Enhancement of Encoder and Attention Using Target Monolingual Corpora in Neural Machine Translation

Kenji Imamura, Atsushi Fujita, and **Eiichiro Sumita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{kenji.imamura,atsushi.fujita,eiichiro.sumita}@nict.go.jp

Abstract

A large-scale parallel corpus is required to train encoder-decoder neural machine translation. The method of using synthetic parallel texts, in which target monolingual corpora are automatically translated into source sentences, is effective in improving the decoder, but is unreliable for enhancing the encoder. In this paper, we propose a method that enhances the encoder and attention using target monolingual corpora by generating multiple source sentences via sampling. By using multiple source sentences, diversity close to that of humans is achieved. Our experimental results show that the translation quality is improved by increasing the number of synthetic source sentences for each given target sentence, and quality close to that using a manually created parallel corpus was achieved.

1 Introduction

In recent years, neural machine translation (NMT) based on encoder-decoder models (Sutskever et al., 2014; Bahdanau et al., 2014) has become the mainstream approach for machine translation. In this method, the encoder converts an input sentence into numerical vectors called "states," and the decoder generates a translation on the basis of these states. Although the encoder-decoder models can generate high-quality translations, they require large amounts of parallel texts for training.

On the other hand, monolingual corpora are readily available in large quantities. Sennrich et al. (2016a) proposed a method using synthetic parallel texts, in which target monolingual corpora are translated back into the source language (Figure 1). The advantage of this method is that the decoder is accurately trained because the target side of the synthetic parallel texts consists of manually created (correct) sentences. Consequently, this method provides steady improvements. However, this approach may not contribute to the improvement of the encoder because the source side of the synthetic parallel texts are automatically generated.

In this paper, we extend the method proposed by Sennrich et al. (2016a) to enhance the encoder and attention using target monolingual corpora. Our proposed method generates multiple source sentences by sampling when each target sentence is translated back. By using multiple source sentences, we aim to achieve the following.

- To average errors in individual synthetic sentences and reduce their harmful effects.
- To ensure diversity as human translations. This is a countermeasure against machine-translated sentences that have less variety.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work that uses monolingual corpora in NMT. Section 3 describes the proposed method, and Section 4 evaluates the proposed method through experiments. In addition, Section 5 proposes the application of our method as a self-training approach. Finally, Section 6 concludes the paper.

2 Related Work

One approach of using target monolingual corpora is to construct a recurrent neural network language model and combine the model with the decoder (Gülçehere et al., 2015; Sriram et al., 2017). Similarly, there is a method of training language models, jointly with the translator, using multi-task learning (Domhan and Hieber, 2017). These

Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 55–63
Melbourne, Australia, July 20, 2018.

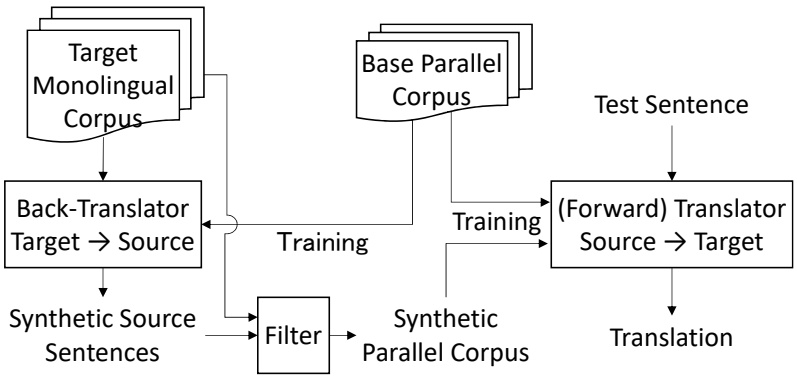


Figure 1: Flow of Our Approach

methods only enhance the decoder and require a modification of the NMT.

Another approach of using monolingual corpora of the target language is to learn models using synthetic parallel sentences. The method of Sennrich et al. (2016a) generates synthetic parallel corpora through back-translation and learns models from such corpora. Our proposed method is an extension of this method. Currey et al. (2017) generated synthetic parallel sentences by copying target sentences to the source. This method utilizes a feature in which some words, such as named entities, are often identical across the source and target languages and do not require translation. However, this method provides no benefits to language pairs having different character sets, such as English and Japanese.

