Improving Japanese-English Patent Claim Translation with Clause Segmentation Models based on Word Alignment

Masato Nishimura[1] **Kosei Buma**[1] **Takehito Utsuro**[1] **Masaaki Nagata**[2]
[1]University of Tsukuba [2]NTT Communication Science Laboratories
{s2320779, s2520812}_@_u.tsukuba.ac.jp utsuro_@_iit.tsukuba.ac.jp
masaaki.nagata_@_ntt.com

Abstract

In patent documents, patent claims represent a particularly important section as they define the scope of the claims. However, due to the length and unique formatting of these sentences, neural machine translation (NMT) systems are prone to translation errors, such as omissions and repetitions. To address these challenges, this study proposes a translation method that first segments the source sentences into multiple shorter clauses using a clause segmentation model tailored to facilitate translation. These segmented clauses are then translated using a clause translation model specialized for clause-level translation. Finally, the translated clauses are rearranged and edited into the final translation using a reordering and editing model. In addition, this study proposes a method for constructing clause-level parallel corpora required for training the clause segmentation and clause translation models. This method leverages word alignment tools to create clause-level data from sentence-level parallel corpora. Experimental results demonstrate that the proposed method achieves statistically significant improvements in BLEU scores compared to conventional NMT models. Furthermore, for sentences where conventional NMT models exhibit omissions and repetitions, the proposed method effectively suppresses these errors, enabling more accurate translations.

1 Introduction

The claims in patent documents are critically important for defining the scope of patent rights. However, due to the length and unique descriptive style of these sentences, neural machine translation (NMT) models often encounter issues such as omissions and repetitions in translation. Figure 1 shows the distribution of subword token lengths for Japanese patent claims included in the Japanese-English patent parallel corpus JaParaPat (Nagata et al., 2024a) and for Japanese sentences commonly used in the ASPEC (Nakazawa et al., 2016) Japanese-English parallel corpus. Comparing the two reveals that the patent parallel corpus used in this study has a higher proportion of long sentences compared to scientific paper's abstract. The divide-and-conquer translation approach is known to be an effective method for addressing challenges in long sentences translation. Sudoh et al. (2010) proposed a method in statistical machine translation (SMT) that segments input sentences into clause units based on syntactic parsing, translates each clause separately, and then reorders them according to their hierarchical structure. This approach was shown to improve translation accuracy. Applying this divide-and-conquer approach to neural machine translation (NMT), Kano (2022) proposed a "divide-and-conquer neural machine translation" method for English-Japanese translation, which divides input sentences into clauses based on syntactic parsing and reassembles them after translation. While this method demonstrated the potential to improve translation accuracy, challenges remained in selecting appropriate clause units and ensuring accurate reassembly after clause translation.

In response, Ishikawa (2024) addressed two challenges highlighted in the document (Kano, 2022): the selection of clause segmentation units and the translation accuracy of clauses after segmentation. They sought to improve translation accuracy by adopting clause segmentation based on conjunctions and utilizing mBART (Liu et al., 2020), a pre-trained model, for both the clause translation model and the reordering/editing model. Additionally, they attempted to enhance clause translation accuracy by fine-tuning the clause translation model with pseudo-parallel data at the clause level. Experiments showed a significant reduction in excessively long translations, as well as suppression of

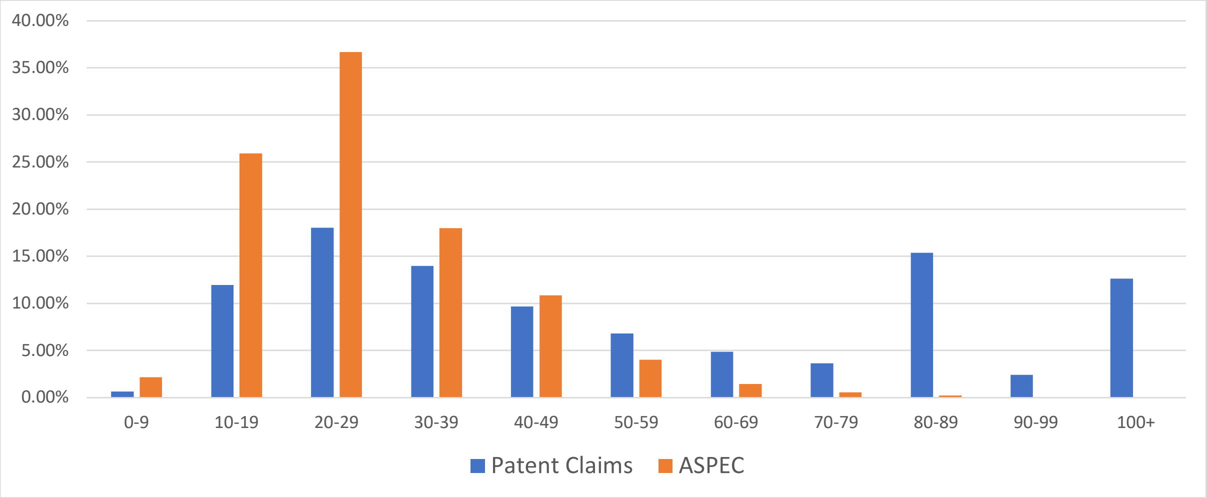


Figure 1: Comparison of Sentence Lengths between ASPEC and Patent Claim Test Data

