Mitigating Domain Mismatch in Machine Translation via Paraphrasing

Hyuga Koretaka[1] koretaka@ai.cs.ehime-u.ac.jp
Tomoyuki Kajiwara[1] kajiwara@cs.ehime-u.ac.jp
Atsushi Fujita[2] atsushi.fujita@nict.go.jp
Takashi Ninomiya[1] ninomiya@cs.ehime-u.ac.jp
[1]Graduate School of Science and Engineering, Ehime University, Ehime, Japan
[2]National Institute of Information and Communications Technology, Kyoto, Japan

Abstract

Quality of machine translation (MT) deteriorates significantly when translating texts having characteristics that differ from the training data, such as content domain. Although previous studies have focused on adapting MT models on a bilingual parallel corpus in the target domain, this approach is not applicable when no parallel data are available for the target domain or when utilizing black-box MT systems. To mitigate problems caused by such domain mismatch without relying on any corpus in the target domain, this study proposes a method to search for better translations by paraphrasing input texts of MT. To obtain better translations even for input texts from unforeknown domains, we generate their multiple paraphrases, translate each, and rerank the resulting translations to select the most likely one. Experimental results on Japanese-to-English translation reveal that the proposed method improves translation quality in terms of BLEU score for input texts from specific domains.

1 Introduction

Despite recent advances in machine translation (MT), translation quality still depends on the characteristics of the data on which the MT system is trained. Therefore, for input texts having significantly different characteristics from the training data, there is a risk that translation quality may be degraded (Koehn and Knowles, 2017). To alleviate mismatches in one of such characteristics, content domain (henceforth, domain), an approach of transfer learning (Chu and Wang, 2018) is commonly used, where a pre-trained MT model is fine-tuned on a parallel corpus in the target domain. However, there are many challenges in supporting a variety of domains. First of all, there are only a limited number of domains that have access to a bilingual parallel corpus sufficient for fine-tuning pre-trained MT models. Even if parallel corpora are available for a large number of domains, then the time required for fine-tuning for each domain and the management cost for resulting models will not be negligible. More importantly, existing domain adaptation methods are not applicable in situations where we target a black-box MT system, such as Google Translate and DeepL, even if they are already superior to pre-trained MT models in the domain of interest.

This paper proposes a method to bridge the domain gap between sentences to be translated and MT training data without a need for additional training for a specific target domain as in existing domain adaptation methods, such as fine-tuning pre-trained MT models on human-made

Proceedings of the 10th Workshop on Asian Translation, pages 29–40
September 4, 2023, Macau SAR, China

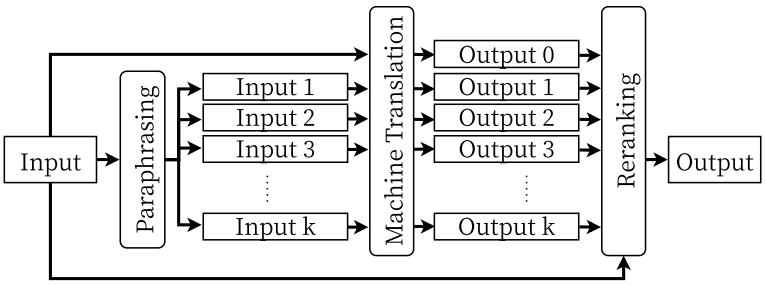

Figure 1: Overview of the proposed method.

and/or synthetic in-domain parallel data. As shown in Figure 1, our method first generates multiple paraphrases from a given input sentence, generates translation candidates using the given black-box MT system (henceforth, target MT system), and finally reranks those candidates to select the best one. We assume that diverse paraphrases of the sentences to be translated could include expressions that are less deviated for the target MT system, and such paraphrase could lead to an improvement of translation quality. The proposed method has the advantage that it requires neither fine-tuning the MT model nor any bilingual parallel corpus in the target domain. Our controlled experiment on Japanese-to-English translation revealed that the proposed method improves translation quality in terms of BLEU score in the domains that are not covered when training the target MT system.