On the other hand, the basis of source monolingual corpora, a pre-training method based on an autoencoder has been proposed to enhance the encoder (Zhang and Zong, 2016). However, the decoder is not enhanced by this method. Cheng et al. (2016) trained two autoencoders using source and target monolingual corpora, while translation models are trained using a parallel corpus. This method enhances both the encoder and decoder, but it requires two monolingual corpora, respectively. Our proposed method enhances not only the decoder but also the encoder and attention using target monolingual corpora.

3 Proposed Method

3.1 Synthetic Source Sentences

The back-translator used in this study is an NMT trained on a small parallel corpus (hereinafter referred to as the base parallel corpus). Each sentence in a target monolingual corpus is translated by the back-translator to generate synthetic source sentences. The back-translator does not output only high-likelihood sentences but generates sentences by random sampling.

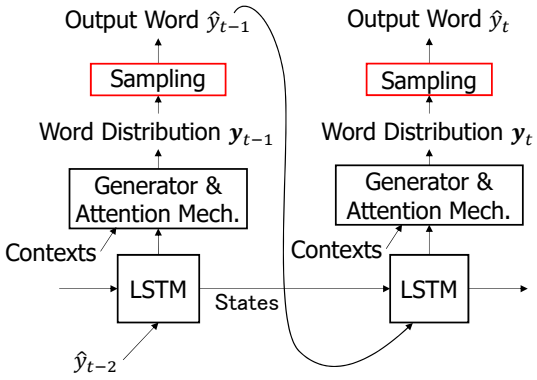


Figure 2: Decoding Process of Back-Translator

Figure 2 illustrates the decoding process of the back-translator. When the decoder generates a sentence word-by-word, it also generates the posterior probability distribution of an output word $\Pr(y_t)$ through the decoding process. We call this a word distribution. In a usual decoding process, the output word $\hat{y}_t$ is determined by selecting a word with the highest probability (if the decoder outputs 1-best translation by greedy search). [1]

$$\hat{y}_t = \underset{y_t}{\operatorname{argmax}} \Pr(y_t | \boldsymbol{y}_{<t}, \boldsymbol{x}), \qquad (1)$$

where $\boldsymbol{y}_{<t}$ and $\boldsymbol{x}$ are the history of the output words and the input word sequence, respectively.

In contrast, the back-translator in this paper determines the output word by sampling based on the

[1] In translation, an output sentence is generally generated from multiple hypotheses using beam search. However, it is the same that the beam search selects high-likelihood words.

Log-likelihood	Synthetic Source Sentence
-2.25	what should i do when i get injured or sick in japan ?
-2.38	what should i do *if* i get injured or sick in japan ?
-5.20	what should i do *if* i get injured or *illness* in japan ?
-5.52	what should *we* do when *we* get injured or sick in japan ?
-13.87	*if i get injured or a sickness in japan , what shall i do ?*
Target Sentence	日本 で 怪我 や 病気 を し た とき は どう すれ ば いい の でしょう か ?
Manual Back-Translation	what should i do when i get injured or sick in japan ?

Table 1: Examples of Synthetic Source Sentences (English-Japanese Translation): The italicized words indicate differences with the manual back-translation.

word distribution.

$$\hat{y}_t = \underset{y_t}{\text{sampling}}(\Pr(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x})), \qquad (2)$$

where $\text{sampling}_y(P)$ denotes the sampling operation of y based on the probability distribution P. The decoding continues until the end-of-sentence symbol is generated.[2] We repeat the above process to generate multiple synthetic sentences. Note that this generation method is the same as that of the minimum risk training (Shen et al., 2016).

In NMT, even if a low-probability word is selected by the sampling, the subsequent word would become fluent because it is conditioned by the history. Table 1 presents examples of the synthetic source sentences produced by the back-translator. Most of the synthetic source sentences are identical, or close to, the manual back-translation (i.e., the reference translation). On the other hand, the last example is quite different from the perspective of word order because the clauses are inverted. Such a synthetic sentence is usually not produced by the n-best translation because of the low likelihood. However, it is possible to generate diverse source sentences by sampling.

