Improving Japanese-to-English Neural Machine Translation by Paraphrasing the Target Language

Yuuki Sekizawa and **Tomoyuki Kajiwara** and **Mamoru Komachi**
{sekizawa-yuuki, kajiwara-tomoyuki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

Abstract

Neural machine translation (NMT) produces sentences that are more fluent than those produced by statistical machine translation (SMT). However, NMT has a very high computational cost because of the high dimensionality of the output layer. Generally, NMT restricts the size of the vocabulary, which results in infrequent words being treated as out-of-vocabulary (OOV) and degrades the performance of the translation. In order to improve the translation quality regarding words that are OOV in the target language, we propose a preprocessing method that paraphrases infrequent words or phrases expressed as OOV with frequent synonyms from the target side of the training corpus. In an evaluation using Japanese to English translation, we achieved a statistically significant BLEU score improvement of 0.55–0.77 over baselines that included the state-of-the-art method.

1 Introduction

Recently, neural-network-based methods have gained considerable popularity in many natural language processing tasks. In the field of machine translation, neural machine translation (NMT) is actively being researched because of the advantage that it can output sentences that are more fluent compared with statistical machine translation (SMT). However, NMT has a problem of high computational cost because it addresses the output generation task by solving a classification problem in vocabulary dimension. Typically, NMT has to restrict the size of the vocabulary to reduce the computational cost. Therefore, the target language vocabulary includes only high-frequency words (e.g., 30,000 high-frequency words) in training; other words are treated as out-of-vocabulary (OOV) and substituted with a special symbol such as "<unk>" in the output. The symbol has no meaning, so the output has reduced quality.

As a previous work attempting to reduce the OOV rate in NMT, Li et al. (2016), replaced OOV words with a translation table using word similarity in the training and test data. In particular, they replaced each OOV word with an in-vocabulary word using word similarity in a parallel training corpus; they reduced the OOV rate in the output and improved the translation quality. However, they sometimes substituted OOV words with a similar word such as a proper noun. In addition, they deleted OOV words that aligned to null, which can result in a loss of sentence content and reduced translation adequacy.

In this work, we present a preprocessing method for improving translation related to OOV words. We paraphrase low-frequency words treated as OOV in the target corpus with high-frequency words while retaining the meaning.

Our main contributions are as follows.

- We propose a paraphrasing-based preprocessing method for Japanese-to-English NMT to improve translation accuracy with regard to OOV words. Our method can be combined with any NMT system.

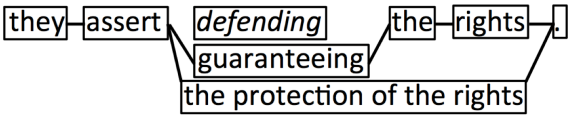
- We show that our method achieved a statistically significant BLEU (Kishore et al., 2002) score improvement of 0.58 and a METEOR (Lavie and Agarwal, 2007) score improvement of 0.52 over the previous method (Li et al., 2016) and reduced the OOV rate in output sentences by approximately 0.20.

Proceedings of the 4th Workshop on Asian Translation, pages 64–69,
Taipei, Taiwan, November 27, 2017.

2 Related Work

There have been studies on improving translation accuracy by reducing the OOV rate using pre- and post-processing for machine translation. Luong et al. (2015) proposed a post-processing method that translates OOV words with a corresponding word in the source sentence using a translation dictionary. This method needs to align training sentence pairs before training to learn correspondences between OOV words and their translations. In the method described in this paper, we need no word alignment, and we retain the meaning of the original word by paraphrasing the target side of the training corpus. Jean et al. (2015) proposed another post-processing method that translates each OOV word with the word that has the largest attention weight in the source sentence using a translation dictionary. Their method does not need word alignment, but it still does not necessarily consider the meaning in the target language, unlike our paraphrasing approach. Sennrich et al. (2016) applied byte pair encoding (BPE) to source and target corpora to split OOV words into units of frequent substrings to reduce the OOV rate. Their method splits words greedily without considering their meaning. Since we use lexical paraphrasing in the training data, we hope to reduce the OOV rate in the translation output while retaining the meaning. Additionally, since ours is a preprocessing method, it can be combined with a post-processing method.

