Evaluation Scheme of Focal Translation for Japanese Partially Amended Statutes

Takahiro Yamakoshi[†], **Takahiro Komamizu**[‡], **Yasuhiro Ogawa**[†♣], and **Katsuhiko Toyama**[†♣]
[†] Graduate School of Informatics, Nagoya University
[‡] Institutes of Innovation for Future Society, Nagoya University
[♣] Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Abstract

For updating the translations of Japanese statutes based on their amendments, we need to consider the translation "focality;" that is, we should only modify expressions that are relevant to the amendment and retain the others to avoid misconstruing its contents. In this paper, we introduce an evaluation metric and a corpus to improve focality evaluations. Our metric is called an Inclusive Score for DIfferential Translation: (*ISDIT*). ISDIT consists of two factors: (1) the n-gram recall of expressions unaffected by the amendment and (2) the n-gram precision of the output compared to the reference. This metric supersedes an existing one for focality by simultaneously calculating the translation quality of the changed expressions in addition to that of the unchanged expressions. We also newly compile a corpus for Japanese partially amendment translation that secures the focality of the post-amendment translations, while an existing evaluation corpus does not. With the metric and the corpus, we examine the performance of existing translation methods for Japanese partially amendment translations.

1 Introduction

In the world's globalized society, governments must quickly announce their statutes worldwide to facilitate international trade, economic investments, legislation support, and so on. The Japanese government addressed this issue in April 2009 by launching the Japanese Law Translation Database System (JLT) (Toyama et al., 2011) where it announces the English translations of Japanese statutes. However, as of January 2020, only 23.4% (163/697) of the translated statutes in JLT correspond to their latest versions (Yamakoshi et al., 2020). After amending a statute, its translation must be promptly updated to avoid creating confusion among international readers. Unfortunately, statutory sentences are much tougher to translate than ordinary sentences because the former are highly technical, complex, and long.

Furthermore, when translating statutory sentences that are partially modified by an amendment, we must consider *focal* translations. That is, we should only modify expressions that are changed by the amendment without changing the others. For example, consider the following sentence: "申立ては、事故の事実を示して、書面でこれをしなければならない。" (The request shall be made in a document stating the facts of the accident.) Its amendment rewrote "事故" (*jiko*; accident) to "海難" (*kainan*; marine accident). The following revision satisfies the focality requirement: "The request shall be made in a document stating the facts of the marine accident" because it contains minimum modifications. On the other hand, although "The petition shall be made in a document describing the facts of the marine accident" is fluent and adequate, it is unsuitable as a revision from the focality perspective because "申立て" (*moshitate*; request) and "示して" (*shimeshite*; stating), which are irrelevant to the amendment, were changed.

Yamakoshi et al. (2020) proposed a machine translation method for Japanese partially amendment translation that generates translation candidates by a Transformer (Vaswani et al., 2017)-based neural machine translation (NMT) model. It selects the best one by comparing the candidates with the output of a template-aware statistical machine translation (SMT) model (e.g., (Koehn and Senellart, 2010; Kozakai et al., 2017)) that only changes the affected expressions. They also proposed an evaluation metric for the focality of the translations.

However, we argue that two matters from their study must be improved: the evaluation metric

Proceedings of the 8th Workshop on Asian Translation, pages 124–132
Bangkok, Thailand (online), August 5-6, 2021.

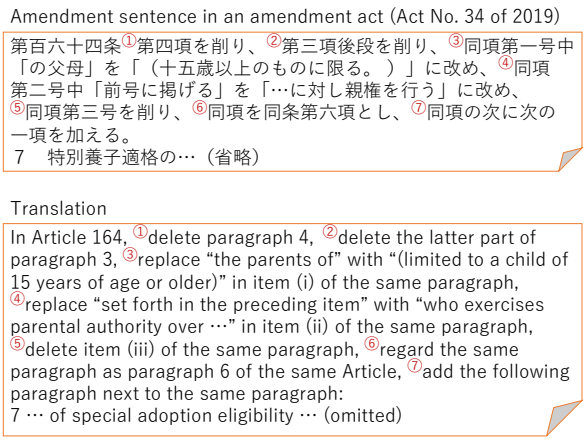

Figure 1: Amendment sentence

and the dataset. Their metric consists of two factors: (1) the n-gram recall of expressions unaffected by amendments and (2) a redundant penalty for lengthy outputs. Although with this metric we can evaluate how completely the method retained expressions irrelevant to the amendment, we cannot evaluate how adequately it translated expressions relevant to the amendment. The second is the dataset they used for their experiments. Their translation examples of partially amended statutory sentences are from amendment-version-controlled bilingual statutes in JLT. However, translations in JLT are not always focal. Therefore, their reported scores do not seem accurate.