The sampling occasionally generates identical sentences as a result. However, we did not remove the duplication to reflect the original probability distribution.

3.2 Training

The synthetic source sentences are paired with the target sentences to construct the synthetic parallel corpus. The NMT model is trained on a mixture of the synthetic corpus and the base parallel corpus.

In the training, we must deal with the two different types of sentence pairs. In addition, if we use multiple source sentences for a given target sentence, the model will be biased toward the synthetic corpus. To avoid this problem, we adjust the learning rate according to the size of the corpora. Specifically, we first configure two mini-batch sets each from the base and synthetic corpora. Thereafter, the learning rate η/N is applied to the mini-batches of the synthetic corpus, in contrast to the learning rate η for those of the base corpus, where N denotes the number of synthetic source sentences per target sentence. Finally, the two sets are shuffled and used for training.

The training time increases along with the increase of data. However, the translation speed does not change because the model structure is not changed.

It must be noted that if the domains of the base parallel and the target monolingual corpora are different, it is better to perform "further training" using the base parallel corpus for domain adaptation (Freitag and Al-Onaizan, 2016; Servan et al., 2016). [3]

3.3 Filtering of Synthetic Parallel Sentences

The synthetic source sentences contain errors. A direct approach to reduce such errors involves filtering the sentence pairs according to their quality. In this paper, we consider the following three methods.

3.3.1 Likelihood Filtering

The first method is filtering by the likelihood output from the back-translator. We consider the likelihood as an indicator of translation quality, and low-likelihood synthetic sentences are filtered

[2]The back-translator does not use the beam search because the sampling is independently performed for each word.

[3]We did not perform "further training" in this paper.

out. Note that the likelihood is corrected with the length of the synthetic source sentence. We call this the length biased log-likelihood ll_{len} (Oda et al., 2017).

$$ll_{\text{len}}(\boldsymbol{y}|\boldsymbol{x}) = \sum_t \log \Pr(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) + WP \cdot T, \quad (3)$$

where the first term on the right-hand side is the log-likelihood, WP denotes the word penalty ($WP \geq 0$), and T denotes the number of words in the synthetic source sentence.

NMTs tend to generate shorter translations than the expectation (Morishita et al., 2017). The word penalty works to increase the likelihood of long hypotheses when it is set to a positive value. With an appropriate value, we can obtain synthetic sentences that are almost of the same length as the manual back-translation. We set the word penalty such that the lengths of the translation and reference translation on the development set are approximately equal, using line search.

3.3.2 Confidence Filtering

The second method involves filtering with the confidence of translation used in the translation quality estimation task. We use the data provided by Fujita and Sumita (2017), which is a collection of manual labels indicating whether the translation is acceptable or not. We train the support vector machines (SVMs) on the sentence-level data and regard the classifier's score as the confidence score.

The features of the SVM classifier include the 17 basic features of QuEst++ (Specia et al., 2015).[4] They are roughly categorized into the following two types.

- Language model features of each of the source and target sentences.
- Features based on the parallel sentences such as the average number of translation hypotheses per word.

In addition, we add the source and target word embeddings. The sentence features are computed by averaging all word embeddings (Shah et al., 2016). The hyperparameters for the training are set using the grid search on the development set.

In the expriments of Section 4, features are extracted from the base parallel corpus.

[4] http://www.quest.dcs.shef.ac.uk/

Type		# Sentences
Parallel	Base	400,000
	Development	2,000
	Test	2,000
Monolingual (Japanese)	GCP Corpus	1,552,475
	BCCWJ	4,791,336

Table 2: Corpus Statistics

3.3.3 Random Filtering

The third method is random filtering. This is identical to the reduction of the number of synthetic source sentences to be generated.

4 Experiments

4.1 Experimental Settings

Corpora The corpus sizes used here are shown in Table 2. We used the global communication plan corpus (the GCP corpus, (Imamura and Sumita, 2018)), which is an in-house parallel corpus of daily life conversations and consists of Japanese (Ja), English (En), and Chinese (Zh). The experiments were performed on English-to-Japanese and Chinese-to-Japanese translation tasks. We randomly selected 400K sentences for the base parallel corpus, and the remaining (1.55M sentences) were used as the Japanese monolingual corpus. The reason for dividing the parallel corpus into two corpora is to measure the upper-bound of quality improvement by using existing parallel texts on the same domain as the manual back-translation.