hallucinations and repetitions. These findings suggest that the divide-and-conquer translation method has the potential to mitigate translation errors commonly caused by conventional NMT models in long sentences translation.

However, the study (Ishikawa, 2024) has some limitations. One issue is that clause segmentation based on conjunctions sometimes fails to divide long sentences into appropriately short clauses. Another issue is that the clause translation model was trained using pseudo-parallel data rather than real parallel data collected from actual sources.

Based on the above, this study proposes a novel approach to the divide-and-conquer translation method, specifically targeting Japanese-English translation of patent claims, which differs from previous studies (Kano, 2022; Ishikawa, 2024). In this method, we introduce a clause segmentation model that divides the source patent claim sentences into clauses optimized for translation by the model. In particular, this study ensures that the clause units, determined based on word alignments in parallel texts, are consistent between the two languages. By doing so, the proposed method suppresses errors such as omissions and repetitions in the final translations, enabling the generation of more accurate translations. Specifically, we propose a method to generate high-quality clause-level parallel data from the original parallel corpus using a word alignment tool. Furthermore, we propose a method to construct the following three models using the generated clause-level parallel corpus:

1. A clause segmentation model that divides the source Japanese sentences into clause units.

2. A clause translation model specialized for clause-level translation.

3. A reordering and editing model that rearranges and edits the translated clauses to generate the final translated text.

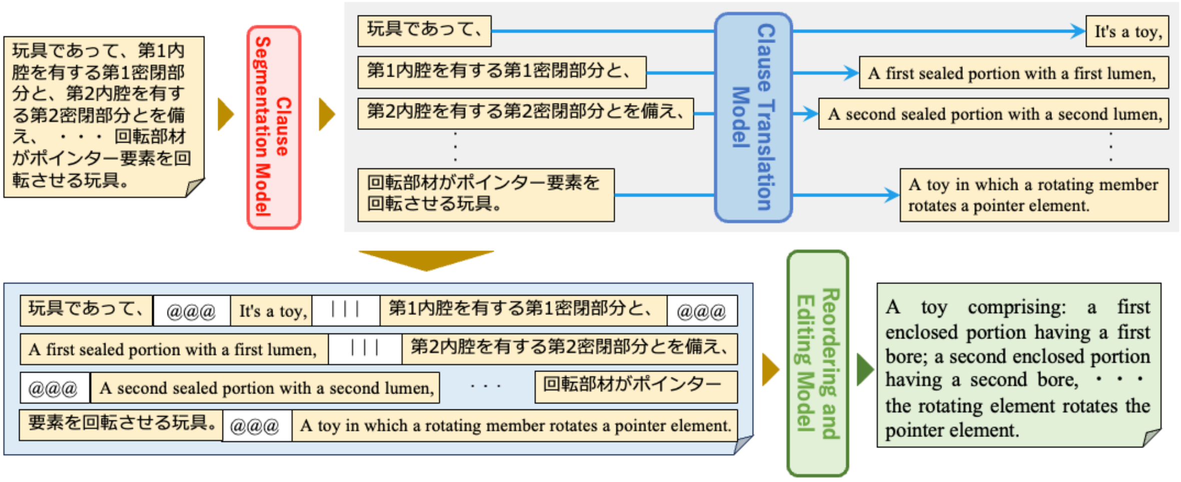
In the experiments, the proposed translation method, which integrates these models, was evaluated using the Japanese-English patent parallel corpus JaParaPat (Nagata et al., 2024a). The results demonstrated that the proposed method achieved statistically significant improvements in BLEU scores compared to conventional NMT models for Japanese-English translation of patent claims. Furthermore, compared to the translation results of conventional models, it was confirmed that the proposed method effectively suppresses omissions, resulting in more accurate translations.

2 Related Work

2.1 Long Sentences Translation

Various approaches have been explored to address the challenges of long sentences translation. Sudoh et al. (2010) adopted a divide-and-conquer translation method in statistical machine translation for translating long sentences. They divided input sentences into clause units based on syntactic parsing, translated them, and reordered the results using the hierarchical structure of the clauses, thereby improving translation accuracy.

