PheMT: A Phenomenon-wise Dataset for Machine Translation Robustness on User-Generated Contents

Ryo Fujii[1], **Masato Mita**[2,1], **Kaori Abe**[1], **Kazuaki Hanawa**[2,1], **Makoto Morishita**[3,1], **Jun Suzuki**[1,2], **Kentaro Inui**[1,2]
[1]Tohoku University, [2]RIKEN, [3]NTT Communication Science Laboratories
{r-fujii, abe-k, jun.suzuki}@ecei.tohoku.ac.jp,
{masato.mita, kazuaki.hanawa}@riken.jp,
makoto.morishita.gr@hco.ntt.co.jp,
inui@ecei.tohoku.ac.jp

Abstract

Neural Machine Translation (NMT) has shown drastic improvement in its quality when translating clean input, such as text from the news domain. However, existing studies suggest that NMT still struggles with certain kinds of input with considerable noise, such as User-Generated Contents (UGC) on the Internet. To make better use of NMT for cross-cultural communication, one of the most promising directions is to develop a model that correctly handles these expressions. Though its importance has been recognized, it is still not clear as to what creates the great gap in performance between the translation of clean input and that of UGC. To answer the question, we present a new dataset, **PheMT**, for evaluating the robustness of MT systems against specific linguistic phenomena in Japanese-English translation. Our experiments with the created dataset revealed that not only our in-house models but even widely used off-the-shelf systems are greatly disturbed by the presence of certain phenomena.

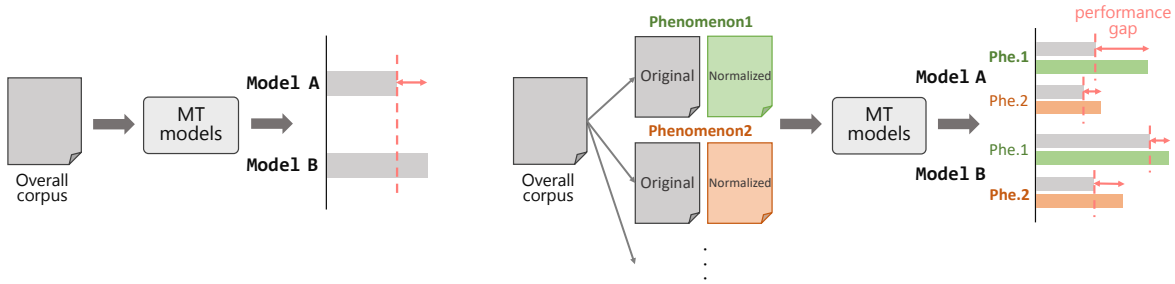
1 Introduction

The advancement of Neural Machine Translation (NMT) has brought great improvement in translation quality when translating clean input, such as text from the news domain (Luong et al., 2015; Vaswani et al., 2017), and it was recently claimed that NMT has even achieved human parity in certain language pairs (Hassan et al., 2018; Barrault et al., 2019). Despite its remarkable advancements, the applicability of NMT over User-Generated Contents (UGC), such as social media text, still remains limited (Michel and Neubig, 2018; Berard et al., 2019a). Since UGC are prevailing in our real-life communication, it is undoubtedly one of the challenges we need to overcome to make MT systems invaluable for promoting cross-cultural communication.

Recently, with the increasing interest in handling UGC, a shared task was organized to measure how well MT systems adapt to those texts (Li et al., 2019). However, the way in which they evaluate systems is just giving an overall score to a dataset, which is the same as traditional MT evaluation (Figure 1a). The overall score does not provide precise information for understanding what leads to the huge performance gap between the translation of clean input and that of UGC. To find a clue for improving the performance of MT systems on UGC, we need a solid basis for more detailed error analysis.

As a first step towards a more refined evaluation of MT systems on UGC, we present a new dataset, **PheMT**: **Phe**nomenon-wise Dataset for **M**achine **T**ranslation Robustness, designed for phenomenon-wise evaluation in Japanese-English translation (Figure 1b). More specifically, we create a set of datasets, each of which provides a focused evaluation on one of four linguistic phenomena commonly seen on UGC, i.e., *Proper Noun*, *Abbreviated Noun*, *Colloquial Expression* and *Variant*. By focusing locally on a specific part of a sentence presenting one of the above phenomena, we directly measure the ability of MT systems to handle the phenomenon with the help of translation accuracy. Moreover, based on the idea of contrastive datasets, we normalize targeted expressions to its canonical form in the dictionary. We feed both original and normalized versions of a source sentence to obtain the difference of arbitrary metrics as our robustness measure. Using our dataset, we analyze the strengths and weaknesses of current NMT

Proceedings of the 28th International Conference on Computational Linguistics, pages 5929–5943
Barcelona, Spain (Online), December 8-13, 2020



(a) Traditional MT evaluation

(b) Phenomenon-wise evaluation (Ours)

Figure 1: Overview of traditional MT evaluation and our proposal: phenomenon-wise evaluation.

systems from the point of available training data size and the way of tokenization. We reveal that some of the phenomena are severely problematic even to widely used, strong off-the-shelf systems.

We made our dataset publicly available for further development in MT systems. We hope our dataset will provide promising directions to future MT systems and move the community one step forward with an additional axis for evaluation.

The contributions of this paper are:

1. We proposed a novel dataset designed for phenomenon-wise evaluation in Japanese-English translation as a protocol for detailed error analysis.
2. We revealed with our dataset that some of the phenomena commonly seen on UGC greatly degrade the performance of current NMT systems, including widely used off-the-shelf systems.

