

# Findings of the First Patent Claims Translation Task at WAT2025

**Toshiaki Nakazawa**

The University of Tokyo

nakazawa@nlab.ci.i.u-tokyo.ac.jp

**Takashi Tsunakawa**

Shizuoka University

tuna@inf.shizuoka.ac.jp

**Isao Goto**

Ehime University

goto.isao.fn@ehime-u.ac.jp

**Kazuhiro Kasada**

Japan Patent Information Organization

kasada.kazuhiro.cn@japio.or.jp

**Katsuhito Sudoh**

Nara Women’s University

sudoh@ics.nara-wu.ac.jp

**Shoichi Okuyama**

Patent Attorney in Japan

okuyama@quon-ip.jp

**Takashi Ieda**

Japan Institute for

Promoting Invention and Innovation

t-ieda@jiii.or.jp

**Masaaki Nagata**

NTT Inc.

masaaki.nagata@ntt.com

## Abstract

This paper presents the results and findings of the first shared task of translating patent claims. We provide training, development, and test data for participants and perform human evaluation of the submitted translations. This time, 2 teams submitted their translation results. Our analysis of the human-annotated translation errors revealed not only general, domain-independent errors but also errors specific to patent translation. We also found that the human annotation itself exhibited some serious issues. In this paper, we report on these findings.

## 1 Introduction

The performance of machine translation using Neural Machine Translation (NMT) and Large Language Models (LLMs) has improved dramatically and in some cases can even surpass human translation depending on the language and domain. However, currently there is no universal method to accurately evaluate the performance of machine translation. Even widely used metrics such as COMET (Rei et al., 2020) have been reported to yield unstable or inaccurate evaluation results (Kocmi et al., 2025) when applied to translations of texts from domains other than those used in COMET’s training.

The same applies to the translation of patent documents. Although the average translation quality has improved significantly, it remains difficult

to accurately evaluate aspects such as appropriate terminology usage and term consistency. In particular, patent claims present additional challenges due to their length and distinctive writing style, making an accurate evaluation even more difficult.

Therefore, we conducted a Shared Task focusing on translating Japanese-English patent claims<sup>1</sup>. The goal is not only to compete on translation quality, but also to ultimately develop an automatic evaluation method that can accurately assess translation results.

For this first iteration, our primary objective is to collect translation outputs produced by various methods and annotate them with human-identified errors, thereby creating training data for future development of models capable of accurately performing automatic evaluation of translations in the patent domain.

## 2 Training Data

We used the publicly available subset of JaParaPat, the Japanese-English Parallel Patent Application Corpus (Nagata et al., 2024), as the training data for the shared task. In August 2025, the authors released a subset of JaParaPat, covering the period from 2016 to 2020, which comprises more than 100 million sentence pairs, for research purposes.<sup>2</sup>

JaParaPat is made from the publication of unex-

<sup>1</sup><https://sites.google.com/view/pat-claims-trans-2025/>

<sup>2</sup><https://www.kecl.ntt.co.jp/icl/lirg/japarapat/>

	jp-us	jp-x-us	us-jp	pct	sum
2016	7,241,502	1,322,124	1,181,150	10,287,313	20,032,089
2017	7,892,204	1,399,012	1,226,177	10,354,135	20,871,528
2018	7,639,692	1,262,972	1,044,728	11,171,128	21,118,520
2019	8,867,148	1,450,851	1,157,361	11,625,720	23,101,080
2020	8,617,540	1,570,684	1,088,832	10,843,470	22,120,526
sum	40,258,086	7,005,643	5,698,248	54,281,766	107,243,743

Table 1: Number of sentence pairs

amined patent applications from the Japan Patent Office (JPO) and the United States Patent and Trademark Office (USPTO) from 2000 to 2021. They are aligned based on patent family information from the DOCDB, a bibliographic database maintained by the European Patent Office (EPO).

Table 1 shows the number of sentence pairs available in the public version of JaParaPat. There are two primary routes for filing international patent applications: the Paris Convention route and the Patent Cooperation Treaty (PCT) route. JaParaPat includes data from both routes. In Table 1, within the Paris route, ‘jp-us’ refers to patent pairs first filed in Japan and subsequently in the United States. ‘us-jp’ refers to those first filed in the United States and then in Japan. ‘jp-x-us’ refers to patents initially filed in a country other than Japan or the United States, and subsequently filed in both Japan and the United States. The public version employs different methods for document alignment, sentence segmentation, and sentence alignment, resulting in a different number of sentence pairs compared to Table 1 in the original JaParaPat paper.

As the training data for the shared task of Patent Claim Translation, one of the most important problems of JaParaPat is its sentence segmentation and alignment for patent claims. It often segments a long claim into segments by a new line and provides segment-level alignment, which makes it difficult to reconstruct claim-level alignment. We are discussing with the authors of JaParaPat how to solve this problem.

### 3 Development Data

This time we focused on claims rather than specification to see how different engines will handle relatively difficult sentence structures, technical terms, non-technical terms, ambiguous language (i.e. phrases that can be interpreted in more ways than one), etc. Claims serving as development

data were selected from existing patent application documents. In the selection, we mainly considered the following factors as elements impacting the difficulty of translation:

- Paragraph length
- Term peculiarity
- Construction
- Structural/semantic ambiguity (e.g. whether a given phrase should be interpreted as "A including B, and C (not included in A)", or "A including both B and C")
- Terminological ambiguity (e.g. whether the term "対向" in Japanese means "opposing", "reverse", "facing", etc.)
- Whether a term has a corresponding term/concept in target language
- Existence/lack of an official translation (e.g. a US application having a corresponding JP application)

Based on these criteria, we selected 13 Japanese documents and 11 English documents for this study. Example of development data is shown in Table 7.

