Identifying Source Language Expressions for Pre-editing in Machine Translation

Norizo Sakaguchi, Yugo Murawaki, Chenhui Chu, Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{n-sakaguchi, murawaki, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Machine translation-mediated communication can benefit from pre-editing source language texts to ensure accurate transmission of intended meaning in the target language. The primary challenge lies in identifying source language expressions that pose difficulties in translation. In this paper, we hypothesize that such expressions tend to be distinctive features of texts originally written in the source language (*native language*) rather than translations generated from the target language into the source language (*machine translation*). To identify such expressions, we train a neural classifier to distinguish native language from machine translation, and subsequently isolate the expressions that contribute to the model's prediction of native language. Our manual evaluation revealed that our method successfully identified characteristic expressions of the native language, despite the noise and the inherent nuances of the task. We also present case studies where we edit the identified expressions to improve translation quality.

Keywords: machine translation, source language expression identification, pre-editing

1. Introduction

With its rapid progress, machine translation (MT) holds the potential to facilitate seamless cross-lingual communication (Yamashita and Ishida, 2006; Robertson and Díaz, 2022). MT-mediated communication involves individuals or groups who speak different languages. Individuals can input their messages in their native language, which is then automatically translated into the target language for the recipient. In turn, the recipient can respond in their native language, which is translated back into the original language of the sender. There is no longer a requirement for each party to possess knowledge of the other party's language.

While MT-mediated communication has been studied within the field of human-computer interaction, it has an important implication for the NLP community: MT-mediated communication deviates from the traditional setup of MT. While an MT system has no right to modify the source language text and aims to translate it into the target language faithfully, MT-mediated communication is more flexible in that a user can pre-edit the text to ensure accurate transmission of the intended meaning in the target language. For instance, in Japanese, it is commonplace to describe the noun "restaurant" with the adjective "delicious," resulting in a literal English translation as "That restaurant is delicious." However, this translation is often perceived as awkward by many English speakers (Honna, 2010). By slightly modifying the original Japanese sentence, we can obtain a more natural-sounding English translation, "The food at that restaurant is delicious."

The example above suggests the necessity of running the cycle of MT, the evaluation of the target language output, and the modification of the source language input. Given the goal of achieving broad adoption of MT-mediated communication, however, it is unrealistic to expect users to possess a sufficient level of proficiency in the target language. For users to operate exclusively in the source language, we need a system capable of identifying source language expressions that potentially pose difficulties in translation.

To that end, we present a key assumption in this paper: Such expressions tend to be distinctive features of texts originally written in the source language (*native language*) rather than translations from the target language (*machine translation*). Note here that the translation direction is reversed. If the goal is successful Japanese-English translation, for instance, we employ an English-Japanese MT system to identify Japanese expressions.

As it is challenging to test this assumption directly, we adopt an indirect approach. We construct a dataset comprising texts in both the native language and machine-translated versions of Japanese. Subsequently, we train a neural classifier that distinguishes native Japanese and Japanese translation with this dataset (Figure 1(top)). We discover a negative correlation between (1) the classification score, which indicates the probability of being in the native language, and (2) the performance of Japanese-English machine translation. This finding provides indirect evidence supporting the validity of the key assumption.

LREC-COLING 2024, pages 8605–8616
20-25 May, 2024.