On the other hand, there are methods similar to ours that paraphrase corpora as a preprocessing step of machine translation to reduce the complexity of source and/or target sentences. Sanja and Maja (2016) paraphrased source sentence vocabulary with a simple grammar as a preprocessing step for machine translation. We attempt to improve translation quality by reducing the OOV rate in the target language using paraphrasing without simplifying the source input sentences. Li et al. (2016) substituted OOV words in training corpora with a similar in-vocabulary word as pre- and post-processing steps. They replaced OOV words with frequent words using cosine similarity and a language model. They obtained word alignment between an OOV word and its counterpart in training corpora. In addition, they deleted OOV words from the training corpus if they aligned to null. However, this leads to a loss of sentence meaning and degrades the adequacy of the translation. They also might replace OOV words with similar but non-synonymous words since they used distributional similarity. For instance, they replaced "surfing" with "snowboard", which leads to rewriting "internet surfing" as "internet snowboard", resulting in a change of meaning. We use a paraphrase score calculated from bilingual pivoting instead of distributional similarity; therefore, we are not likely to paraphrase OOV words with inappropriate expressions. In the aforementioned example, we paraphrase "surfing" as "browser", which preserves the original meaning to some extent.



original: the *pedagogues* had *quarrels*.
paraphrase, first round: the educators had discussions.
paraphrase, second round: the teachers had discussions.

Figure 1: Examples of paraphrasing. Original word is shown in italics. Upper: paraphrase lattice; lower: iterative paraphrasing of OOV word.

3 Proposed Method

In this paper, we propose a preprocessing method that paraphrases infrequent words or phrases with frequent ones on the target side of the training sentences in order to train a better NMT model by reducing the number of OOV words while keeping their original meaning. We paraphrase infrequent words using a paraphrase dictionary that has paraphrase pairs annotated with a paraphrase score. We employ three scores: (1) paraphrase score, (2) language model (LM) score, and (3) a combination of these scores. The paraphrase score is meant to reflect translation adequacy, and the language model score is sensitive to fluency. We combined the paraphrase score and the language model score by linear interpolation[1] as follows:

$$\begin{aligned} paraphrase\_score &= \\ \lambda(PPDBscore) &+ (1-\lambda)(LMscore) \end{aligned}$$

Figure 1 shows an example of paraphrasing with a paraphrase lattice and the Viterbi algorithm. Suppose "defending" is OOV. We can paraphrase the OOV word "defending" with a frequent word,

[1] In a preliminary experiment, normalization of these scores was not found to yield any improvements.

method	BLEU	METEOR	OOV
baseline	25.70†	31.06	1,123
Luong et al.	25.87†	31.04	567
Sennrich et al.	25.92*	31.50	0
Li et al.	25.89*	31.10	832
proposed (multi. word + phrase)	**26.47**	**31.62**	668

Table 1: Japanese-to-English translation result of each method. † and * indicate that the proposed method significantly outperformed the other methods at $p<0.01$ and $p<0.05$, respectively, using bootstrap resampling.

"guaranteeing", or we can paraphrase the OOV phrase "defending the rights" with another phrase, "the protection of the rights", which has no OOV words. In addition to calculating the paraphrase score, our paraphrase algorithm calculates the 2-gram language model score in "assert guaranteeing the rights .", "assert the", and "rights ." and chooses the highest scoring paraphrase, thus generating "they assert the protection of the rights.". We do not calculate the 2-gram language model score in phrases.[2]

In addition, our method can paraphrase OOV words iteratively until a paraphrase with frequent words is reached. In the lower example in Figure 1, suppose that "pedagogues" and "quarrels" are OOV. The latter word in the original sentence is paraphrased with a frequent word, "discussions", whereas the former is paraphrased with an infrequent word, "educators", in the first round. We can then paraphrase the infrequent word "educators" again, this time with a frequent word, "teachers", in the second round. If we allow only the first round of paraphrasing, the infrequent word "pedagogues" will not be paraphrased with the frequent word "teachers" because the paraphrase dictionary does not have this entry, and the infrequent word "pedagogues" will not be paraphrased with the infrequent word "educators". In this paper, we express one-pass paraphrasing as "single", and iterative paraphrasing as "multi.". In addition, we use "word" when we paraphrase words, and "word + phrase" when paraphrasing words and phrases.

4 Experiment

4.1 Settings

In this study, we used the Japanese–English portion of the Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016). For training, we used one million sentence pairs ranked by alignment accuracy. We deleted sentence pairs longer than 41 words. The final training corpus contained 827,503 sentence pairs. We followed the official development/test split: 1,790 sentence pairs for development, and 1,812 sentence pairs for testing. We used the development dataset to select the best model and used the test dataset to evaluate BLEU scores.

