Neural Machine Translation System using a Content-equivalently Translated Parallel Corpus for the Newswire Translation Tasks at WAT 2019

Hideya Mino [1,3] **Hitoshi Ito** [1] **Isao Goto** [1] **Ichiro Yamada** [1]
Hideki Tanaka [2] **Takenobu Tokunaga** [3]
[1]NHK Science & Technology Research Laboratories
[2]NHK Engineering System
[3] Tokyo Institute of Technology
{mino.h-gq,itou.h-ce,goto.i-es,yamada.i-hy}@nhk.or.jp,
tanaka.hideki@nes.or.jp, take@c.titech.ac.jp

Abstract

This paper describes NHK and NHK Engineering System (NHK-ES)'s submission to the newswire translation tasks of WAT 2019 in both directions of Japanese→English and English→Japanese. In addition to the JIJI Corpus that was officially provided by the task organizer, we developed a corpus of 0.22M sentence pairs by manually, translating Japanese news sentences into English content-equivalently. The content-equivalent corpus was effective for improving translation quality, and our systems achieved the best human evaluation scores in the newswire translation tasks at WAT 2019.

1 Introduction

We participated in the newswire translation tasks with JIJI Corpus, one of the tasks in WAT 2019 (Nakazawa et al., 2019). JIJI Corpus, a Japanese-English news corpus, comes from Jiji Press news, which has various categories including politics, economy, nation, business, markets, and sports. The newswire official tasks of WAT started in 2017, and some participants and organizer had already submitted their translation results before WAT 2019. Their quality, however, has not been equivalent with that in other tasks, such as scientific paper tasks and patent tasks. This is because of not only the small size (0.20M) of the JIJI Corpus but also a significant amount of noise for the neural machine translation (NMT) system training. The English news articles, which are generated as news-writing, not as translating, are mainly targeted at native English speakers, so information is often omitted or added. Figure 1 shows an example from JIJI Corpus. The omitted and added phrases become noise for the NMT training, and we consider this is one of the reasons for the low translation quality. To solve this problem and improve the translation quality of an NMT system, we are making a corpus with content-equivalent English translations of Japanese Jiji Press news, i.e. translations that do not omit and add information. We called the corpus Equivalent-JIJI Corpus[1].

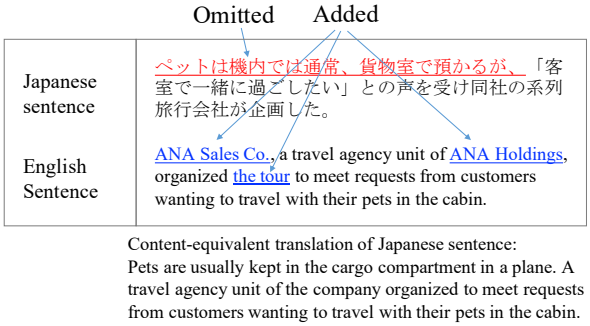

Figure 1: Example of a Japanese-English parallel sentence pair in JIJI Corpus. Contents of the underlined parts are not contained in the other language.

In this system description paper, we focus on these two styles of news parallel data, called the JIJI Corpus and the Equivalent-JIJI Corpus, and we named their styles the JIJI-style and the Equivalent-style, respectively. For WAT 2019, we submitted two translation results using translation systems adapted to the JIJI-style. In addition, to confirm the effectiveness of the content-equivalent translation, we submitted two more translation results using translation systems adapted to the Equivalent-style. Results showed that although our NMT systems adapted to the Equivalent-style scored lower than that adapted to the JIJI-style in the automatic evaluation, their scores reversed in the human evaluation.

[1]Equivalent-JIJI Corpus is still under construction and will be completed at the end of March 2021.

Proceedings of the 6th Workshop on Asian Translation, pages 106–111
Hong Kong, China, November 4, 2019.

Corpus name	Construction method	External data	Japanese	English	Size
JIJI Corpus	Alignment	No	Jiji Press news	Jiji Press news	0.20 M
Equivalent-JIJI Corpus	Manual translation	Yes	Jiji Press news	Content-equivalent translation	0.22 M
Aligned-JIJI Corpus	Alignment	Yes	Jiji Press news	Jiji Press news	0.29 M
BT-JIJI Corpus	Back translation	Yes	NMT output	Jiji Press news	0.53 M
Aligned-Yomiuri Corpus	Alignment	Yes	Yomiuri Shimbun	Yomiuri Shimbun	0.61 M

Table 1: Dataset.

2 Corpus Description

JIJI Corpus, which is extracted from Japanese and English Jiji Press news, is relatively small compared with those used in other Japanese→English or English→Japanese tasks of WAT 2019. To alleviate this low-resource translation problem, Morishita et al. (2017) used other resources for pre-training and fine-tuned with JIJI Corpus. We also used the external resources to improve the translation quality of the newswire tasks. For this purpose, we developed four types of corpora apart from JIJI Corpus. The first one was constructed through content-equivalent manual translation of Japanese Jiji Press news into English and is named Equivalent-JIJI Corpus. The second one was obtained through automatic sentence alignment between Japanese and English Jiji Press news using a sentence similarity score and is named Aligned-JIJI. The official JIJI Corpus is constructed in the same way. JIJI Corpus and Aligned-JIJI Corpus include noise as training data for an NMT system. The third corpus was constructed by back-translating monolingual English news sentences into Japanese (Sennrich et al., 2016b). This corpus is used for Japanese→English translation only. We named this parallel data BT-JIJI Corpus. For the back-translation, we used our best English→Japanese system adapted to the JIJI-style. Finally, we used another newspaper parallel corpus originating from the Yomiuri Shimbun, which we named Aligned-Yomiuri Corpus. Aligned-Yomiuri Corpus is made with a parallel sentence similarity score, as is the case of JIJI Corpus. Table 1 summarizes the detail of each corpus.

3 Domain Adaptation Techniques

In this paper, we used a domain-adaptation technique to train a model adapted to the JIJI- and Equivalent-style. The multi-domain method (Chu et al., 2017; Sennrich et al., 2016a) is one of the most effective approaches to leverage out-of-domain data. Chu et al. (2017) proposed training an NMT system with multi-domain parallel corpora using domain tags such as "<domain-name>" attached to the respective corpora. We used domain adaptations with the names of the styls as domain tags. We used a "<JIJI-style>" tag for the JIJI, Aligned-JIJI, and BT-JIJI corpora and a "<Equivalent-style>" tag for Equivalent-JIJI Corpus. In addition, we used a "<YOMIURI-style>" tag for Aligned-Yomiuri Corpus because it comes from a newspaper other than Jiji Press news.

4 Experiments

In this study, we verified the effectiveness of the Equivalent-style translation through the following procedures. Firstly, we trained the multiple NMT models with different combinations of five corpora as shown in Table 1, and evaluated these NMT models with an official test-set, in which the number of data was 2000. Then, we evaluated these NMT models with a further test-set, in which the number of data was 1764, extracted from Equivalent-JIJI Corpus of Equivalent-style in contrast to the official test-set extracted JIJI Corpus in JIJI-style. Finally, we evaluated the effectiveness of the Equivalent-style translation.

4.1 Data Processing and System Setup

All of the datasets were preprocessed as follows. We used the Moses toolkit [2] to clean and tokenize the English data and used KyTea (Neubig et al., 2011) to tokenize the Japanese data. Then, we used a vocabulary of 32K units based on a joint source and target byte-pair encoding (BPE) (Sennrich et al., 2016c). For the translation model,

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

Training corpus	Num. of data	Domain adaptation	Tag (Style)	JIJI-style test-set	Equivalent-style test-set
JIJI (Official data)	0.2M	No	-	18.14	7.76
Equivalent-JIJI	0.22M	No	-	9.15	21.36
JIJI, Equivalent-JIJI	0.42M	No	-	20.56	20.8
JIJI, Equivalent-JIJI	0.42M	Yes	JIJI-style	21.69	12.65
JIJI, Equivalent-JIJI	0.42M	Yes	Equivalent-style	12.01	23.16
JIJI, Equivalent-JIJI, Aligned-JIJI, Aligned-Yomiuri	1.32M	No	-	23.44	20.95
JIJI, Equivalent-JIJI, Aligned-JIJI, Aligned-Yomiuri	1.32M	Yes	JIJI-style	24.50	14.65
JIJI, Equivalent-JIJI, Aligned-JIJI, Aligned-Yomiuri	1.32M	Yes	Equivalent-style	13.40	24.78
JIJI, Equivalent-JIJI, Aligned-JIJI, Aligned-Yomiuri, BT-JIJI	1.85M	Yes	JIJI-style	25.54	15.03
JIJI, Equivalent-JIJI, Aligned-JIJI, Aligned-Yomiuri, BT-JIJI	1.85M	Yes	Equivalent-style	13.47	24.80

Table 2: BLEU scores for Japanese→English translation tasks.

Training corpus	Num. of data	Domain adaptation	Tag (Style)	JIJI-style test-set	Equivalent-style test-set
JIJI (Official data)	0.20M	No	-	18.46	15.9
Equivalent-JIJI	0.22M	No	-	17.92	36.67
JIJI, Equivalent-JIJI	0.42M	No	-	25.07	39.97
JIJI, Equivalent-JIJI	0.42M	Yes	JIJI-style	24.63	36.75
JIJI, Equivalent-JIJI	0.42M	Yes	Equivalent-style	24.52	39.93
JIJI, Equivalent-JIJI, Aligned-JIJI, Aligned-Yomiuri	1.32M	No	-	28.49	43.52
JIJI, Equivalent-JIJI, Aligned-JIJI, Aligned-Yomiuri	1.32M	Yes	JIJI-style	28.14	43.82
JIJI, Equivalent-JIJI, Aligned-JIJI, Aligned-Yomiuri	1.32M	Yes	Equivalent-style	27.77	43.68

Table 3: BLEU scores for English→Japanese translation tasks.

we used the encoder and decoder of the transformer model (Vaswani et al., 2017), which is a state of the art NMT model. The transformer model uses a multi-headed attention mechanism applied as self-attention and a position-wise fully connected feed-forward network. The encoder converts the received source language sentence into a sequence of continuous representations, and the decoder generates the target language sentence. We implemented our systems with the Sockeye toolkit (Hieber et al., 2018), and trained them on one Nvidia P100 Tesla GPU. While training our models, we used the stochastic gradient descent (SGD) with Adam (Kingma and Ba, 2015) as the optimizer, using a learning rate of 0.0002, multiplied by 0.7 after every eight checkpoints. We set the batch size to 5000 tokens and maximum sentence length to 99 BPE units. For the other hyperparameters of our models, we used the default parameter values of Sockeye. We used early stopping with a patience of 32. Decoding was performed with a beam search with a beam size of 5, and we did not apply an ensemble decoding with multiple models, although this could possibly improve the translation quality, though we used a beam search with a beam size of 30 and an ensemble of ten models when submitting the official results. To evaluate translation quality, we used BLEU (Papineni et al., 2002). BLEU is calculated using multi-bleu.perl [3]. We report case-sensitive scores.

4.2 Results

Tables 2 and 3 show the experimental results. The Training corpus column shows the corpora used for training. The Style column shows the tag used for translation, i.e. the JIJI- or Equivalent-style. The JIJI-style test-set is equal to the official test-set in the newswire task of WAT 2019.

4.2.1 Trained with Different Combinations of Five Corpora

The JIJI-style test-set column of Tables 2 and 3 shows the translation quality of the JIJI-style test-sets with the BLEU metric for different combinations of the five corpora. For the models without domain adaptation, where Domain adaptation column is "No," the BLEU scores are improved by adding the other domains' data into the JIJI

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

Task	Tag (Style)	BLEU	Rank	RIBES	Rank	AMFM	Rank	Pairwise	Rank	Adequacy	Rank
JIJI-JE	JIJI-style	26.83	1/4	0.70	1/4	0.55	1/4	72.00	2/4	-	-
	Equivalent-style	14.23	4/4	0.61	4/4	0.53	3/4	89.00	1/4	4.55	1/2
JIJI-EJ	JIJI-style	29.76	1/4	0.74	1/4	0.65	2/4	81.25	2/4	-	-
	Equivalent-style	28.75	2/4	0.73	2/4	0.66	1/4	87.75	1/4	4.11	1/2

Table 4: Official results for the newswire translation tasks of WAT 2019: For JIJI-EJ task, we show the BLEU, RIBES, and AMFM scores with KyTea tokenizer.