We translated the development data using two types of translation engines: an NMT model trained on JaParaPat and an open-weight LLM, and conducted a preliminary human evaluation using this data. The purpose was to determine appropriate evaluation procedures and the feasible level of granularity prior to performing the main evaluation using the test data.

Figure 1 shows the excel interface of the human evaluation. We instructed the annotators to perform the following three tasks:

1. Highlight segments containing translation errors or input issues and specify the corresponding error category within the cell.



Figure 1: Human evaluation interface.

2. Assign a holistic quality score to the translation on a 100-point scale.
3. Post-edit the translation.

The post-edited translations were used as reference translations to form parallel data, which we provided as development data.

#### 4 Test Data

Source texts in Japanese and English were selected from existing patent applications. We have considered the following factors when selecting source texts.

- Type of machine translation: The type of translation was estimated to be neural machine translation (NMT) or large language model (LLM)-based translation.
- Length/construction: It is known that a longer single text without a line break may result in poorer translation quality (Kondo et al., 2021). The primary purpose of this research was not to examine how different engines would deal with length, but to see if general claim wording, which may contain one or more of the factors mentioned above or below, will be handled. As such, we selected source texts that generally contained no more than about 220 English words or 500 Japanese characters with or without one or more line breaks in them. The purpose of including a few longer texts was to see how a relatively long text would be processed.

- Existing translation: A patent application may have a family including a corresponding application in another language; for example, an application filed to the Japan Patent Office (JPO) may have a corresponding application filed to the United States Patent and Trademark Office (USPTO). Applications in the same family are linked in some search engines including Google Patents. An LLM may be able to locate an official translation of an application, i.e. correct solution, through such search engines if the application has a family. We therefore selected source texts from applications that did not have a corresponding application in the target language at least at the time when the source texts were distributed to the participants. Because of this factor, we cannot automatically collect reference translations from publicly available data. In addition, we do not have sufficient budget to create reference translations for the test data. Therefore, as described in Section 6, we conducted reference-free automatic evaluation (i.e., quality estimation).
- Field: The source texts come from applications in a variety of fields including information processing, communication, electric engineering, chemistry, etc.
- Ambiguity/parsing: Machine translation is processing that is based essentially or entirely on natural language information. The processing is not expected to rely on visual

or other non-natural language-based information. Meanwhile, claim wording sometimes requires reference to information based on other than natural language, a typical example of which is drawings that patent applications often contain. As the USPTO Patent Application Filing Guide states "*a patent application is required to contain drawings if drawings are necessary to understand the subject matter*", natural language *per se* could be insufficient to arrive at a correct interpretation of claim wording. In addition, there are also cases where reference to the specification is necessary to fully understand the meaning of a claim. For instance, with the phrase "a device comprising a controller that has an analyzer, a processor, and a memory", it may be necessary to refer to the specification to determine whether the "processor" and "memory" are part of the "device" or the "controller".

For the current project, we have selected source texts, the content of which was — at least to the persons in charge of the selection — comprehensible on its own without additional information. The selected source texts contain ambiguous terms such as 区間 in Japanese, which can be interpreted as a temporal concept (interval: period between two times) or dimensional concept (interval: space between two points). We allowed for the inclusion of such terms only where it was possible to ascertain the meaning of a term from the context. For example, the aforementioned 区間 is stated in the claim in which the term appears to be a section of a road (a physical interval within a road) on which a vehicle travels. So, it should be obvious that the term does not mean a temporal concept.

Selecting a source text that does not require additional information to interpret is also beneficial from the perspective of evaluating the translation: A satisfactory evaluation by either a human or non-human evaluator should be possible without additional information. This means that the respective evaluation abilities of a human evaluator and a non-human evaluator can be put to comparison essentially on the basis of their abilities to process natural language without additional information.

Team ID	Organization	Country	J-E	E-J
UTSK25	University of Tsukuba	Japan	1	3
EHIME-U	Ehime University	Japan	12	0
Commercial 1	online service	n/a	1	1
Commercial 2	closed system	n/a	1	1
Commercial 3	free LLM model for MT	n/a	1	1

Table 2: List of participants and the number of submissions for each direction. For the commercial systems, the organizers collected the translations.

Taking the above factors into consideration, we prepared 26 documents with 70 claims for the Japanese–English direction and 30 documents with 81 claims for the English–Japanese direction as the test data.

## 5 Participants

Table 2 shows the list of participants and the number of submissions from each system. The organizers collected the translations of the commercial systems. Whereas the UTSK25 conducted continual pretraining of an open-weight LLM on JaPara-Pat, Ehime University performed prompt tuning on a closed/proprietary LLM. For Commercial 1 we used a standard translation prompt. For Commercial 3 we performed translation using the chat template provided in its accompanying documentation. Commercial 2 is a closed system.

We selected 1 submission for each translation direction for all the systems except EHIME-U for the human evaluation. For EHIME-U, we selected 2 submissions for Ja-En because they did not submit any result for En-Ja.