In NMT, Pouget-Abadie et al. (2014) proposed an automatic segmentation method, which splits long sentences into clauses, translates each clause



玩具であって、第1内腔を有する第1密閉部分と、第2内腔を有する第2密閉部分とを備え、・・・ 回転部材がポインター要素を回転させる玩具。
Clause Segmentation Model
玩具であって、
第1内腔を有する第1密閉部分と、
第2内腔を有する第2密閉部分とを備え、
回転部材がポインター要素を回転させる玩具。
Clause Translation Model
It's a toy,
A first sealed portion with a first lumen,
A second sealed portion with a second lumen,
A toy in which a rotating member rotates a pointer element.
玩具であって、 @@@ It's a toy, ||| 第1内腔を有する第1密閉部分と、 @@@ A first sealed portion with a first lumen, ||| 第2内腔を有する第2密閉部分とを備え、 @@@ A second sealed portion with a second lumen, · · · 回転部材がポインター要素を回転させる玩具。 @@@ A toy in which a rotating member rotates a pointer element.
Reordering and Editing Model
A toy comprising: a first enclosed portion having a first bore; a second enclosed portion having a second bore, · · · the rotating element rotates the pointer element.

Figure 2: The Prediction Framework of Proposed Model

individually, and then reassembles them sequentially. This method utilizes an RNN to predict the optimal segmentation points for dividing long sentences into parts that are easier for the model to translate. However, this approach was designed for English-French translation. When applied to language pairs with significantly different word order, such as Japanese-English translation, it often resulted in unnatural word order during reassembly.

To address this issue, Kano (2022) developed a neural network model that divides long sentences into smaller segments for translation and rearranges the translated clauses into the appropriate order in English-Japanese translation. Furthermore, Ishikawa (2024) proposed a novel segmentation method for English clauses based on coordinating conjunctions, along with a training method for a model that references the context of the sentence during the translation of the clause. This approach achieved improvements in translation accuracy for long English-Japanese sentences.

2.2 Translation of Patent Claims

Fuji et al. (2015) proposed a method for translating English, Chinese, and Japanese patent claims using statistical machine translation (SMT). Their approach involved manually constructing synchronous context-free grammar rules for sentence structure transformation. These rules were then used to convert the sentence structure of the source language into that of the target language, addressing the unique descriptive style commonly found in patent claims. However, this method has a limitation: the need for manual rule creation makes it difficult to flexibly adapt to new descriptive styles.

3 Method

Figure 2 illustrates the overall structure of the proposed method. The source Japanese patent claim is first divided into multiple clauses using the clause segmentation model. Each clause is then translated by the clause translation model, and finally, the translated clauses are integrated by the reordering and editing model to generate the English patent claim. This method aims to suppress omissions and repetitions that are commonly encountered in conventional NMT models.

It is worth noting that the term *clause* used in this study does not refer to syntactic clauses in the traditional linguistic sense. As shown in Figure 2, in our approach, Japanese sentences are first segmented at punctuation marks, and adjacent segments are then grouped based on word alignments to ensure semantic correspondence with the English side. Thus, we define clauses as semantically coherent segments that preserve consistency between source and target languages. This operational definition aims to support alignment quality rather than adhere strictly to syntactic boundaries.

3.1 Clause-Level Parallel Corpus

In this study, we propose a method for automatically generating a clause-level parallel corpus, inspired by the approach of Zhang and Matsumoto (2019), which generates parallel sub-sentences from long parallel sentence data. This method obtains word alignment information from sentence-level parallel corpora using the word alignment tool WSPAlign (Wu et al., 2023). Based on the word alignment information, corresponding clauses within sentences are extracted to generate clause-level parallel data.

The clause-level parallel corpus is created using the following procedure. Following the report by Zhang and Matsumoto (2019), we set the word inclusion ratio threshold to 0.5.

1. Use WSPAlign to obtain word alignments for the parallel sentences in the patent parallel data.

2. Split the Japanese and English sentences into multiple clauses at the positions of delimiters such as “、”, “，”, “。”, “．”, “；”, and “：”.

3. Calculate the word inclusion ratio for each pair of parallel clauses based on the word alignment information. If the ratio exceeds 0.5, the clauses are determined to have a alignment. The word inclusion ratio is defined as the proportion of words in a Japanese clause $s\text{-}seg_i$ that are aligned, based on word alignment, to words in the corresponding English clause $t\text{-}seg_j$. In cases where none of the Japanese clauses have a word inclusion ratio larger than 0.5 with any English clause, no clause pairs are created from that sentence pair.

4. For cases where clause alignments are one-to-many or many-to-many, merge the multiple clauses into a single clause on one side to ensure a one-to-one alignment.