In this paper, we solve these two matters. For the first, we introduce another metric for focality called the Inclusive Score for DIfferential Translation (ISDIT), which incorporates n-gram precision between the output and the reference instead of a redundant penalty. With this modification, the metric simultaneously evaluates the translation quality of both the changed and unchanged expressions that indicate the quality of the focal translation. For the second, we compile a corpus that secures focality between pre- and post-amendment translations and achieve it by asking professional human translators to translate focal post-amendment translations.

This paper makes the following contributions to amended statutory sentence translation tasks:

- introduces a new metric that more adequately reflects the focality of translations;
- compiles a translation corpus that ensures the focality of post-amendment translations;
- examines the translation performance of relevant methods with a metric and a corpus.

This paper is organized as follows. In Section 2, we clarify the background of our study. In Section 3, we explain related work. In Section 4, we describe our proposal and present our evaluation experiments and discussions in Section 5. Finally, we summarize and conclude in Section 6.

2 Background

In this section, we clarify the background of our study. First, we introduce the partial amendment process in Japanese legislation from the viewpoint of document modification and then we identify our study objective in the process.

2.1 Partial Amendments in Japanese Legislation

In Japanese legislation, a partial amendment is created by “patching” modifications to a target statute. Such modifications are prescribed as amendment sentences in an amendment statute. Based on their functions, Ogawa et al. (2008) categorized such modifications as follows:

1. Modification of part of a sentence: (a) replacement, (b) addition, and (c) deletion.
2. Modification of such structural elements as sections, articles, items, sentences, etc.: (a) replacement, (b) addition, and (c) deletion.
3. Modification of element numbers: (a) renumbering, (b) attachment, and (c) shifts.
4. Combined modification of element renumbering and replacement of its title string.

For modifying part of a sentence, Japanese legislation rules (Hoseishitsumu-Kenkyukai, 2018) mandate that the target expressions must be unique and form a chunk of meaning.

Figure 1 shows an example of an amendment sentence prescribed by an amendment act. Any of the seven modifications in the sentence can be assigned to one of the categories described above: Modifications ①, ②, and ⑤ respectively belong to category 2. (c) of a paragraph, a sentence, an item; modifications ③ and ④ belong to category 1. (a); modification ⑥ belongs to category 3. (c); modification ⑦ belongs to category 2. (b).

Most statutes enacted in recent years are amendment statutes. According to Nihon Horei Sakuin (Index of Japanese Statutes) [1], 78% (73/94) of acts enacted in 2019 are amendment ones. After

[1] https://hourei.ndl.go.jp/

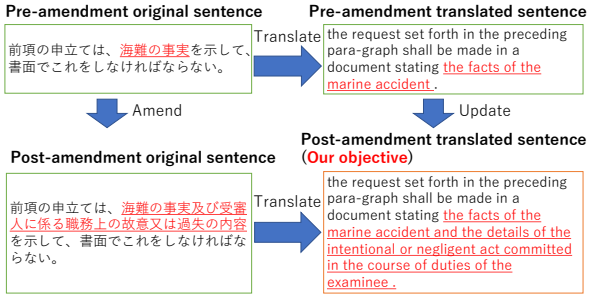

Figure 2: Differential translations in an amended statutory sentence

amending statutes, we should update their translations provided in JLT promptly. However, regarding the discussion in the introduction, many statutes available in JLT are out of date, which can provide wrong legal facts to international readers.

2.2 Objective

To solve the problem discussed in the previous section, our study focuses on translating partially amended statutes automatically. More specifically, it adopts a task declared by Yamakoshi et al. (2020). Among the categories described in the previous section, the task focuses on categories that modify the parts of an existing statutory sentence (i.e., category 1). In Fig. 1, modifications ③ and ④ are the targets. It also targets category 2, especially modifications that insert an additional sentence (e.g., a proviso) into an existing element or delete a sentence since such additions and deletions affect the main sentence. Modification ② in Fig. 1, which removes the latter part, is a case.