## 6 Automatic Evaluation

Automatic evaluation of MT has been studied for a long time, along with the evolution of MT technologies. It faces new challenges, such as very long and complex claim sentences in our task. For the first attempt, we conducted the automatic evaluation in a reference-free manner using MetricX-24-Hybrid-XL<sup>3</sup> (Juraska et al., 2024) and WMT23-CometKiwi-DA-XL<sup>4</sup> (Rei et al., 2023) because the corresponding translations of the test set were not available, as mentioned above. We had two variants of automatic evaluation: *segment-level* (claim-by-claim) and *document-level*. The document-level evaluation

<sup>3</sup><https://github.com/google-research/metricx>

<sup>4</sup><https://github.com/Unbabel/COMET>

System	ja-en		en-ja	
	MetricX ↓	CometKiwi ↑	MetricX ↓	CometKiwi ↑
UTSK25	3.761 $\pm$ 1.654	0.544 $\pm$ 0.122	3.623 $\pm$ 1.474	0.641 $\pm$ 0.111
EHIME-U 1	2.882 $\pm$ 1.614	0.560 $\pm$ 0.134	n/a	n/a
EHIME-U 2	2.978 $\pm$ 1.607	0.568 $\pm$ 0.131	n/a	n/a
Commercial 1	2.792 $\pm$ 1.416	0.572 $\pm$ 0.133	2.916 $\pm$ 0.842	0.681 $\pm$ 0.088
Commercial 2	3.879 $\pm$ 2.454	0.567 $\pm$ 0.139	3.126 $\pm$ 1.031	0.676 $\pm$ 0.093
Commercial 3	2.920 $\pm$ 1.107	0.573 $\pm$ 0.127	2.581 $\pm$ 0.780	0.707 $\pm$ 0.078

Table 3: Segment-level automatic evaluation results

System	ja-en		en-ja	
	MetricX ↓	CometKiwi ↑	MetricX ↓	CometKiwi ↑
UTSK25	4.669 $\pm$ 1.439	0.313 $\pm$ 0.128	4.577 $\pm$ 1.605	0.489 $\pm$ 0.118
EHIME-U 1	3.827 $\pm$ 1.392	0.308 $\pm$ 0.110	n/a	n/a
EHIME-U 2	4.071 $\pm$ 1.613	0.305 $\pm$ 0.106	n/a	n/a
Commercial 1	3.471 $\pm$ 1.003	0.279 $\pm$ 0.123	3.435 $\pm$ 0.817	0.539 $\pm$ 0.093
Commercial 2	5.303 $\pm$ 2.153	0.259 $\pm$ 0.139	4.022 $\pm$ 1.025	0.525 $\pm$ 0.126
Commercial 3	3.568 $\pm$ 0.871	0.298 $\pm$ 0.127	3.183 $\pm$ 0.751	0.567 $\pm$ 0.098

Table 4: Document-level automatic evaluation results

considered the whole document as a single segment.

Tables 3 and 4 show average segment- and document-level scores, respectively.

## 7 Human Evaluation

Due to budget constraints, human evaluation was conducted only on a subset of the test data. The selection of evaluation files followed the same Diversity Sampling procedure used in the WMT25 General Machine Translation Shared Task(Kočmi et al., 2025), resulting in 13 files per translation direction.

Table 8 and 9 in Appendix A.2 shows the human evaluation criteria we used. We made several modifications to Freitag’s metric (Freitag et al., 2021) to better adapt it to the patent-translation domain. We also referred to the MQM website<sup>5</sup> for the descriptions and examples. Categories shown with a gray background were deemed unnecessary for patent translation and were therefore excluded.

## 8 Official Results

Table 5 shows the average score of the human evaluation. There was no system that achieved the best accuracy in both translation directions. On

System	ja-en	en-ja
UTSK25	63.04	79.29
EHIME-U 1	81.61	n/a
EHIME-U 2	86.07	n/a
Commercial 1	87.68	70.00
Commercial 2	66.96	60.71
Commercial 3	67.50	54.11

Table 5: Average score of the human evaluation.

average, Commercial 1 exhibited the highest accuracy.

Table 6 shows the correlation coefficients between human evaluation and each automatic evaluation measure. Surprisingly, none of the metrics showed substantial correlation with the human evaluation. Several factors may account for this outcome:

1. Both automatic evaluation methods used in this study are reference-free, which may limit their ability to accurately assess translation quality.
2. These automatic evaluation methods may not function effectively in the patent domain.
3. The human evaluations themselves may contain inaccuracies (we discuss this in detail in

<sup>5</sup><https://themqm.org/the-mqm-full-typology/>

Measure	ja-en	en-ja
MetricX (seg)	-0.235	-0.121
MetricX (doc)	-0.230	-0.023
CometKiwi (seg)	0.288	0.186
CometKiwi (doc)	0.029	-0.079

Table 6: Correlation coefficients between human evaluation and each automatic evaluation measure.

the Discussion section).

## 9 Discussion

Our analysis of the translation outputs and human annotations revealed various issues on both the translation side and the annotation side. In this section, we discuss several of these problems.

The selected source texts contained several phrases which could be interpreted or rendered in more ways than one yet the correct meaning or valid rendition of which could be derived from the context. A few examples of such phrases will be observed below along with annotations they were marked with. In view of the following examples, we shall focus on two issues that are broadly applicable to translation in general and more specifically to patent translation, namely "use of generic or specific terms" and "differences in routines/legal restrictions between Countries/intellectual property (IP) offices".

### 9.1 Use of Generic or Specific Terms

Source: "前記信頼度情報が予め設定された閾値よりも小さい状態が継続している区間を補正対象区間として検知して"  
 "... 前記運動状態推定部は、…  
 前記補正対象区間を走行している前記  
 他車両の運動状態を推定し"

A technically correct translation should be something along the lines of:

TR: "... detects, as a correction target section, a section in which the confidence information continues to remain below a preset threshold"  
 "the motion-state estimation unit estimates the motion state of the other vehicle traveling through the correction target section"

Note that the discussion below focuses on the term "区間", which can be rendered into a number of terms including "section", "interval", "segment", "portion" or the like as long as it is clear

that the term refers to a physical segment of a road, not to a time interval. From the second phrase above stating that the other vehicle travels through this section, it should be obvious that the section is not a time interval.