2 Related Work

Paraphrasing input sentences has improved the performance of various natural language processing tasks such as document summarization (Siddharthan et al., 2004) and information extraction (Evans, 2011). Paraphrasing has been studied also for MT. Miyata and Fujita (2017, 2021) investigated manual pre-editing, including paraphrasing, to push the limit of existing MT services, and identified diverse types of pre-editing that can improve translation quality. However, there are two issues in automating paraphrasing for MT. One is that we lack a method for producing diverse (and accurate) paraphrases. Past work on automatic paraphrasing for MT (Štajner and Popović, 2016, 2018; Mehta et al., 2020) has examined only a limited variation, i.e., lexical and/or syntactic simplification, and observed quality improvement only in limited settings. Another issue is that the effect of each particular paraphrasing is unpredictable due to the sensitivity of neural MT to input sentences (Miyata and Fujita, 2021). We thus need to assess the quality of MT outputs in a post-hoc manner rather than the quality of paraphrased sentences before translating them with the target MT system.

3 Proposed Method

To automatically bridge the gap between the domains of input sentences and MT training data, we propose a framework consisting of three steps shown in Figure 1. Given an input sentence, we generate multiple paraphrases for it, translate each of the input sentence and multiple paraphrases using the target MT system, and select one candidate translation through reranking.

In this section, we describe the first paraphrasing step and the third reranking step. For the second step, i.e., MT, we primarily assume a black-box system, such as online MT services. However, to explore better reranking, we also consider a glass-box setting, assuming that some information can be drawn from the target MT system.

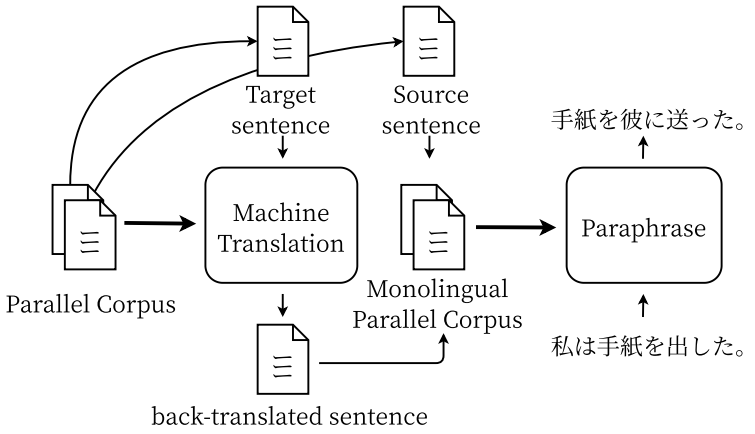

Figure 2: Sentence-level paraphrase.

3.1 Paraphrasing

Given a sentence to be translated, we can assume that there must be a paraphrase of it which can be better translated by the target MT system. However, it is difficult to know in advance which paraphrases are more likely to be better translated by the target MT system (Miyata and Fujita, 2021). Instead of determining the best paraphrase of the input sentence for the MT system, we leave the selection for the post MT step.

We consider generating multiple paraphrases both at sentence and word levels. Whereas sentence-level paraphrasing aims to paraphrase entire sentences using a sequence-to-sequence model, word-level paraphrasing focuses on altering single words relying on a masked language model and word embeddings.

Sentence-level Paraphrasing: Following previous work, we regard sentence-level paraphrasing as a sequence-to-sequence task. There are two conceivable options: back-translation-based approach (Wieting and Gimpel, 2018) illustrated in Figure 2 and monolingual translation-based approach (Thompson and Post, 2020b). The back-translation-based approach relies on a bilingual parallel corpus of the source language and an arbitrary target language. First, we train a translation model from the target language to the source language on the parallel corpus. Through translating the target side of the parallel corpus into the source language and coupling them with the source sentences in the parallel corpus, we automatically obtain a monolingual parallel corpus in the source language. Using this monolingual parallel corpus, we train a paraphrasing model[1], placing the back-translated sentences at the source side. On the other hand, the monolingual translation-based approach uses a pre-trained multilingual MT model that covers the source language. The model has been trained only on the bilingual parallel data, but no parallel data on the same language for paraphrasing; nevertheless, it is inherently capable of generating paraphrases provided that the language appears both in the source and target side during pre-training. We generate sentence-level paraphrases by specifying the same language for both source and target. For both approaches, given an input sentence, we generate its k paraphrases with the paraphrase models by performing a beam search with a beam size of k.