2 Related Work

Michel and Neubig (2018) created the MTNT dataset with the increasing interest in creating noise-robust MT systems. They collected comments from the social discussion website, Reddit[1], and added translations by professional translators. They also provided statistics of the dataset, showing that the source side of the dataset is much noisier than previous benchmarks for MT systems. Their results with the baseline systems demonstrated the difficulty of properly translating UGC. The dataset was also used as in-domain data for the first shared task on machine translation robustness held at WMT 2019.[2]

However, the dataset is still miscellaneous in the degree of politeness, domain of the conversations, and even in the quality of translations. Though it is still a question whether we actually need to develop any UGC-specific techniques or not, we do not even know with such a many-sided dataset that how much the improvement in some metrics, such as BLEU score (Papineni et al., 2002), actually contributes to improve robustness on various noise. In fact, Berald et al. (2019b), the winning team in the shared task, reported that none of the techniques specifically designed for UGC was more effective in improving BLEU score than corpus filtering. Though there is no doubt that corpus filtering is one of the essential techniques for data-driven MT systems (Koehn et al., 2018; Junczys-Dowmunt, 2018), this is rather aimed at removing inappropriate *sentence pairs* generated during the process of creating corpora, not at handling noisy input. The way of current evaluation definitely prevents us from developing truly robust systems, and motivated us to create a new dataset for focused evaluation.

A range of studies have aimed to elucidate the cause of mistranslations from the viewpoint of linguistic phenomena, such as typographical errors (Heigold et al., 2018; Belinkov and Bisk, 2018; Karpukhin et al., 2019; Niu et al., 2020), grammaticality (Sennrich, 2017), presence of named entities (Ugawa et al., 2018), and identification of anaphoric pronouns (Bawden et al., 2018; Müller et al., 2018).

One of the pioneering works to analyze the behavior of NMT is the challenge set approach proposed by Isabelle et al. (2017). They defined various subcategories of structural differences between English and French to evaluate how well models can handle them in detail. Though the approach has the potential of

[1] https://www.reddit.com

[2] http://www.statmt.org/wmt19/robustness.html

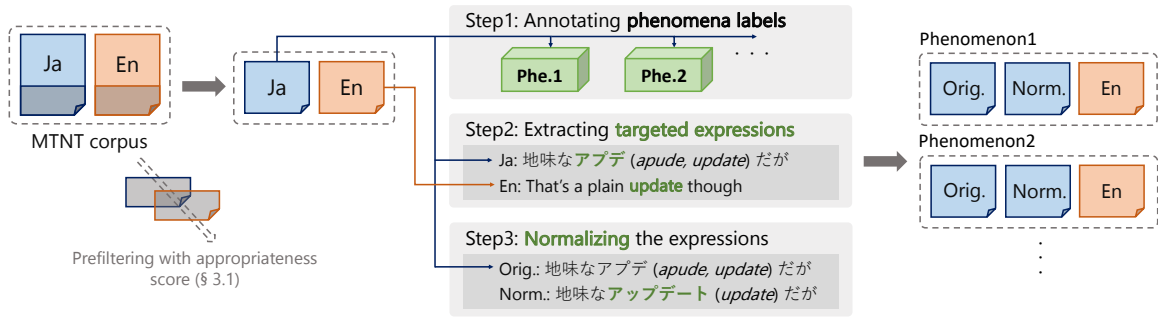


Figure 2: Entire flow of our phenomenon-wise dataset creation.

accelerating our understanding of MT systems, there lies a problem that their way of evaluation requires human evaluators with highly advanced knowledge on linguistics.

In response to the problem, Sennrich (2017) proposed the contrastive dataset approach to automatically evaluate the grammaticality of a model in a comparative manner. The author added an error-introduced contrastive version of reference to each source sentence by minimally modifying gold reference translations. They defined the accuracy of a model as the number of times the model assigned a higher conditional probability to the original reference. The approach was later followed by Bawden et al. (2018) to evaluate models' ability to exploit preceding contexts. However, as the authors pointed out, the evaluation does not guarantee that the most probable translation by the model is free from errors even if the model ranked two references correctly.

Similar but different way of contrastive evaluation is performed on a clean input and its noisy counterpart. Heigold et al. (2018) introduced rule-based character replacement noise to imitate misspellings found in a variety of real-world applications. Following work by Karpukhin et al. (2019) and Belinkov and Bisk (2018) extended its scope to natural noise by using edit histories from online websites. However, instead of giving translations to raw noisy sentences, they relied on a noisy version of input artificially created from the clean text. Anastasopoulos et al. (2019) is similar to our work in that they explored the effect of errors naturally created by humans. They focused on the effect of grammatical errors against NMT by adding translations to the JFLEG corpus (Napoles et al., 2017), one of the common benchmarks for grammatical error correction. Their results demonstrated that even a very small perturbation could significantly drop the performance of MT systems while exposure to similar noise during training time alleviates the problem.

However, these aspects are only a small subset of possible reasons to explain why current models are still not good at handling UGC. To the best of our knowledge, there is no previous work aimed at investigating the effect of UGC-specific challenges in a fine-grained manner. Also, behavioral analysis of NMT in dissimilar language pairs such as Japanese-English has not been studied extensively. We expect a brand-new solution in this challenging language pair to be developed in the future with our dataset.

3 Creating Phenomenon-wise Dataset

3.1 Data selection for quality assurance

The entire flow of our dataset creation is described in Figure 2. As the methodology to create brand-new, high-quality parallel data for UGC is not trivial, we started with the existing dataset for machine translation robustness, the MTNT dataset (Michel and Neubig, 2018). The number of sentences originally created for evaluation was not enough to be further classified into several categories, so we have decided to utilize the train and development data as well to scale out our dataset. However, such data might not be of sufficient quality to be adopted as evaluation data. To confirm how much low-quality data it actually contains, we manually assessed the *appropriateness* of source-target sentence pairs in the dataset as a preliminary experiment.[3] Figure 3 shows the distribution of annotated scores for the MTNT dataset. We filtered out sentences by the threshold of 4.0 to assure the quality of our phenomenon-wise dataset.