The following is a machine translation produced by one of the six engines.

sys: "... detect, as a correction-target section, a section during which the reliability information remains less than the predetermined threshold"  
 "the motion state estimation unit is configured to estimate the motion state of the other vehicle traveling in the correction-target section"

Renditions of the underlined phrase by other engines include:

- detects, as a correction-target section, a section in which
- detect, as a correction target section, a period during which
- detects the interval during which ... as a correction target interval

Both nouns "section" and "interval" on their own could be either a physical or temporal concept. In the above context, the preposition (plus relative pronoun), i.e. "during (which)" or "in (which)", is decisive in whether the preceding noun will be interpreted as a physical or temporal concept. For the example above, it can be said that while "during" is incorrect, "in" is ambiguous (i.e. can be interpreted in more ways than one) yet potentially correct (i.e. encompasses the correct meaning). Choosing a specific term is preferable if the concept including the term is unambiguous, but if a concept is ambiguous, choosing a generic term may increase the chance of the concept being interpreted correctly.

Multiple human annotators, who must have been exposed to the concept that the "section" is a segment of a road on which a vehicle travels, did not leave any annotation to the expression "a section during which".

The following are a few examples of ambiguous terms that are often used in patent-related documents.

- 挿まれる (*hasamareru*): The term means an either physical or conceptual entity being located, interposed, or held between two or more other physical or conceptual entities. It

is often rendered as "sandwiched" but incorrectly in some contexts. A generic term suggesting a location between two or more entities, e.g. simply "between", may be more suitable in some cases.

- (～である) が ((dearu) ga): This is a highly context-sensitive particle and could mean "but", "and", "whereas", "yet", "thus", "in this regard/respect", etc. connecting the phrases before and after it to some degree and in some way. It is often rendered as "but/however", but expressions such as "in this regard/respect" may be a better option in some contexts. Moreover, the term can often be omitted entirely.
- 対象 (taisho): One of the most ambiguous yet convenient terms to refer to something that the writer of a text wants to refer to. "… in question" should be one of the most generic English equivalents, but it can make the translation vague. In some cases, it may be necessary to explicitly say what the writer wants to refer to by converting the term into a more specific concept.

See Appendix A.3 for more details.

## 9.2 Differences in Routines/Legal Restrictions between Countries/IP offices

Source: …プログラムであって、  
コンピュータを、  
…クリアデッキを記憶する記憶手段、  
…一のクリアデッキを編成できるか否かを判定する判定手段、  
…コンテンツを特定コンテンツとして特定する特定手段、  
…取得画面を表示させる制御手段、  
として機能させる、  
プログラム。

A more or less literal/mirror translation would be something along the lines of:

TR: A program …, the program causing a computer to function as  
…a storage means that stores a clear deck …,  
…a determination means that determines whether one clear deck can be organized …,  
…a specifying means that specifies, as specific content, content that is …, and  
…a control means that causes an acquisition screen to be displayed ….

The following is a machine translation provided by one of the engines.

sys: A program … causing a computer to: store … a clear deck …; determine … whether at least one clear deck … can be organized …; identify …, as specific content, content that is …; and display … an acquisition screen ….

The term "手段 (means)" is not reproduced in this translation. From a technical point of view, "causing a computer to function as a storage means that stores information" is equivalent to "causing a computer to store information". From the perspective of patent prosecution, some patent practitioners choose not to use the term "means" or any equivalent thereof (unit, portion, etc.) to avoid means-plus-function language (see, e.g., 35 U.S.C. 112(f)), a potential cause of rejection by a US examiner. The presence of the term "means" would probably not produce any benefit in patent prosecution in other IP offices where an application can be filed in English. Thus, since the use of the term "means" does not seem to add any value to this claim and may cause an issue in the US, it may be better to omit the term.

If omission, or addition in some cases, of certain terms or concepts can improve the quality of translation from the perspective of patent prosecution in the target country/region without distorting the content of the source text more than allowed, it should be considered an appropriate "adjustment".

The annotators marked the aforementioned omission of "means" as an error, namely "omission; major". From the reasons explained above, the omission may be beneficial. Although it may be possible to mark the omission, it should not be marked as a major error.

Other examples of appropriate adjustments are as follows:

- Addition/omission: "特徴とする (characterized)" is a good example of a term/concept that may be added or omitted according to the IP office the application is filed to.
- Inconsistency vs consistency: In Japan, translating a term into multiple equivalents is generally regarded as careless inconsistency. Outside Japan, in some cases, rendering a term into multiple terms in the target language can be beneficial. For instance, the applicant can let the examiner at some IP office choose a most suitable term for them to allow the claim.

### 9.3 Annotation Issues

As noted above, human annotations contained significant issues. The following are examples of numerous issues we found in the annotations, which were provided by one of Japan's most well-known patent translation companies.

#### Failure to detect errors

- "the first electronic device comprises a thermostat" was translated as "前記第1電子デバイスがサーモスタットである (the first electronic device is a thermostat)"
- "to the motion state estimation unit, wherein"

#### Failure to detect relatively minor error

- Inconsistency between "operate" and "travel" as equivalents of "走行". "A vehicle traveling" in a segment of a road suggests any type of vehicle running through that segment. "A vehicle operating" in a segment of a road may suggest a more specific type of vehicle (e.g. truck) operating in that segment for a specific purpose (e.g. moving goods).

#### Failure to detect relatively major error

- See above discussion on "section during which".
- The source text states "characterized in that" in one place; the translation strongly suggests a different place for it.

#### Error detected by annotator is not an error

- Stating the subject matter of a claim twice, i.e. at the beginning and end of the claim, was marked as a major error. This is a common claim structure in Japanese patent applications.