The task takes a triple of sentences (*a pre-amendment original sentence*, *a post-amendment original sentence*, and *a pre-amendment translated sentence*) as input and generates a translation for the post-amendment original sentence called *a post-amendment translated sentence*. Pre- and post-amendment original sentences are statutory sentences in a statute before and after an amendment, respectively. A pre-amendment translated sentence is a translation of the pre-amendment original sentence. Figure 2 illustrates this task.

In generating post-amendment translated sentences, Yamakoshi et al. advocated the *focality* of translations. This idea argues for only modifying expressions that are changed by the amendment without changing the others based on two reasons from the viewpoint of precise publicization. First, such sentences clearly represent the amendment contents, which helps international readers understand them. On the other hand, non-focal translations contain unnecessary modifications, which blur the amendment contents. Second, since the expressions in the pre-amendment translated sentences are assumed to be reliable, reusing them ensures translation quality.

For example, assume that an amendment statute instructs that we should replace "海難の事実" (*kainan no jijitsu*; the facts of the marine accident) with "海難の事実及び…の内容"(*kainan no jijitsu oyobi ... no naiyo*; the facts of the marine accident and the details of ...)" as depicted in Figure 2. In this case, we should replace "the facts of the marine accident" in the pre-amendment translated sentence with "the facts of the marine accident and the details of ..." and retain the other expressions to comply with the focality.

We define our task as follows:

Input:
Pre-amendment original sentence W_{PrO};
Post-amendment original sentence W_{PoO};
Pre-amendment translated sentence W_{PrT}.

Output: Generated post-amendment translated sentence $\widehat{W}_{\mathrm{PoT}}$.

Requirements:

Focality: $\widehat{W}_{\mathrm{PoT}}$ should reflect amendment W_{PrO} to W_{PoO} and preserve the expressions in W_{PrT} that are irrelevant to the amendment;

Fluency: $\widehat{W}_{\mathrm{PoT}}$ should have natural phrasing and syntax;

Adequacy: $\widehat{W}_{\mathrm{PoT}}$ should have W_{PoO}'s contents without excesses or inadequacies.

3 Related Work

We describe related work in this section. We overview the suitable machine translation methods for partially amended sentences in Section 3.1. We discuss metrics and data in Sections 3.2 and 3.3.

3.1 Method

We consider the focality of translations, which is uncommon in ordinary machine translation tasks. To achieve focal translations, the unchanged expressions must be retained as they appear in the pre-amendment translation. One solution is using a template-aware SMT method. Koehn and Senellart (2010)'s method is a choice, which can retain the unchanged expressions in the pre-amendment translations by copying them to the post-amendment translations.

Kozakai et al. (2017) optimized this method to

Japanese partially amendment translation by applying the following two modifications. First, they used pre-amendment original sentences and their translations instead of a relevant pair from the translation memory. Second, to determine objective expressions, they used the underlined information in a comparative table instead of the edit distance. Such underlined information is more reasonable as a translation unit than edit distance since sentence modification is done by a chunk of meaning in Japanese legislation.

Both methods can meet the focality requirement by copying the unchanged expressions in the pre-amendment translated sentences. However, the translation quality, especially fluency, suffers for the following three reasons. First, they use SMT for the translation model, which is typically outperformed by NMT. Second, their methods completely lock the unchanged expressions, which may strongly restrict the translations. Third, they use word alignment to find English expressions that correspond to Japanese ones, perhaps weakening their performance due to alignment error.

Yamakoshi et al. (2020)'s method solved these problems by incorporating NMT with a template-aware SMT. Their method, which uses an NMT model and a template-aware SMT model, allows the former to output n-best translations as candidates by applying Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) to improve the output diversity. It then chooses the candidate that most resembles the *interim* reference translation generated from a template-aware SMT model.

3.2 Metrics

Kozakai et al. (2017) used BLEU (Papineni et al., 2002) and RIBES (Hirao et al., 2014) as automatic evaluation metrics in their experiment. BLEU's calculation is based on n-gram precision between the system output and references; RIBES's calculation is based on word-order correlation. Therefore, RIBES is more sensitive to drastic structural modifications. However, both metrics are indifferent to whether an expression in the system output is a changed part in the amendment, and thus both fail to indicate the quality of the focality.