[3]See Appendix A for detailed experimental settings.

Annotation label	Examples
Proper Noun	安倍首相 (*abeshushō*, Prime Minister Abe), 平昌 (*Pyongchang*)
Abbreviated Noun	アプデ (*apude*, update), WHO (World Health Organization)
Colloquial Expression	ねむーーい (*nemūi*, sleepy), かなちい (*kanachii*, sad)
Variant	アリガトウ (*arigatou*, thank you), いいよ (*iiyo*, no problem)

Table 1: List of annotation labels and examples for each phenomenon.

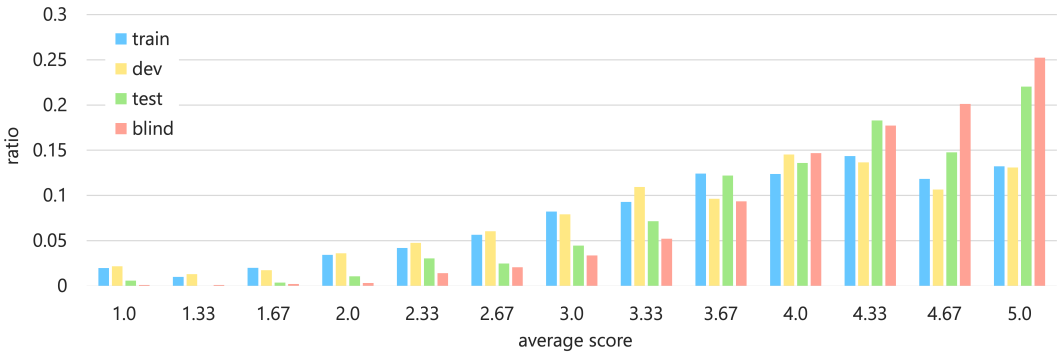


Figure 3: Distribution of appropriateness scores for the MTNT dataset. Human evaluators answered the question on the basis of a 5-point scale: 1 (very poor) – 5 (excellent).

3.2 Annotation of linguistic phenomena

(i) Definition of phenomena labels To define the labels, we first investigated what kind of UGC-specific phenomena cause significant errors in other NLP applications. Sasano et al. (2013) and Saito et al. (2014) focused on the presence of unnormalized orthographic variations in Japanese morphological analysis. They introduced some handcrafted derivation rules, such as inserting prolonged sounds and substitutions to lowercased characters, to simulate alternate forms typically seen on the Internet. Ikeda et al. (2016) also applied similar rules to create synthetic data for text normalization task and demonstrated its effectiveness in improving the robustness of neural-based models. However, the impact of those expressions has yet to be explored in a variety of cross-lingual tasks including machine translation. Thus in this paper, we defined two types of linguistic phenomena, namely *Colloquial Expression* and *Variant*, by following their derivation rules.

Additionally, we defined *Proper Noun* and *Abbreviated Noun*, two phenomena commonly seen across various domains including UGC. To estimate how many of the sentences in UGC actually contains these phenomena, we randomly selected 500 sentences from the training data of the MTNT dataset and annotated them in our preliminary experiment. The result showed that more than 40% of the sentences included one or more proper nouns, and more than 10% of the sentences had abbreviated nouns. Also, the effect of named entities over machine translation is receiving more and more attention in the context of transliteration (Shao and Nivre, 2016; Rosca and Breuel, 2016).

To summarize, we targeted four phenomena as described below in our phenomenon-wise dataset (see Table 1 for examples) ;

- *Proper Noun* ; the name of a person, company, country and others, something that is unique.
- *Abbreviated Noun* ; nouns made by abbreviating its canonical form, including acronyms.
- *Colloquial Expression* ; words deviated from its canonical form by inserting/dropping/replacing vowels, consonants, prolonged sounds (“ー”), or geminate consonants (“っ”).
- *Variant* ; words deviated from its canonical form by lowercasing characters, or by using unusual *hiragana*, *katakana* notation.

1 :	*Abbreviated Noun*
Orig. (Ja)	地味なアプデ(*apude*, update)だが
Norm. (Ja)	地味なアップデート(*update*)だが
Ref. (En)	That's a plain **update** though

2 :	*Colloquial Expression*
Orig. (Ja)	ここまで描いて飽きた、かなちい (*kanachii*, sad)
Norm. (Ja)	ここまで描いて飽きた、かなしい (*kanashii*)
Ref. (En)	Drawing this much then getting bored, how **sad**.

Table 2: Examples of original sentence (Orig.), normalized sentence (Norm.), and reference sentence (Ref.) in our dataset.

Dataset	# sent.	# unique expressions (ratio)	average edit distance
Proper Noun	943	747 (79.2%)	-
Abbreviated Noun	348	234 (67.2%)	5.04
Colloquial Expression	172	153 (89.0%)	1.77
Variant	103	97 (94.2%)	3.42

Table 3: Basic statistics of our phenomenon-wise dataset.

(ii) Extraction and normalization of targeted expressions We used crowdsourcing to add annotations to the MTNT dataset. Considering the difficulty and inter-annotator agreement of the task, we divided the whole process into three subtasks: (i) annotating phenomena labels, (ii) extracting targeted expressions, and (iii) normalizing the expressions. To ensure the quality, we assigned five workers per sentence for all tasks and retained the result only if a majority of workers answered the same.