We also used the Balanced Corpus of Contemporary Written Japanese (BCCWJ)[5] as a monolingual corpus from a different domain. We used approximately 4.8M sentences, each of which contains less than 1024 characters. We assume practical situations in which the domains of parallel and monolingual corpora are not identical.

All sentences were segmented into words using an in-house word segmenter. The words were further segmented into 16K sub-words based on the byte-pair encoding rules (Sennrich et al., 2016b) acquired from the base parallel corpus for each language independently.

Translation System The translation system used in this study was OpenNMT (Klein et al., 2017). We modified it to accept Sections 3.1 and 3.2.

[5] http://pj.ninjal.ac.jp/corpus_center/bccwj/en/

The encoder was comprised of a two-layer Bi-LSTM (500 + 500 units), the decoder included a two-layer LSTM (1,000 units), and the stochastic gradient descent was used for optimization. The learning rate for the base parallel corpus was 1.0 for the first 14 epochs, followed by the annealing of 6 epochs while decreasing the learning rate by half. The mini-batch size was 64.

At the translation stage, we generated 10-best translations and selected the best among them on the basis of the length reranking (Morishita et al., 2017). Equation 3 was used as the score function for the reranking. By correcting the translation length, the translation quality can be compared without the effect of the brevity penalty of the BLEU score.

The back-translator was comprised of the same system. We generated 10 synthetic source sentences per target sentence using the method described in Section 3.1, and filtered them to create synthetic parallel sentences.

Competing Methods In this paper, we consider the case in which only the base parallel corpus is used as the baseline, and the case in which the manual back-translation of the GCP corpus is added as the upper-bound of the translation quality. Thereafter, we compare the following methods and settings:

- Various numbers of synthetic source sentences for a given target sentence
- The methods for generating synthetic source sentences: sampling vs. n-best generation
- The three filtering methods described in Section 3.3

Evaluation BLEU (Papineni et al., 2002) was used for the evaluation. The multeval tool (Clark et al., 2011)[6] was used for statistical testing at a significance level of 5% ($p < 0.05$).

4.2 Results with GCP Corpus

Figures 3 and 4 depict the relationship between the number of synthetic source sentences and the BLEU score on the GCP corpus of En-Ja and Zh-Ja translation tasks, respectively. The graphs and tables in the figures present the same data for overviews and for analyzing the data in detail. Note that the method of Sennrich et al. (2016a) corresponds to the case of one synthetic source sentence of the n-best generation (i.e., 1-best generation).

In both En-Ja and Zh-Ja translation, the score was improved when multiple synthetic sentences were given. Even though the method of Sennrich et al. (2016a) achieved improvements of $+2.42$ and $+2.38$ BLEU points from the base corpus only for En-Ja and Zh-Ja translations, respectively, further improvements were observed by using multiple synthetic sentences. Since the target sentences were the same in all cases except for the base corpus only, we can conclude that providing multiple source sentences is effective for improving the encoder and attention.[7]

The improvements from the base corpus only to the manual back-translation reached $+4.86$ and $+5.29$ BLEU points in En-Ja and Zh-Ja translations, respectively. When we focus on the case in which the number of synthetic source sentences is 6, for example, the improvements in the proposed methods (the likelihood, confidence, and random filtering) were achieved at least $+4.08$ and $+5.01$ BLEU points. This means that more than 80% of improvements with the manual back-translation were achieved using only monolingual corpora. Nevertheless, all methods did not reach the BLEU score of the manual back-translation; thus, we cannot substitute parallel corpora with monolingual corpora.

When we compared the three filtering methods, the BLEU scores were almost equivalent in most cases. In fact, there were no significant differences among filtering methods in all cases of Zh-Ja translation. In En-Ja translation, there were some significantly different cases, but the significance was not consistently derived.

