A Cross-Lingual Multi-Jurisdiction Dataset for Patent Drafting from Patent Cooperation Treaty Applications

Phuong Trung Le[1], Hirofumi Nonaka[2], Koji Marusaki[3],
Katsuhito Sudoh[4], Seiya Kawano[1,5]
[1]Kyoto Institute of Technology [2]Aichi Institute of Technology
[3]Nagaoka Institute of Technology [4]Nara Women's University [5]RIKEN GRP
{kawano}@kit.ac.jp

Abstract

Patent claim drafting is a crucial task in intellectual property practice, requiring both technical expertise and legal knowledge. While prior work has explored automated claim refinement, these efforts remain confined to single languages and jurisdictions, overlooking the inherently international nature of patent practice under the Patent Cooperation Treaty (PCT) system. We present a cross-lingual, multi-jurisdiction parallel dataset constructed from PCT applications at the Japan Patent Office (JPO) and the United States Patent and Trademark Office (USPTO), linking pre-grant and granted claims across Japanese and English through PCT application numbers. This dataset enables cross-lingual patent drafting tasks: transforming patent claim drafts written in one language into claims in another language adapted to the requirements of the target patent office. Our experiments show that machine translation alone fails to capture the required transformation, and that different alignment categories (e.g., pre-grant to pre-grant vs. pre-grant to granted) exhibit distinct patterns, reflecting jurisdiction-specific adaptation, prosecution history, and claim drafting strategies inherent to international patent practice.

1 Introduction

Patent claims define the legal scope of protection for an invention and play a crucial role in intellectual property strategy [1]. The process of refining patent claims to obtain stronger rights or to satisfy examination requirements is essential for successful patent prosecution [2, 3]. With the increasing globalization of innovation, inventors frequently seek patent protection in multiple jurisdictions through the Patent Cooperation Treaty (PCT) system, which provides a unified procedure for filing patent applications in over 150 countries.

While research has been conducted on basic patent information processing tasks such as patent retrieval and classification [4, 5], as well as on supporting the understanding of patent claims [6], automated patent claim refinement has only recently begun to be explored. Several studies have proposed datasets and models for claim refinement using large language models, targeting major patent offices including the JPO [7], USPTO [8], and EPO [9]. However, these efforts have focused on monolingual refinement within single jurisdictions, neglecting the international nature of patent practice.

Cross-lingual patent processing has primarily focused on machine translation, with parallel corpora constructed under the assumption that corresponding documents are faithful translations [10, 11]. However, this assumption does not account for jurisdiction-specific adaptations in international patent practice, where claims are revised to satisfy office-specific drafting conventions and independent examination outcomes. In practice, applicants filing through the PCT system must adapt claims beyond translation, reflecting jurisdictional drafting conventions, legal terminology, and examination practice. For instance, claims drafted for the JPO may require substantive restructuring when prosecuted at the USPTO. For background on patent translation corpora and drafting studies, see Appendix A.1.

In this paper, we present a cross-lingual, multi-jurisdiction parallel dataset for patent drafting constructed from PCT applications. Our dataset links JPO and USPTO documents at pre-grant and granted stages, and enables cross-lingual drafting (e.g., JA-pre → EN-granted) rather than literal translation. By leveraging PCT application numbers as the alignment key, we construct the dataset and demonstrate through baseline experiments that machine translation is insufficient for cross-lingual patent drafting, with distinct patterns observed across alignment categories.

2 Dataset Construction

We link JPO-side and USPTO-side documents originating from the same PCT application using PCT application numbers, select stage-specific English counterparts, normalize claim text, and export aligned pairs for train/dev/test splits. See Appendix A.2 for the detailed procedure.

2.1 Overview of PCT application system

The Patent Cooperation Treaty (PCT) enables applicants to seek patent protection across multiple countries through a single international application. After filing, applicants enter the ''national phase'' in selected countries, where each patent office examines the application independently. Publications are categorized by kind codes: A-kind for pre-grant and B-kind for granted patents. By combining PCT application numbers with kind codes, we link documents at corresponding stages across jurisdictions, which forms the basis of our dataset construction.

