NTT Neural Machine Translation Systems at WAT 2017

Makoto Morishita, Jun Suzuki and **Masaaki Nagata**
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan
{morishita.makoto,suzuki.jun,nagata.masaaki}@lab.ntt.co.jp

Abstract

In this year, we participated in four translation subtasks at WAT 2017. Our model structure is quite simple but we used it with well-tuned hyper-parameters, leading to a significant improvement compared to the previous state-of-the-art system. We also tried to make use of the unreliable part of the provided parallel corpus by back-translating and making a synthetic corpus. Our submitted system achieved the new state-of-the-art performance in terms of the BLEU score, as well as human evaluation.

1 Introduction

In this paper, we describe our systems submitted to this year's translation shared tasks at WAT 2017 (Nakazawa et al., 2017). For this year, we focused on scientific paper (ASPEC Japanese-English, English-Japanese) and newspaper (JIJI Corpus Japanese-English, English-Japanese) translation subtasks.

We use a simple Neural Machine Translation (NMT) model with an attention mechanism (Luong et al., 2015). In addition, for ASPEC, we made a synthetic corpus for the unreliable part of the provided corpus, in a way similar to that reported by Sennrich et al. (Sennrich et al., 2016a). This technique and the well-tuned hyper-parameters led to new state-of-the-art results in all the subtasks in which we participated.

2 Common Settings

2.1 Model Structure

Our model is based on the encoder-decoder with a global attention model proposed by Luong et al. (2015), with a general scoring function and input feeding. The original model uses a uni-directional encoder, but we changed it to a bi-directional one proposed by Bahdanau et al. (2015). After running the bi-directional encoder, we simply added each state and used it for a decoder.

We implemented this model with Chainer toolkit (Tokui et al., 2015), and the implementation is now open for further experiments[1].

2.2 Data Preprocessing

First, we tokenize the provided corpus using KyTea (Neubig et al., 2011) for the Japanese side, and Moses tokenizer[2] for the English side. We remove the sentences over 60 words to clean the corpus. Then we further split it into sub-words using joint byte pair encoding (joint-BPE) (Sennrich et al., 2016c) with applying 16,000 merge operations.

For ASPEC subtasks, though the provided training data contained over 3.0M sentences, we only used the first 2.0M sentences, in the same way as the previous participants (Neubig, 2014). ASPEC was collected by aligning parallel sentences automatically and sorting them on the basis of the alignment confidence score (Nakazawa et al., 2016). This means that the latter side of the corpus may contain noisy parallel sentences, which would have a negative impact on training. We used the latter 1.0M sentences as a monolingual corpus and made a synthetic corpus (see section 3.1.1 for details).

2.3 Training

Table 1 shows the settings of hyper-parameters we used and tested. We tried several combinations

[1] https://github.com/nttcslab-nlp/wat2017

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

Proceedings of the 4th Workshop on Asian Translation, pages 89–94,
Taipei, Taiwan, November 27, 2017.

Hyper-parameter	Used	Tested
Vocabulary size	16,000	1,000, 5,000
Embedding dimension	512	—
Hidden dimension	512	—
Attention dimension	512	—
Encoder layer	2	4, 1
Decoder layer	2	4, 1
Optimizer	SGD	—
Initial learning rate	1.0	0.5
Gradient clipping	5.0	6.0
Dropout rate	0.3	0.2, 0.0
Mini-batch size	128 sent	64, 256

Table 1: Hyper-parameter settings

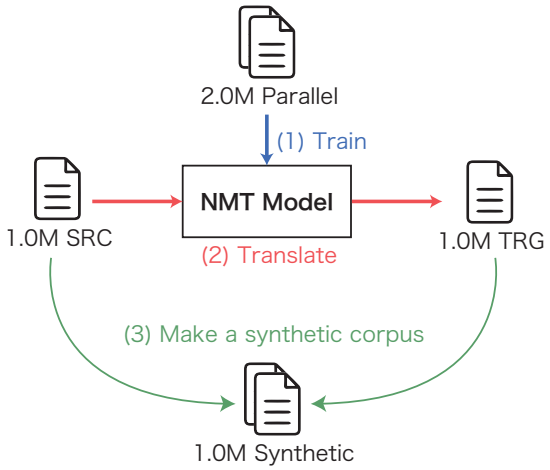

Figure 1: Overview of making a synthetic corpus. First, we make an NMT model with a reliable parallel corpus, then translate the unreliable part of the corpus to make a synthetic parallel corpus.