#### Minor error detected should be relatively major error

- "said first electronic device being adapted to respond to user instructions by changing device state" was translated to mean "said first electronic device being adapted change device state in response to user instructions (この第1の電子装置はユーザ指示に応じて装置の状態を変化させる)". While the Japanese translation was marked as "awkward: minor" for some reason, the error is obviously a major error significantly distorting the meaning of the source text.

#### Major error detected should be relatively minor error (or no error)

- "A sensor, comprising" at the beginning of the English claim was rendered as "以下の構成要素からなるセンサー: (A sensor comprising the following constituents:)" at the beginning of the Japanese claim. Although it is not a common claim structure in Japan, a JPO examiner would probably accept it.

Human annotation may serve as training data for developing automatic annotation technology. Using erroneous annotations as training data will have negative consequences. If the annotations above, provided by a major translation company, represent a typical quality of human annotation in Japan, developing accurate automatic annotation technology in this country may encounter difficulties.

## 10 Conclusion and Future Perspective

This paper summarizes the first shared tasks of the patent claims translation. This year, we had 2 participants who submitted their translation results. Based on the human evaluation results, no system achieved consistently strong performance in any translation direction. However, comparisons with automatic evaluation results and analyses of human annotations revealed various issues, as reported in this paper.

In subsequent years, building on the insights obtained here, we aim to define a framework for more stable and higher-quality human evaluation, as well as to use the human annotations as training data to develop highly accurate automatic evaluation methods for patent translation.

## Acknowledgments

The human evaluation fee is supported by Asia-Pacific Association for Machine Translation (AAMT) and AAMT/Japio<sup>6</sup> Special Interest Group on Patent Translation.

## References

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

<sup>6</sup>Japan Patent Information Organization

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Seiichiro Kondo, Kengo Hotate, Toshio Hirasawa, Masahiro Kaneko, and Mamoru Komachi. 2021. [Sentence concatenation approach to data augmentation for neural machine translation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 143–149, Online. Association for Computational Linguistics.

Masaaki Nagata, Makoto Morishita, Katsuki Chousa, and Norihito Yasuda. 2024. [JaParaPat: A large-scale Japanese-English parallel patent application corpus](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9452–9462, Torino, Italia. ELRA and ICCL.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

## A Appendix

### A.1 Example of Development Data

Table 7 shows examples of development data. Each document may contain one or more claims. Each claim is basically composed of only one sentence, but it may contain line breaks for readability.

### A.2 Human Evaluation Criteria

Table 8 and 9 shows the human evaluation criteria we used.

### A.3 Extensive Discussion

The discussion in the body text on the use of generic or specific terms and differences in routines/legal restrictions between countries/IP offices will be presented below with more details. Note that some of the content below is a reproduction of Section 9.

#### Use of Generic or Specific Terms

Source: "前記信頼度情報が予め設定された閾値よりも小さい状態が継続している区間を補正対象区間として検知して"  
"…前記運動状態推定部は、…前記補正対象区間を走行している前記他車両の運動状態を推定し"

A technically correct translation should be something along the lines of:

TR: "... detects, as a correction target section, a section in which the confidence information continues to remain below a preset threshold" "the motion-state estimation unit estimates the motion state of the other vehicle traveling through the correction target section"

Note that the discussion below focuses on the term "区間", which can be rendered into a number of terms including "section", "interval", "segment", "portion" or the like as long as it is clear that the term refers to a physical segment of a road, not to a time interval. From the second phrase above stating that the other vehicle travels through this section, it should be obvious that the section is not a time interval.

The following are the machine translations produced by the six engines.

sys1: "... detect, as a correction-target section, a section during which the reliability information remains less than the predetermined threshold" and "the motion state estimation unit is configured to estimate the motion state of the other vehicle traveling in the correction-target section"

sys2: "... detects, as a correction-target section, a section in which the state that the reliability information is less than the predetermined threshold continues" and "the motion state estimation unit is configured to estimate the motion state of the other vehicle traveling in the correction-target section"

sys3: "... detect, as a correction target section, a period during which reliability information ... is smaller than a predetermined threshold and such a state continues longer than a predetermined time" and "the motion state estimation unit estimates the motion state of the other vehicles traveling in the correction target section"

Japanese

[請求項 1]

ポリエーテルポリオール（a 1）と有機ポリイソシアネート（a 2）を反応させて得られるイソシアネート基末端ウレタンプレポリマーを含有する主剤（A）、並びにポリオール（b 1）、導電剤（b 2）、及び制電剤（b 3）を含有する硬化剤（B）からなる半導電性ウレタンエラストマー形成性組成物において、

導電剤（b 2）が導電性カーボン、制電剤（b 3）が炭素数10～20の脂肪族系不飽和炭化水素基及び1個のヒドロキシアルキル基を有する第4級アンモニウムカチオンとビス（トリフルオロメタンスルホニルイミド）アニオンとからなるイオン性塩であって、導電剤（b 2）及び制電剤（b 3）の、主剤（A）と硬化剤（B）の総和量における各々の含有量が（b 2）：0.1～1質量%、（b 3）：0.001～10質量%であることを特徴とする半導電性ウレタンエラストマー形成性組成物。

[請求項 2]

ポリエーテルポリオール（a 1）と有機ポリイソシアネート（a 2）を反応させて得られるイソシアネート基末端ウレタンプレポリマーを含有する主剤（A）、並びにポリオール（b 1）、導電剤（b 2）、及び制電剤（b 3）を含有する硬化剤（B）からなる半導電性ウレタンエラストマー形成性組成物において、