By applying the above procedure to parallel sentences extracted from a patent parallel corpus, we generate the clause-level parallel corpus.

architecture	transformer_wmt_en_de_big
enc-dec layers	6
optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98$)
learning rate schedule	inverse square root decay
warmup steps	4,000
max learning rate	0.001
dropout	0.3
gradient clip	0.1
batch size	1M tokens
max number of updates	60K steps
validate interval updates	1K steps
patience	5

Table 1: List of hyperparameters for the Transformer

3.2 Clause Segmentation Model

In this study, we developed a clause segmentation model based on ERSATZ, a sentence segmentation model proposed by Wicks and Post (2021). The model was trained to perform segmentation at the clause level. ERSATZ formulates sentence segmentation as a binary classification task, predicting whether periods (e.g., “。” or “．”) indicates the “middle of a sentence” or the “end of a sentence”. To extend this functionality for clause segmentation, we modified the model to use commas (e.g., “、” or “，”) as candidate punctuation marks for clause boundaries.[1] The training data for the model utilized the clause-level parallel corpus proposed in Section 3.1.

The training data was prepared by extracting Japanese clauses from the clause-level parallel corpus and labeling punctuation marks at clause boundaries (e.g., commas) with end-of-clause labels. This enabled the creation of a model capable of segmenting Japanese patent claims into clauses based on word alignment information.

3.3 Clause Translation Model

In this study, to create a clause translation model specialized for clause-level translation, we fine-tuned a pre-trained Japanese-English translation model, initially built using a patent parallel corpus as was also the case in prior studies, with the clause-level parallel corpus generated by the method described in Section 3.1. The experimental settings for the clause translation model, summarized in Table 1, follow those used in JaParaPat (Nagata et al., 2024b). The clause translation model aims to suppress the tendency to infer or supplement contextual information that may be lost due to segmentation, thereby enabling more accurate translations of the segmented clauses.

3.4 Reordering and Editing Model

The purpose of using a reordering and editing model is to reconstruct multiple translated clauses produced by the clause translation model into a single English sentence as the target language sentence. Since the word order in Japanese and English differs significantly, merely dividing a Japanese sentence into clauses and connecting them would not adequately handle the word order

[1] In practice, to perform clause segmentation at positions within parentheses that represent supplementary explanations, the model utilizes both commas (e.g., “、” or “，”) and sentence-ending punctuation marks (e.g., “。” or “．”).

Model	**Data Used**	**Number of Data**
Baseline Model	JaParaPat2016-2020	61,364,685 sentence pairs
Clause Segmentation Model	Clause-Level Parallel Corpus(claims)	200,462 sentence
Clause Translation Model	JaParaPat2016-2019 Clause-Level Parallel Corpus	49,474,547 sentence pairs 5,480,682 clause pairs

Reordering and Editing Model	JaParaPat2016-2020(Bidirectional)
	JaParaPat2016-2020(claims)
		109,028,682 sentence pairs
		2,613,107 sentence pairs

Table 2: Overview of Data Used for the Baseline Model and Proposed Method

Evaluation Target	Overall		Long Sentences	
	BLEU ↑	MetricX-24 ↓	BLEU ↑	MetricX-24 ↓
Baseline Model	55.5	2.90	50.1	4.77
Ishikawa	56.3	2.89	51.1	4.76
Proposed Method	**56.6****	**2.84**	**51.6****	**4.69**

Table 3: BLEU Scores and MetricX-24 Scores for Each Evaluation Target. ** indicates a significant difference (p<0.01) in BLEU Scores between the Baseline Translation Model and the Proposed Method.

transformation between these languages. Therefore, the reordering and editing model is expected to rearrange the translated clauses into the appropriate word order during the process of connecting them. An example of reordering and editing is shown at the bottom of Figure 2.

The training data for the reordering and editing model is prepared by segmenting sentences in the corpus using the clause segmentation model. The segmented Japanese clauses, along with their translated English clauses, are concatenated to form the input data, while the original English sentences from the corpus are used as the target data. Special tokens, "@@@" and "|||", are added to the model's vocabulary. The token "@@@" is used to connect a Japanese clause with its corresponding translated English clause, while "|||" is used to link pairs of these clause segments. The reason for structuring the input data this way is to preserve information about the relationships between the translated English clauses by including the original Japanese sentence. If only the translated English clauses were used as input, information about the relationships between the clauses would be lost. Adding the Japanese text provides additional context.

Since the input to the reordering and editing model contains words in both Japanese and English, it requires an understanding of both languages. Therefore, the reordering and editing model is created by fine-tuning a Japanese-English bidirectional translation model using the training data prepared as described above.

4 Experiments

4.1 Experimental Setup

In this study, experiments on Japanese-English translation were conducted using the JaParaPat Japanese-English patent parallel corpus (Nagata et al., 2024a). The data used for the experiments consisted of full-text patent parallel data from 2016 to 2020 as the training data, and patent claim parallel data from the first half of 2021 as the test data.