and we found that these settings were the best. For the vocabulary, we only included the most frequent 16,000 sub-words in the training set[3]. After 13 epochs, we multiplied the learning rate by 0.7 for every epoch, then continued training till 20 epochs.

2.4 Testing

2.4.1 Length Normalized Re-ranking

Naive beam searches with a large beam size may tend to output shorter sentences, leading to a drop in performance (Tu et al., 2017). To reduce this negative effect, we re-ranked the candidate output sentences $\boldsymbol{t}$ by using the following score function once we finished the beam search (Cromieres et al., 2016):

$$\hat{t} = \arg\max_{t \in \boldsymbol{t}} \left\{ \frac{p(t)}{|t|} \right\}, \quad (1)$$

where $p(t)$ is the predicted log-probability of a candidate output sentence t and $|t|$ is the length of t.

With this length normalized re-ranking, we can use a large beam size without taking the above explained negative effect into account. Through preliminary experiments, we found that a beam size of 20 was sufficient.

2.4.2 Ensembling

It has been reported that ensembling several different models together significantly improves performance. In an ensembling process, several models are run at each time step and an arithmetic mean of predicted probability is obtained, which is used to determine the next word. In our settings, we trained eight models independently and used them for the ensemble.

[3]Applying joint-BPE with 16,000 merge operations should make the vocabulary size under 16,000 sub-words, but for Japanese, it may contain some unknown characters (kanji). The actual vocabulary size for each corpus was the following: ASPEC Ja:11271, En:10942, JIJI Ja:16000, En:15795

3 Task-Specific Settings

3.1 ASPEC

3.1.1 Synthetic Corpus

As we mentioned in section 2.2, ASPEC contains some unreliable sentence pairs. For SMT, we can use these sentences as monolingual data to train a language model. However in the current NMT model architecture, the model cannot be trained with monolingual data, so the previous participants with NMT models simply ignored these parts of the data (Neubig, 2016; Eriguchi et al., 2016).

In a way similar to that reported by Sennrich et al. Sennrich et al. (2016b), we tried to use the unreliable part of the corpus by making a synthetic corpus. Figure 1 illustrates the overview of how we made the synthetic corpus. First, we made an NMT model with the reliable part of the provided data (in our case, the first 2.0M sentences), then translated the unreliable part of the corpus by using it to make a synthetic corpus. Finally, we made a corpus of 3.0M sentences by concatenating this

synthetic corpus and the reliable part of the data. With this corpus, we continued the training of the model for a further 10 epochs.

It should be noted that the target side of the synthetic corpus should be the original sentences (not those generated by the NMT model). This is because an NMT model includes a target side language model and uses it to generating a natural sentence, so it would be better to keep the target side original to train an NMT model effectively. Thus, the synthetic corpus used for Japanese-English training is made with an English-Japanese NMT model, and vice versa.

3.2 JIJI

3.2.1 Model Fine-tuning

We thought the JIJI corpus was too small to train an NMT model, so we tried to train the model with other large parallel corpora and then fine-tune it with the JIJI corpus (Luong and Manning, 2015). In our settings, we first trained the model with ASPEC (2.0M) and Japan Patent Office Patent Corpus (JPC) (1.0M). We learned BPE codes with the JIJI corpus and applied them to ASPEC and JPC. We trained the model with ASPEC and JPC for 20 epochs, then continued training with the JIJI corpus for a further 20 epochs.

4 Official Results

Tables 2 and 3 show the official results of our submissions[4]. Our system achieved the best BLEU scores and adequacy for all the subtasks in which we participated. For pairwise crowdsourcing evaluations, our system also obtained the best evaluations except for the ASPEC Ja-En subtask. Even in this case, it obtained the second best evaluation.