Yamakoshi et al. (2020) proposed focality scores to solve this issue. A focality score quantizes the focality of the system output by calculating the recall of the n-grams shared by both the pre- and post-amendment translations. With pre-amendment translated sentence W_{PrT} and actual post-amendment translated sentence W_{PoT} written by humans, we calculate focality score $\mathrm{Foc}(\widehat{W}_{\mathrm{PoT}}; W_{\mathrm{PrT}}, W_{\mathrm{PoT}})$ of generated sentence $\widehat{W}_{\mathrm{PoT}}$ as follows:

$$\mathrm{Foc}(\widehat{W}_{\mathrm{PoT}}; W_{\mathrm{PrT}}, W_{\mathrm{PoT}}) \quad (1)$$
$$= \mathrm{RP}(W_{\mathrm{PoT}}, \widehat{W}_{\mathrm{PoT}}) \cdot \mathrm{Rec}(\widehat{W}_{\mathrm{PoT}}; W_{\mathrm{PrT}}, W_{\mathrm{PoT}}),$$
$$\mathrm{RP}(W_{\mathrm{PoT}}, \widehat{W}_{\mathrm{PoT}})$$
$$= \min(1, \exp(1 - |\widehat{W}_{\mathrm{PoT}}|/|W_{\mathrm{PoT}}|)), \quad (2)$$

where RP avoids overestimating the scores of the redundant sentences. $|W|$ is the word count of W. Rec is the recall of the n-grams shared by W_{PrT} and W_{PoT}, calculated as follows:

$$\mathrm{Rec}(\widehat{W}_{\mathrm{PoT}}; W_{\mathrm{PrT}}, W_{\mathrm{PoT}}) = \quad (3)$$
$$\frac{\sum_{s \in \mathrm{CN}(\mathcal{W}_1)} \min(c_{\widehat{W}_{\mathrm{PoT}}}(s), c_{W_{\mathrm{PrT}}}(s), c_{W_{\mathrm{PoT}}}(s))}{\sum_{s \in \mathrm{CN}(\mathcal{W}_2)} \min(c_{W_{\mathrm{PrT}}}(s), c_{W_{\mathrm{PoT}}}(s))},$$
$$\mathcal{W}_1 = \{\widehat{W}_{\mathrm{PoT}}, W_{\mathrm{PrT}}, W_{\mathrm{PoT}}\} \quad (4)$$
$$\mathcal{W}_2 = \{W_{\mathrm{PrT}}, W_{\mathrm{PoT}}\}, \quad (5)$$

where $c_W(s)$ is the number of occurrences of the n-gram s in W, and $\mathrm{CN}(\mathcal{W})$, where $\mathcal{W} = \{W_1, W_2, \cdots, W_m\}$, returns common n-grams of $W_1, W_2, \cdots, W_m$:

$$\mathrm{CN}(\mathcal{W}) = \left\{ s \;\middle|\; s \in \bigcap_{W_i \in \mathcal{W}} \mathrm{ngrams}(W_i) \right\}, \quad (6)$$

where $\mathrm{ngrams}(W)$ returns all n-grams in W for a given n. We use multiple lengths of n-grams:

$$\mathrm{ngrams}(W) = \bigcup_{i=1}^{N} i\text{-gram}(W), \quad (7)$$

where $i\text{-gram}(W)$ returns the i-grams of W.

3.3 Data

Kozakai et al. (2017) used JLT bilingual resources to compile corpora for their experiment. For training data, they gathered 158,928 Japanese-English sentence pairs from 407 statutes provided in JLT. For test data, they selected 17 amendments available in JLT [2] from which they compiled 158 examples of sentence amendments, each of which consists of W_{PrO}, W_{PrT}, W_{PoO}, and W_{PoT}. Yamakoshi et al. (2020) also used this corpus for their experiment.

[2] JLT has a function to browse statutes and the translations of different amendment versions.

Sort	Content
W_{PrO}	前項の申立ては、海難の事実を示して、書面でこれをしなければならない。
W_{PrT}	The request set forth in the preceding paragraph shall be made in a document stating the facts of the marine accident.
W_{PoO}	前項の申立ては、海難の事実及び受審人に係る職務上の故意又は過失の内容を示して、書面でこれをしなければならない。
W_{PoT}	The petition set forth in the preceding paragraph shall be made in writing describing the facts of the marine accident and the details of the intentional or negligent act committed in the course of duties of the examinee.
Focal W_{PoT}	The request set forth in the preceding paragraph shall be made in a document stating the facts of the marine accident and the details of the intentional or negligent act committed in the course of duties of the examinee.