Word-level Paraphrasing: For word-level paraphrasing, we propose a method based on a pre-trained masked language model and pre-trained word embeddings both covering the source

[1]We refer to the back-translation-based paraphrase model as Denoiser.

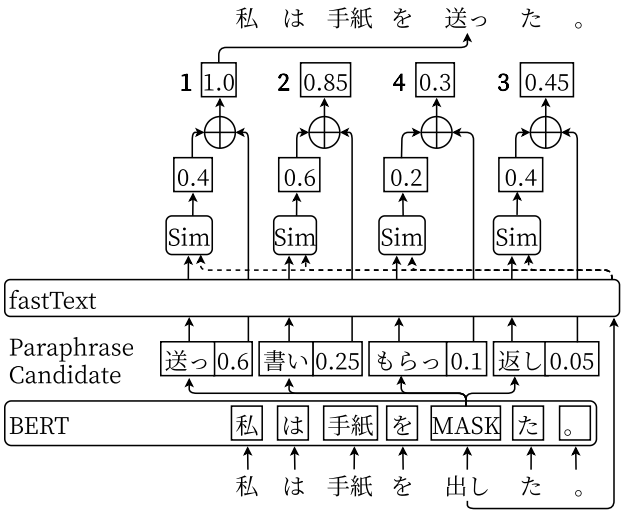

Figure 3: Word-level paraphrase.

language. As illustrated in Figure 3, our method consists of two steps: candidate generation and (word-level) reranking. Let $X = x_1, \ldots, x_{|X|}$ be an input sentence consisting of $|X|$ words. First, we generate the top n paraphrase candidates for each word x_i ($1 \leq i \leq |X|$): the input sentence with each word x_i masked is input to the masked language model and n-best words[2] according to their probability are output. Then, among $(|X| \times n)$ paraphrase candidates, we select top k candidates. For this reranking, we use the sum of the probability of the masked language model and the cosine similarity of static word embeddings of the original word x_i and the paraphrase candidate.

3.2 Reranking

As shown in Figure 1, the reranking step selects one translation among $(k + 1)$ candidates generated by the target MT system for each of the input sentence and its k paraphrases. Various features have been proposed for reranking translation candidates (Marie and Fujita, 2018; Kiyono et al., 2020), but most of them are not available when we target a black-box MT system, such as Google Translate and DeepL. Therefore, we implement a reranking method for such an MT system relying only on a simple translation likelihood score that can be drawn from freely available MT models, such as mBART (Tang et al., 2020) and M2M-100 (Fan et al., 2021). We call this black-box reranking, since no information is retrieved from the target MT system. On the other hand, if the target MT system can give a score for a given pair of source sentence and translation candidate, such score should better help reranking. We therefore also examine this glass-box reranking.

Black-box Reranking: In black-box reranking, we use a multilingual MT model that covers both the source and target languages. As in previous work (Thompson and Post, 2020a; Kiyono et al., 2020), we compute the forced-decoding score[3] from the input sentence to each translation candidate. Additionally, forced-decoding score from each candidate to the input sentence is computed and the candidates are reranked simply by the average of the forced decoding scores for the two directions.

[2]This may include x_i itself.

[3]e.g., log probability normalized by the length.

Glass-box Reranking: In glass-box reranking, we use the target MT model and another MT model for the backward translation direction, i.e., the target language to the source language. Given a pair of the input sentence and the translation candidate, we compute forced-decoding scores for both translation directions using these models in the same manner with the black-box reranking, and the candidates are simply reranked by the average of these scores. Note that the score from the input sentence to each candidate should have already been given during the previous MT decoding step.

4 Experiments

We evaluated the effectiveness of the proposed method on three Japanese-to-English translation tasks. For the sake of fair comparison, in our work, we built an MT system by ourselves instead of using a truly black-box one.

4.1 Settings

Data: To train the target MT system, we randomly extracted 10 million sentence pairs for training and another 2,000 sentence pairs for validation from JParaCrawl[4] (Morishita et al., 2022). To train a sentence-level paraphrasing model with the back-translation-based approach, we randomly extracted 10 million sentence pairs for training and another 2,000 sentence pairs for validation from the remaining part of JParaCrawl.