5 Analysis

5.1 Synthetic Corpus

From Table 2, we can see that the synthetic corpus has a positive impact on the performance, especially for the En-Ja subtask, and contributes to achieving better performance. We also tried to use the original 3.0M corpus for training, but could not see any improvements over the model that uses only the first 2.0M sentences.

Manually comparing the synthetic corpus and the originally provided corpus, we found that the quality of the synthetic corpus was much better than the original one. The original corpus often includes noisy pairs where the contents are different on each side. Table 4 shows an example sentence of the original parallel corpus and our synthetic corpus. The original Japanese sentence does not contain the words for "*a nonlinear least squares method*" and "*the method of steepest decent*", but the synthetic sentence contains these words and improves the quality of the parallel corpus. Using a synthetic corpus makes it possible to alleviate the noisy sentences and helps to achieve better performance.

5.2 Model Fine-tuning

We thought that training with a larger amount of data would enable the model to use more sentences and that this would be beneficial for further training. However, as is clear from Table 3, we couldn't find any improvements over fine-tuning. We suspect that the parallel corpus used to initialize the model is quite out-of-domain, so the model couldn't get any benefits from it.

5.3 JIJI Corpus Quality

In the JIJI corpus subtasks, we were only able to see a small correlation between BLEU scores and human evaluation. To find out the reason for this, we manually looked into the JIJI corpus. In doing so, we found that it was too noisy for efficient learning. It contained a lot of parallel sentences with different content, which can be noise for NMT training. The JIJI corpus originally comes from Japanese news articles that were translated into English. During this process, translators often add or remove the content of the article to make it easy to understand for English readers. However, this makes it hard to find clean one-by-one sentence alignment and leads to make the parallel corpus dirty. As a result, the trained model learns to generate a sentence with a different meaning, and it leads to a higher BLEU score but lower human evaluations. To deal with this problem, it would be better to consider how to train a cleaner model from a noisy parallel corpus.

5.4 BLEU Scores and Tokenizer

After the evaluation period finished, we found that our BLEU scores tended to be better with KyTea tokenizer. In the English-Japanese subtasks, participants de-tokenize system outputs and the submission system will re-tokenize them with JU-

[4]In these tables, we exclude the organizer's submissions for ranking.

	System	BLEU	Rank	Pairwise	Rank	Adequacy	Rank
En-Ja	Single (3.0M)	37.15	—	—	—	—	—
	Single (2.0M)	37.90	7/14	—	—	—	—
	Single (2.0M + 1.0M Synthetic)	38.87	4/14	—	—	—	—
	8 Ensemble (2.0M)	39.80	3/14	72.250	3/11	—	—
	8 Ensemble (2.0M + 1.0M Synthetic)	**40.32**	**1/14**	**75.750**	**1/11**	**4.41**	**1/4**
Ja-En	Single (3.0M)	26.07	—	—	—	—	
	Single (2.0M)	27.43	6/13	75.000	4/10	—	—
	Single (2.0M + 1.0M Synthetic)	27.62	4/13	—	—	—	—
	8 Ensemble (2.0M)	**28.36**	**1/13**	**77.250**	**2/10**	**4.14**	**1/2**
	8 Ensemble (2.0M + 1.0M Synthetic)	28.15	2/13	—	—	—	—

Table 2: Official results of our submitted systems for ASPEC subtasks. For the En-Ja subtask, we show the BLEU scores with JUMAN tokenizer.

	System	BLEU	Rank	Pairwise	Rank	Adequacy	Rank
En-Ja	Single	19.13	3/4	14.500	2/3	—	—
	8 Ensemble	**20.37**	**1/4**	**17.750**	**1/3**	**2.03**	**1/2**
Ja-En	Single	19.44	2/8	**32.000**	**1/6**	**2.05**	**1/2**
	Fine-Tuning	15.77	7/8	—	—	—	—
	8 Ensemble	**20.90**	**1/8**	26.750	2/6	—	—

Table 3: Official results of our submitted systems for JIJI corpus subtasks. For En-Ja subtasks, we show the BLEU scores with JUMAN tokenizer.