Table 1: Non-focal amendment example

However, some of these examples are not focal because they contain modifications irrelevant the amendment. Table 1 describes such an example. The straight lines in its sentences depict modifications that correspond to the amendment, and the wavy lines depict modifications irrelevant to the amendment. "Request," "a document," and "stating" in W_{PrT} are replaced with "petition," "writing," and "describing" in W_{PoT}, respectively, although corresponding Japanese expressions "申立て" (*moshitate*), "書面" (*shomen*), and "示して" (*shimeshite*) was retained throughout the amendment. An ideal translation for W_{PoT} is shown in the table's last row that retains all the expressions irrelevant to the amendment.

4 Proposal

In this section, we propose an evaluation scheme for Japanese partially amendment translations. Our evaluation scheme includes a new evaluation metric *ISDIT* and a differential translation corpus that secures the focality of its examples.

4.1 ISDIT Scores

The focality score in Section 3.2 assesses only the retention rate of the unchanged expressions in W_{PrT}. That is, it is unaware of the adequacy of expressions that are relevant to the amendment. Therefore, we update the focality scores so that they assess both factors. Our metric, Inclusive Score for DIfferential Translation (*ISDIT*), is calculated as follows:

$$\mathrm{ISDIT}(\widehat{W}_{\mathrm{PoT}}; W_{\mathrm{PrT}}, W_{\mathrm{PoT}}) = \mathrm{Pre}(\widehat{W}_{\mathrm{PoT}}; W_{\mathrm{PoT}}) \cdot \mathrm{Rec}(\widehat{W}_{\mathrm{PoT}}; W_{\mathrm{PrT}}, W_{\mathrm{PoT}}), \tag{8}$$

where Rec is the recall defined in Eq. 3. Pre is the precision of system output $\widehat{W}_{\mathrm{PoT}}$ compared to reference W_{PoT}, which is calculated as follows:

$$\mathrm{Pre}(\widehat{W}_{\mathrm{PoT}}; W_{\mathrm{PoT}}) = \frac{\sum_{s \in \mathrm{CN}(\mathcal{W})} \min(c_{\widehat{W}_{\mathrm{PoT}}}(s), c_{W_{\mathrm{PoT}}}(s))}{\sum_{s \in \mathrm{CN}(\{\widehat{W}_{\mathrm{PoT}}\})} c_{\widehat{W}_{\mathrm{PoT}}}(s)}, \tag{9}$$

$$\mathcal{W} = \{\widehat{W}_{\mathrm{PoT}}, W_{\mathrm{PoT}}\}. \tag{10}$$

For example, we consider the example shown in Table 2. Case 1 contains an unnecessary modification, and Case 2 fails to translate "四十万" (*yonjuman*; four hundred thousand) that is relevant to the amendment. The focality score penalizes the first case, but not the second case. ISDIT penalizes both. From the viewpoint of focal translations that should reflect the amendment contents, penalizing both the unnecessary modification errors and amended phrase translation errors is preferable.

4.2 Focal Differential Translation Corpus

As discussed in Section 3.3, the differential translation corpus compiled by Kozakai et al. (2017) includes non-focal examples. To provide a fairer evaluation, we compiled a new corpus that secures the focality of every translation example. We applied the following instructions for the corpus compilation:

1. Compile the versions of statutes provided in JLT;
2. Compile those provided in e-LAWS[3];
3. Compile statutes whose JLT version lags behind its e-LAWS version;

[3] https://elaws.e-gov.go.jp/ e-LAWS (e-Legislative Activity and Work Support System) provides a governmental open database of national statutes which are original (i.e., written in Japanese) and most recent.

Sort	Content	ISDIT	Foc.
W_{PrO}	解職請求は、八十万人を超える者の連署を要する。	—	—
W_{PrT}	A request for recall requires joint signatures of more than eight hundred thousand people.	—	—
W_{PoO}	解職請求は、四十万人を超える者の連署を要する。	—	—
W_{PoT}	A request for recall requires joint signatures of more than four hundred thousand people.	—	—
Case 1	A petition for recall requires joint signatures of more than four hundred thousand people.	0.82	0.70
Case 2	A request for recall requires joint signatures of more than forty hundred thousand people.	0.85	1.00

Table 2: Example for ISDIT calculation ("Foc." stands for focality score)