We used three test sets on specific domains. One is ASPEC (Nakazawa et al., 2016), an excerpt from scientific papers, consisting of 1,812 sentence pairs. Second one is the test set used in WMT20 Shared Task on News (Barrault et al., 2020), an excerpt from news domain consisting of 993 sentence pairs. Last one is MTNT2019, the test set used in WMT 2019 Machine Translation Robustness Shared Task (Li et al., 2019) excerpted from the Reddit discussion website, consisting of 1,100 sentence pairs.[5] For reference, we randomly extracted 2,000 sentence pairs from JParaCrawl as an general-domain[6] test set. Note that they do not overlap with the training and validation sets used for the MT and sentence-level paraphrasing models.

As a preprocessing step, we first removed duplicates and split the data from JParaCrawl. We then applied NFKC normalization to all train/validation data and the source side of test data,[7] and then trained unigram-based subwording models (Kudo, 2018) on the training data using SentencePiece[8] (Kudo and Richardson, 2018). For MT, we obtained two separate vocabularies of 32,000 subwords for Japanese and English, respectively, and then applied the model to the training and validation data in their respective language. For sentence-level paraphrasing model with the back-translation-based approach, we first generated a monolingual parallel corpus by back-translating the English side of the sampled parallel data into Japanese, and then trained a single model covering 32,000 subwords on the training part of the monolingual parallel corpus. Both sides of the training and validation data were tokenized with the obtained SentencePiece model. We set the character coverage option of sentencepiece to 1.0 and 0.9998 for English and Japanese, respectively.

When inputting the source side of the test data to our sentence-level paraphrasing model, it was tokenized using the corresponding model. Sentence-level paraphrases were once detokenized with the same model, and again tokenized using the model trained on the Japanese side of the sampled bilingual parallel data before the succeeding MT step.

[4] `https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/`, v3.0

[5] `https://pmichel31415.github.io/mtnt/`, with an empty line 1,033 excluded.

[6] JParaCrawl can be considered as a general-domain parallel corpus, because it covers various domains seen on the Internet.

[7] We left the target side of the test data unprocessed, i.e., reference translations, following Post (2018).

[8] `https://github.com/google/sentencepiece/`

Models: We trained a Transformer Base model (Vaswani et al., 2017) for MT, sentence-level paraphrasing, and the backward MT for glass-box reranking with Fairseq[9] (Ott et al., 2019). We trained these models using mini-batch size of $60,000$ tokens, a learning rate of 5×10^{-4}, dropout of 0.1, label smoothing of 0.1, and Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We computed the cross-entropy loss for the validation data every 1,500 steps and stopped the training after 10 consecutive times without an improvement of the best cross-entropy loss. We ran the model training only once on three A6000 GPUs, which consumed 22 and 26 hours for the MT and sentence-level paraphrasing models, respectively. For inference of MT in our framework, we used 1-best output generated by beam search with a beam size of 5 and the length penalty of 1.0.

To implement sentence-level paraphrasing with the back-translation-based approach, we generated back-translated sentences using a multilingual MT model called M2M-100[10] (Fan et al., 2021). On the other hand, to implement the monolingual translation-based approach, we used M2M-100 and mBART fine-tuned for multilingual translation[11] (Tang et al., 2020) separately for the sake of comparison. Note that none of these models have been fine-tuned on any paraphrase-specific data.

To implement the word-level paraphrasing model, we exclusively used Japanese model of BERT (Devlin et al., 2019) (JaBERT[12]) and multilingual model of BERT (mBERT[13]) as the masked language model, and the Japanese model of fastText[14] (Bojanowski et al., 2017) as the word embeddings. We set the number of candidate paraphrases n for each word to 10.

To implement black-box reranking, we used mBART.[9]

Comparison Method: As a comparison method without paraphrasing, we output one translation for each input sentence using beam search with a beam size of $(k+1)$. Also, we output $(k+1)$ candidate translations for each input sentence using beam search with a beam size of $(k+1)$, assuming such a functionality in the given black-box MT system, and reranked them in the same manner as the third step in our framework.

Evaluation Metric: To evaluate the quality of translation, we computed BLEU score (Papineni et al., 2002) using SacreBLEU[15] (Post, 2018). To determine if differences in BLEU scores are significant, we performed statistical significance testing ($p < 0.05$) based on paired bootstrap resampling implemented in SacreBLEU.