MAN, KyTea or MeCab tokenizers, then calculate the BLEU scores. In our experiments, we first pre-tokenized sentences with KyTea tokenizer, and then further split them into sub-words by applying BPE. Therefore, we suspect that our systems are likely to be optimized with KyTea, so we carried out experiments using JUMAN as a pre-tokenizer. Table 5 shows the BLEU scores of our systems pre-tokenized with KyTea or JUMAN. From the results, we found that if we used JUMAN as a pre-tokenizer, we achieved better BLEU scores calculated with JUMAN tokenizer.

5.5 Beam Size and Length Normalized Re-ranking

Figure 2 shows the BLEU score changes in terms of increasing the beam size with the length normalized re-ranking described in section 2.4.1 (w/ LN), and without it (w/o LN). In the case of w/ LN, the BLEU score tends to gradually get better by increasing the beam size. In contrast, the BLEU score dropped as we enlarge the beam size from the highest score at the beam size of 3 in the case of w/o LN.

The reason behind these observations is that the BLEU score is strongly penalized if the length of

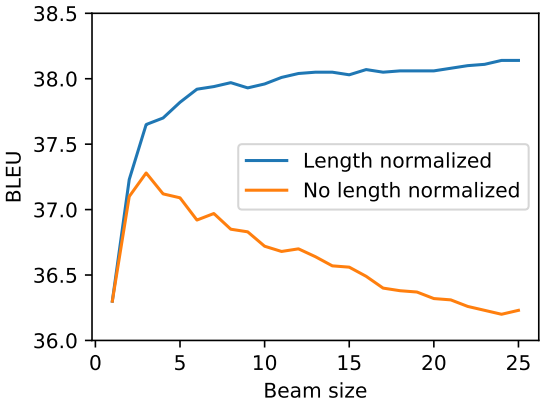

Figure 2: Relations between beam size and BLEU score on ASPEC En-Ja. With length normalization, we achieved better BLEU scores as the beam size became larger.

the hypothesis sentence is shorter than the corresponding reference sentence. This penalty is referred to as "Brevity Penalty (BP)". Figures 3 (a) and (b) respectively show the BP and the "raw BLEU score" (BLEU score while discarding the BP term) changes in w/ LN and w/o LN in terms of increasing the beam size. Clearly, the BP in-

Source	The search procedure utilizes a nonlinear least squares method coupled with the method of steepest descent.
Original	また，具体的な探索の手順を示した。 (We also show the specific search procedure.)
Synthetic	探索手順は最急降下法と結合した非線形最小二乗法を用いた。 (The search procedure utilizes a nonlinear least squares method coupled with the method of steepest descent.)

Table 4: An example sentence pair in the original and synthetic corpus.

	System	BLEU (JUMAN)	BLEU (KyTea)	BLEU (MeCab)
En-Ja	Single (KyTea pre-tokenized)	37.90	**40.48**	38.61
	Single (JUMAN pre-tokenized)	**38.12**	40.22	**38.80**

Table 5: Experimental results of ASPEC En-Ja subtask with different pre-tokenziers.

creasingly penalized the raw BLEU scores as the beam size increased in the case of w/o LN, while for w/ LN it maintained the BP. This observation reveals that the length normalized re-ranking (w/ LN) effectively works to keep the length of the best hypothesis sentences even if we enlarge the beam size. This is basically good behavior for actual use since we do not need to pay much attention to tuning the beam size.

5.6 Ensemble

Figure 4 shows the relation between the number of model ensembles and the BLEU score[5]. As we increased the number of models used, the BLEU scores improved but the impact gradually decreased. We only ensembled eight models for our submissions due to time and computational cost limitations but it would be more effective to ensemble more models.

6 Conclusion

In this paper, we described the systems we submitted to WAT 2017 shared translation tasks. We tried to make a synthetic corpus for an unreliable part of the provided corpus, and found it effectively improves the translation performance. Even though we achieved the highest BLEU score on JIJI corpus subtasks, the human evaluation of our system was worse than we had expected. We suspect that this is due to the noise on the JIJI corpus, so for future work, it would be beneficial to find out how to train the model with the noisy parallel corpus.