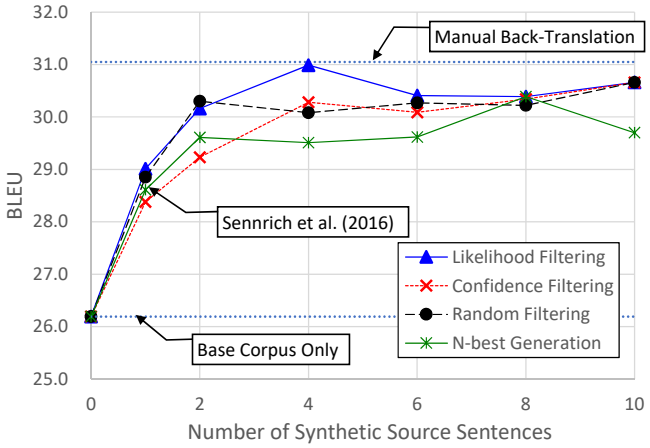
When the synthetic source generation was changed to the n-best generation, the BLEU scores were visibly degraded relative to the proposed method (i.e., sampling). We speculate that the likelihood and confidence filtering were ineffective because of the high-quality back-translator, and the diversity of the synthetic source sentences contributed considerably to quality improvement.

4.3 Results with BCCWJ

Table 3 shows the results using BCCWJ as a monolingual corpus (the results of the GCP cor-

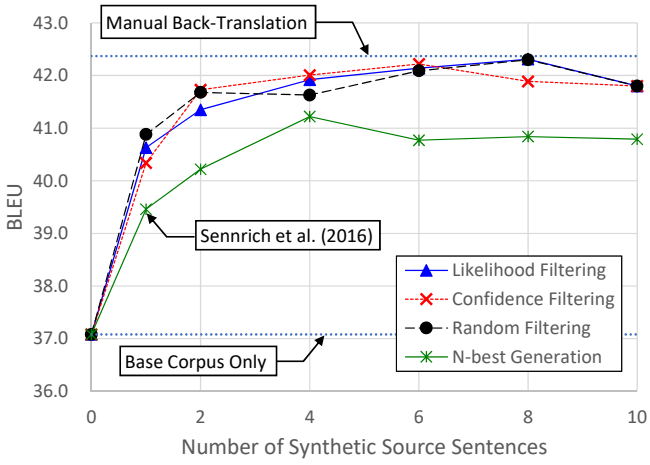
[6] https://github.com/jhclark/multeval

[7] Unfortunately, it is unknown in this experiment whether the encoder or attention were enhanced. We plan to investigate which module is enhanced by freezing parameters (Zoph et al., 2016) of the encoder and attention through the training.



# of Synthetic Sentences	En-Ja Likelihood Filtering	Confidence Filtering	Random Filtering	N-best Generation
Base Corpus Only	26.19			
Sennrich et al. (2016a)	28.61 (+2.42)			
1	29.01 (+2.82)	28.49 (+2.30)	28.85 (+2.66)	28.61 (+2.42)
2	30.16 (+3.97)	29.26 (+3.07)	30.30 (+4.11)	29.61 (+3.42)
4	30.99 (+4.80)	30.26 (+4.07)	30.08 (+3.89)	29.51 (+3.32)
6	30.41 (+4.22)	30.59 (+4.40)	30.27 (+4.08)	29.62 (+3.43)
8	30.39 (+4.20)	30.53 (+4.34)	30.22 (+4.03)	30.39 (+4.20)
10	30.66 (+4.47)	30.66 (+4.47)	30.66 (+4.47)	29.70 (+3.51)
Manual Back-Translation	31.05 (+4.86)			

Figure 3: The BLEU scores using the GCP corpus (English-Japanese translation) represented by a graph and table. The bracketed values of the table indicate differences from those of the base corpus only.