導電剤（b 2）が導電性カーボン、制電剤（b 3）が炭素数10～20の脂肪族系不飽和炭化水素基及び2個のヒドロキシアルキル基を有する第4級アンモニウムカチオンとビス（トリフルオロメタンスルホニルイミド）アニオンとからなるイオン性塩であって、導電剤（b 2）及び制電剤（b 3）の、主剤（A）と硬化剤（B）の総和量における各々の含有量が（b 2）：0.1～1質量%、（b 3）：0.001～10質量%であることを特徴とする半導電性ウレタンエラストマー形成性組成物。

English

1. An aftermarket vehicle communication device engageable to a vehicle for providing location information associated with the vehicle to a V2X data stream, the device comprising:
  - a housing configured to be detachably engageable to the vehicle;
  - a GPS circuit disposable in communication with a GPS system to receive a GPS signal therefrom, the received GPS signal being representative of a location of the vehicle when the housing is engaged to the vehicle; and
  - an antenna circuit coupled to the housing and in communication with the GPS circuit, the antenna circuit being configured to receive the GPS signal from the GPS circuit and communicate the GPS signal to the V2X data stream;
  - a micro computing unit (MCU) coupled to the housing and in communication with the GPS circuit and the antenna circuit, the MCU being configured to generate an alert signal communicable to the V2X data stream via the antenna circuit, the alert signal being receivable by autonomous vehicles via the V2X data stream to facilitate assigning a prescribed margin of separation to the vehicle to which the housing is engaged;
  - the GPS circuit and the antenna circuit being configured to facilitate both the receipt of the GPS signal from the GPS system and communication of the GPS signal to the V2X data stream independent of receiving information or data from the vehicle.

Table 7: Example of development data.

sys4: "… reliability information … is smaller than a preset threshold value and the state continues for a longer time than a preset time" and "the motion state estimation unit estimates a motion state of the other vehicle traveling in the correction target section"

(Note: There was a significant omission in this translation and the term "区間" was not reproduced in the first clause.)

sys5: "… detects the interval during which the reliability information remains below the predetermined threshold as a correction target interval" and "the motion state estimation unit estimates the motion state of the other vehicle while it is traveling through the correction target interval"

sys6: "… detects, as a correction target section, a section in which reliability information … is smaller than a preset threshold value and the state continues for longer than a preset time" and "the motion state estimation unit estimates the motion state of the other vehicle traveling in the correction target section"

Both nouns "section" and "interval" on their own could be either a physical or temporal concept. In the above context, the preposition (plus relative pronoun), i.e. "during (which)" or "in (which)", is decisive in whether the preceding noun will be interpreted as a physical and/or temporal concept. In this context, the preposition "in" can be said to be a more generic preposition than "during". In other words, while "a section/interval during which" represents a temporal concept, "a section/interval in which" can represent both temporal and physical concepts. For the

example above, it can be said that while "during" is incorrect, "in" is ambiguous (i.e. can be interpreted in more ways than one) yet potentially correct (i.e. encompasses the correct meaning). Although the latter clause "…前記運動状態推定部は、…前記補正対象区間を走行している前記他車両の運動状態を推定し (the motion-state estimation unit estimates the motion state of the other vehicle traveling through the correction target section)" provides enough information to determine if the "section" is a physical or temporal concept, such a determination cannot be made solely from the former phrase "前記信頼度情報が予め設定された閾値よりも小さい状態が継続している区間を補正対象区間として検知して (… detects, as a correction target section, a section in which the confidence information continues to remain below a preset threshold)".

From the above, it is conceivable that if a set of information comprising one or more words is ambiguous and remains ambiguous even with reference to other information that are processed together with said information, it is better for the engine to choose one or more generic terms that keep the interpretation of the information open-ended. Moreover, if an engine is equipped with auto-correct function, it is also conceivable that the engine flags such ambiguous information while temporarily providing a generic term to it, then after

processing other sections, refers back to it to examine if a more specific, context-suited term can be provided.

Let us examine the present case from the perspective patent rights. Even if the preposition "in" may render the first instance of "section" ambiguous, the latter clause will clarify the meaning of the term. Thus, an examiner, or a judge or opponent in a court case, will understand the meaning of the term and there will be no clarity-related rejection (35 U.S.C. 112 (b)) or dispute due to the ambiguity of the term. The meaning of a term or concept in a claim is often interpreted in view of the overall technical feature that is set forth by the claim as a whole. Choosing a specific term is preferable if the concept including the term is unambiguous, but if a concept is ambiguous, choosing a generic term may increase the chance of the concept being interpreted correctly.

More than one human annotators, who were exposed to the concept that the "section" is a segment of a road on which a vehicle travels, did not leave any annotation to the expression "a section during which". The expression "as a correction target section, a period during which" was marked with the annotation "inconsistency: major", but this probably refers to the inconsistency between "section" and "period", not to the semantic/technical inaccuracy.

The following are just a few examples of ambiguous terms that are often used in patent-related documents.

- **挟まる (hasamareru):** The term means an either physical or conceptual entity being interposed or held between one or more other physical or conceptual entities. A patty held by a bun, an interval between the first and second halves of a concert, Chomsky 's thoughts between Marks ' s and Fodor ' s, an insulator between and in contact with or with a gap to two layers, Jupiter in relation to Saturn and Mars or even in relation to Uranus and Earth, a river flowing between banks, or any such concept can be described using **挟まる**. The term is often translated as "sandwiched (between …)", but obviously the expression can be misleading or incorrect in some context. In a context in which the specific manner of interposition can be, or intended to be, interpreted in more ways than one, a specific term such as "sandwiched"

should be avoided.

(Needless to say, however, that a specific term such as "held (between)" should be chosen if **挟まる** focuses on the concept of an entity being physically held by other entity/entities. Inappropriate ambiguity may lead to abstract ideas, hence to clarity-related issues in patent prosecution or litigations.)