The machine translation software used fairseq (Ott et al., 2019), and the Transformer big (Vaswani et al., 2017) architecture was employed for the baseline model, clause translation model, and reordering and editing model. Sentence tokenization was performed using sentencepiece (Kudo and Richardson, 2018). The model was trained on 10M randomly sampled sentence pairs from the patent parallel data. The vocabulary size was set to 32K for both Japanese and English. Additionally, the clause segmentation model was trained using ERSATZ[2].

Table 2 provides an overview of the data used to train the baseline model and the three proposed models. The clause-level parallel corpus was created by obtaining word alignment information us-

[2] https://github.com/rewicks/ersatz

(a) the Entire Test Set

		Baseline Model		
		2 or more	2–0.5	0.5 or less
Proposed Method	2 or more	1,055	376(*)	4
	2–0.5	273(**)	234,489	900(##)
	0.5 or less	2	515(#)	1,217

(b) the Subset of Inputs with less than 100 Tokens

		Baseline Model		
		2 or more	2–0.5	0.5 or less
Proposed Method	2 or more	651	279(*)	3
	2–0.5	194(**)	202,298	528(##)
	0.5 or less	2	236(#)	695

(c) the Subset of Inputs with 100 to 150 Tokens

		Baseline Model		
		2 or more	2–0.5	0.5 or less
Proposed Method	2 or more	137	41(*)	0
	2–0.5	32(**)	16,155	97(##)
	0.5 or less	0	62(#)	95

(d) the Subset of Inputs with more than 150 Tokens

		Baseline Model		
		2 or more	2–0.5	0.5 or less
Proposed Method	2 or more	265	53(*)	1
	2–0.5	47(**)	15,823	274(##)
	0.5 or less	0	217(#)	424

Table 4: Omission and Repetition Analysis (Proposed Method vs. Baseline Model) on the Entire Test Set

ing WSPAlign[3] for half of the 2020 data (5,976,295 sentence pairs) and following the method described in Section 3.1. This process resulted in a clause-level parallel corpus containing 5,480,682 clause pairs. For training the clause segmentation model, Japanese clause data was created by segmenting Japanese patent claims in the clause-level parallel corpus. The clause translation model was pre-trained on the full-text patent parallel data from 2016 to 2019 and fine-tuned using the entire clause-level parallel corpus. The reordering and editing model was pre-trained on bidirectional full-text patent parallel data from 2016 to 2020 and fine-tuned using training data created by applying the methods described in Section 3.4 to Japanese patent claims from 2016 to 2020, segmented and translated using the clause segmentation and translation models.

To compare with conventional divide-and-conquer neural machine translation methods, we reproduced Ishikawa (2024)'s approach, which involves clause segmentation based on conjunctions and fine-tuning a clause translation model with pseudo-parallel data at the clause level, adapting it for Japanese-to-English translation of patent claims. The training data used for this reproduction was within the same range as the data used to train the three models in the proposed method. This comparison allows us to evaluate the effectiveness of using the clause segmentation model adopted in the proposed method and the clause-level parallel corpus created using word alignment information.

For evaluation, BLEU (Papineni et al., 2002) was

[3] https://github.com/qiyuw/WSPAlign

used as the primary metric, calculated with sacreBLEU[4] (Post, 2018). Since accurate translation of technical terms is critical in patent translation, BLEU was selected as the main evaluation criterion in this study.

To evaluate whether the proposed method can suppress translation errors such as omissions and repetitions, we conducted an assessment using MetricX-24[5] (Juraska et al., 2024). MetricX-24 is a machine translation evaluation metric developed by Google based on a regression model that predicts MQM (Multidimensional Quality Metrics) scores (Lommel et al., 2014). Traditional machine translation evaluation metrics, such as COMET (Rei et al., 2022a) and CometKiwi (Rei et al., 2022b), are trained on Direct Assessment (DA) scores and are highly effective in measuring semantic adequacy. However, MQM scores allow for weighting different types of translation errors, making them more suitable for evaluating issues such as omissions and repetitions. Furthermore, MetricX-24 has demonstrated high robustness against translation errors, including omissions and repetitions, by leveraging mixed training on synthetic error data and DA/MQM data. In this study, we used the MetricX-24-Hybrid-XL[6] model for evaluation.

4.2 Results

4.2.1 Accuracy Evaluation

In this study, the performance of the proposed method was evaluated using a test set consisting of patent claims (238,902 sentences) extracted from 2021 patent data. Table 3 showed that the proposed method achieved a BLEU score of 56.6, which statistically significantly outperformed the baseline model's score of 55.5 ($p < 0.01$). This confirmed that the proposed method improves overall translation accuracy.

Additionally, the performance was evaluated on a subset of the test set containing only long sentences with more than 100 subword tokens in the source Japanese text. For this subset, the proposed method recorded a BLEU score of 51.6, statistically significantly exceeding the baseline model's score of 50.1. While the overall test set showed an improvement of 1.1 points, the improvement for long sentences was 1.5 points, indicating that the proposed method achieved greater improvement for longer sentences.

The results using MetricX-24 showed that the proposed method achieved a score of 2.84, compared to 2.90 for the baseline model. In MetricX-24, lower scores indicate fewer translation errors, such as omissions and repetitions. This suggests that the proposed method effectively suppresses translation errors in patent claim translations, including omissions and repetitions.

Next, a comparison was made between the proposed method and conventional divide-and-conquer neural machine translation methods. For the conventional method, the BLEU score was 56.3 points, and the MetricX-24 score was 2.89 points. In contrast, the proposed method achieved a BLEU score of 56.6 and a MetricX-24 score of 2.84, outperforming the conventional method in both metrics. Both the baseline and our proposed method use parallel data extracted from the same portion of JParaPat.Our method differs from previous divide-and-conquer approaches in a key aspect: whereas prior methods typically rely solely on the syntactic structure of the source language—often segmenting at coordinating conjunctions—our proposed approach leverages word alignments to identify clause boundaries based on source–target correspondence. This alignment-based segmentation results in divisions that are more suitable for translation. These results confirm that, compared to the conventional divide-and-conquer neural machine translation method, the clause segmentation model and the clause-level parallel corpus leveraging word alignment information employed in the proposed method contribute to improved accuracy in divide-and-conquer neural machine translation.

4.2.2 Analysis of Omissions and Repetitions

To further analyze whether the proposed method can produce more accurate translations with fewer errors such as omissions and repetitions, the sentence length ratios between the translated text and the reference text were calculated for both the baseline model and the proposed method. These ratios were categorized into three groups: "2 or more," "2–0.5," and "0.5 or less," and their trends were observed. The classification results for the entire test set (238,902 sentences) are shown in Table 4 (a). Additionally, the test set was grouped by the token length of the input sentences into "less than 100," "100–150," and "more than 150," with the classification results for each group shown in Ta-

[4] https://github.com/mjpost/sacrebleu

[5] https://github.com/google-research/metricx

[6] https://huggingface.co/google/metricx-24-hybrid-xl-v2p6

bles 4 (b), 4 (c), and 4 (d), respectively. Within these tables, special attention was given to the four categories where one model produced translations with omissions or repetitions while the other model performed well: "middle column, upper row(*)," "left column, middle row(**)," "middle column, lower row(#)," and "right column, middle row(##)."

The analysis showed that in all four tables, the "left column, middle row(**)," which represents cases where the proposed method successfully avoided repetitions that the baseline model did not, occurred less often than the "middle column, upper row(*)." This suggests that the baseline model had fewer cases of repetition overall. On the other hand, for omissions, the "middle column, lower row(#)," where the proposed method avoided omissions, occurred less frequently than the "right column, middle row(##)." This indicates that the proposed method was better at reducing omissions compared to the baseline model.

Examples of improved translations addressing omissions by the proposed method are shown in Table 6. The baseline model in Table 6 (a), parts of the input sentence, such as "preferably by a length of the heat exchanger" and "finned tube shape, coiled shape, and/or fin shape", were not translated despite being present in the original Japanese sentence. Additionally, in patent claims, reference numerals in drawings are typically enclosed in parentheses, as seen in "The cryogenic refrigeration system (1)". However, in the baseline model's translation, the number inside the parentheses was omitted. In contrast, the proposed method not only translates the entire input Japanese sentence without missing any information but also correctly retains the numerical references within parentheses. As a result, it produces a more appropriate translation for patent claims. Similarly, in Table 6 (b), the baseline model fails to translate some words in input sentence such as "such as methanol, ethanol," whereas the proposed method correctly translates all examples. These results indicate that, compared to the baseline model, the proposed method preserves all necessary information in patent claim translations and produces more accurate outputs.

Examples of improved translations addressing repetitions by the proposed method are shown in Table 7. In the baseline model, the term "cantilever shaped" was excessively repeated, whereas no such repetition occurred with the proposed method.

4.2.3 Impact of Pre-training the Reordering and Editing Model

In this study, bidirectional Japanese-English parallel data was used for pre-training the reordering and editing model. Experiments were conducted to evaluate the effect of this pre-training on the accuracy of the final reordering and editing model. The parallel data used for pre-training consisted of JaParaPat data from 2016 to 2020, and two models were created: one trained with Japanese-English parallel data and the other with bidirectional parallel data. These models were fine-tuned using the same reordering and editing model training data, and their performance was compared.

The BLEU evaluation results, obtained using test data comprising patent claims extracted from 2021 patent data, are shown in Table 5. The results show that the reordering and editing model pre-trained with bidirectional Japanese-English parallel data achieved a BLEU score of 56.6, statistically significantly outperforming the model pre-trained only in the Japanese-to-English direction, which scored 55.0 ($p < 0.01$). The results suggest that understanding both Japanese and English is critical for the reordering and editing model. Furthermore, using bidirectional Japanese-English parallel data for pre-training improves the accuracy of reordering and editing.

Table 5: BLEU scores of the Reordering and Editing Model with different pre-training data: comparison between Unidirectional (Japanese-English) and Bidirectional (Japanese-English) parallel data. ** indicates a significant difference (p<0.01) in BLEU scores.

Data used for Pre-Training	**BLEU**
Unidirectional	55.0
Bidirectional	**56.6****

4.2.4 Evaluation of the Clause Segmentation Model

The clause segmentation model developed in this study was evaluated to determine its ability to accurately segment Japanese patent claim sentences. For the evaluation, 2,000 sentences were sampled from 238,902 patent claim sentences extracted from 2021 patent data. First, word alignment information was obtained for the 2,000 sentences using WSPAlign, and the sentences were segmented into clauses based on the method described in Section 3.1. Ground truth data was then created by assigning end-of-sentence labels to punctuation

(a) Example 1

Omissions
Input Sentence 前記温度因子及び/又は前記NTUが、前記熱交換器(3)の伝熱面積によって、好ましくは前記熱交換器の長さによって提供され、前記熱交換器(3)が、好ましくは、フィン付きチューブ形状、コイル形状、及び/又はフィン形状であり、前記流路(2)の円周を少なくとも部分的に取り囲む、 請求項2に記載の極低温冷凍システム(1) 。
Reference Translation Cryogenic refrigeration system (1) according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger (3), preferably by a length of the heat exchanger, wherein the heat exchanger (3) is preferably of a finned tube shape, coiled shape, and/or fin shape and at least partially surrounds a circumference of the conduit (2).
Baseline Model (BLEU: 15.65, COMET: 68.94, MetricX-24: 2.82) The cryogenic refrigeration system according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger.
Proposed Method (BLEU: 78.24, COMET: 84.31, MetricX-24: 2.12) The cryogenic refrigeration system (1) according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger (3), preferably by a length of the heat exchanger, wherein the heat exchanger (3) is preferably finned tube-shaped, coil-shaped and/or fin-shaped and at least partially surrounds a circumference of the flow channel (2).

(b) Example 2

Omissions
Input Sentence 有機溶媒1が、アルコール溶媒、例えば、メタノール、エタノール、n−プロパノール、イソプロパノール、n−ブタノール、イソブタノール;エステル溶媒、例えば、酢酸メチル、酢酸エチル、酢酸プロピル、酢酸イソプロピル、酢酸ブチル;ケトン溶媒、例えば、アセトンおよびブタノン;またはその混合物である、請求項9に記載の方法。
Reference Translation The method according to claim 9, wherein the organic solvent 1 is an alcohol solvent, such as methanol, ethanol, n-propanol, isopropanol, n-butanol, isobutanol; an ester solvent, such as methyl acetate, ethyl acetate, propyl acetate, isopropyl acetate, butyl acetate; a ketone solvent, such as acetone and butanone; or a mixture thereof.
Baseline Model (BLEU=21.11, COMET=60.85, MetricX-24=2.74) The method according to claim 9, wherein the organic solvent 1 is an alcohol solvent; an ester solvent; a ketone solvent; or a mixture thereof.
Proposed Method (BLEU=88.99, COMET=90.98, MetricX-24=1.77) The method according to claim 9, wherein the organic solvent 1 is an alcohol solvent such as methanol, ethanol, n-propanol, isopropanol, n-butanol, isobutanol; an ester solvent such as methyl acetate, ethyl acetate, propyl acetate, isopropyl acetate, butyl acetate; a ketone solvent such as acetone and butanone; or a mixture thereof.

Table 6: Examples of Omission Improvements by the Proposed Method

marks at the segmentation points, which served as the test set for evaluating the accuracy of the clause segmentation model. Next, the clause segmentation model created in Section 4.1 was applied to the test set's Japanese patent claim sentences to perform clause segmentation. The segmentation points predicted by the model were compared to the ground truth segmentation points, and the F1

Repetitions

Input Sentence
前記流路遮断バルブは、 内部に前記閉鎖部材を収容し、カンチレバー形状からなる少なくとも1つの片持ちばりを備え、円筒形に形成されて、前記連通流路の流入口に挿入されるように設置されるボディー;及び、 一側は前記片持ちばりから突出形成される係止部により支持され、他側は前記閉鎖部材と接触するように設置されるリング部材;を含み、 前記リング部材は設定された温度以上になると、前記閉鎖部材が中心部を通過するように変形されて、前記閉鎖部材を前記連通流路の内部に向けて移動させる、 ことを特徴とする、請求項17に記載のバルブアセンブリ。

Reference Translation
The valve assembly of claim 17, wherein the flow path blocking valve includes:a body for containing the blocking member therein, and providing at least one cantilever portion formed in a cantilever shape, the body formed in a cylindrical shape and disposed to be inserted into the inlet of the communication flow path; anda ring member having one side supported by a locking portion protruding from the cantilever portion and the other side disposed to contact the blocking member,wherein the ring member moves the blocking member towards the inside of the communication flow path by deforming the blocking member to pass through a central part of the ring member when the internal temperature exceeds the preset temperature.