We used the Moses script as an English tokenizer and MeCab[3] (using IPAdic) as a Japanese tokenizer. We employed KenLM[4] to build a 2-gram language model trained with all sentences from ASPEC. We utilized the XXXL-size PPDB 2.0 (Pavlick et al., 2015) as the English paraphrase dictionary and PPDB:Japanese (Mizukami et al., 2014) as the Japanese paraphrase dictionary. Neither of these dictionaries contains the ASPEC corpus. We paraphrased either the target side of the training corpus only or both the source and target sides of the training corpus to conduct a fair comparison. We experimented with $\lambda = 0.0$, 0.25, 0.50, 0.75, and 1.0.

We used OpenNMT-py[5] as the NMT system, which is a Python implementation of OpenNMT (Klein et al., 2017). We built a model with settings as described below. We used bi-recurrent-neural-network, batch size 64, epoch 20, embedding size 500, vocabulary size of source and target 30,000, dropout rate 0.3, optimizer SGD with learning rate 1.0, and number of RNN layers 2 with an RNN size of 500. Our baseline was trained with these settings without any paraphrasing. We re-implemented previous methods described in this paper (Luong et al., 2015; Li et al., 2016; Sennrich et al., 2016) using the underlying NMT with the abovementioned settings. We

[2]Calculating language model scores of phrases does not improve NMT.

[3]https://github.com/taku910/mecab

[4]http://kheafield.com/code/kenlm/

[5]https://github.com/OpenNMT/OpenNMT-py

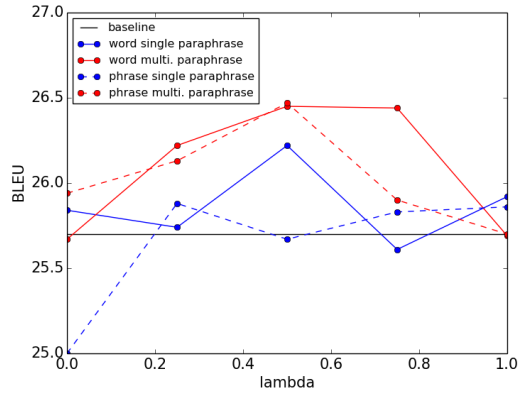


Figure 2: BLEU score of the proposed method in Japanese-to-English translation using various weightings for the paraphrase score and the language model score.

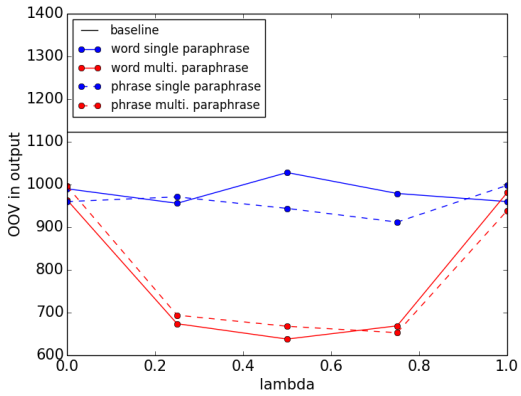


Figure 3: Number of OOV terms in the output of the proposed method in Japanese-to-English translation.

used BLEU (Kishore et al., 2002) and METEOR (Lavie and Agarwal, 2007) for extrinsic evaluation. We also analyzed the number of OOV words in the translated sentences as an intrinsic evaluation.

English PPDB 2.0 achieved higher quality than PPDB 1.0 by using a supervised regression model to estimate paraphrase scores. However, because there are no training data to build a supervised regression model for PPDBs other than in English, the quality of PPDBs in other languages may affect the quality of the proposed method. To investigate whether PPDB quality relates to translation quality, we performed English to Japanese translation by the proposed method using PPDB:Japanese (Mizukami et al., 2014).

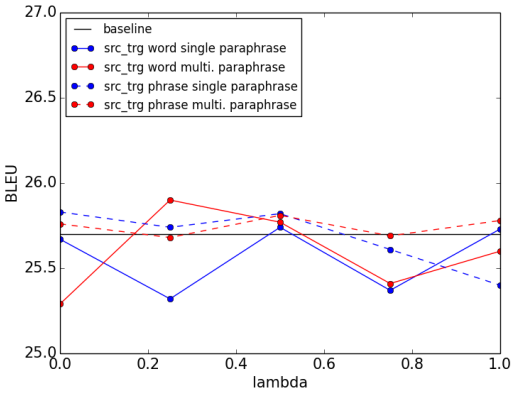


Figure 4: BLEU score in Japanese-to-English translation using source and target paraphrasing.

method	BLEU	OOV
baseline	33.91	1,003
single (word)	33.97	915
multi. (word)	**34.09**	966
single (word + phrase)	33.65	938
multi. (word + phrase)	33.86	**902**

Table 2: English-to-Japanese translation results with variations in the number of paraphrasings and the unit used.