Firstly, we asked crowdworkers to classify the source (Japanese) sentences with the above definitions (Figure 2 Step1). A question consists of four yes-no questions, each of which corresponds to one of the four phenomena. We asked if there exist one or more expressions presenting each phenomenon for each sentence in the dataset.[4]

Then, we associated the labels with corresponding expressions in a sentence. More specifically, we designed a task to extract up to five expressions from a source sentence for each (source sentence, label) pair. Also, we asked crowdworkers to extract translation of the targeted expressions, i.e., the alignment from the target language (Figure 2 Step2). To avoid some sentences from being overrated, we discarded sentences having more than one expression with the same label.

Finally, to create contrastive input from raw noisy sentences, we asked crowdworkers to normalize the extracted expressions in the source language (Figure 2 Step3). The process of normalization stands for rewriting an expression to its canonical form in the dictionary, namely applying an inverse transformation to remove the reason for the classification. For instance, the workers are to normalize an expression アプデ (*apude*, an example of *Abbreviated Noun* in Table 1) to アップデート (*update*) by resolving abbreviation. Another example from *Colloquial Expression* is to normalize ねむーーい (*nemūi*, sleeeepy) to ねむい (*nemui*, sleepy) by removing unnecessary prolonged sound. Here, the word is more commonly written in *kanji* characters as 眠い (*nemui*, sleepy) than in *hiragana* characters (as in the example), however, the workers are instructed not to normalize the word in two stages because it is outside the scope of *Colloquial Expression*. On the other hand, if the given expression, ねむい (*nemui*, sleepy) was originally in the text, it is counted as a *Variant* and will be normalized to its *kanji* notation. We skipped this step for *Proper Noun* because there is no concept of *canonical form* for proper nouns.

We created our phenomenon-wise dataset in the form of quadruple consists of (original source sentence, normalized source sentence, alignment, target sentence) by replacing the extracted expressions with their normalized counterparts. Table 2 shows some examples from our final dataset. Also, we provide basic statistics of the dataset in Table 3.

4 Translation Models

We prepared five in-house models with different size of training data and preprocessing methods for our experiments. The smaller model (SMALL) was trained on the data offered in the WMT 2019 shared

[4]Note that a sentence could be given more than one label. These sentences are treated differently according to the label which we focus on.

task for machine translation robustness, namely TED talks, KFTT (Kyoto Free Translation Task), and JESC (Japanese-English Subtitle Corpus). The MTNT dataset was also available in the task, but we didn't include any of the sentence pairs to train our models. We replaced *emojis* and emoticons with placeholders following a previous study by Murakami et al. (2019). In addition, we replaced possible usernames with regular expressions. We offered this model to see whether or not the phenomena would become less problematic with increasing amount of training data.

For the other four models, we additionally used JParacrawl v2.0 (Morishita et al., 2020), one of the largest parallel corpora available in Japanese-English. The larger model (LARGE), is only different in the size of training data from the SMALL. We applied Byte-Pair-Encoding (BPE) models (Sennrich et al., 2016) with a joint vocabulary of 32,000 for these models using the `sentencepiece` toolkit (Kudo and Richardson, 2018).

The character-based model (CHAR) is different from the two models in the way of segmentation. The model translates a sequence of characters in the source language into another sequence of characters in the target language (Wang et al., 2015). Durrani et al. (2019) pointed out that character-based models are more robust to noisy text than BPE-based models. We revisit the issue of segmentation to see if the model is also good at handling UGC. We shared the vocabulary between two languages in this setting as well to expect the model to capture copying behavior.

For the pronunciation-based model (PRON), we applied a unique preprocessing method to the source (Japanese) sentence. More specifically, we first applied the `MeCab` toolkit (Kudo et al., 2004), a Japanese morphological analyzer, with naist-jdic for the dictionary to obtain the pronunciation of each morpheme in the sentences. We can transliterate any words in Japanese by using phonetic symbols such as *hiragana* and *katakana* characters. Since the `MeCab` toolkit outputs the pronunciation in *katakana* characters by default, we simply concatenated them to create a fully pronunciation-based corpus. We prepared this model with the expectation to improve the robustness against *Variant* expressions. More specifically, we aimed at absorbing the orthographic variations caused by *hiragana-katakana* confusion, which is a part of *Variant*. Also, previous study suggests that phonetic information is highly useful to resolve homophone noise (Liu et al., 2019).

Finally, we prepared the concatenated model (CAT), trained on a joined corpus for the LARGE and the PRON.[5] In this setting, we converted the transliterated part into *hiragana* characters and applied the same BPE model as used in the LARGE to the whole corpus. We expect the model to learn robust representations by forcing it to produce the same translation from the original source sentence and its transliterated counterpart. We used transformer-base architecture (Vaswani et al., 2017) implemented in the `fairseq` toolkit (Ott et al., 2019) and hyperparameters proposed by Murakami et al. (2019) for all models. The size of the training data was 3.9 M for the SMALL, 14.0 M for the LARGE, CHAR and PRON, and 28.0 M for the CAT.

In addition to the in-house models, we also investigated the impact of the phenomena on two widely used MT systems, namely, Google Translate[6] and DeepL Translator. [7,8] These systems are expected to be more robust against UGC because they are by nature exposed to user input. By conducting experiments on such systems, we reveal the presence of phenomena with impending needs for improvement, and also confirm the usefulness of normalization as one of the tricks users can do.

5 Phenomenon-wise Evaluation

We provide an overview of the current state of NMT by evaluating the performance of both in-house models and off-the-shelf systems on the proposed phenomenon-wise dataset. We fed both the original and normalized sentences to the models and measured the difference of single reference BLEU between them. Since the only difference between the two sentences is the presence of the corresponding phenomenon, our dataset ensures that a phenomenon degrades the models more significantly if there is a

[5]We also tried combining two sentences with delimiter tokens `<sep>` like the *paste* command in the Unix-like operating systems, but we could not see any meaningful results from the model.