- **(～である) が ((dearu) ga):** This is a highly context-sensitive particle and could mean "but/however", "and", "whereas", "yet", "so/thus", "in this regard/respect", etc. connecting the phrases before and after it to some degree and in some way. The particle is often used in office actions issued by the JPO in the context of, for example: "文献 1 には～が記載されていないが、文献 2 には記載されている (Document 1 does not disclose … but document 2 does)"; "文献 1 は AA を記載しているが、文献 2 は BB を記載しており、両者を組み合わせることは容易である (Document 1 discloses AA, whereas document 2 discloses BB, and it would be easy to combine the two)"; or "本願は CCA と記載しているが、文献 1 は CCB と記載しており、文献 1 は本願発明を開示しているに等しい (The present application sets forth CCA. In this regard, document 1 discloses CCB and can be regarded as disclosing an equivalent of the invention of the present application)". Note that the generic concept "in this regard" may replace "but" and "and" in the first two example sentences. Moreover, it may be possible to entirely omit "が" and say "Document 1 does not disclose … Document 2 does", "Document 1 discloses AA; document 2 discloses BB; it would be easy to combine the two", and "The present application sets forth CCA. Document 1 discloses CCB and can be regarded as disclosing an equivalent of the invention of the present application."

- **対象 (taisho):** This is probably one of the most ambiguous yet convenient terms to refer to something that the writer of a text wants to refer to. The term could mean "target", "… in question", "destination", "… to be", "subject", "object", etc. In this research, an engine translated "補正対象区

間" as "correction target section". Although the translation is not erroneous, a more accurate and natural rendition would be "a section to be corrected" or "a segment subject to correction".

In some cases, it may be necessary to explicitly say what the writer wants to refer to by converting the term into a more specific concept. For instance, in an invention in which a tune is differentiated from the tune being analyzed and the analyzed tune is referred to as **対象楽曲** (literally, e.g. "the tune in question"), it may be better to refer to this tune as "the tune being analyzed". This is the case where use of a generic term does not work and it is better to use a more specific term, which may involve some additional/supplemental/complementary concepts.

### Differences in Routines/Legal Restrictions between Countries/IP offices

Source: …プログラムであって、  
コンピュータを、  
…クリアデッキを記憶する記憶手段、  
…一のクリアデッキを編成できるか否かを判定する判定手段、  
…コンテンツを特定コンテンツとして特定する特定手段、  
…取得画面を表示させる制御手段、  
として機能させる、  
プログラム。

A more or less literal/mirror translation would be something along the lines of:

A program …, the program causing a computer to function as  
…a storage means that stores a clear deck …,  
…a determination means that determines whether one clear deck can be organized …,  
…a specifying means that specifies, as specific content, content that is …, and  
…a control means that causes an acquisition screen to be displayed ….

The following are the machine translations produced by the six engines.

sys1: A program … causing a computer to: store … a clear deck …; determine … whether at least one clear deck … can be organized …; identify …, as specific content, content that is …; and display … an acquisition screen ….

sys2: A non-transitory computer-readable medium storing instructions … causing a computer to: store … a clear deck …; determine … whether one clear deck … can be organized; identify …, as specific content, content …; and cause an acquisition screen … to be displayed ….

sys3: A program …  
causing a computer to function as:  
a storage means for storing … a clear deck;  
a determination means for determining …  
whether a clear deck … can be organized;  
a specifying means for specifying …, as specific  
contents, contents that are …; and  
a control means for displaying … an acquisition  
screen ….

sys4: Program …, wherein a storage means  
for storing a clear deck … the computer …; a  
determination means for determining whether or  
not one clear deck … can be organized …; and,  
the control unit causes (the player) to function  
as: a specifying unit that specifies content … as  
specific content; and a control unit that causes  
… to display an acquisition screen ….

sys5: A program … (comprising):  
a computer configured to function as:  
a memory means for associating (each quest)  
with a cleared deck …;  
a judgment means for determining … whether  
their owned content is sufficient to assemble …;  
a specification means for identifying … the  
content items … as specified content items; and  
a control means for displaying … an acquisition  
screen ….

sys6: (Omission …) storing a clear deck …;  
determining whether or not a clear deck … can  
be organized …; identifying, as specific content,  
content that is …; and causing an acquisition  
screen … to be displayed ….

In Japanese patent-drafting routines, it is common to repeat the subject matter of a claim at the end of the claim, as it can be seen in the above text where the term "プログラム (program)" appears at the beginning and the end of the claim. In view of how applications are drafted in English-speaking countries/regions, this repetition should not be reproduced in an English translation. In this regard, most of the engines seem to have managed to adopt a relatively correct sentence construction without such repetition.

Some engines (see Sys 1, Sys2, and Sys6) omitted the term "手段 (means)" from the translation. This omission may be evaluated from a technical point of view as well as from the perspective of patent prosecution. From a technical point of view, "causing a computer to function as a storage means that stores information" is equivalent to "causing a computer to store information". Both expressions mean that a computer having a memory is caused to store information in the memory. From the perspective of patent prosecution, some patent practitioners choose not to use the term "means" or any equivalent thereof (unit, portion, etc.) to avoid means-plus-function language (see, e.g., 35 U.S.C. 112(f)). Means-plus-function language may benefit the applicant under certain

conditions but may also narrow the scope of the claim, especially in the US. In other countries and regions where it is possible to file an application in English, the omission of the term "means" would probably not result in any disadvantage for the applicant. So, for the current case, since the use of the term "means" does not seem to add any technical value to the claim, it may be better to omit the term at least in terms US drafting routines. From the above, it can be said that omission of certain terms or concepts, which may be called an appropriate "adjustment", may enhance the quality of translation from the perspective of patent prosecution. Similar adjustments can often be seen in more general writing. For example, a meaningful translation of the phrase "I am all ears" will be distant from a literal/mirror translation. Transition of a phrase from one sprachbund to another may require an appropriate adjustment. The value of a patent application is bound to the routines and legal restrictions existing in the country/region in which the application is filed. When evaluating the quality of patent translation, the value of an appropriate adjustment should be taken into account in view of the routines and legal restrictions in the country/region to which the translation is destined to.