# of Synthetic Sentences	Zh-Ja Likelihood Filtering	Confidence Filtering	Random Filtering	N-best Generation
Base Corpus Only	37.08			
Sennrich et al. (2016a)	39.46 (+2.38)			
1	40.63 (+3.55)	40.34 (+3.26)	40.88 (+3.80)	39.46 (+2.38)
2	41.35 (+4.27)	41.73 (+4.65)	41.68 (+4.60)	40.22 (+3.14)
4	41.92 (+4.84)	42.01 (+4.93)	41.63 (+4.55)	41.22 (+4.14)
6	42.14 (+5.06)	42.22 (+5.14)	42.09 (+5.01)	40.77 (+3.69)
8	42.31 (+5.23)	41.89 (+4.81)	42.30 (+5.22)	40.84 (+3.76)
10	41.80 (+4.72)	41.80 (+4.72)	41.80 (+4.72)	40.79 (+3.71)
Manual Back-Translation	42.37 (+5.29)			

Figure 4: The BLEU scores using the GCP corpus (Chinese-Japanese translation) represented by a graph and table. The bracketed values of the table indicate differences from those of the base corpus only.

# of Synthetic Source Sentences	BCCWJ BLEU	GCP Corpus BLEU
0 (Base Corpus Only)	26.19	
1	29.84	29.01
2	29.94	30.16
4	30.66	30.99
Manual Back-Translation	-	31.05

Table 3: The BLEU scores of the BCCWJ and GCP corpora according to the number of synthetic source sentences (En-Ja, the random filtering).

	Sampling	N-best Gen.
BLEU	15.05	21.55
Edit Distance		
A) between SYN and MAN	9.73	8.52
B) among SYNs	9.34	3.90

Table 4: The BLEU scores and the edit distances of synthetic source sentences based on 10 synthetic sentences and manual back-translation for the same 1,000 target sentences in the GCP corpus.

pus are also shown for reference). In this study, we only performed random filtering experiments on En-Ja translation due to resource limitations.

In the case of BCCWJ, the BLEU scores increased with the number of synthetic source sentences, similar to the GCP corpus. We cannot directly compare the scores of the two corpora; however, similar improvement was achieved when we used a several-fold size of the different domain monolingual corpus.

4.4 Analysis

The above experiments consider diversity under the following two assumptions.

- The number of synthetic source sentences indicates the diversity.
- The diversity of the synthetic sentences by sampling is higher than that of the n-best generation.

In this section, we quantify the diversity using the edit distance among the systhetic source sentences to compare the generation methods.

We sampled 1,000 Japanese sentences from the GCP corpus in En-Ja translation with their ten corresponding back-translations generated by each method. Table 4 shows the results. The BLEU scores were computed regarding the 10,000 sentences as a document. The edit distances were computed for the following two cases, setting the insertion, deletion, and substitution costs to 1.0.

A) The average distance between a synthetic sentence (SYN) and the manual back-translation (MAN; i.e., reference translation). Note that this value also indicates translation quality because it is a source for computing the word error rate (smaller value represents better quality).

B) The average distance among synthetic source sentences of a target sentence (${}_{10}C_2 = 45$ combinations per target sentence).

As for the BLEU scores in Table 4, the sampling method achieved a lower score than that of the n-best generation. Similarly, the edit distance A of the sampling had a larger value than that of the n-best generation. These results imply that the sampling generates poor synthetic sentences. However, these scores are influenced by the diversity because they naturally become worse along with the variety of synthetic sentences when they are computed using a single reference.

On the other hand, as for edit distance B, the distance of the n-best generation was less than half of that of the sampling, even though sentences of the sampling generation can include identical sentences. Intuitively, the n-best generation generates similar sentences where only few words are different. As shown in Table 4, the distances of the synthetic sentences by sampling were almost the same as those from the manual back-translation, and the distances by the n-best generation were not. This result verifies that the generation by sampling increases the diversity of the synthetic source sentences.

5 Application to Self-Training Using Parallel Corpora

In this paper, we enhanced the encoder and attention using target monolingual corpora. Our proposed method can be applied to a self-training method only using parallel corpora. Specifically, we train a back-translator using a given parallel corpus, and the target side of the parallel corpus is translated into the source. Then, the original and synthetic parallel corpora are mixed. We finally train the forward translator using this corpus to enhance the encoder.

# of Source Sentences	BLEU
1 (Manual Bitexts Only)	31.05
2 (Manual Bitexts + 1 Syn. Sentence)	31.29
4 (Manual Bitexts + 3 Syn. Sentences)	31.65
6 (Manual Bitexts + 5 Syn. Sentences)	31.75
8 (Manual Bitexts + 7 Syn. Sentences)	32.25
10 (Manual Bitexts + 9 Syn. Sentences)	32.28

Table 5: The effect of self-training (En-Ja translation)

5.1 Settings

We confirm whether the quality can be improved from the upper-bound of the experiments in Section 4.