[6]`https://translate.google.co.jp`

[7]`https://www.deepl.com/translator`

[8]The results are as of June 10, 2020.

	SMALL		LARGE		CHAR		PRON		CAT	
	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.
Abbreviated Noun	10.4 / 10.8	+0.4	14.5 / 14.4	-0.1	11.8 / 12.0	+0.2	10.2 / 10.9	+0.7	13.8 / 13.6	-0.2
Colloquial Expression	11.9 / 12.7	+0.8	13.8 / 14.9	**+1.1**	12.3 / 11.7	-0.6	10.4 / 11.5	**+1.1**	13.9 / 14.7	+0.8
Variant	10.4 / 10.9	+0.5	13.7 / 15.3	**+1.6**	13.2 / 16.0	**+2.8**	11.1 / 11.9	+0.8	13.3 / 15.7	**+2.4**

Table 4: BLEU scores measured with our dataset (in-house models). * Orig. for original, Norm. for normalized sentences.

	SMALL		LARGE		CHAR		PRON		CAT	
	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.
Proper Noun	34.3 / -	-	49.7 / -	-	47.1 / -	-	43.2 / -	-	48.0 / -	-
Abbreviated Noun	24.1 / 30.5	+6.4	33.6 / 33.0	-0.6	34.2 / 34.8	+0.6	30.2 / 31.3	+1.1	34.2 / 33.0	-1.2
Colloquial Expression	18.0 / 23.8	+5.8	14.5 / 24.4	+9.9	17.4 / 21.5	+4.1	8.7 / 30.2	**+21.5**	15.7 / 32.6	**+16.9**
Variant	15.5 / 35.0	**+19.5**	13.6 / 38.8	**+25.2**	13.6 / 34.0	**+20.4**	**25.2** / 35.9	**+10.7**	**26.2** / 35.0	+8.8

Table 5: Accuracy (%) measured with our dataset (in-house models). * Orig. for original, Norm. for normalized sentences.

larger gain of BLEU score after replacement. We also calculated the ratio of correctly translated expressions, i.e., the accuracy, before and after normalization. While the BLEU-based method enables us to measure the relative change in fluency from the sentence level, the accuracy-based method is rather aimed at evaluating the models locally. We used these two measures supplementarily to investigate more closely what becomes an obstacle to current MT systems.

5.1 Quantitative analysis

In-house models Table 4 and 5 show the BLEU scores and the accuracy, respectively, for the in-house models. The results showed that the scores were constantly improved after normalization for the SMALL, which indicates that all of the targeted phenomena may adversely affect the model to some extent. However, there seems to be a clear difference in the trend between *Proper Noun*, *Abbreviated Noun* and the other two UGC-specific phenomena. First, the accuracy with original sentences for the *Proper Noun* and *Abbreviated Noun* increased with the size of training data, while we observed a slight drop for the other two. Also, the gain from normalization for the *Abbreviated Noun* was exceptionally high in the SMALL. It is also notable that the difference scores for the *Colloquial Expression* and *Variant* were even larger in the LARGE than in the SMALL. These figures support that we need special treatment beyond collecting massive training data to further improve MT systems on UGC.

From the point of tokenization, the CHAR could not outperform the LARGE in all phenomena with the BLEU scores. However, we could see an improvement of 5.8 points in the difference of accuracy for the *Colloquial Expression*, showing its high robustness against the phenomenon. We speculate that this might result from the small edit distance in the *Colloquial Expression* dataset. Similar to typographical errors in alphabetical languages, character-based models seem to prove their true worth with phenomena for which surrounding characters become an important clue. On the other hand, it was surprising that *Variant*, which is an instance of orthographic variations, was not treated well by the model. This might result from the fact that *Variant* is rather a word-level phenomenon unlike typographical errors, which are in most cases limited within several characters.

The PRON also performed poorly with the BLEU scores. However, it is notable that the model showed the smallest difference score for the *Variant* among four models trained on the larger data. Here, we refer to Table 5 for the translation accuracy of the models. While the accuracy for the *Variant* after normalization stayed almost the same for all five models, the accuracy for the original sentences attained by the PRON went more than 10 points higher than that by the LARGE and the CHAR. The results indicate that the decrease in difference is not brought by the limited expressiveness of phonetic symbols but by the increasing capacity to handle non-standard input. We might discard the model for its low BLEU scores without our dataset, but our phenomenon-wise dataset provides a new axis to the evaluation, discovering the possibility of the model.

The CAT seems to be a better alternative to the PRON. The model showed a relatively small drop in the BLEU scores from the LARGE (Table 4), and also benefited from the robust representations of the pronunciation-based corpus. The model reached 26.2% accuracy for the *Variant*, which is significantly

	BLEU				Accuracy (%)			
	Google Translate		DeepL Translator		Google Translate		DeepL Translator	
	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.	Orig. / Norm.	Diff.
Proper Noun	- / -	-	- / -	-	55.2 / -	-	50.5 / -	-
Abbreviated Noun	14.6 / 15.0	+0.4	16.3 / 16.2	-0.1	41.1 / 36.8	-4.3	39.1 / 37.9	-1.2
Colloquial Expression	14.4 / 16.0	**+1.6**	15.6 / 15.8	+0.2	19.2 / 26.2	+7.0	22.7 / 28.5	+5.8
Variant	15.3 / 17.6	**+2.3**	14.4 / 15.2	+0.8	23.3 / 37.9	**+14.6**	18.4 / 35.0	**+16.6**