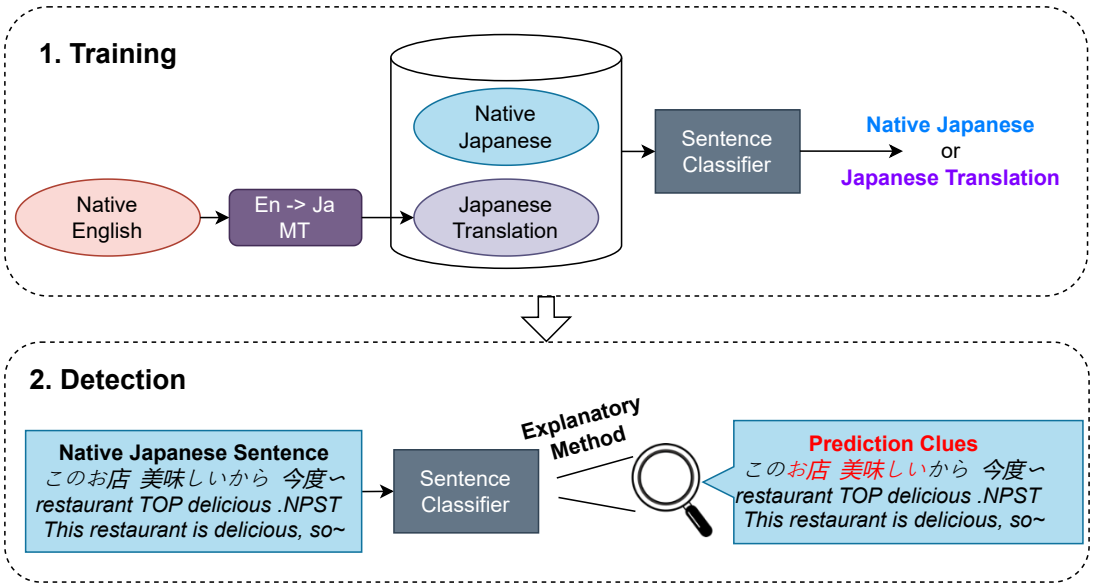

Figure 1: Overview of the proposed approach. The goal is to identify Japanese expressions that may present challenges in English translation, although the proposed method is applicable to other language pairs. (top) We begin by training a neural classifier that distinguishes native Japanese texts from texts machine-translated from English. (bottom) We subsequently analyze the internal behavior of the classifier using an explanatory method, enabling the identification of expressions that contribute to the prediction of native Japanese.

Using the classifier, we then identify the expressions that contribute to the prediction of native language (Figure 1 (bottom)). To achieve this, we expand upon the contextual decomposition approach (Murdoch et al., 2018; Jin et al., 2020) and enhance it in a way that allows for the efficient identification of multiple phrases within a given text.

Our manual evaluation of a sample of the identified expressions leads to the following findings. (1) The identification was noisy, but nevertheless, we were able to find expressions characteristic expressions of the native language. (2) Typically, identified expressions were neither ungrammatical nor completely absent in machine translation but exhibited gaps in frequency.

Lastly, we present case studies of manual pre-editing and quantitative evaluation using GPT-4 as a substitute for human pre-editing. We have shown that translation can be improved through pre-editing by an individual who possessed proficiency in the target language although our ultimate goal is to enable people who only understand the source language to perform pre-editing. The automated evaluation also suggests that the proposed method yields more natural translations while preserving the original meanings.

2. Related Work

2.1. Machine Translation Pre-editing

Pre-editing is a technique of modifying the source text prior to translation to improve the quality of machine translation outputs. While pre-editing has garnered continued interest in NLP (O'Brien, 2003; Aikawa et al., 2007; Seretan et al., 2014; Marzouk and Hansen-Schirra, 2019; Mehta et al., 2020; Miyata and Fujita, 2021), its focus is primarily on the professional translation process.

Although MT-mediated communication appears to be largely out of the scope of previous studies, it expands the possibilities of pre-editing. In contrast to translators, who do not have the authority to alter the source language text, users in MT-mediated communication have the flexibility to make adjustments based on system feedback, recognizing that the original source text may not always perfectly and succinctly capture their intentions.

Controlled language has been a focal point of interest in pre-editing (O'Brien, 2003; Aikawa et al., 2007). It is characterized by a small set of rules that includes limitations on vocabulary usage, restrictions on certain coordination constructions, and avoidance of the passive voice (O'Brien, 2003). Although these rules have shown effectiveness in rule-based and statistical MT systems, they either have no impact or can even yield negative effects in neural MT (Marzouk and Hansen-

Schirra, 2019).

Our focus markedly differs from that of traditional pre-editing research. As exemplified by the example of "delicious" above, we prioritize diverse linguistic phenomena that resist reduction to a narrow set of rules. For this reason, we take a fully data-driven approach in this paper.

2.2. Translationese Studies

Translationese refers to the distinctive features found in the text that was translated into a given language, setting it apart from the text originally composed in that language (Gellerstam, 1986). There are artifacts depending on the source language and general effects of the process of translation that are independent of source language (Baker, 1993).

When reading a translated text, one can often develop an intuition that it was not originally written in that particular language. However, elucidating this intuition by identifying concrete traces of translationese is notoriously challenging (Tirkkonen-Condit, 2002).

Previous studies working on automatically identifying translationese rely on aggregate statistics such as type-token ratio (Toral, 2019) and the weights of hand-crafted features of an SVM classifier that distinguishes translationese from native language (Baroni and Bernardini, 2005; Volansky et al., 2013). Although these studies shed light on the general characteristics of translationese, it remains challenging to attribute these findings to specific instances within texts (Amponsah-Kaakyire et al., 2022).

A common aspect between our study and translationese studies is the absence of clear boundaries between the expressions to be identified and other expressions. On the other hand, a distinctly different aspect between the two is that translationese implies the intention to eliminate it if possible, while we are open to the source language text becoming translationese-like if it leads to an improved translation.

2.3. Explaining Text Classification

There is a growing body of interest in explaining neural networks. Among numerous approaches proposed to date, our approach can be categorized as prediction-level explanation of post hoc analysis, as opposed to dataset-level explanation (Murdoch et al., 2019). This particular subcategory is still in the developmental phase, and various methods have been proposed (Simonyan et al., 2013; Li et al., 2016; Sundararajan et al., 2017; Jin et al., 2020). Although our work builds upon a class of methods named contextual decomposition (Murdoch et al., 2018), we do not claim it is the definitive choice for our purpose.