The annotators marked the aforementioned omission of "means" as an error in the form of "omission; major". From the reasons explained above, the omission may be beneficial, and although it may be possible to mark the omission as an error, the error should not be marked as "major".

Other examples of appropriate adjustments are as follows:

- Addition/omission: 特徴 (characteristic feature) is a good example of a term that may be added or omitted according to the IP office the application is filed to. The term means the characteristics of an invention that make the invention novel and inventive over prior art. Some IP offices may request that the characterizing portion (e.g. novel engine) of a claim be distinguished from the part of the claim adopting prior-art (e.g. any automobile) by using the term "特徴".
- Inconsistency vs consistency: As a general rule, a term used in a claim should be used consistently throughout the claim and in its dependent claims. In Japanese practice, a

term that is used in the specification (e.g. 音響 (e.g. audio)) and that corresponds to the term in the claim (e.g. 音信号 (sound signal)) is often also used consistently throughout the specification. Translating 音響 into two or more terms (audio, acoustic, voice, etc.) may be considered careless inconsistency. However, 音響 encompasses a wide range of concepts and different examiners in certain IP offices may have different word choice preferences. Translating 音響 into different equivalents and amending the claim according to the examiner's preferred word choice may render the prosecution smoother.

Top Category	Mid Category	Sub Category	Description	Example
Accuracy	Addition		Translation includes information that is not present in the source and that is not supposed to be included.	A translation includes portions of another translation that were inadvertently pasted into the document.
	Omission		Translation is missing content from the source and the omission is inappropriate.	A paragraph present in the source is missing in the translation.
	Untranslated text		Source text has been left untranslated.	A sentence in a Japanese document translated into English is left in Japanese.
	Mistranslation		Translaiton does not accurately represent the source.	A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should be administered in doses greater than 200 mg (i.e., negation has been omitted).
	(Mistranslation)	Numerals	/ Translation errors related to numerals and symbols.	3000 is translated as 30000
	(Mistranslation)	Symbols		
	(Mistranslation)	Article	Incorrect use of articles	A translation uses "a" for the item which appears for the second time.
	(Mistranslation)	Incorrect dependency	dependency	A of B, and C is translated as A of B and C (the dependency of C is incorrect)
	(Mistranslation)	Unknown dependency	dependency	said drive link being formed of one integral metallic piece = 駆動リンクにおいて、一体成形の金属片からなり
	(Mistranslation)	Ambiguity		
Fluency	Punctuation		Incorrect punctuation (for locale or style, including improper sentence division, since patent claims must be written in one sentence).	1) An English text uses a semicolon where a comma should be used. 2) A two-digit year reference begins with an open single quote instead of a close single quote (apostrophe). 3) A Greek text uses a question mark instead of the anticipated semicolon to express a question. 4) German quotation marks are carried over into English or French target content.
	Spelling		Incorrect spelling or capitalization.	The German word Zustellung is spelled Zustetlugn.
	Grammar		Problems with grammar, other than orthography.	An English text reads “The man was seeing the his wife.
	Register		Wrong grammatical register (eg, inappropriately informal pronouns).	A formal letter uses contractions, colloquialisms, and expressions characteristic of spoken rather than written language, and those elements come across as less serious than intended.
	Inconsistency		Internal inconsistency (not related to terminology)	1) One part of a text is written in a clear, “terse” style, while other sections are written in a more wordy style. 2) The same text recurs at several points in a large document that has been divided up and submitted to multiple translators, with the result that that text is translated in three different ways, which can involve different style as well as terminology or register differences.
	Character encoding		Characters are garbled due to incorrect encoding.	A text document in UTF-8 encoding is opened as ISO Latin-1, resulting in all “upper ASCII” characters being garbled.

Table 8: Human Evaluation Criteria.

Top Category	Mid Category	Sub Category	Description	Example
Terminology	Inappropriate for context		Terminology is non-standard or does not fit context.	The word 'river' in an English source text is translated into French as 'rivière'. But the river in question flows into the sea, not into a lake or another river, so the correct French translation should have been 'fleuve'.
	Inconsistent use		Terminology is used inconsistently.	The text refers to a component as the 'brake release lever', 'brake disengagement lever', 'manual brake release', and 'manual disengagement release'.
Style	Awkward		Translation has stylistic problems.	A text is written with many embedded clauses and an excessively wordy style. While the intended meaning can be understood, and the text is grammatically correct, the text is very awkward and difficult to follow. "However, a personal language variety (in such approaches called "idiolect") usually is internally heterogeneous (it varies in particular according to different situations and/or media) and therefore not suitable to serve as the smallest unit of linguistic variation, whereby in contrast, idiolects according to the framework developed in this document, are homogeneous by definition, whereas personal varieties are sets of idiolects."
Locale convention	Address format Currency format Date format Name format Telephone format Time format		Wrong format for addresses. Wrong format for currency. Wrong format for dates. Wrong format for names. Wrong format for telephone numbers. Wrong format for time expressions.	
Other			Any other issue.	
Source error			An error in the source.	
Non-translation			Impossible to reliably characterize distinct errors.	

Table 9: Human Evaluation Criteria (contd.).