2.2 Analysis of dataset

The dataset covers PCT applications published/granted between **2004 and 2022**, and contains **1,138,204** aligned claim pairs across all categories. To characterize the dataset beyond total scale, Table 1 reports per-category sample counts, average Japanese character length, average English word length, and the average number of claims on each side.

Several characteristics are noteworthy. First, in same-stage translation-aligned categories such as `en_pre→ja_pre`, the average claim counts are nearly identical (25.5 vs. 25.6), indicating limited structural divergence in claim enumeration. Second, in `ja_pre→en_granted`, the average claim count differs substantially (26.2 vs. 17.9), quantitatively reflecting the drafting gap between an original filing-stage claim set and a granted claim set. Third, granted Japanese texts are longer on average than pre-grant counterparts (e.g., 4,589 chars in `ja_granted→en_granted` vs. 3,447 chars in `ja_pre→en_pre`), suggesting increased textual density at the granted stage.

Unlike conventional patent translation corpora, even pre-grant correspondences across jurisdictions can differ substantially in structure and content due to jurisdiction-specific practice. Therefore, our dataset preserves these naturally occurring differences rather than filtering them out as noise.

2.3 Qualitative examples

To illustrate the difference between translation-style correspondences and drafting-stage changes, we present Table 2 side-by-side comparison of the **same invention** (Family ID: `52449909`) across pre-grant and granted stages. For readability, we show Claim 1 excerpts.

This example clarifies the intended distinction in our dataset: while pre-grant JA–EN pairs tend to preserve the overall claim structure, the granted-stage claim may introduce additional constraints or specific elements that are not explicit in the pre-grant Claim 1 excerpt, highlighting the challenge of `ja_pre→en_granted`-type drafting.

3 Experimental Settings

We conduct baseline experiments using machine translation to reveal the gap between simple translation and cross-jurisdictional patent drafting. By translating source claims and comparing against target-jurisdiction references, we quantify the divergence arising from jurisdiction-specific drafting conventions and independent examination processes. We also observe that evaluation outcomes are sensitive to document length; see Appendix A.3 for details.

Baseline system. We use Google Cloud Translation API (Advanced) v3[1] as a commercial baseline for cross-lingual generation. For long claim texts, we split inputs into multiple chunks when necessary and translate them via multiple API calls.

Evaluation implementation. We compute metrics using standard open-source libraries. BLEU and chrF are computed with `sacrebleu`. SARI is computed with the Hugging Face `evaluate` library. COMET is computed using the `Unbabel/wmt22-comet-da` checkpoint via the `comet` library.

[1] https://docs.cloud.google.com/translate/docs/api-overview#advanced_ed

Table 1 Detailed dataset statistics by alignment category.

Category	# Pairs	Avg. JA Chars	Avg. EN Words	Avg. Claims (JA/EN)
ja_pre→en_pre	112,791	3,447	1,127	24.4 / 25.7
ja_granted→en_granted	17,296	4,589	1,072	21.2 / 16.0
ja_pre→en_granted	308,962	3,775	1,099	26.2 / 17.9
en_pre→ja_pre	552,112	3,608	1,199	25.5 / 25.6
en_pre→ja_granted	20,208	3,267	1,261	22.3 / 22.7

Table 2 Qualitative comparison for the same invention (Family ID: 52449909). Claim 1 excerpts are shown.

Pre-grant stage (translation-style)	Granted stage (drafting-stage)
JA Pre (Claim 1) ‘‘長尺形状であって、第 1 表示領域とこの第 1 表示領域と... タッチパネル式ディスプレイと...’’	**JA Granted (Claim 1)** ‘‘電子機器であって、画像を表示する長尺形状のディスプレイと... プロセッサによる実行制御は... メインプログラムとして特定し...’’
EN Pre (Claim 1) ‘‘An electronic device comprising: a display having an elongated shape and configured to display images... composed of a first region and a second region...’’	**EN Granted (Claim 1)** ‘‘An electronic device comprising: a display having an elongated shape... a memory configured to store a plurality of programs and for each of the programs, an icon image...’’

Table 3 Dataset split for each alignment category.