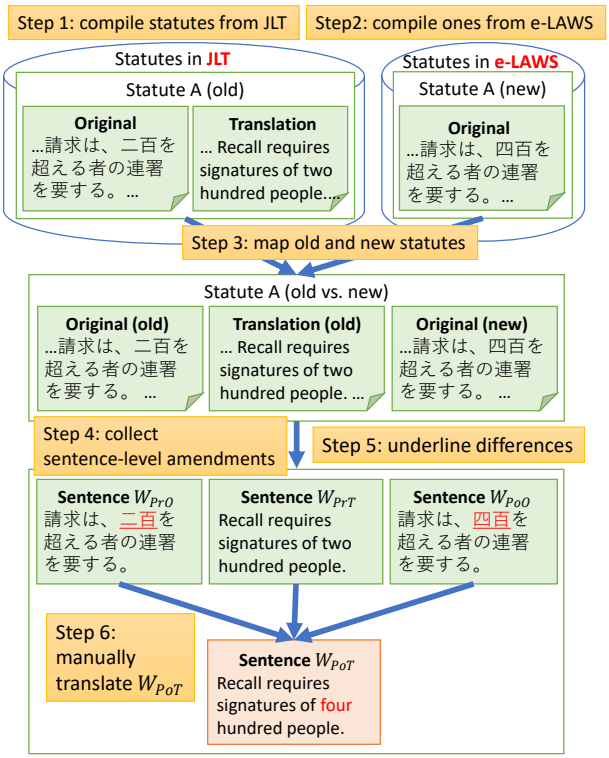

Figure 3: Compilation procedure for a focal corpus

4. Collect sentence-level amendments of such statutes;
5. Underline the modified expressions in W_{PrO} and W_{PoO} as if they were highlighted in an actual amendment statute;
6. Manually translate the W_{PoT} of the amendments by the following instructions:
   (a) Correct W_{PrT} in advance if it includes inadequate expressions;
   (b) Use W_{PrT} as a template of W_{PoT};
   (c) Edit only expressions relevant to the underlining in W_{PrO} and W_{PoO}.[4]

[4]We allow grammatical modifications (e.g., number agreement, tense agreement, article selection) in the expressions outside the amendment if they are triggered by it.

Figure 3 depicts this procedure.

As of April 2021, we compiled 1,483 differential translation examples from 62 amendment cases. These examples include the following modification instances:

- Phrase-level modifications: 786 replacements, 201 additions, and 89 deletions;
- Sentence-level modifications: 8 replacements, 11 additions, and 2 deletions.

5 Experiment

We experimentally evaluated the machine translation methods with our new resources.

5.1 Outline

For training data, we mixed two bilingual-statutory sentence corpora. One was made by Kozakai et al. (2017) from JLT. This corpus consists of 158,928 sentence pairs from 407 statutes. We compiled the other one from statutes in JLT that we collected in Step 1 in Section 4.2. Our corpus consists of 232,830 sentence pairs from 462 statutes.

We split our differential translation corpus into development data and test data by the statutes. The development and test data respectively consisted of 745 examples from 30 amendments and 738 examples from 32 amendments.

We used Transformer (Vaswani et al., 2017) for the NMT model under the following settings: six encoder/decoder hidden layers, eight self-attention heads, 512 hidden vectors, a batch size of eight, and an input sequence length of 256. We implemented the training and prediction codes based

Model	BLEU	RIBES	ISDIT	Focality
Naive Moses	47.93	61.75	29.32	51.54
Naive Koehn model	83.00	92.05	77.31	**91.20**
Naive Kozakai model	82.79	92.04	77.53	90.62
Naive Transformer	80.72	94.16	71.32	83.64
Transformer + Koehn model	82.39	94.70	75.05	86.66
Transformer + Kozakai model	82.46	94.75	74.69	86.42
Transformer + Koehn model + MC dropout	**84.43**	**96.04**	**79.33**	90.36
Transformer + Kozakai model + MC dropout	84.37	95.80	78.31	89.45
Transformer + W_{PoT} + MC dropout	86.62	96.72	81.95	90.92

Table 3: Experimental results

on the TensorFlow official model [5]. We used SentencePiece (Kudo and Richardson, 2018) as a tokenizer and set the vocabulary size to 8,192. We chose a dropout rate of 0.1 for training, which is the default setting of the official Transformer implementation. In the prediction phase, we executed the model with two dropout rates, 0.0 and 0.1, where a 0.0 dropout means that no dropout was applied. We investigated the optimal number of iterations from $\{100\mathrm{k}, 200\mathrm{k}, \cdots, 2{,}000\mathrm{k}\}$ using the development data.