The experimental settings were the same as those of Section 4 except for the corpora. We considered the mixture of the base and GCP corpora (including the manual back-translation) in Table 2 as the original parallel corpus, with 1.95M sentences. The monolingual corpus was the target side of the entire parallel corpus. The back-translator generated nine synthetic source sentences, and they were randomly filtered. The original and synthetic parallel corpora were concatenated to train the forward translator. Namely, the number of source sentences per target sentence was at most ten.

In this experiment, we used the learning rate η for the original parallel corpus and η/N for the synthetic parallel corpus, where N denotes the number of synthetic source sentences per target sentence. The learning rate was $\eta = 0.5$, which means 1.0 for a target sentence in total.

5.2 Results

Table 5 shows the BLEU scores in the En-Ja translation according to the number of source sentences. Similar to the results in Section 4, the BLEU scores increased along with the increase in the number of source sentences. When we added nine synthetic source sentences, the BLEU score was improved by $+1.23$ points in comparison to the manual bitext only. Therefore, by increasing the diversity of the manual translation using synthetic sentences, we can further enhance the encoder and attention.

6 Conclusions

In this paper, we enhanced the encoder and attention by using multiple synthetic source sentences, in which target monolingual corpora were translated by sampling. During the training, we used different learning rates for the base and synthetic parallel corpora to avoid overfitting to the synthetic corpus. As a result, the translation quality was improved by increasing the number of synthetic source sentences for a given target sentence, and the quality approached that of the manual back-translation. In addition, we confirmed the generation by sampling synthesized diverse source sentences and consequently improved the translation quality in comparison with the n-best generation. We also attempted some filtering methods on the synthetic source sentences to obtain improved parallel sentences, but we could not confirm their effectiveness in our experiments.

Our future work is to clarify the other conditions where the proposed method is effective, such as the relationship between qualities of the backward and forward translations, experiments on public data sets, and comparison with the number of synthetic sentences and monolingual corpus size at the same training time. In addition, we plan to consider other applications, such as applying our methods to smaller parallel corpora and using source monolingual corpora.

Acknowledgments

The authors appreciate anonymous reviewers for their helpful comments.

This work was supported by "Promotion of Global Communications Plan — Research and Development and Social Demonstration of Multilingual Speech Translation Technology," a program of the Ministry of Internal Affairs and Communications, Japan.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, pages 176–181, Portland, Oregon, USA.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Atsushi Fujita and Eiichiro Sumita. 2017. Japanese to English/Chinese/Korean datasets for translation quality estimation and automatic post-editing. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 79–88, Taipei, Taiwan.

Çaglar Gülçehere, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Ling, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. In *CoRR, abs/1503.03535*.

Kenji Imamura and Eiichiro Sumita. 2018. Multilingual parallel corpus for global communication plan. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3453–3458, Miyazaki, Japan.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2017. NTT neural machine translation systems at WAT 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 89–94, Taipei, Taiwan.

Yusuke Oda, Katsuhito Sudoh, Satoshi Nakamura, Masao Utiyama, and Eiichiro Sumita. 2017. A simple and strong baseline: NAIST-NICT neural machine translation system for WAT2017 English-Japanese translation task. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 135–139, Taipei, Taiwan.

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 (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

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-2016, Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Christophe Servan, Josep Maria Crego, and Jean Senellart. 2016. Domain specialization: a post-training domain adaptation for neural machine translation. *CoRR*, abs/1612.06141.

Kashif Shah, Fethi Bougares, Loïc Barrault, and Lucia Specia. 2016. Shef-lium-nn: Sentence level quality estimation with neural network features. In *Proceedings of the First Conference on Machine Translation*, pages 838–842, Berlin, Germany.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2017. Cold fusion: Training seq2seq models together with language models. *CoRR*, abs/1708.06426.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 3104–3112.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-2016)*, pages 1535–1545, Austin, Texas.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas.