NHK's Lexically-Constrained Neural Machine Translation at WAT 2021

Hideya Mino [1,2] **Kazutaka Kinugawa** [1] **Hitoshi Ito** [1]
Isao Goto [1] **Ichiro Yamada** [1] **Takenobu Tokunaga** [2]
[1] NHK Science & Technology Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo 157-8510, Japan
{mino.h-gq,kinugawa.k-jg,itou.h-ce,
goto.i-es,yamada.i-hy}@nhk.or.jp
[2] Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
take@c.titech.ac.jp

Abstract

This paper describes the system of our team (NHK) for the WAT 2021 Japanese↔English restricted machine translation task. In this task, the aim is to improve quality while maintaining consistent terminology for scientific paper translation. This task has a unique feature, where some words in a target sentence are given in addition to a source sentence. In this paper, we use a lexically-constrained neural machine translation (NMT), which concatenates the source sentence and constrained words with a special token to input them into the encoder of NMT. The key to the successful lexically-constrained NMT is the way to extract constraints from a target sentence of training data. We propose two extraction methods: proper-noun constraint and mistranslated-word constraint. These two methods consider the importance of words and fallibility of NMT, respectively. The evaluation results demonstrate the effectiveness of our lexical-constraint method.

1 Introduction

Our team (NHK) participated in the restricted machine translation task[1] using the Japanese-English dataset of the Asian scientific paper excerpt corpus (ASPEC-JE) (Nakazawa et al., 2016) at WAT 2021 (Nakazawa et al., 2021). In this task, the aim is to improve translation quality while preserving consistent terminology for translating scientific papers that include technical terms and proper nouns. In this task, a list of target words is given for each source sentence to appear in a target sentence. Figure 1 shows the overview of this task. There are two evaluation criteria: the translation accuracy via bilingual evaluation understudy (Papineni et al., 2002) (BLEU score) and the consistency score of the ratio of sentences satisfying an exact match of given constraints (consistency score). The final ranking is determined by the combined score of both: calculating BLEU with only the exact match sentences[2].

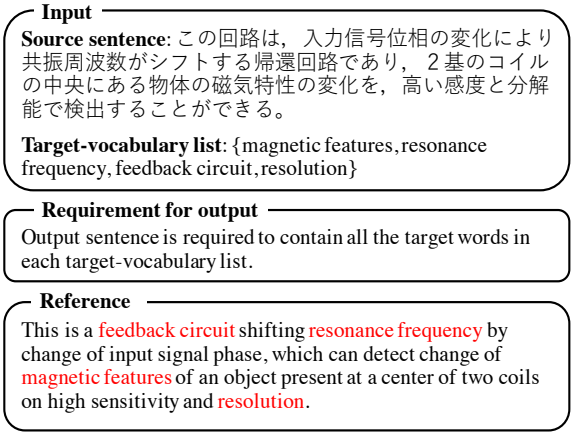

Figure 1: Overview of the restricted translation task (Japanese→English).

In related work (Chen et al., 2020a; Song et al., 2019; Wang et al., 2019; Post and Vilar, 2018; Hokamp and Liu, 2017), since it does not require higher computational complexity than the other methods using the grid beam search (GBS) decoding algorithm (Hokamp and Liu, 2017; Post and Vilar, 2018), we use the lexical-constraint method of Chen et al. (2020a). This method concatenates a source sentence and constrained words with a special token to input them into an encoder of the neural machine translation

[1] https://sites.google.com/view/restricted-translation-task/

[2] If the translation does not satisfy the constraint, replace the translation with an empty string.

Proceedings of the 8th Workshop on Asian Translation, pages 46–52
Bangkok, Thailand (online), August 5-6, 2021.

(NMT). In addition to the merit of reducing the computational cost compared with GBS decoding, this method has two other merits: no need to modify the architecture of the NMT system or prepare any word alignment data. In this method for this task, one of the main problems is how to extract constraints from training data since only constrained word lists for dev, devtest, and test sets are provided to participants.

In this paper, we propose extracting constraints from target sentences on the basis of proper-noun and mistranslated-word constraints considering the importance of words and fallibility of NMT. The former constraint is a list of proper nouns extracted with named-entity recognition. The latter constraint is a list of words mistranslated or under-translated with vanilla NMT compared with a target sentence. We conducted experiments to evaluate the NMT using the proposed method and found that the proposed method outperformed a baseline lexical-constraint method.

2 Restricted Translation Task Description

2.1 Official Dataset

The main dataset of the restricted translation task is the Japanese-English paper abstract corpus (ASPEC-JE) and the target vocabulary list as constraints. In addition to the main dataset, participants can use any other resources by mentioning their details. The ASPEC-JE dataset consists of training, dev, devtest, and test data. The training data contains 3.0 million bilingual pairs provided with similarity scores automatically calculated by DP matching (Utiyama and Isahara, 2007). The target vocabulary list for restricted translation is attached to the dev, devtest, and test data dedicated for this task. Participants are not told the detailed way to select constraints. Table 1 shows statistics of each data.

2.2 Official Evaluation

In this task, four distinct metrics are calculated: BLEU, RIBES (Isozaki et al., 2010), AMFM (Banchs et al., 2015), and consistency scores. The BLEU, RIBES, and AMFM scores are calculated in accordance with the WAT convention. The consistency score is the ratio of the number of sentences satisfying the exact match of given constrained words over the whole test corpus. The final score is calculated using both BLEU and consistency scores by WAT 2021 organizers as below:

Language pair	Number of sentences			
	Train	Dev	Devtest	Test
JA-EN	3.0M	1,790	1,784	1,812
		(2.8/2.9)	(3.2/3.2)	(3.2/3.3)

Table 1: Statistics of official data including ASPEC-JE and target vocabulary lists. Average numbers of constrained words per sentence (Left:Japanese / Right:English) are shown for the dev, devtest, and test data. There are no vocabulary lists for the training data.

1. Check whether the translation satisfies the given constraints or not.
2. If the translation does not satisfy the constraint, replace the translation with an empty string.
3. Calculate BLEU with modified translations.

Furthermore, bilingual human annotators evaluate the top-ranked submitted systems based on source-based direct assessment (Federmann, 2018; Cettolo et al., 2017) and source-based contrastive assessment (Federmann, 2018; Sakaguchi and Van Durme, 2018).

3 NMT with Lexical Constraint

Borrowing Chen et al. (2020a)'s idea, we implemented a lexically-constrained NMT with encoder and decoder modules. We concatenated a source sentence and constrained words with a special token to input into the encoder, as illustrated in Figure 2. The key to the successful lexically-constrained NMT is the way to extract constraints from a target sentence. Though the constraints are given for the dev, devtest, and test data, they are not given for the training data. In this paper, we focus on the way to extract a constraint from the target sentence in training data for the training phase.

The simplest method of extracting a lexical constraint is randomly sampling words from the target sentence, as Chen et al. (2020a) did. Beyond the random sampling method, we propose two other directions with a focus on proper nouns and mistranslated words to extract the constrained words automatically from the target sentence.

- **Proper-Noun Constraint.** Though participants were not told the detailed way to se-

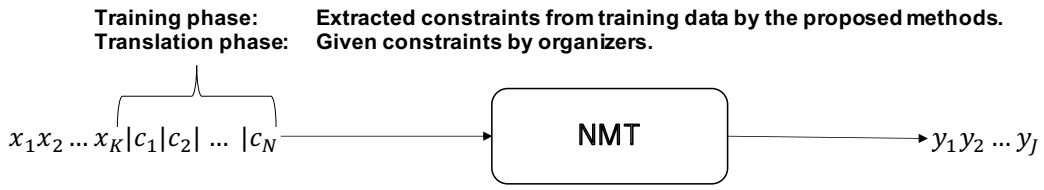

Figure 2: Overview of NMT using lexical-constraint method. $\boldsymbol{x} = (x_1, x_2, ..., x_K)$, $\boldsymbol{c} = (c_1, c_2, ..., c_N)$, and $\boldsymbol{t} = (y_1, y_2, ..., y_J)$ show source-, constraint-, and predicted-sequences, respectively. K and J are the lengths of source and target sentences. N is a number of constrained words. "|" is a special token for delimiter. During the training phase, constraints are extracted from training data by the proposed methods. During the translation phase, constraints are given by WAT 2021 organizers.

lect constraints, we found that the vocabulary list in dev data includes many technical terms and proper nouns. Supposing that the important words such as technical terms and proper nouns tend to be selected as constraints, we extract proper nouns on the basis of the named-entity recognition.

- **Mistranslated-Word Constraint.** The proper-noun constraint is not enough to be sufficient to cover all constraints in this task. Given constrained words including the proper-noun constraints accounted for 21% of the Japanese dev data. To increase the number of appropriate constrained words, we extract mistranslated or dropped words by NMT as constraints. First, we trained an NMT model on parallel training data, and translated the source sentences in training data with this model. We then picked out the words that do not appear in the translated sentence but appear in the target sentence. Both proper-noun and mistranslated-word constraints could cover 38% of constraints for the dev data. The remaining 62% constrained words could be translated correctly without adding them as constraints.

- **Both the Proper-Noun and Mistranslated-Word Constraints.** Both constraints are made by concatenating the proper-noun and mistranslated-word constraints and removing duplicates.

4 Experiments

4.1 Data

In this paper, we used only the first 2.0 million bilingual pairs[3] in the official dataset, i.e., ASPEC-JE, with high similarity scores for training the models. We did not use any other resources.

4.2 System Setup

We used the KyTea (Neubig et al., 2011) to tokenize Japanese sentences and the Moses toolkit[4] to clean and tokenize English sentences. We then used a vocabulary of 48K tokens on the basis of joint byte-pair encoding (BPE) (Sennrich et al., 2016) for the source and target. We used the encoder and decoder of the transformer model (Vaswani et al., 2017), which is a state-of-the-art NMT model. The encoder converts a source sentence into a sequence of continuous representations, and the decoder generates a target sentence. We implemented this system with the Sockeye 2 toolkit (Hieber et al., 2020). All models were trained within at most three days on four Nvidia V100 Tesla GPUs with 16-GB memory in parallel. In training the model, we applied stochastic gradient descent with Adam (Kingma and Ba, 2015) as the optimizer, using a learning rate of 0.0002, multiplied by 0.7 after every 8 checkpoints. We set the batch size to 5000 tokens and the maximum sentence length to 150 BPE tokens. We applied early stopping with a patience of 32. Dropout was set to 0.1 for encoder, decoder, attention layer, and feed-forward layer after testing with 0.1, 0.3, and 0.5 using development data. For the other hyperparameters of the models, we used the default Sockeye 2 parameters[5].

Translation was carried out through a beam search with a beam size of 30, and we used an ensemble of 5 models with different seeds.

We used three types of constraints for the pro-

[3]The remaining 1.0 million bilingual pairs were often noisy as described in Neubig (2014). We found the performance degraded when using all data in this work.

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

[5] Sockeye 2 uses a transformer model with 6 encoder and decoder layers, 8 parallel attention heads, model dimensionality of 512, and a feed-forward layer size of 2048 as default.

Task	Method	Average of constrained words	Consistency rate (word)	BLEU
Japanese→English	Baseline	N/A	52.3	29.3
	Random-word	4.99	78.6	29.5
	Proper-noun	1.25	78.8	36.3
	Mistranslated-word	4.68	86.1	39.2
	Prop. & Mistrans.	5.63	**96.0**	**43.9**
English→Japanese	Baseline	N/A	61.7	45.9
	Random-word	4.89	77.8	37.4
	Proper-noun	1.91	96.2	48.2
	Mistranslated-word	2.74	85.7	48.3
	Prop. & Mistrans.	4.48	**97.4**	**53.2**

Table 2: Experimental results for each task. Baseline is trained without any constraint, Random-word is trained with the randomly extracted constraint, Proper-noun is trained with the proper-noun constraint, Mistranslated-word is trained with the mistranslated-word constraint, and Prop. & Mistrans. is trained with both the proper-noun and mistranslated-word constraints. "Average of constrained word" shows the average number of constrained words per sentence.

posed method: the proper-noun constraint, the mistranslated-word constraint, and both, called "Proper-noun," "Mistranslated-word," and "Prop. & Mistrans.," respectively. For extracting the proper nouns from the target sentence, we used GiNZA 4.0[6] for Japanese and *en_core_web_sm* model of spaCy 2.3[7] for English. We used at most five words from candidates sorted on the basis of term-frequency inverse document frequency (TF-IDF) scores (Chen et al., 2020b) in each constraint.

To evaluate translation quality separately from the official evaluation, we calculated case-insensitive BLEU (Papineni et al., 2002) scores by using multi-bleu.perl[8] and a consistency rate of words, which is the ratio of the number of words appearing in the output of given constrained words.

4.3 Baselines

We trained two types of baselines using the transformer model.

1. **Baseline**: The model trained on the parallel data (2.0 million bilingual pairs) without any constraint.

2. **Random-word**: The model trained on the parallel data with constraints of five words randomly extracted from the target sentence. We extracted different constraints randomly for each epoch.

[6]https://megagonlabs.github.io/ginza/

[7]https://spacy.io/usage/v2-3

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

4.4 Experimental Results

Table 2 shows the experimental results for Japanese↔English tasks. Compared with the Baseline method, our proposed methods improved both consistency rates of words and BLEU scores for Japanese↔English tasks.

Though models using the Random-word method improved the consistency rate compared with Baseline, there is no or little improvement in BLEU scores. For the Japanese→English task, though the consistency rates of the Random-word and Proper-noun methods are almost same, the BLEU scores of the Proper-noun performed better than the Random-word method. The average number of constrained words of the Random-word method is higher than the Proper-noun method. This result indicates that translation quality highly depends on the way to extract constraints rather than the number of constraints.

From comparing among the versions of our proposed method using three types of constraints, the model using the Prop. & Mistrans. method performed the best for both the Japanese↔English tasks.

From comparing the use of the proper-noun and mistranslated-word constraints, the "Mistranslated-word" method performed better for Japanese→English, whereas the "Proper-noun" method performed better for

Task	Method	BLEU	RIBES	AMFM	HUMAN DA	HUMAN CA	Final score
Japanese→English	Baseline	29.25	0.77	0.62	N/A	N/A	N/A
	Random-word	29.49	0.68	0.52	N/A	N/A	N/A
	Prop. & Mistrans. + rule	42.94	0.80	0.66	74.1	77.2	33.9
English→Japanese	Baseline	45.93	0.85	0.76	N/A	N/A	N/A
	Random-word	37.43	0.80	0.70	N/A	N/A	N/A
	Prop. & Mistrans. + rule	52.69	0.82	0.80	73.9	73.5	37.5

Table 3: Official results (Japanese tokenizer:KyTea and English tokenizer:moses). HUMAN DA and CA is source-based direct assessment and source-based contrastive assessment. See 2.2 for the details of each evaluation criterion.

English→Japanese. In addition, there is no significant difference in the consistency rate of the mistranslated-word constraint between English→Japanese and Japanese→English. The proper-noun constraint for English→Japanese appears likely to be more similar to constraints of the test data than that for Japanese→English.

For the average number of constrained words, though the Random-word method has the most constrained words, it did not perform the best for either the consistency rate or BLEU score. The results indicate that the quality of the model using constraints relies on whether constraints are suitable for the task or not.

As a whole, we found that the using both the proper-noun and mistranslated-word constraints is effective for the restricted machine translation task.

4.5 Official Results

Table 3 lists the official results. For "Prop. & Mistrans. + rule" method, we input the unsatisfied constrained word, which does not appear in the output with the following procedure:

1. extracts unsatisfied words, which do not appear in the output, from the constrained words.
2. calculates Levenshtein distance between each unsatisfied word and each word in the output.
3. swaps the word of the output with the closest distance for the unsatisfied word.

The outputs of the "Prop. & Mistrans. + rule" method satisfy all given constraints. The official results indicate the effectiveness of using the proposed constraints in terms of the human evaluation since the rankings of "BLEU," "HUMAN DA," "HUMAN CA," and "Final score" are the same as among participants of this task at WAT 2021.

5 Conclusion

We described our proposed method using lexical constraints for a Japanese↔English restricted machine translation task with the Asian scientific paper excerpt corpus (ASPEC). We proposed a method to extract appropriate constraints of the lexically-constrained neural machine translation (NMT) for this task. Our proposed method using the proper-noun and mistranslated-word constraints improved translation performance compared with random-word constraint.

For future work, we plan to apply the proposed constraints into NMT with a grid beam search decoding algorithm (Hokamp and Liu, 2017; Post and Vilar, 2018) to compare the performance.

References

Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating MT in the continuous space model framework. *IEEE ACM Trans. Audio Speech Lang. Process.*, 23(3):472–482.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *2017 International Workshop on Spoken Language Translation, IWSLT 2017, Tokyo, Japan*, pages 2–14.

Guanhua Chen, Yun Chen, Yong Wang, and Victor O.K. Li. 2020a. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3587–3593. International Joint Conferences on Artificial Intelligence Organization. Main track.

Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020b. Content word aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 358–364, Online. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, Lisboa, Portugal. European Association for Machine Translation.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

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

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. Association for Computational Linguistics.

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'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

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

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.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

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

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. *Proceedings of Machine Translation Summit XI, Copenhagen, Denmark, 2007*, pages 457–482.

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.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. One model to learn both: Zero pronoun prediction and translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 921–930, Hong Kong, China. Association for Computational Linguistics.