4.2 Results of Individual Paraphrasing Methods

Table 1 shows the BLEU scores of our methods and baseline methods that do not use any paraphrasing method, where the oracle results based on the best sentence-level BLEU scores are also presented.

First, we focus on the paraphrase generation method. In ASPEC and WMT20, word-level paraphrasing (Models (3)–(4) and (9)–(10) in Table 1) consistently performed better than sentence-level paraphrasing (Models (5)–(7) and (11)–(13)). In contrast, in MTNT2019 and JParaCrawl, the word-level and sentence-level paraphrasing methods were not superior or inferior to each other. Overall, the oracle results show that word-level paraphrasing (Models (15)–

[9] https://github.com/facebookresearch/fairseq/
[10] https://huggingface.co/facebook/m2m100_418M/
[11] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt/
[12] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking/
[13] https://huggingface.co/bert-base-multilingual-cased/
[14] https://fasttext.cc/docs/en/crawl-vectors.html
[15] https://github.com/mjpost/sacrebleu/
Short signature: BLEU|#:1|c:mixed|e:no|tok:13a|s:exp|v:2.3.1

(16)) has a larger potential compared to sentence-level paraphrasing (Models (17)–(19)), and in fact word-level paraphrasing methods have performed more effectively than sentence-level paraphrasing.

Next, we focus on the domain of evaluation data. In ASPEC and WMT20, word-level paraphrasing (Models (3)–(4) and (9)–(10)) consistently showed the best performance. On the other hand, in MTNT2019 and JParaCrawl, paraphrasing caused either no change or a degradation in translation quality. As for MTNT2019, the oracle BLEU scores (Models (14)–(19)) show that the gains are comparably large with those for ASPEC and WMT20 tasks. We therefore consider that the poor results for MTNT2019 are attributed to the versatility of the pre-trained MT models used for reranking, i.e., mBART and our own models based on JParaCrawl; they have not been trained on translations of less formal texts, such as user-generated contents in MTNT2019, and thus were not capable of selecting appropriate translations among candidates. In contrast, the oracle BLEU scores for JParaCrawl are substantially lower than those based on beam search, indicating the poor potential of our paraphrasing methods. One possible explanation for this is that there is no domain gap between the input sentences and the MT model, and paraphrasing may make the sentences unsuitable for the MT model.

Finally, we focus on the number of paraphrases, i.e., k. In ASPEC and WMT20, translation quality improved as k increased in the black-box reranking of word-level paraphrases (Models (3)–(4)) and glass-box reranking (Models (9)–(13)). In contrast, in MTNT2019 and JParaCrawl, translation quality decreased or remained the same even when k was increased.

4.3 Results of Combinations of Paraphrasing Methods

Having evaluated the sentence-level and word-level paraphrasing methods, we explored whether their combinations further boost the translation quality. Combination here means the merger of two sets of $(k + 1)$ translation candidates, which results in $(2k + 1)$ candidates since two sets contain an identical one, i.e., the translation for the original input sentence.

Table 2 shows the BLEU scores of all the combinations of word-level and sentence-level paraphrasing methods and the baseline methods that do not use any paraphrasing method but rely on $(2k+1)$ translation candidates obtained by beam search. Compared to the results for the word-level methods only (Models (3)–(4) and (9)–(10) in Table 1), most of the combinations resulted in either no change or degeneration on BLEU scores. Even though some conditions in WMT20 and MTNT2019 tasks have some benefits from the combination, we recommend to use the word-level paraphrasing methods alone rather than combining them with sentence-level paraphrasing methods.

4.4 Analysis of Translations for Paraphrases

Figure 4 shows the percentage of candidate translations for paraphrases that have an increased sentence-level BLEU score compared to the translation for the original sentence. For all paraphrase generation methods, about 20–30% of candidate translations for paraphrases improved the BLEU score compared to the translations for the original sentences. For word-level paraphrasing, which was the most effective paraphrase generation method, the percentages increased as k increased. Therefore, we consider that increasing k is effective in obtaining better candidate translations, even though it increases the cost for generating candidates.