Table 6: BLEU scores and accuracy (%) measured with our dataset (off-the-shelf systems). * Orig. for original, Norm. for normalized sentences.

higher than 13.6% by the LARGE (Table 5). Also, the accuracy for the *Colloquial Expression* showed 8.2 points gain after normalization as compared to the LARGE. This implies that the model could be more adaptive to the phenomenon with proper preprocessing. We speculate that one reason for the improvement comes from the increasing capacity of the CAT to treat unexpected segmentation caused by *hiragana* characters. In Japanese, most of the highly-frequent function words consist of a few *hiragana* characters. Sasano et al. (2013) pointed out that expressions in *hiragana* characters are more likely to combine each other to produce these function words than kept as single words. The idea of mixing a pronunciation-based corpus forces a model to produce correct output from unexpectedly segmented, difficult sequences. The results suggest the importance of deep consideration for possible perturbations from the viewpoint of linguistic phenomena to better handle UGC.

Overall, *Proper Noun* was handled relatively well by all in-house models as compared with the other three phenomena. The results also showed that BPE-based models (LARGE and PRON) performed slightly better with proper nouns than character-based models. On the other hand, the scores for the *Abbreviated Noun* were rather inconsistent: the differences even went into minus in some models. However, the result does not necessarily mean that the phenomenon is less important to cope with. To investigate the effect of *Abbreviated Noun* more deeply, we conducted an additional experiment to subdivide the dataset into several groups.[9] The result showed that there were roughly two types of expressions for the phenomenon, namely the alphabetical acronyms and the others, and the behavior of the models was completely different from each other. The process of normalization unnecessarily led a model to explain the acronyms redundantly to induce a drop in accuracy. It might be better to exclude these expressions from our *Abbreviated Noun* dataset for more precise evaluation.

Off-the-shelf systems It is worth surprising that even the off-the-shelf systems performed poorly with our *Variant* dataset (Table 6). The systems dropped more than 10 points in accuracy when faced with the original sentences, and showed large differences in the BLEU scores as well. Also, the result is quite suggestive in that a system better at BLEU scores is not always better at handling specific phenomena. For instance, the accuracy for the *Abbreviated Noun* dataset with DeepL Translator was 2 points lower than Google Translate, but the system showed 1.7 points higher BLEU score. We speculate that this might have been caused by the different behaviors of the two systems. In our experiments, DeepL Translator tended to ignore uncommon phrases to keep the overall translation fluent, but Google Translate endeavored to provide some output even if phrases were confusing to the model. Practically, the preference over high-precision systems or high-recall systems depends on the application for which the translation is used. The two-way evaluation, from the BLEU scores and the accuracy, could be of great help for us to make a decision about what models to deploy.

5.2 Qualitative analysis

We also analyzed qualitatively how translations generated by the models were changed after normalization. Table 7 shows some examples of the output from our in-house models.

Example (a) is a case where the output was improved by replacing the *hiragana* expression ぎゃくたい (*gyakutai*, abuse) with its common notation in *kanji*, 虐待 (*gyakutai*). In this case, the LARGE mistakenly output the phrase *want to* when we fed the original source sentence. This might have resulted

[9] See Appendix B for the detail of the experiment.

(a) *Variant*	
Source	{ぎゃくたい / 虐待 (*gyakutai*, abuse)} だ
LARGE$_{orig}$	I want to do it!
CAT$_{orig}$	It's **abuse**!
LARGE$_{norm}$	It's **abuse**!
Reference	It's abuse!
(b) *Abbreviated Noun*	
Source	進化する {サバゲー (*sabagē*, survival game) / サバイバルゲーム (*survival game*)}
LARGE$_{orig}$	The evolving mackerel game.
CAT$_{orig}$	Evolving sabage.
LARGE$_{norm}$	Evolving **Survival Game**
Reference	An evolving survival game.
(c) *Proper Noun*	
Source	平昌 (*Pyongchang*) で「米日VS南北」の戦いが始まる
SMALL	In the Heisho era, the battle of 'South and South America' began.
LARGE	The Battle of 'America-Japan VS North-South' begins in **Pyeongchang**
Reference	The "US and Japan vs. North and South Korea" battle has begun in Pyeongchang.

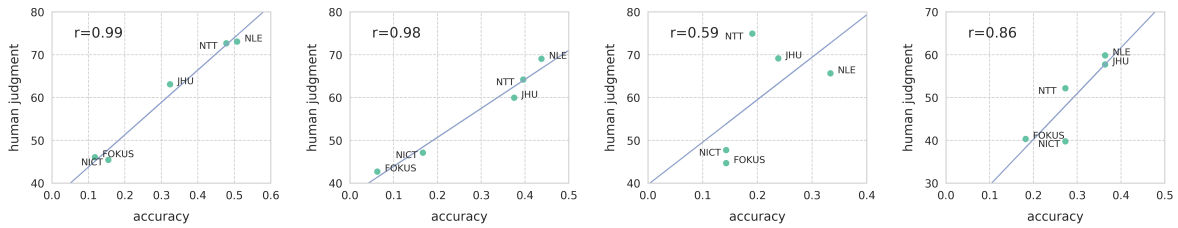
Table 7: Output examples from our in-house models. *{original expression / normalized expression}

from the fact that the original expression was overly segmented into four parts with our BPE model. Here, the presence of a segmented prefix たい (*tai*), a highly-frequent auxiliary verb often combined with other verbs to show one's desire, possibly worked badly to produce the wrong output. On the other hand, though the input was the same, the CAT could produce a correct translation, *abuse* for the original expression. In most case, the character preceding the auxiliary verb たい (*tai*) generates *i* or *e* sound, such as したい (*shitai*, want to do) and 食べたい (*tabetai*, want to eat). The pronunciation-based corpus might have provided enough false examples to learn this rule, resulted in the improvement.