4.2 Results

Table 1 shows the experimental results compared with those in previous work. The proposed method is multi. word + phrase paraphrasing. In the BLEU evaluation, our method significantly outperformed not only the baseline and Luong et al. ($p<0.01$) but also Sennrich et al. and Li et al. ($p<0.05$). We improved the BLEU score by 0.77 and the METEOR score by 0.56 as well as reducing the number of OOV words in the output by approximately 40% compared with the baseline.

Figure 2 reports the BLEU score of our method under variations in the linear interpolation coefficient and the number of paraphrasings. Figure 3 shows the number of OOV words in the output. The best BLEU score was achieved by multi-round paraphrasing and $\lambda = 0.50$, which means that the paraphrase score is balanced by the PPDB score and the LM score.

Table 2 shows the BLEU score of the proposed method on English to Japanese translation. The best model improved the BLEU score by 0.18 over the baseline, and the number of OOV words in the output decreased slightly.

In the last experiment, we paraphrased the

method	translation
source	ロックインアンプを使用すれば, ノイズを著しく減少できることを期待できる。
reference	with the lock - in amplifier used , significant reduction of the noise is expected .
baseline	it is expected that the noise can be reduced remarkably , if the <unk> is used .
multi. (word)	it is expected that the noise can be remarkably decreased , if the amplifier is used .
multi. (phrase)	it is expected that the noise can be remarkably reduced by using the lock-in amplifier .

Table 3: Translation example in Japanese-to-English translation.

infrequent word	frequent word
megahertz	mhz
deflagration	combustion
cone-shaped	conical
revalued	examined
titrated	measured
teleportation	transport

Table 4: Iterative paraphrasing example of domain-specific words with frequent words.

source and target sides of the training corpora to compare the effect of target-only paraphrasing. Figure 4 shows that the method paraphrasing both source and target sentences does not improve the translation quality over the baseline.

5 Discussion

Figures 2 and 3 show that a multi-round paraphrasing method is better than a single-round paraphrase in terms of BLEU score and OOV rate. In multi-round paraphrasing, however, a paraphrased word does not necessarily retain its original meaning in successive paraphrases. The number of OOV words is negatively correlated with the BLEU score, demonstrating that our hypothesis is correct.

On English-to-Japanese translation, the improvement is not statistically significant; however, we believe that our system does not rely on PPDB quality, although the degree of improvement will depend on the quality of the PPDB.

Table 3 is an example of a translation result. This table indicates that the baseline system outputs "<unk>" instead of "amplifier". In contrast, a paraphrasing system can output "amplifier" because a number of words corresponding to "amplifier" are paraphrased into "amplifier" in the proposed method. As a result, the proposed systems can correctly output the word "amplifier".

Table 4 is an example of iterative paraphrasing on special words in ASPEC. This shows that we can paraphrase domain-specific words and that these paraphrases can improve the translation. The paraphrases shown in the upper half of the table preserve meaning, whereas those in the lower half lose a little of the original meaning.

6 Conclusion

This paper has proposed a preprocessing method that paraphrases infrequent words with frequent words in a target corpus during training to train a better NMT model by reducing the OOV rate. An evaluation using the Japanese-to-English part of the ASPEC corpus showed a decrease in the OOV rate in the translation result and a significant improvement in the BLEU score over state-of-the-art methods. We expect that our method can be effective not only in NMT but also in other text generation tasks using neural networks, such as abstractive summarization, which solves the classification problem of vocabulary dimension.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for WMT'15. In *Proc. of the Tenth Workshop on Statistical Machine Translation*. pages 134–140.

Papineni Kishore, Roukos Salim, Ward Todd, and Zhu Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*. pages 311–318.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:* 1701.02810.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc. of the Second Workshop on Statistical Machine Translation*. pages 228–231.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *Proc. of IJCAI*. pages 2852–2858.

Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proc. of ACL-IJCNLP*. pages 11–19.

Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Building a free, general-domain paraphrase database for Japanese. In *Proc. of O-COCOSDA*. pages 1–4.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proc. of LREC*. pages 2204–2208.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proc. of ACL*. pages 425–430.

Štajner Sanja and Popovic Maja. 2016. Can text simplification help machine translation? *Baltic Journal of Modern Computing* 4(2):230–242.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL*. pages 1715–1725.