Harust et al. (2020) employ contextual decomposition to identify expressions characteristic of native English speakers, as opposed to L2 speakers. They assume that at most one phrase is predominantly responsible for the classification, given the infrequent occurrence of native-like expressions. In contrast, we have to abandon this assumption because we often encounter multiple traits of the native language in a single text.

3. Proposed Method

3.1. Overview

Figure 1 illustrates the overview of our approach using the Japanese-English language pair. The goal is to identify Japanese expressions that may present challenges when translating into English. A straightforward approach would involve translating a Japanese source text into English. While it might be possible to identify unnatural portions in the English text, neural MT makes it hard to map them back to the corresponding Japanese text fragments. For this reason, we resort to English-to-Japanese MT and focus on classifying Japanese texts.

We prepare two comparable corpora that are written in Japanese and English, respectively. The English corpus is translated into Japanese using an English-to-Japanese MT system. We then build a sentence-level neural classifier, which is trained to distinguish the two types of Japanese texts, *native language* and *machine translation*.

Given the classifier, we proceed to analyze its internal behavior using an explanatory method. For a native language sentence classified as native language, we identify expressions that contribute to the prediction.

3.2. Contextual Decomposition for a Neural Classifier

Given a sequence of tokens representing a sentence, the neural classifier outputs one of the two labels, *native language* and *machine translation*. We build the classifier by fine-tuning a pre-trained RoBERTa model (Liu et al., 2019).

For a native language sentence classified as native language, we want to decompose the prediction score into two components: one based on S, a subset of the token sequence, and the other based on the rest of the input. The key idea of contextual decomposition (CD) (Murdoch et al., 2018) is that by defining a decomposition operation for each neural network layer, we can trace the forward computation to propagate the decomposed

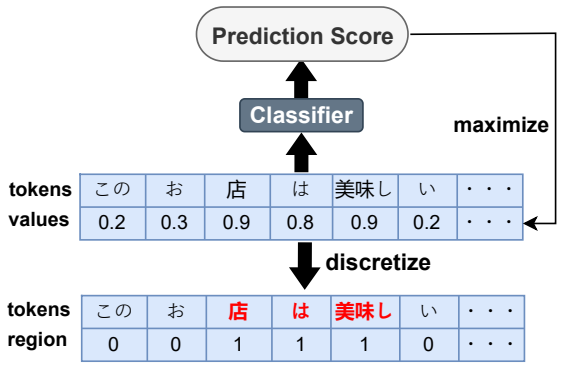

Figure 2: Continuously relaxed contextual decomposition.

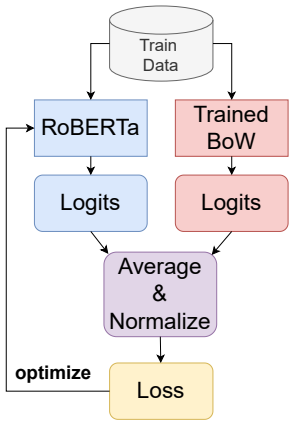

Figure 3: Use of an bag-of-words classifier as an auxiliary model.

input to a decomposed output. Using CD, our aim is to identify a subset S that significantly contributes to the prediction score.

The classifier $y = f(x)$ can be represented as a recursive application of L operations in the form of $y = (g_L \circ g_{L-1} \circ \cdots \circ g_1)(x)$. For each operation $g_l(x)$, CD defines an approximate decomposition such that $g_l^{\mathrm{CD}}(\beta_{l-1}(x), \gamma_{l-1}(x)) = (\beta_l(x), \gamma_l(x))$, where β represents the contribution of the given subset S while γ represents the contribution of the remaining part. For each l, $\beta_l(x) + \gamma_l(x) = g_l(x)$.

The decomposition is trivial for the embedding layer: $\beta(e_i) = e_i$ and $\gamma(e_i) = \mathbf{0}$ if $i \in S$; otherwise $\beta(e_i) = \mathbf{0}$ and $\gamma(e_i) = e_i$.

For a linear layer with a weight matrix W and a bias b, the input $v_{l-1} = \beta(v_{l-1}) + \gamma(v_{l-1})$ is transformed into the output $v_l = \beta(v_l) + \gamma(v_l)$ as follows.

$$\beta(v_l) = W\beta(v_{l-1}) + \frac{|W\beta(v_{l-1})|}{|W\beta(v_{l-1})| + |W\gamma(v_{l-1})|} b$$

$$\gamma(v_l) = W\gamma(v_{l-1}) + \frac{|W\gamma(v_{l-1})|}{|W\beta(v_{l-1})| + |W\gamma(v_{l-1})|} b$$

The first terms are the linear decomposition of Wv_{l-1}. The partition of the bias term is an approximation based on Singh et al. (2019), who found that partitioning the bias in proportion to the absolute values of the first terms empirically worked well.