Corpus. For the JIJI-style Japanese→English test-set, the BLEU scores are higher with the use of tags for the domain adaptation. However, the use of the domain adaptation is not effective for the JIJI-style English→Japanese test-set. This seems to be due to the different origins of the target-side sentences in the Equivalent-JIJI Corpus. Japanese sentences in the Equivalent-JIJI Corpus come from the Japanese Jiji Press news, the same as for JIJI Corpus. In contrast, the English sentences in Equivalent-JIJI Corpus does not come from Jiji Press news. It appears that the use of different domain tags is less effective when using the same-origin data for the target-side as shown in the fourth and fifth columns of Table 1. In the case of Japanese→English task with JIJI Corpus and Equivalent-JIJI Corpus, the origin of the target-side English sentences differs between the two corpora (JIji Press news and Content-equivalent translation) despite the origin of the source-side is being the same (Jiji Press news), so the NMT system cannot decide which style, JIJI- or Equivalent-style, should be output. In contrast, no choice is necessary for the English→Japanese task because the target-side Japanese sentences is the same origin (Jiji Press news).

The Equivalent-style test-set column in Tables 2 and 3 shows translation quality of the Equivalent-style test-sets. For the models without domain adaptation, the BLEU scores are not improved by adding the other domains' data into the Equivalent-JIJI Corpus in case of the Japanese→English task. The domain adaptation using tags is extremely effective for the Japanese→Engish task. Although the amount of Equivalent-style data is much smaller than that of JIJI-style data, the BLEU scores for the Equivalent-style test-set are higher than those for the JIJI-style test-set. In particular, the BLEU scores of the Equivalent-style test-set for the English→Japanese are over 43. It appears that it is more difficult to improve the translation quality for the JIJI-style test-set than for the Equivalent-style test-set because the JIJI-style test-set includes noise for training the NMT system.

4.2.2 Translation with Different Types of Systems

Supposing that JIJI Corpus includes noise, the NMT system adapted to the Equivalent-style seems to be a better system to translate news generally. However, the BLEU scores for the JIJI-style test-set trained with Equivalent-JIJI Corpus are 9.15 for Japanese→English and 17.92 for English→Japanese and they are lower than the scores for the test-set trained with JIJI Corpus, as shown in Tables 2 and 3. To determine whether or not the translation systems adapted to the Equivalent-style are better for human evaluation than those adapted to the JIJI-style, we submitted the translated results with both of the translation systems adapted to JIJI- and Equivalent-style.

4.3 Official Results

We used the bottom translation systems of Tables 2 and 3 for submitting to WAT 2019. These systems can be adapted to each style by attaching domain tags, "<JIJI-style>" for JIJI-style translation and "<Equivalent-style>" for Equivalent-style translation, at the top of the source sentence. To improve the translation quality further, we submitted the translation results with an ensemble decode of ten models and a beam search with a beam size of 30. Table 4 shows the official results of our submission to WAT 2019. Our systems adapted to JIJI-style achieved the best BLEU and RIBES scores. In contrast, for the pairwise crowdsourcing evaluation and the JPO adequacy evaluation[4], our systems adapted to the Equivalent-style achieved the best evaluation. For the AMFM, our system

[4]WAT 2019 organizer selected submissions for JPO adequacy evaluation, and only one submission for each task was evaluated.

	Example sentence	BLEU
Source Content-equivalent translation	終了後、赤松氏は記者団に「公平、公正な形で国民の意見を聴く」と強調し、江田氏は「客観的に調査した結果をそのまま受け入れる」と語った。 After the meeting, Akamatsu emphasized to the press, "We will hear opinions of the public in a fair and equitable manner," and Eda said, "We will accept the results of the objective investigation."	
Reference (JIJI Corpus)	After the meeting, Akamatsu told reporters, "We will seek the views of the public in a fair and equitable manner."	
NMT output adapted to JIJI-style	After the meeting, Akamatsu told reporters that he will listen to public opinions in a fair and equitable manner.	51.61
NMT output adapted to Equivalent-style	After the meeting, Akamatsu emphasized to the press, "We will listen to the opinions of the people in a fair and fair manner," and Eda said, "We will accept the results of the investigation objectively."	22.70
Source Content-equivalent translation	同日午後の参院本会議で可決、成立する見通し。 It is expected to be passed and enacted at a plenary session of the House of Councilors in the afternoon of the same day.	
Reference (JIJI Corpus)	The House of Councillors, the upper chamber of the Diet, approved the spending program at a plenary meeting on Monday afternoon after the House of Representatives, the lower chamber, passed it earlier in the day.	
NMT output adapted to JIJI-style	The House of Councillors, the upper chamber, is expected to approve the bill at a plenary meeting later in the day.	26.33
NMT output adapted to Equivalent-style	It is expected to be passed and enacted at the plenary session of the House of Councilors in the afternoon of the same day.	0.00

Table 5: Example results of JIJI-JE tasks translated with JIJI- and Equivalent-style NMT

Task	Tag (Style)	Omitted-words	Added-words
JIJI-JE	JIJI-style	18.40	9.39
	Equivalent-style	4.27	1.44
JIJI-EJ	JIJI-style	28.19	13.16
	Equivalent-style	8.22	2.55

Table 6: Further human evaluation results, which is the average number of words per 100 words.

adapted to the JIJI-style achieved the best evaluation for the Japanese→English task, whereas our system adapted to the Equivalent-style achieved the best evaluation for the English→Japanese task. These results show that the NMT systems adapted to the Equivalent-style are generally better systems for translating the news. The overview paper for WAT 2019 gives the details of our submission including the other WAT participants' results.

4.4 Further Human Evaluation

Apart from the official pairwise crowdsourcing evaluation and JPO adequacy evaluation, we also evaluated our official submission independently with a translation company to analyze deeply the official results. We randomly selected 300 and 50 sentences from the Japanese→English and English→Japanese official test-sets respectively, and three evaluators counted the number of omitted and added words in the NMT outputs adapted to the JIJI- and Equivalent-styles. Table 6 shows the average number of words per 100 words of the three evaluators. These results indicate that the NMT systems adapted to the Equivalent-style can prevent the omission and addition of information. Table 5 shows the examples of NMT outputs adapted to the JIJI- and Equivalent-styles in the official tasks. The first example shows omitted information, and the second example shows added information in the NMT output adapted to the JIJI-style. The references also include omitted and added information. The NMT output adapted to the Equivalent-style is translated without omitted and added information. The sentence BLEU scores of outputs adapted to the JIJI-style NMT are higher than those of outputs adapted to the Equivalent-style NMT. These results indicate that NMT outputs adapted to the JIJI-style often include the omission and addition of information, and these cause the worse human evaluation. This seems to be a reason that our systems adapted to

the Equivalent-style, which prevent the omission and addition of information, achieved the best human evaluation in spite of the lower BLEU scores.

5 Conclusions

In this description paper, we presented our NMT systems adapted to the JIJI- and the Equivalent-styles. In addition to the JIJI Corpus in the JIJI-style that was officially provided by the WAT 2019 organizer, we developed a corpus of 0.22M sentence pairs in the Equivalent-style by manually, content-equivalently translating Japanese news sentences into English. We obtained the state-of-the-art results for the newswire tasks of WAT 2019. In our four submissions, the translation models adapted to the JIJI-style achieved the best results for the BLEU evaluation. In contrast, the translation models adapted to the Equivalent-style achieved the best results for the pairwise crowdsourcing evaluation and JPO adequacy evaluation. We showed that the content-equivalently translated data is effective for the widespread news translation from the perspective of a human evaluation.

Acknowledgments

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

References

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

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

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

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. Asian Federation of Natural Language Processing.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

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

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

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