The following are the settings of these template-aware SMTs: GIZA++ (Och and Ney, 2005) for the word alignment, SRILM (Stolcke, 2002) for the language model generation, and Moses (Koehn et al., 2007) for the decoder. We used MeCab (Kudo et al., 2004) for the Japanese tokenizer.

We evaluated the fluency and adequacy with BLEU and RIBES. For the focality evaluation, we utilized the focality scores (Yamakoshi et al., 2020) and our ISDIT. We set the maximum n-gram length N to 4 in calculating the focality scores, ISDIT, and BLEU. Using the four metrics, we compared the following translation models:

- Naive Moses (Koehn et al., 2007);
- Naive Koehn model (Koehn and Senellart, 2010);
- Naive Kozakai model (Kozakai et al., 2017);
- Naive Transformer;
- Transformer + Koehn model;
- Transformer + Kozakai model;
- Transformer + Koehn model + MC dropout (Yamakoshi et al., 2020);
- Transformer + Kozakai model + MC dropout (Yamakoshi et al., 2020);
- Transformer + W_{PoT} + MC dropout.[6]

"+" expresses a combination of techniques.

[5] https://github.com/tensorflow/models/

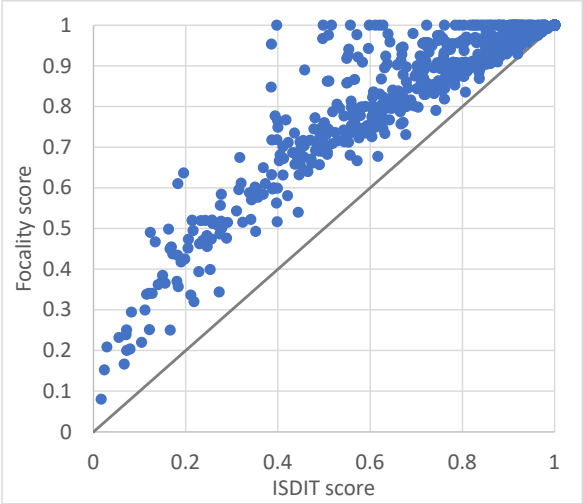

Figure 4: ISDIT and focality scores of each example

	RIBES	ISDIT	Focality
BLEU	0.724	0.960	0.833
RIBES	—	0.693	0.611
ISDIT	—	—	0.927

Table 4: Correlation coefficients between evaluation metrics

5.2 Results

Table 3 shows our experimental results. We achieved the same findings as those reported by Yamakoshi et al. (2020). The combination of Transformer, a template-aware SMT model, and MC dropout achieved the best performance in BLEU and RIBES among the comparisons; the naive template-aware SMT methods achieved the best performance in the focality scores; the

[6] We used W_{PoT} as an "oracular" interim reference.

Sort	Output
W_{PrO}	（ 火災共済協同組合の地区）
W_{PrT}	(district of a fire mutual aid cooperative)
W_{PoO}	（ 火災等共済組合等の地区）
W_{PoT}	(district of a fire and fire-related disaster mutual aid association , etc .)
Output	(district of a fire mutual aid cooperative , etc .)

Table 5: Example with distant ISDIT and focality scores

Model	Output
(W_{PrO})	協会及びその子会社から成る集団における業務の適正を確保するための体制
(W_{PrT})	A system to ensure the appropriateness of the operations in the group forming NHK and its subsidiary company
(W_{PoO})	次に掲げる体制その他の協会及びその子会社から成る集団の業務の適正を確保するための体制
(W_{PoT})	The systems listed below and a system to ensure the appropriateness of the operations of a group consisting of NHK and its subsidiary companies
Yamakoshi	The following systems and any other system to ensure the appropriateness of the operations of the group comprised of NHK and its subsidiary company:
Kozakai	A system to ensure the appropriateness of the operations of the group forming the following systems and any other association and its subsidiary company

Table 6: Translation example in our corpus

template-aware SMT and MC dropout were also both effective. One different finding from their report is that using the Koehn model generally worked more effectively than the Kozakai model. For our ISDIT metric, the combination methods of Yamakoshi et al. (2020) outperformed the naive template-aware SMT methods.