As seen above, we need to define a decomposition operation for every neural network layer. Murdoch et al. (2018) define decomposition operations required to build an LSTM classifier while Harust et al. (2020) present a simple extension to a BiLSTM classifier. Jin et al. (2020) implement decomposition operations for BERT (Devlin et al., 2019) as their baseline method. Since RoBERTa is BERT with a collection of minor improvement techniques, we can readily apply their method to our RoBERTa classifier.

3.3. Continuously Relaxed Contextual Decomposition

A limitation of CD is that we need to specify a subset of the input token sequence to run it. For a token sequence of length n, we need to perform $O(2^n)$ runs of CD, which is prohibitively expensive even for a relatively small n.

To address this problem, we propose a method named continuously relaxed contextual decomposition (CRCD). The discrete notion of the presence or absence of each token in the subset is relaxed into a continuous value between 0 and 1. The overview of CRCD is shown in Figure 2. We exploit the fact that the decomposed forward computation $(g_L^{\mathrm{CD}} \circ g_{L-1}^{\mathrm{CD}} \circ \cdots \circ g_1^{\mathrm{CD}})$ is also differentiable. Using backpropagation, CRCD iteratively optimizes the continuous values by maximizing the β of the predicted score, with a regularization term penalizing overdetection. After that, CRCD discretizes the continuous values into 0 and 1 to select a subset of the input token.

For numerical stability, we optimize auxiliary variables that take values from negative infinity to positive infinity. These auxiliary variables are converted using the sigmoid function, resulting in values ranging from 0 to 1. Since the inference can get stuck in local optima, we try multiple runs using different random initializations and choose the best result. They can easily be parallelized using a mini-batch.

3.4. Auxiliary Bag-of-words Classifier

Since the classifier simply aims to minimize the classification loss, there is a potential danger in relying on spurious cues such as minor domain mismatches. To alleviate this problem, we incorporate an auxiliary bag-of-words (BoW) classifier.

The overview of the method is shown in Figure 3. Before fine-tuning the RoBERTa classifier, we sep-

arately train the weak BoW classifier. We freeze its parameters and use the logits it outputs when fine-tuning RoBERTa. For a given input, let l_{RoBERTa} be the logits output by RoBERTa, and let l_{BoW} be the logits output by the BoW classifier. The loss function, loss, used for training RoBERTa is computed as follows:

$$\text{loss} = \text{H}(\text{expit}((l_{\text{RoBERTa}} + l_{\text{BoW}})/2), R), \quad (1)$$

where H is the cross-entropy loss, expit is the logistic function, and R is the ground truth label. Only the parameters of RoBERTa are updated using backpropagation.

The BoW classifier encourages RoBERTa to focus on cues that go beyond the word level. If words alone provide sufficient cues for classification, RoBERTa has no need to override l_{BoW}. RoBERTa proves its worth when there is a need to capture intricate word-to-word interactions for classification.

4. Experiments

4.1. Dataset

4.1.1. Native Language Corpora

We require both native Japanese texts and native English texts, with the latter being translated into Japanese using machine translation. While our primary objective is to enhance MT-mediated communication, the absence of readily available corpora compelled us to prioritize domain comparability. The preparation of comparable corpora is strongly encouraged, as classifiers often tend to exploit spurious cues.

We chose Japanese and English Wikipedia for experiments. We only used articles with titles composed of common nouns, given that divergence in the occurrence patterns of proper nouns could have a stronger impact on the classifier. We relied on the Shinra Project[1] to identify such articles. The Shinra Project assigned extended named entities to the titles of Japanese Wikipedia, and one of the hierarchically-organized labels, 0, indicated common nouns. To further alleviate the impact of culture-specific topics, we selected articles that had corresponding counterparts in 35 or more other language versions of Wikipedia. We anticipated that articles on these popular topics would garner significant attention, to the extent that even if they were originally translated from another language, the articles had been adequately edited. We finally segmented the main text of each article into sentences using SpaCy.[2] As a result, we obtained $1,073,431$ English sentences and $648,507$ Japanese sentences.

Parameter	BoW	RoBERTa
lerning rate	0.001	3e-5
optimizer	Adam ($\beta = (0.9, 0.999)$)	
scheduler	-	linear warmup
batch-size	1,024	128

Table 1: Hyperparameters of classifiers.

	BoW	RoBERTa (+BoW)
Accuracy	0.88	0.95

Table 2: Classification accuracy on the test set.

4.1.2. English-to-Japanese MT

We prepared an English-to-Japanese MT system to translate the native English corpus. As an initial model, we obtained an English-to-Japanese MT model pre-trained on JPara Crawl (Morishita et al., 2020), a parallel corpus created by crawling the web and automatically aligning parallel sentences. We fine-tuned it using the Japanese-English portions of WikiMatrix (Schwenk et al., 2021). A total of $479,315$ sentences were used as training data, while $1,000$ sentences were used as test data. The hyperparameters for training were taken from the original paper of JPara Crawl.

Using SacreBLEU (Post, 2018), we achieved the test BLEU score of 21.82. For comparison, DeepL,[3] a popular commercial MT service, gave 16.75 on the same test data, indicating that our system outperformed DeepL. We attribute the system's superior performance to the fine-tuning process, which compelled the system to adapt to the specific writing style found in Wikipedia.

The English-to-Japanese MT system generated $1,721,938$ Japanese sentences from the native English corpus. For the purpose of classification, we labeled these sentences as *machine translation*, while the native Japanese sentences were given the label *native language*. We randomly extracted $10,000$ sentences as the test data while the remaining portion was used for training.

4.2. Training the Classifier

The Japanese dataset we described above was used to train a neural classifier. Specifically, we fine-tuned a pre-trained Japanese RoBERTa base model.[4]

As an auxiliary BoW classifier, we trained a neural network consisting of an embedding layer, a

[1] http://shinra-project.info/
[2] https://spacy.io/
[3] https://www.deepl.com/translator
[4] https://huggingface.co/nlp-waseda/roberta-base-japanese

Label	Sentence
Native	また、平面上、空間上の座標を示す方法もある。 (There are also ways to indicate coordinates on a plane or in space.)
Translation	秋葉原にはアニメ、マンガ、レトロビデオゲーム、小像、カードゲームなどを専門とする多数の店舗がある。 (Akihabara also has dozens of stores specializing in anime, manga, retro video games, figurines, card games, and other collectibles.)

Table 3: Four examples where the predictions of BoW were incorrect, but those of RoBERTa+BoW were correct.

mean pooling layer, and two linear layers. The embeddings were initialized with those of the pre-trained RoBERTa model. Table 1 shows the hyper-parameters used for training the classifiers.

Table 2 shows the accuracy of two models. We can see that even the weak BoW classifier achieved high accuracy. The fact that RoBERTa brought a further performance gain suggests that the sentences in the dataset contained abundant cues beyond the word level. Note that our goal is not to maximize classification accuracy. In fact, lower performance is even preferable if it stems from the classifier's indifference to spurious cues.

Table 3 shows examples that were incorrectly predicted by BoW but correctly predicted by RoBERTa+BoW. There are no native-like words in the first sentence, but its structure, such as "また" (*also*, *furthermore*) on the sentence-initial position, is native-like (we revisit this specific pattern for further discussion in Section 4.4). By contrast, the second sentence is about Japanese land and culture and includes words that appear more often in the native language than translation from English. These features make BoW's prediction erroneous and RoBERTa+BOW's correct.

4.3. Correlation between Classification Score and Translation Accuracy

Is it difficult to accurately translate Japanese sentences, which the classifier considers to be *native language*, into English? To test this, we investigated the correlation between classification scores and Japanese-to-English translation accuracy.

We chose the same test set, consisting of $1,000$ Japanese sentences taken from WikiMatrix. The classification score was determined by calculating the difference between the two logits output by the classifier. A higher classification score indicates a higher likelihood of being a native Japanese sentence.

To assess the translation accuracy, we built a Japanese-to-English MT system. We followed the procedure described in Section 4.1.2 and simply reversed the translation direction: We used a Japanese-to-English MT model pre-trained on JPara Crawl and fine-tuned it using WikiMatrix. We obtained the test BLEU score of 27.76 with SacreBLEU. Finally, we used BLEURT (Sellam et al., 2020) to calculate sentence-level translation accuracy because the widely used BLEU was unreliable at the sentence level.

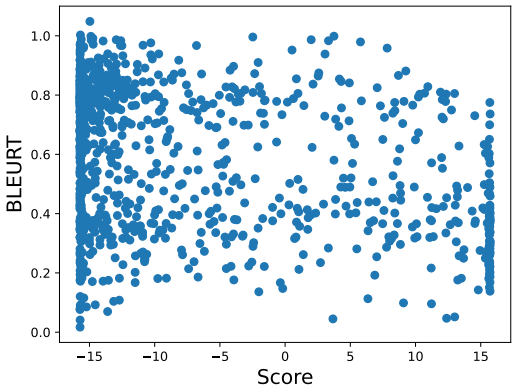

Figure 4: Correlation between classification scores and translation accuracy.

Evaluation	neutral	kind-of-native	native
Frequency	102	66	37

Table 4: Manual evaluation for 205 identified phrases in a sample of 100 sentences.