Alignment Category	Train	Dev	Test
ja_pre→en_pre	110,934	857	1,000
ja_granted→en_granted	15,854	442	1,000
ja_pre→en_granted	177,502	1,000	1,000
en_pre→ja_pre	177,502	1,000	1,000
en_pre→ja_granted	18,208	1,000	1,000

Evaluation protocol. We evaluate in an **all-documents** setting, including long claim texts rather than filtering them out. Inputs are truncated only when they exceed the effective maximum length supported by the evaluation pipeline, to reflect practical long-context conditions.

Dataset. Table 3 shows the data split for each alignment category. Our dataset supports both intra-lingual drafting (e.g., JP-pre → JP-granted) and cross-lingual drafting (e.g., JP-pre → EN-pre). In this paper, we focus on cross-lingual drafting.

4 Experimental Results

4.1 Baseline performance

Table 4 reports baseline performance of Google Cloud Translation API (Advanced) v3 across five alignment categories under the all-documents evaluation setting. These results quantify how well a strong off-the-shelf translation system can model the cross-lingual correspondences present in our dataset.

Key finding 1: same-stage correspondences are not faithful translations. Even for same-stage categories that are often treated as parallel data in patent translation research (e.g., pre-grant JA → pre-grant EN), BLEU remains far from saturation (e.g., 35.81 for ja_pre→en_pre and 29.57 for ja_granted→en_granted). This indicates that cross-jurisdiction correspondences in PCT-derived documents are not necessarily literal translations, but can include structural and content-level differences. Such differences are consistent with practical international prosecution: claims are adapted to local drafting conventions and legal requirements even when they originate from the same PCT application. This observation implies that constructing ‘‘clean’’ parallel corpora by filtering out mismatches (a common practice in translation-oriented corpus construction) would remove precisely the cases that matter for cross-lingual drafting.

Key finding 2: cross-stage correspondences are substantially harder than same-stage ones. Performance further decreases when the target is a granted claim set (e.g., BLEU 26.68 for ja_pre→en_granted). This aligns with our dataset analysis in Table 1, where ja_pre→en_granted exhibits a large average claim-count discrepancy (26.2 vs. 17.9), suggesting that prosecution can merge, cancel, or narrow claims during examination. Because granted claims reflect examination outcomes that are jurisdiction-dependent, cross-stage cross-lingual drafting requires more than translation: it must model legally valid amendments that arise from independent examination and drafting practice at the target office.

Implications and relation to prior work. Prior cross-lingual patent research has primarily focused on machine translation using sentence-level aligned patent corpora, typically assuming meaning-preserving parallelism and filtering out divergences [10, 11]. Our baseline results demonstrate that this assumption does not hold for PCT-derived multi-jurisdiction correspondences, even at the same publication stage. Accordingly, cross-lingual patent drafting should be framed as a transformation task that includes jurisdiction-specific optimization rather than as translation alone, motivating dedicated datasets and models.

Comparison with translation-focused patent corpora. Prior work that treats patent documents as meaning-preserving parallel data reports substantially higher BLEU than what we observe in our setting. For example, JaParaPat trains Japanese→English MT models on hundreds of millions of aligned **sentence pairs** and reports SacreBLEU of 55.6–56.5 on Paris-route test sets and 52.7–57.3 on PCT-route test sets, compared with 31.9–35.6 BLEU when training only on the general-domain JParaCrawl baseline (Table 6 in [11]). In contrast, our commercial baseline (Google Translate v3) achieves 35.81 BLEU for `ja_pre→en_pre` and 29.57 BLEU for `ja_granted→en_granted` under **all-documents** evaluation. This gap is expected because our aligned claim sets are not curated to be literal sentence-level translations: they preserve jurisdiction-specific drafting conventions and prosecution-driven amendments, which introduce structural and content-level divergence even at the same stage. Therefore, BLEU in our experiments should be interpreted as a measure of **cross-jurisdiction correspondence difficulty** rather than standard MT fidelity.

4.2 Metric divergence in EN→JA

In EN→JA tasks, BLEU scores are low (approximately 9–11) while COMET scores are high (above 0.90). This divergence suggests that n-gram overlap metrics can be sensitive to surface-form differences for Japanese references (e.g., segmentation or phrasing variation), whereas embedding-based evaluation indicates that outputs may still be semantically close to references. Accordingly, chrF and COMET provide complementary signals for EN→JA evaluation in our setting. This observation further suggests that off-the-shelf translation systems may be suboptimal for patent-claim style and terminology, motivating domain adaptation and fine-tuning as an important future direction.