5.3 Discussion

First, we identified the characteristics of ISDIT. The plots in Fig. 4 indicate the focality and ISDIT scores of the Transformer + Kozakai model + MC dropout method (hereinafter "Yamakoshi method") for each translation example. The focality score of every example is higher than or equal to its ISDIT score. This result is natural because both these metrics share n-gram recall calculation, and ISDIT introduces n-gram precision that is more severe than the redundant penalty in the focality scores. We can observe many examples that have high focality scores but low ISDIT scores. Table 5 shows such an example. Yamakoshi method's output evaluated 100.0 focality scores and 39.74 ISDIT scores. In this example, however, their system failed to translate "等" in "火災等," which denotes a "fire-related disaster." This mistake greatly changed the system output from the reference, which suffered a low ISDIT score. On the other hand, since expressions shared by W_{PrT} and W_{PoT} were retained in the system output with no redundant generation, it received the maximum focality score.

Table 4 shows the correlation coefficients among the evaluation metrics. ISDIT and the focality scores have a high correlation coefficient of 0.927. ISDIT has also a strong relationship with BLEU, which is 0.960. High coefficients among them seem to come from a shared calculation strategy that utilizes the n-gram match rate.

Next we conducted a short qualitative analysis of our corpus. Table 6 shows a translation example. In this example, we replace "協会" (*kyokai*) with "次に掲げる体制その他の協会" (*tsugi ni kakageru taisei sonotano kyokai*). Its translation is divided into two parts: "the systems listed below and" (corresponding to "次に掲げる体制その他の") and "NHK" (corresponding to "協会"), which generally happens in Japanese partially amendments. The Kozakai method (also the Koehn method) cannot cope with this kind of examples: They put all the translation of the changed expression in W_{PoO} to the position where such changed expression appears in W_{PrT}.

Another tricky point in this case is the translation of "協会," which generally means "association." However, here it denotes "NHK" (Japan

Broadcasting Corporation). The Kozakai method failed to appropriately translate this word, possibly because it did not use the context of the translation target, "次に掲げる体制その他の協会." On the other hand, the Yamakoshi method successfully placed the new expression and adequately translated "協会." Its success reflects its use of the whole sentence in the translation.

6 Summary

We proposed a better evaluation scheme for Japanese partially amendment translations and developed a new metric called ISDIT that assesses the translation quality of both changed and unchanged expressions. We also compiled a corpus that secures the focality of translation. Using our corpus, we observed the characteristics of translation methods and ISDIT.

Our future work will increase the size of our corpus so that it can be used for neural network training, considering the publicization of the corpus. We will also identify the best weighting of the two factors in ISDIT. Third, we will consider applications of ISDIT to other domains of version-controlled documents such as contracts, technical documents, and product manuals.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 18H03492 and 21H03772.

References

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1–10.

Tsutomu Hirao, Hideki Isozaki, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2014. Evaluating translation quality with word order correlations. *Journal of Natural Language Processing*, 21(3):421–444. (In Japanese).

Hoseishitsumu-Kenkyukai. 2018. *Workbook Hoseishitsumu (newly revised second edition)*. Gyosei. In Japanese.

Phillip Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180.

Phillip Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.

Tadao Kozakai, Yasuhiro Ogawa, Tomohiro Ohno, Makoto Nakamura, and Katsuhiko Toyama. 2017. Shinkyutaishohyo no riyo niyoru horei no eiyaku shusei. In *Proceedings of NLP2017*, pages 1–4. (In Japanese).

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.

Franz Josef Och and Hermann Ney. 2005. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Yasuhiro Ogawa, Shintaro Inagaki, and Katsuhiko Toyama. 2008. Automatic consolidation of japanese statutes based on formalization of amendment sentences. *New Frontiers in Artificial Intelligence: JSAI 2007 Conference and Workshops, Revised Selected Papers, Lecture Notes in Computer Science*, 4914:363–376.

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.

Andreas Stolcke. 2002. SRILM — an extensible language modeling toolkit. In *Proceedings of ICSLP 2002*, pages 901–904.

Katsuhiko Toyama, Daichi Saito, Yasuhiro Sekine, Yasuhiro Ogawa, Tokuyasu Kakuta, Tariho Kimura, and Yoshiharu Matsuura. 2011. Design and Development of Japanese Law Translation System. In *Law via the Internet 2011*, pages 1–12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems 30*, pages 6000–6010.

Takahiro Yamakoshi, Takahiro Komamizu, Yasuhiro Ogawa, and Katsuhiko Toyama. 2020. Differential translation for japanese partially amended statutory sentences. In *Proceedings of the Fourteenth International Workshop on Juris-informatics*, pages 1–14.