Though *Variant* is one of the phenomena specific to languages with various writing systems, similar problems are actually observed in other languages as well. For example, the negative effects caused by some types of typographical errors can be explained in the same way as the example above. It is a challenge how we obtain correct translation in case that an erroneous expression is partially associated with other words.

The next example (b) is from our *Abbreviated Noun* dataset. In this example, the LARGE could not produce the correct translation for the original expression サバゲー (*sabagē*, survival game), and mistakenly treated the word as a combination of the two words, サバ (*saba*, mackerel) and ゲー (*gē*, game). The CAT also suffered from translating the expression, but it instead transliterated the word into the alphabet. The result implies that the model captures character-level cooccurrence inside a word better than naive models: mackerels usually do not appear in a game. Also, we found an interesting example where an abbreviated word could be interpreted differently according to the context (生保, *seiho*, life insurance or life security). It is important to capture the context not only inside but outside a word to further improve the models.

Finally, in the example (c), the expression 平昌 (*Pyongchang*) was correctly handled by the LARGE, while the SMALL could not. Though it is not unnatural to conclude that the increasing capacity of treating *Proper Noun* resulted from the large corpus on which the model was trained, we believe it is not a sufficient condition to explain the consequence. An observation behind is that the term *Pyongchang* became popular after the Olympics was held there in 2018. The corpora we used for training the SMALL were no newer than 2018, and that possibly resulted in fewer occurrences of the term. To create truly robust systems against *Proper Noun*, we believe it is necessary to divide corpora chronologically to measure the generalization ability against nouns that appear only in the test data. However, we believe our dataset could be of some help to evaluate models' performance against the phenomenon, considering the fact that it is quite unrealistic to keep a test data always newer than any training data.



(a) *Proper Noun* (b) *Abbreviated Noun* (c) *Colloquial Expression* (d) *Variant*

Figure 4: Correlation between the accuracy and human judgment scores for each phenomenon (WMT submitted systems). The r-value is Pearson's correlation coefficient.

6 Correlation with Human Evaluation

To demonstrate a potential use case of our phenomenon-wise dataset, we conducted an additional experiment, where we reassessed the systems submitted to the WMT 2019 robustness shared task in a phenomenon-wise manner. We downloaded five official submissions for the blind test[10] portion of the MTNT dataset.[11] We then extracted the intersection between the blind test data and our phenomenon-wise dataset, obtaining 136 sentences for *Proper Noun*, 48 for *Abbreviated Noun*, 21 for *Colloquial Expression*, and 11 for *Variant*. The task organizer also provides the results of human judgment of all submissions for each sentence (Li et al., 2019), where three human raters were instructed to rate each translation on a scale from 1 (completely incorrect) to 100 (accurate). For each of the five submissions, we averaged all the human ratings for each sentence in the phenomenon subset and investigated the correlation between the averaged human ratings and the phenomenon-wise accuracy.

From the results in Figure 4, we could see that the accuracy for our *Proper Noun* and *Abbreviated Noun* dataset strongly correlated to the human judgment scores with $r > 0.9$. This is worth surprising because we have no access to the whole sentences but only to the targeted expressions in our accuracy-based method. The result suggests that the two phenomena are key factors for humans in evaluating overall translation quality. One reason might be that humans can easily tell whether the translated sentences include these nouns or not. This implies that undertranslation of words for these two phenomena could bring a more serious impact on human judgment. We believe that the accuracy can be used as a strong signal for estimating human judgment scores, when combined with traditional evaluation metrics such as BLEU.

7 Conclusion

We proposed a novel dataset designed for phenomenon-wise evaluation in Japanese-English translation. In this research, we focused on four linguistic phenomena commonly seen on User-Generated Contents, namely *Proper Noun*, *Abbreviated Noun*, *Colloquial Expression*, and *Variant*.

Using our dataset, we analyzed how current MT systems are negatively affected by the presence of the phenomena. The result showed that *Variant* is one of the phenomena that significantly degrade the model's performance including widely used, strong off-the-shelf systems. This implies that collecting massive training corpora is not a sufficient condition to handle these peculiar inputs, and we need special treatment against them to further improve MT systems.

We also analyzed the correlation between human judgment and translation accuracy scores from our dataset by using official submissions from the WMT 2019 shared task. From the experiments, we confirmed that the accuracy-based scores from our dataset strongly correlated with human judgment, showing its potential to reduce the cost of the evaluation.

We made our dataset publicly available for further development in MT systems[12]. As future work, we would like to consider new model architectures or data preprocessing methods to improve performance against specific phenomena using our dataset.

[10]The data used for ranking systems in the shared task. It was kept blind to participants until the evaluation period ends.
[11]http://matrix.statmt.org/matrix/systems_list/1917
[12]https://github.com/cl-tohoku/PheMT

References

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. 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 3070–3080.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in 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 1304–1313.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018*.

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019a. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019b. Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. 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 1504–1516.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv*, abs/1803.05567.

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 68–80.

Taishi Ikeda, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Japanese text normalization with encoder-decoder model. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 129–137.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739.

Taku Kudo and John Richardson. 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.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 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.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Paul Michel and Graham Neubig. 2018. MTNT: A Testbed for Machine Translation of Noisy Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3603–3609.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72.

Soichiro Murakami, Makoto Morishita, Tsutomu Hirao, and Masaaki Nagata. 2019. NTT's machine translation systems for WMT19 robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 544–551.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 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.

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.

Mihaela Rosca and Thomas Breuel. 2016. Sequence-to-sequence neural network models for transliteration. *arXiv*, abs/1610.09565.