[5]In this figure, we simply ensembled the models in random order. However, it may be more effective to fix the order in accordance with the BLEU score on the dev set .

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Fabien Cromieres, Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2016. Kyoto university participation to WAT 2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT)*, pages 166–174.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Character-based decoding in tree-to-sequence attention-based neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT)*, pages 175–183.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 76–79.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Graham Neubig, Hideto Kazawa, Yusuke Oda, Jun Harashima, and Sadao Kurohashi. 2017. Overview of the 4th Workshop on Asian Translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT)*.

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 10th International Conference on Language Resources and Evaluation (LREC)*.

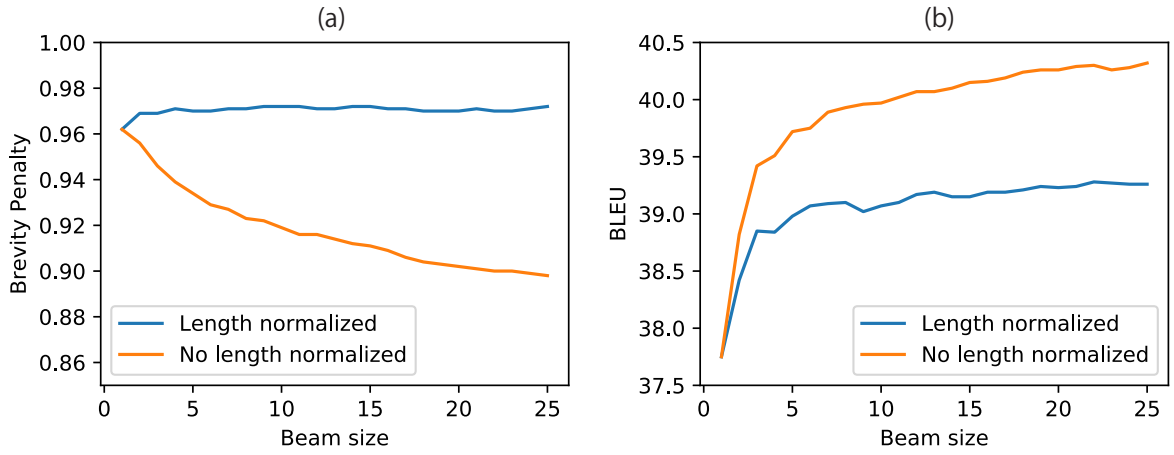

Figure 3: Experimental results on ASPEC En-Ja. (a) Relations between beam size and Brevity Penalty (BP). (b) Relations between beam size and BLEU score without BP. BP penalizes the score if the length of the hypothesis is shorter than the reference. BLEU score gets worse as BP goes down.

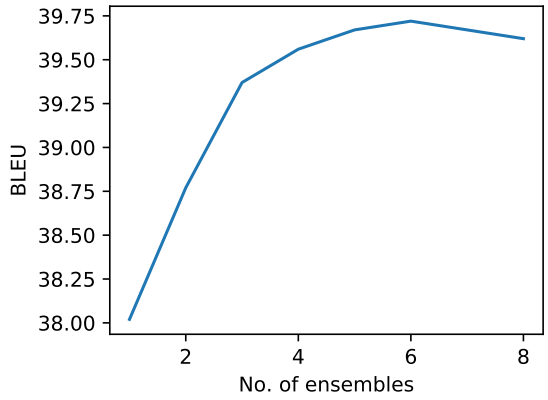

Figure 4: Relation between number of model ensembles and BLEU score on ASPEC En-Ja.

Graham Neubig. 2014. Forest-to-string SMT for asian language translation: NAIST at WAT2014. In *Proceedings of the 1st Workshop on Asian Translation (WAT)*, pages 20–25.

Graham Neubig. 2016. Lexicons and minimum risk training for neural machine translation: NAIST-CMU at WAT2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT)*, pages 119–125.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the 1st Conference on Machine Translation (WMT)*, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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. 2016c. 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.

Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pages 3097–3103.