	Paraphrasing		Reranking	ASPEC			WMT20			MTNT2019			JParaCrawl		
ID	Level	Model	Model	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
(1)	-	-	-	20.5	20.7	20.6	20.6	20.9	20.8	15.5	15.6	15.9	34.6	34.6	34.5
(2)	-	-	mBART (black-box)	20.6	20.8	21.1*	20.8	21.2*	21.4*	**15.8**	**15.9**	**15.9**	**34.1***	**33.9***	**33.4***
(3)	Word	JaBERT		20.9*	21.0	20.8	**21.5***	**21.8***	**21.9***	15.6	15.6	15.5	33.3*	32.9*	32.6*
(4)	Word	mBERT		**21.1***	**21.2***	**21.2***	21.2*	21.4*	21.8*	15.4	15.3	15.4	33.3*	33.1*	32.7*
(5)	Sentence	mBART		20.3	20.1*	20.1*	20.3	20.4*	20.4*	14.8*	14.6*	14.7*	33.1*	32.9*	32.2*
(6)	Sentence	M2M-100		19.9*	19.8*	19.7*	20.8	20.6	20.8	15.3	15.2	14.5*	32.9*	32.5*	32.0*
(7)	Sentence	Denoiser		20.4	20.3*	20.4	20.7	20.9	20.9	15.5	15.3	15.4	33.8*	33.1*	32.9*
(8)	-	-	MT (glass-box)	20.7	21.1*	21.2*	20.8	20.9	21.2*	15.8*	**15.9**	**15.9**	**34.7**	**34.7**	34.3
(9)	Word	JaBERT		**21.2***	**21.4***	21.2*	**21.6***	**22.0***	21.9*	**15.9***	15.7	15.5	34.5	34.5	34.3
(10)	Word	mBERT		**21.2***	21.3*	**21.6***	**21.6***	21.8*	**22.1***	15.7	15.4	15.5	34.5	34.4	34.0*
(11)	Sentence	mBART		20.8*	20.7	21.0*	21.0*	20.9	21.1	15.4	15.5	15.4	**34.7**	34.6	34.6
(12)	Sentence	M2M-100		20.6	20.6	20.6	21.2*	21.4*	21.5*	15.8	**15.9**	15.7	34.6	**34.7**	**34.7**
(13)	Sentence	Denoiser		20.6	20.7	20.8	20.9*	20.9	21.2*	15.8	15.7	15.8	**34.7**	**34.7**	**34.7**
(14)	-	-	Oracle	**24.3**	26.0	**27.9**	23.6	25.3	26.8	19.0	20.2	**22.0**	**38.6**	**40.1**	**41.5**
(15)	Word	JaBERT		**24.3**	25.9	27.4	24.9	**26.3**	27.6	**19.3**	**20.4**	21.8	37.8	38.7	39.8
(16)	Word	mBERT		**24.3**	**26.1**	**27.9**	**25.0**	26.2	**27.7**	19.0	**20.4**	21.7	37.6	38.6	39.7
(17)	Sentence	mBART		23.4	24.4	25.4	23.1	23.8	24.8	17.8	18.3	19.2	36.4	37.1	37.6
(18)	Sentence	M2M-100		23.6	24.5	25.5	23.8	24.5	25.5	18.2	19.0	19.8	36.6	37.1	37.7
(19)	Sentence	Denoiser		23.5	24.6	25.8	23.7	24.7	25.4	18.4	19.3	20.3	36.7	37.4	38.1

Table 1: BLEU scores (k: number of paraphrases, **bold**: the highest BLEU score of each reranking result by column, *: statistically significant difference ($p < 0.05$) over the baseline method in the first row.)