The scatter plot of the classification scores and the translation accuracy is presented in Figure 4. We found a moderate negative correlation of -0.33 for the Pearson correlation coefficient between the two metrics. This can be interpreted as indirect evidence supporting our key assumption: Expressions that potentially pose difficulties in translation tend to be distinctive features of texts originally written in the source language.

The exact reason for the correlation staying within the moderate range is not entirely clear. A possible explanation is that the utilization of reverse-direction translation represents is a rough approximation with a moderate level of noise. Another possible explanation is that reference-based metrics for MT inherently introduce noise at the sentence level.

Phrases	Frequencies in the training data	
	Native Language	Machine Translation
[CLS] また (*Also*, sentence-initial)	18,956	8,658
また (*also*, anywhere in a sentence)	30,334	65,589
のことである。 (*It is about …*)	1,161	555
の (genitive marker)	321k	786k
こと (*thing*)	74k	152k
である (copula)	54k	139k

Table 5: Two examples of Japanese phrases identified by the proposed method. The first row of each block shows the detected phrase, followed by unconditional or word-by-word counterparts for comparison.

4.4. Evaluation of the Identified Phrases

We proceeded to evaluate the expressions identified by the proposed method. We focused on the $10,000$ sentences used for evaluating the classifier. With the threshold of the classification score of 14, we obtained $2,371$ sentences that contained identified expressions. From this set of sentences, we randomly selected a sample of 100 sentences for manual evaluation. Because the subset of input tokens our method identified were generally disconnected, we conveniently refer to each subsequence as a *phrase*. There were 205 phrases identified in the sample.

Manual evaluation posed significant challenges and was susceptible to subjectivity. Nevertheless, we asked a native Japanese speaker to do that based on his intuition. Each identified phrase was manually classified into one of three categories: *neutral*, *kind-of-native*, and *native*. In this classification, *neutral* indicates that the phrase in question can naturally occur in both native Japanese and translations, while *native* indicates that an MT system would dare not select the phrase in question for translation.

Table 4 shows the evaluation results. We can see that more than half of the phrases were evaluated as either kind-of-native or native, indicating the effectiveness of the proposed method.

As case studies, we showcase two identified phrases in Table 5. The first one is the conjunction "また" (*also*, *furthermore*). The identified phrase included the special `[CLS]` token, indicating that the classifier placed importance on the sentence-initial position. While the conjunction was present frequently in both types of data, the odds ratio of 4.73 indicated a strong association between the sentence-initial position and native Japanese texts. Given that the auxiliary BoW classifier ignored word positions, it is reasonable to suggest that RoBERTa demonstrated its effectiveness. Although not grammatically incorrect, the overuse of sentence-initial conjunctions seems to be a characteristic feature of native Japanese texts.

Similarly, the sentence-final phrase "のことである。" (*It is about …*) revealed a disparity in frequency. It, again, appeared frequently in both types of data. Compared with its constituents, "の" (a genitive marker), "こと" (*thing*, a grammaticalized noun), and "である" (the non-past form of a copula), this particular combination was strongly associated with native Japanese texts, with the odds ratios of 5.12, 4.30, and 5.38, respectively. This again suggests that The RoBERTa classifier successfully identified patterns to which the BoW classifier was insensitive.

Nevertheless, we encountered numerous cases that were difficult to explain. Many of them seemed to be spurious cues picked up by the classifier, but some others might have been instances that represented genuine patterns requiring further analysis.

It is important to note that our method is primarily exploratory in nature. While the two phrases presented in Table 5 were amenable to frequency-based *post hoc* analysis, conducting an *a priori* analysis is challenging due to the combinatory nature of phrases and the ad hoc conditions such as word positions.

4.5. Case Studies of Manual Pre-editing

We investigated whether manually editing the identified expressions could improve the translation. Considering our ultimate goal, it is desirable for users who only comprehend the source language to have the ability to pre-edit. However, our current method solely identifies expressions without offering specific improvement suggestions. Therefore, as an oracle setting, we have employed a native Japanese speaker who possesses proficiency in the target language. Specifically, we asked him to edit identified expressions to make them much easier to be translated into English. In addition, we applied the classifier described in Section 4.2 to the Japanese sentence before and after editing to monitor the changes in classification scores.