Baseline Model (BLEU=7.63, COMET=34.41, MetricX-24=8.48)
The valve assembly of claim 17, wherein the flow path shut-off valve comprises:a body configured to accommodate the closing member therein, having at least one cantilever shaped

Proposed Method (BLEU=43.72, COMET=81.22, MetricX-24=4.82)
The valve assembly of claim 17, wherein the passage shutoff valve comprises:a body accommodating the closing member therein, having at least one cantilever formed in a cantilever shape, formed in a cylindrical shape, and installed to be inserted into the inlet of the communication passage; anda ring member having one side supported by a locking portion formed to protrude from the cantilever and the other side installed to be in contact with the closing member,wherein the ring member is deformed such that the closing member passes through a central portion and moves the closing member toward the inside of the communication passage when a set temperature or higher is reached.

Table 7: Examples of Repetition Improvements by the Proposed Method

score was calculated. The model achieved an F1 score of 98.16. These results demonstrate that the clause segmentation model developed in this study can accurately reproduce clause segmentation by effectively utilizing word alignment information.

5 Conclusion

In this study, we proposed a translation method to address the translation errors of "omissions" and "repetitions" which are common challenges in Japanese-English translation of patent claims. The method focuses on the fact that patent claims are often long and have unique structures, utilizing a clause segmentation model to divide patent claims into more translatable units.

The experimental results demonstrated that the proposed method achieved statistically significant improvements over the baseline model in BLEU scores. Notably, it showed remarkable improvements even for sentences prone to omissions and repetitions. These results confirm the effectiveness of the proposed method in resolving issues of omissions and repetitions in the translation of patent claims.

An important direction for future work is to extend the evaluation to include state-of-the-art large language models (LLMs), which we plan to pursue in our subsequent research.

References

Masaru Fuji, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita, and Yuji Matsumoto. 2015. Patent claim translation based on sublanguage-specific sentence structure. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

R. Ishikawa. 2024. Divide-and-conquer neural machine translation with insentence context. *Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology*.

J. Juraska, D. Deutsch, M. Finkelstein, and M. Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proc. 9th WMT*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Y. Kano. 2022. Improving neural machine translation by syntax-based segmentation. *Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology*.

T. Kudo and J. Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *In Proc. EMNLP*, pages 66–71.

A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Y. Liu et al. 2020. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742.

A. Lommel, A. Burchardt, and H. Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

M. Nagata, M. Morishita, K. Chousa, and N. Yasuda. 2024a. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In *Proc. LREC-COLING*, pages 9452–9462.

M. Nagata, M. Morishita, K. Chousa, and N. Yasuda. 2024b. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In *Proc. LREC-COLING 2024)*, pages 9452–9462.

T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. *In LREC2016*, pages 2204–2208.

M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*, pages 48–53.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pages 311–318.

M. Post. 2018. A call for clarity in reporting BLEU scores. In *Proc. 3rd WMT*, pages 186–191.

J. Pouget-Abadie, D. Bahdanau, B. van Merrien-Boer, K. Cho, and Y. Bengio. 2014. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. *In Proc. 8th SSST*, pages 78–85.

R. Rei, J. C. de Souza, D. Alves, C. Zerva, A. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proc. 7th WMT*, pages 578–585.

R. Rei et al. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proc. 7th WMT*, pages 634–645.

K. Sudoh, K. Duh, H. Tsukada, T. Hirao, and M. Nagata. 2010. Divide and translate: Improving long distance reordering in statistical machine translation. In *Proc. 5th WMT*, pages 418–427.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proc. 30th NIPS*, pages 5998–6008.

R. Wicks and M. Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *In Proc. 59th ACL*, pages 3995–4007.

Q. Wu, M. Nagata, and Y. Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proc. 61st ACL*, pages 11084–11099.

J. Zhang and T. Matsumoto. 2019. Corpus augmentation for neural machine translation with chinese-japanese parallel corpora. *Applied Sciences*, 9(10).

A Sustainability Statement

A.1 CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO_2eq/kWh. A cumulative of 1,000 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W).

Total emissions are estimated to be 129.6 kgCO_2eq of which 0 percents were directly offset.

Estimations were conducted using the Machine-Learning Impact calculator presented in Lacoste et al. (2019).