	Paraphrasing		Reranking	ASPEC			WMT20			MTNT2019			JParaCrawl		
ID	Level	Model	Model	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
(1)	-	-	-	20.7	20.6	20.8	20.9	20.8	20.7	15.6	15.9	15.7	34.6	34.5	34.6
(2)	-	-		**20.8**	**21.1***	**21.2***	**21.2***	21.4*	21.2*	**15.9**	**15.9**	**16.1**	**33.9***	**33.4***	**33.0***
(3)	Word + Sentence	JaBERT + mBART		20.5	20.5	20.7	21.1	21.5*	21.4*	14.8*	14.8*	15.0*	32.4*	32.1*	31.6*
(4)	Word + Sentence	JaBERT + M2M-100	mBART	20.2*	20.4	20.2*	**21.2**	21.4*	21.5*	15.3	15.3	14.7*	32.1*	31.6*	31.3*
(5)	Word + Sentence	JaBERT + Denoiser	(black-	20.7	20.8	20.8	**21.2**	**21.7***	**21.7***	15.5	15.5	15.6	33.0*	32.4*	32.1*
(6)	Word + Sentence	mBERT + mBART	box)	20.6	20.5	20.9	21.0	21.3*	21.5*	14.5*	14.7*	15.2	32.3*	32.2*	31.5*
(7)	Word + Sentence	mBERT + M2M-100		20.5	20.6	20.7	21.1	21.2	21.5*	15.3	15.2*	15.0*	32.3*	31.9*	31.5*
(8)	Word + Sentence	mBERT + Denoiser		**20.8**	20.8	20.8	21.0	21.2	21.6*	15.3	15.4	15.4	33.1*	32.4*	32.0*
(9)	-	-		21.1*	21.2*	21.3*	20.9	21.2*	21.0	15.9	**15.9**	15.7	**34.7**	**34.3**	**34.4**
(10)	Word + Sentence	JaBERT + mBART		**21.2***	21.2*	21.3*	21.7*	21.9*	21.7*	15.8	15.6	15.6	34.3	34.2	34.0*
(11)	Word + Sentence	JaBERT + M2M-100	MT	**21.2***	**21.3***	21.1	**21.8***	**22.2***	21.9*	16.0	15.8	15.5	34.2	34.2	34.0*
(12)	Word + Sentence	JaBERT + Denoiser	(glass-	21.0	**21.3***	21.2*	21.5*	22.0*	21.8*	**16.2**	15.8	15.7	34.4	**34.3**	34.2
(13)	Word + Sentence	mBERT + mBART	box)	**21.2***	21.2*	**21.6***	21.6*	21.8*	21.9*	15.7	15.5	15.7	34.4	34.2	33.9*
(14)	Word + Sentence	mBERT + M2M-100		**21.2***	21.2*	21.4*	**21.8***	21.9*	**22.1***	16.0	15.7	15.6	34.2	34.2	33.8*
(15)	Word + Sentence	mBERT + Denoiser		21.1*	**21.3***	21.4*	21.5*	21.7*	**22.1***	15.8	15.6	**15.9**	34.5	34.2	34.0*
(16)	-	-		**26.0**	**27.9**	**29.8**	25.3	26.8	28.2	20.2	22.0	**23.8**	**40.1**	**41.5**	**42.9**
(17)	Word + Sentence	JaBERT + mBART		25.8	27.4	29.0	25.8	27.2	28.5	20.5	21.6	23.1	38.6	38.6	40.7
(18)	Word + Sentence	JaBERT + M2M-100		25.9	27.5	29.1	26.1	**27.6**	28.8	20.7	22.0	23.5	38.6	39.5	40.7
(19)	Word + Sentence	JaBERT + Denoiser	Oracle	25.8	27.4	28.9	26.1	27.5	28.7	**20.9**	21.9	23.5	38.7	39.7	40.7
(20)	Word + Sentence	mBERT + mBART		25.7	27.5	29.2	25.9	27.1	28.6	20.1	21.6	23.0	38.5	39.6	40.6
(21)	Word + Sentence	mBERT + M2M-100		25.9	27.7	29.4	**26.3**	27.5	**28.9**	20.5	**22.1**	23.4	38.4	39.4	40.5
(22)	Word + Sentence	mBERT + Denoiser		25.8	27.5	29.3	26.2	27.5	28.8	20.6	22.0	23.4	38.6	39.6	40.7

Table 2: BLEU scores (k: number of paraphrases per word and sentence level respectively, **bold**: the highest BLEU score of each reranking result by column, *: statistically significant difference ($p < 0.05$) over the baseline method in the first row, underline: improvement over the component word-level paraphrasing method.)