Table 6 shows the results of our case studies. We can see that the editing of the identified expressions significantly decreased the nativeness indicated by the classification scores. We translated

		Score
Source	古くは鹿の角などを 用いて作成した。	15.7
Translation	It was built in ancient times with an ancient deer corner.	-
Edited Source	鹿の角を用いて古い時代には 作成した。	4.16
Translation	We have done it with a deer corner, in old times.	-
Source	地図サイト構築用ソフトとして販売されているものは この形式が多い。	15.7
Translation	Many are sold as map-site construction software.	-
Edited Source	販売されている 地図サイト構築用のソフトの多くが この形式だ。	12.1
Translation	Many software for building maps they sell is in this form.	
Source	プラズマ振動はプラズマ波動の一種であり、プラズマが電気 的中性を保とうとする傾向をもつために生まれる波動 である。	15.7
Translation	Plasma oscillations are a type of plasma wave wave, created because of a tendency for a plasma to maintain its electrical neutrality.	-
Edited Source	プラズマ振動はプラズマ波動の一種であり、プラズマの電気 的中性を保とうとする傾向が生んだ波動 である。	7.81
Translation	Plasma oscillations are a type of plasma wave, a tendency to maintain the electrical neutrality of a plasma.	-

Table 6: Three sample source sentences before and after editing according to the identified phrases, and their corresponding machine translation results. Underlines indicate the identified phrases in the source sentence and their corresponding editing in the edited source sentence. The score indicates the classification score for the source and edited source sentences, where higher scores indicate higher nativeness judged by the classifier.

MT systems		Ours		TexTra	
Metrics		BLEURT(↑)	PPL(↓)	BLEURT(↑)	PPL(↓)
Original		0.588	205	**0.631**	159
All	Non-specified	**0.583**	**130**	0.619	125
	Specified	0.582	146	0.618	**124**
Positive	Non-specified	0.590	199	0.630	159
	Specified	0.589	202	0.630	159

Table 7: The effect of automatic pre-editing on machine translation in the Wikipedia dataset.

the source sentences before and after editing using the Japanese-to-English MT system described in Section 4.3. All the translations have been improved after the editing.

4.6. Automatic Pre-editing with GPT-4

As a surrogate for manual pre-editing, we employed GPT-4 (OpenAI, 2023). To assess the translation quality both before and after automatic pre-editing, we utilized the MT test set outlined in Section 4.1.2 and the BSD Corpus (Rikters et al., 2019). The BSD Corpus is a parallel corpus consisting of conversations in business scenes, with each conversation labeled with its respective source language. In this experiment, we exclusively utilized conversations in Japanese as per the experimental objective. By utilizing BSD Corpus, we can expect to encounter more linguistically authentic expressions that occur only in conversational contexts.

We instructed GPT-4 to edit these sentences while preserving their original meanings. Specifically, we conducted tests under the following two conditions: (1) **Non-specified**: Editing the entire sentence without specifying the identified expressions. (2) **Specified**: Editing the identified expressions.

In addition, the following two conditions were set to limit the sentences to be edited. (1) **All**: Editing all of the sentences. (2) **Positive**: Editing only sentences classified as positive, judged as *native language*, by the classifier.

The original and pre-edited texts were translated into English using our MT model (4.3) and TexTra.[5] We assessed the translation quality with BLEURT and measured the relative naturalness of the translated texts using perplexity (PPL) based on GPT-2 Large (Radford et al., 2019).

The results of the experiments conducted on the

[5]https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

MT systems		Ours		TexTra	
Metrics		BLEURT(↑)	PPL(↓)	BLEURT(↑)	PPL(↓)
Original		0.502	84.4	**0.694**	**33.5**
All	Non-specified	0.520	85.9	0.685	36.0
	Specified	0.513	84.7	0.685	35.4
Positive	Non-specified	**0.521**	85.4	0.686	35.7
	Specified	0.513	**84.2**	0.686	35.4

Table 8: The effect of automatic pre-editing on machine translation in BSD Corpus.

MT systems			Ours		TexTra	
Metrics			BLEURT(↑)	PPL(↓)	BLEURT(↑)	PPL(↓)
Source	Original		0.650	279	**0.687**	219
	Edited	Non-specified	**0.684**	**100**	0.672	137
		Specified	0.634	135	0.662	120
		Specified + Manual Evaluation	0.642	158	0.675	**116**

Table 9: The effect of manual filtering of detected phrases and automatic pre-editing on machine translation. "Specified + Manual Evaluation" denotes a manual judgment as to whether or not to edit for each detected phrase.

Wikipedia dataset are presented in Table 7. In both MT models, pre-editing resulted in an improvement in PPL. Although there are some settings where the BLEURT have slightly decreased, considering the significant decrease in PPL, it can be inferred that the translations have been improved.

The results of the experiments conducted on the BSD dataset are presented in Table 8. The translations improved when pre-editing was applied to sentences with positive scores using our MT model. However, with TexTra, no improvement was observed from the original text. This outcome suggests that TexTra achieves sufficiently high translation accuracy even without pre-editing, leaving little room for improvement.

