1$ 20s were , **her**$_2$ early life had been even harder .
sentence	**she**$_3$ often cried in our sessions, but then would collect herself by saying, "you can't pick your family, but you can pick your friends."
1-to-1	**I**'ve had tears in my doctor's office, and **I**'ve said, "I don't have a family, but I've got a friend," and **I** calmed down every time.
1-to-1 +2M pseudo data	**I** cried a lot during my examination, but every time **I** said, "I can't choose a family, but I can choose a friend," **I** said calmly.
2-to-2	during **my** diagnosis, **I** ran a lot of tears, and **I** said, "no family can choose," but every time I said, "I can choose a friend," **I** kind of calmed down.
2-to-2 +2M pseudo data	**she** cried many times during her examination, but each time she said, "I can't choose a family, but I can choose a friend," **she** said calmly.

Table 8: Example of translated sentences; zero pronoun Φ_3 is successfully restored in by the 2-to-2 model trained using 2M pseudo data. The corresponding pronouns in the source and target are modified with the same subscripts.

en→fr models Table 6 shows the results of 2-to-2 models with the data augmentation and the best performing models excerpted from (Bawden et al., 2018). 2-TO-2 is a single-encoder model using seq2seq (Bahdanau et al., 2015) instead of Transformer we have adopted, while S-HIER-TO-2 is a multi-encoder model. These models are trained from OpenSubtitles2016 corpus, which has 29M sentence pairs in the same domain as the test set.

When trained on a larger pseudo parallel data, 2-to-2 models achieved a higher accuracy for both coreference and coherence/cohesion datasets. Our 2-to-2 model trained using 2M pseudo parallel data outperforms by 15.0% and 6.0% on the coreference test set against the best-performing single and multi-encoder models trained with 29M in-domain parallel data. A possible explanation for this is that the coreference test is less domain-specific compared to the coherence/cohesion test set. To answer a typical question in the coreference test set, models need to recognize the pronouns in the source sentence, next find the antecedents of them in the source/target contexts, and then check if the gender agrees between them. This process, in most cases, does not require deep knowledge of the antecedent words because gender of a French word tends to be identified by its surface or the article and adjectives attached to it. On the other hand, the coherence/cohesion test includes questions imposing domain-specific tasks like lexical disambiguation, which require more knowledge about particular words specific to the domain. This explains the limited accuracy of our models trained in the domain of IWSLT2017 and Europarl, in contrast to the multi-encoder model S-HIER-TO-2 which is trained on OpenSubtitles2016, the same domain as the test set, achieving larger improvement.

ja→en models Table 7 lists the results of context-aware models. The models trained with larger pseudo parallel data achieve higher accuracy, as we have observed in the en→fr test set.

5.4 Qualitative Analysis

Table 8 shows examples of ja→en translation where the use of contexts and additional pseudo training data help improve the translation quality. Adding 2M pseudo data for training to the 1-to-1 model makes the translation much more fluent although the model cannot restore the correct pronoun "she." On the other hand, 2-to-2 without additional data cannot restore the correct pronoun either, and its translation is as awkward as that of the 1-to-1 model. By extending the sentence-level model with contexts (from both source and target) and adding pseudo data (2-to-2 + 2M pseudo training data), we obtain the best translation.

6 Related Work

Sennrich et al. (2016a) introduce a basic framework to exploit monolingual data (data augmentation by back-translation) for NMT. Imamura et al. (2018) show that back-translation using sampling instead of beam search generates more diverse

synthetic source sentences which are effective for enhancing the encoder. Edunov et al. (2018) further investigate the optimal back-translation procedure by comparing several methods such as beam search, random sampling and adding filter noise that randomly masks words in the synthetic source sentences. They focus on back-translation for sentence-level NMT whereas our interest lies in back-translation for context-aware models.

Although we have used simple beam search with the beam size of 5 for back-translation, those randomized back-translation strategies, if adopted, should strongly boost our baseline (sentence-level translation), as reported in (Imamura et al., 2018). These strategies can be applicable to the data augmentation for context-aware NMT, and would also improve the context-aware models' ability to capture contexts because they, especially adding filler noise (Edunov et al., 2018), produce source/target pairs in which some useful information for disambiguation is lost and the models need to try to find alternative hints in the context.

7 Conclusions

In this study, based on our hypothesis that the performance of context-aware models is more affected by the lack of the training data than sentence-level NMT models, we investigated the impact of large-scale parallel data on the translation quality of context-aware models. We conduct experiments of data augmentation based on back-translation, on four language directions en→ja, ja→en, en→fr and fr→en using IWSLT2017 datasets. The results of BLEU evaluation for ja→en and en→fr support our hypothesis. Through evaluation using the existing en→fr test set and our new ja→en test set, which are specialized in evaluating context-aware NMT models, we demonstrate that pseudo parallel data enhance context-aware NMT models in terms of the ability to capture contextual information.

In the future, we plan to assess the effectiveness of our approach on stronger baselines: multi-encoder models and the randomized back-translation strategies.

Acknowledgments

We deeply thank Dr. Shonosuke Ishiwatari for giving valuable comments on the early draft of this paper. This work was supported by JST CREST Grant Number JPMJCR19A4, Japan.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the third International Conference on Learning Representations (ICLR)*.

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 (NAACL-HLT)*, pages 1304–1313.

Alastair Butler, Kei Yoshimoto, Shota Hiyama, Stephen Wright Horn, Iku Nagasaki, and Ai Kubota. 2018. The keyaki treebank parsed corpus, version 1.1. http://www.compling.jp/keyaki/ accessed on 2019/06/01.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of European Association for Machine Translation (EAMT)*, pages 261–268.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 489–500.

Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. 2018. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the Second Workshop on Neural Machine Translation and Generation (WNMT)*, pages 55–63.

Kenji Imamura and Eiichiro Sumita. 2019. Incorporating long-distance contexts into dialogue translation. In *Proceedings of the 25th Annual meeting of the Association for Natural Language Processing*, pages 550–553. (in Japanese).

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics (TACL)*, 5:339–351.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the the third International Conference for Learning Representations (ICLR)*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The tenth Machine Translation Summit (MT Summit X)*, pages 79–86.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC)*, pages 923–929.

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 (ACL)*, pages 1275–1284.

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

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 (EMNLP)*, pages 2947–2954.

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 (WMT)*, pages 61–72.

Masaaki Nagata and Makoto Morishita. 2019. An evaluation metric for Japanese to English context-aware machine translation. In *Proceedings of the 25th Annual meeting of the Association for Natural Language Processing*, pages 1–4. (in Japanese).

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC)*, pages 2204–2208.

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 on association for computational linguistics (ACL)*, pages 311–318.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 1133–1137.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation (DiscoMT)*, pages 82–92.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association of Computational Linguistics (TACL)*, 6:407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1198–1212.

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

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

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.