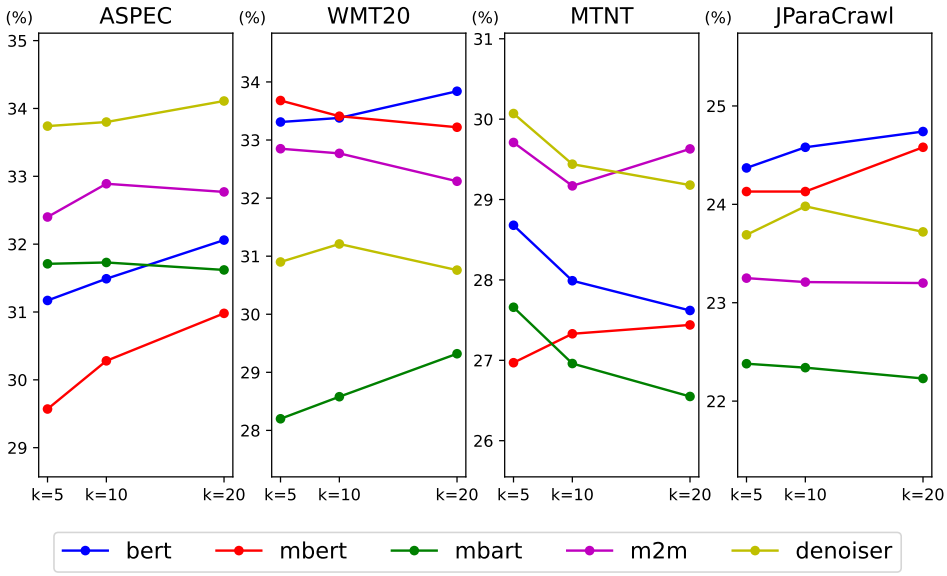

Figure 4: Percentage of candidate translations for paraphrases that have an increased sentence-level BLEU score compared to the translation for the original sentence. (k: number of paraphrases)

5 Conclusion

In this study, to mitigate the domain mismatch between the domains of the input sentence and the training data of the target MT system, we proposed a framework that combines paraphrase generation and reranking. In particular, the combination of word-level paraphrase generation and glass-box reranking consistently improved translation quality in the two specific domains most significantly.

Our future work will focus on improving reranking and filtering word-level paraphrases to further improve performance.

Acknowledgments

We would like to thank the reviewers for their insightful comments and suggestions. This work was supported by JST, ACT-X Grant Number JPMJAX1907. These research results were obtained from the commissioned research (No.22501) by National Institute of Information and Communications Technology (NICT), Japan.

References

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 Conference on Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chu, C. and Wang, R. (2018). A Survey of Domain Adaptation for Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Evans, R. J. (2011). Comparing Methods for the Syntactic Simplification of Sentences in Information Extraction. *Literary and Linguistic Computing*, 26(4):371–388.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2021). Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(1):4839–4886.

Kiyono, S., Ito, T., Konno, R., Morishita, M., and Suzuki, J. (2020). Tohoku-AIP-NTT at WMT 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 145–155.

Koehn, P. and Knowles, R. (2017). Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Li, X., Michel, P., Anastasopoulos, A., Belinkov, Y., Durrani, N., Firat, O., Koehn, P., Neubig, G., Pino, J., and Sajjad, H. (2019). Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102.

Marie, B. and Fujita, A. (2018). A Smorgasbord of Features to Combine Phrase-Based and Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, pages 111–124.

Mehta, S., Azarnoush, B., Chen, B., Saluja, A., Misra, V., Bihani, B., and Kumar, R. (2020). Simplify-Then-Translate: Automatic Preprocessing for Black-Box Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8488–8495.

Miyata, R. and Fujita, A. (2017). Dissecting Human Pre-Editing toward Better Use of Off-the-Shelf Machine Translation Systems. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*, pages 54–59.

Miyata, R. and Fujita, A. (2021). Understanding Pre-Editing for Black-Box Neural Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1539–1550.

Morishita, M., Chousa, K., Suzuki, J., and Nagata, M. (2022). JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2204–2208.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation*, pages 186–191.

Siddharthan, A., Nenkova, A., and McKeown, K. (2004). Syntactic Simplification for Improving Content Selection in Multi-Document Summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 896–902.

Štajner, S. and Popović, M. (2016). Can Text Simplification Help Machine Translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Štajner, S. and Popović, M. (2018). Improving Machine Translation of English Relative Clauses with Automatic Text Simplification. In *Proceedings of the 1st Workshop on Automatic Text Adaptation*, pages 39–48.

Tang, Y., Tran, C., Li, X., Chen, P., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *CoRR*, abs/2008.00401.

Thompson, B. and Post, M. (2020a). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121.

Thompson, B. and Post, M. (2020b). Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wieting, J. and Gimpel, K. (2018). ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.