Itsumi Saito, Kugatsu Sadamitsu, Hisako Asano, and Yoshihiro Matsuo. 2014. Morphological analysis for Japanese noisy text based on character-level and word-level normalization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1773–1782.

Ryohei Sasano, Sadao Kurohashi, and Manabu Okumura. 2013. A simple approach to unknown word processing in Japanese morphological analysis. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 162–170.

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.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382.

Yan Shao and Joakim Nivre. 2016. Applying neural networks to English-Chinese named entity transliteration. In *Proceedings of the Sixth Named Entity Workshop*, pages 73–77.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Ling Wang, Trancoso Isabel, Dyer Chris, and Black Alan W. 2015. Character-based neural machine translation. *arXiv*, abs/1511.04586.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 193–205.

Score	Definition
5	translations that conveys the meaning completely and fluent as target language sentence
4	translations that does not show any lack of information, but highly Translationese (verbatim)
3	translations that has locally untranslated / mistranslated parts, but acceptable
2	translations that has phrase, sentence-level mistranslation, or based on different interpretation
1	translations that is complete nonsense

Table 8: Criterion for appropriateness score annotation

A Preliminary experiment of appropriateness score annotation

To ensure the quality of the resulting dataset, we first applied some basic rule-based filtering to the corpus. More specifically, we removed (i) sentences including inappropriate words using a predefined word list[13], (ii) pairs having identical source and target sentences, (iii) duplicates, and (iv) sentences consisting of 1 word, or more than 80 words. Then, we designed a task to annotate the *appropriateness* of the translation for each sentence. The task was aimed to classify the source and target sentence pairs on the Likert scale having scores ranging from 1 (very poor) to 5 (excellent). To define the criterion, we followed the common practice of assessing machine-translated output from two perspectives: adequacy and fluency (White et al., 1994). However, we added some modifications because the translations to be evaluated were human-generated. The criterion for each grade is given in Table 8.

Since it requires a highly advanced understanding of the source language (Japanese) to correctly capture the meaning of sentences in UGC, we asked ten native speakers of Japanese with high English proficiency to annotate scores in this task.[14] We allocated three different workers per sentence, and averaged these scores to obtain the final score. We filtered out sentences by the threshold of 4.0 and retained only one reference with the highest appropriateness score per source sentence to prevent negative effects caused by single reference BLEU: high precision for one reference may lower the precision for other references.

Figure 3 shows the distribution of annotated appropriateness scores for each portion of the MTNT dataset. There were 4152 sentences with an average score of 4.0 or more out of the 7273 annotated sentences. The number of sentences discarded was large enough to support the necessity of pre-filtering by translation quality to assure the quality of our phenomenon-wise dataset. The result also showed that the train and development portion of the dataset (blue and yellow bars in the figure) included more sentences in lower quality compared to the test and blind portion (green and red bars). The difference was particularly clear in the range lower than the average score of 3.0 and higher than 4.0. The number of sentences we kept for phenomena annotation was 3896, after retaining only one reference with the highest appropriateness score per source sentence.

B Subdivision of the Abbreviated Noun Dataset

The results from Table 5 and 6 showed that *Abbreviated Noun*, unlike other phenomena, did not affect the models in a negative way. To further investigate the effect of the phenomenon, we additionally subdivided the dataset into six groups. Table 9 shows the criterion for each group and the difference in accuracy before and after normalization. In the first three groups, wherein the original expressions were written in alphabetical acronyms, there was a severe drop of up to over 60% accuracy with the LARGE after normalizing the expressions. One reason to explain the result is that those acronyms are usually kept intact in the reference as they tend to be originally imported from the target language (English) to the source language (Japanese). The process of normalization led models to unnecessarily explain terms redundantly, resulted in a drop in the accuracy, which is based on the exact match. However, the output with normalized, expanded expression is not always a wrong translation. For instance, we could see from the result that an expression *DM* was translated as *direct mail* after normalization. It might be better to

[13] https://github.com/1never/open2ch-dialogue-corpus. The list was created for dialog corpus filtering.

[14] We set the standard reward for each worker to 20,000 yen, approx. 185 dollars with the exchange rate as of June 2020. We selected workers who had rich experience in translation or had equivalent skills, from more than 80 applicants.

Group	Orig.	Norm.	Example	# sents.	ΔAcc. (SMALL)	ΔAcc. (LARGE)
1	alphabetical	alphabetical	AI / artificial intelligence	9	-22.2	-88.9
2		*katakana*	PC / パーソナルコンピューター (*personal computer*)	41	-26.9	-61.0
3		others	EU / 欧州連合 (*oushūrengou*, Europe Union)	23	-52.2	-60.9
4	mixed	first / last n characters	サンタ (*Santa*) / サンタクロース (*Santa Claus*)	104	+13.4	+21.2
5		combination of two first-n characters	アニオタ (*aniota*, Anime nerds) / アニメオタク (*animeotaku*)	132	+18.2	+14.3
6		others	マック (*makku*, McDonald's) / マクドナルド (*makudonarudo*)	39	+23.1	+10.2
	overall			348	+6.4	-0.6

Table 9: Criterion and results of subdivided Abbreviated Noun dataset

exclude these sentences from our *Abbreviated Noun* dataset for precise evaluation.

On the other hand, expressions classified into the latter half of the groups seem to harm models significantly as normalization brought great improvement in the translation accuracy. As we discussed in Section 5.2, this might result from increasing ambiguity caused by abbreviation. We observed many expressions classified in these groups written in *katakana* characters. Among the four main types of characters used in Japanese, *hiragana* and *katakana* are less informative because of their characteristics as phonetic symbols. The presence of abbreviations limits the number of accessible characters, and we believe it eventually imposes a deeper understanding of intra-sentential context on the models.