4.7. Manual Filtering of Detection Phrases

The manual evaluation in Section 4.4 suggests that the proposed method for detecting editing points could be recall-oriented. We replicated the experiment described in Section 4.6 using 100 sentences in the Wikipedia dataset, but with additional manual evaluations performed for each detected phrase. Only the expressions that a human evaluator judged as either *kind-of-native* or *native* were edited.

Table 9 shows the results. In our model, manual evaluation resulted in an improvement in BLEURT but a decrease in PPL. On the other hand, TexTra showed improvements in both metrics through manual evaluation, particularly in the specified condition, where a combination of specified condition and manual evaluation yielded the best results.

Why did TexTra fare better with manual evaluation than our model did? The consistent domain of the training data for our model, classifier training data, and the data used in this experiment—all derived from Wikipedia articles—aligned the linguistic elements learned by the classifier in a manner that made translation difficult for the in-house model. Conversely, TexTra, trained on a more diverse set of domains, likely had fewer phrases inherently difficult to translate, leading to a more effective narrowing down of detection points.

5. Conclusions

In this paper, we presented a combination of a neural classifier and an explanatory method to identify expressions that are characteristic of a native language, as opposed to translations from another language. We expected these expressions to pose difficulties when translating into that language. We provided indirect evidence in favor of this assumption and presented several case studies.

We selected Wikipedia as the experimental dataset due to its relative ease in mitigating domain mismatch between the two languages. In the future, we aim to switch to conversational data to advance our ultimate goal of facilitating MT-mediated communication. Investigating language pairs beyond Japanese and English also presents an intriguing avenue for further research. Lastly, reversing the translation direction is a technical compromise, and therefore, it is worthwhile to explore direct identification on the target language side.

Limitations

In line with the recent trends and developments in NLP, we are also addressing a challenging problem where the automatic evaluation of system outputs proves to be difficult. Consequently, we had to depend on manual evaluations using small samples.

Although native expressions may occur discontinuously in a sentence, it is difficult to judge whether multiple discontinuous phrases can compose one native expression with the proposed continuously relaxed contextual decomposition method. Therefore, we only evaluated the nativeness of identified consecutive phrases. We only verified the proposed method with Japanese and English-to-Japanese translations, leaving the method's effectiveness for other language pairs a question.

Ethical Statements

Our ultimate goal is to facilitate machine translation-mediated communication, but adapting to the target language carries the risk of promoting cultural assimilation.

6. Bibliographical References

Takako Aikawa, Lee Schwartz, Ronit King, Mo Corston-Oliver, and Carmen Lozano. 2007. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Genabith, and Cristina España-Bonet. 2022. Explaining translationese: why are neural classifiers better and what do they learn? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 281–296, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and technology: In honour of John Sinclair*, pages 233–250. John Benjamins, Amsterdam & Philadelphia.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, page 88–95. CWK Gleerup.

Oleksandr Harust, Yugo Murawaki, and Sadao Kurohashi. 2020. Native-like expression identification by contrasting native and proficient second language speakers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5843–5854, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nobuyuki Honna. 2010. That restaurant is delicious. [Japan]. *Asian Englishes*, 13(2):64–65.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

S. Marzouk and S. Hansen-Schirra. 2019. Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. *Machine Translation*, 33:179–203.

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Singh Saluja, Vinith Misra, Ballav Bihani, and Ritwik K. Kumar. 2020. Simplify-then-translate: Automatic preprocessing for black-box translation. In *AAAI Conference on Artificial Intelligence*.

Rei Miyata and Atsushi Fujita. 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: Main Volume*, pages 1539–1550, Online. Association for Computational Linguistics.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *Proceedings of 6th International Conference on Learning Representations, ICLR 2018*.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Sharon O'Brien. 2003. Controlling controlled English. In *EAMT Workshop: Improving MT through other language technology tools: resources and tools for building MT*, Budapest, Hungary. European Association for Machine Translation.

OpenAI. 2023. GPT-4 technical report. arXiv 2303.08774.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Samantha Robertson and Mark Díaz. 2022. Understanding and being understood: User strategies for identifying and recovering from mistranslations in machine translation-mediated chat. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2223–2238, New York, NY, USA. Association for Computing Machinery.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Violeta Seretan, Pierrette Bouillon, and Johanna Gerlach. 2014. A large-scale evaluation of pre-editing strategies for improving user-generated content translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1793–1799, Reykjavik, Iceland. European Language Resources Association (ELRA).

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Sonja Tirkkonen-Condit. 2002. Translationese —a myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target. International Journal of Translation Studies*, 14(2):207–220.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, CSCW '06, page 515–524, New York, NY, USA. Association for Computing Machinery.

7. Language Resource References

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the Twelfth Language Resources*

and Evaluation Conference, pages 3603–3609, Marseille, France. European Language Resources Association.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2019. Designing the business conversation corpus. In *Proceedings of the 6th Workshop on Asian Translation*, pages 54–61, Hong Kong, China. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.