Table 4 Google Translate (v3) baseline metrics (full dataset evaluation).

Category	BLEU	SARI	COMET	chrF
ja_pre→en_pre	35.81	75.32	0.86	62.06
ja_granted→en_granted	29.57	77.47	0.87	55.48
ja_pre→en_granted	26.68	70.04	0.84	59.51
en_pre→ja_pre	9.59	54.38	0.90	46.46
en_pre→ja_granted	11.36	54.24	0.91	48.01

5 Conclusion

We presented a cross-lingual, multi-jurisdiction dataset for patent drafting constructed from PCT applications filed at the JPO and USPTO. Our dataset links pre-grant publications and granted patents across Japanese and English, enabling research on cross-lingual patent drafting tasks.

Our key finding is that cross-lingual correspondences derived from PCT applications are not well-modeled by translation alone: even same-stage transfer exhibits non-trivial divergence, and cross-stage settings (pre-grant → granted) are substantially harder due to jurisdiction-specific prosecution and amendments. In addition, we observe that evaluation outcomes are sensitive to long claim texts, underscoring the need for long-context drafting models and evaluation protocols that reflect realistic patent document lengths.

Future work includes: (1) extending the dataset to cover additional jurisdictions such as the EPO and other major patent offices, (2) developing specialized models for cross-lingual patent drafting that can learn jurisdiction-specific adaptation patterns, and (3) incorporating examination history information to guide the drafting process across languages.

Acknowledgement

This study was supported in part by JST ACT-X JPMJAX22A4 and NEDO Wakasapo Project.

References

[1] Alan C Marco, Joshua D Sarnoff, and AW Charles. Patent claims and patent scope. **Research Policy**, 48(9):103790, 2019.

[2] Robert C Faber. **Landis on mechanics of patent claim drafting**. Practising Law Institute New York, 1990.

[3] Greg Reilly. Amending patent claims. **Harv. JL & Tech.**, 32:1, 2018.

[4] Mihai Lupu, Atsushi Fujii, Douglas W Oard, Makoto Iwayama, and Noriko Kando. Patent-related tasks at ntcir. **Current Challenges in Patent Information Retrieval**, pages 77–111, 2017.

[5] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of the patent retrieval task at the ntcir-6 workshop. In **NTCIR**, 2007.

[6] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama. Patent claim processing for readability-structure analysis and term explanation. In **Proc. of the ACL-2003 workshop on Patent corpus processing**, pages 56–65, 2003.

[7] Seiya Kawano, Hirofumi Nonaka, and Koichiro Yoshino. ClaimBrush: A Novel Framework for Automated Patent Claim Refinement Based on Large Language Models . In **2024 IEEE International Conference on Big Data (BigData)**, pages 6594–6603, Los Alamitos, CA, USA, December 2024. IEEE Computer Society.

[8] Mirac Suzgun, Luke Melas-Kyriazi, Suproteem Sarkar, Scott D Kominers, and Stuart Shieber. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. **Advances in neural information processing systems**, 36:57908–57946, 2023.

[9] Lekang Jiang, Chengzu Li, and Stephan Goetz. Enriching patent claim generation with european patent dataset. **arXiv preprint arXiv:2505.12568**, 2025.

[10] Masao Utiyama and Hitoshi Isahara. A Japanese-English patent parallel corpus. In Bente Maegaard, editor, **Proceedings of Machine Translation Summit XI: Papers**, Copenhagen, Denmark, September 10-14 2007.

[11] Masaaki Nagata, Makoto Morishita, Katsuki Chousa, and Norihito Yasuda. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pages 9452–9462, Torino, Italia, May 2024. ELRA and ICCL.

[12] Mihai Lupu, John Tait, Jimmy Huang, and Jianhan Zhu. Trec-chem 2010: Notebook report. **Proceedings of TREC 2010**, 2, 2010.

[13] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. Clef-ip 2011: Retrieval in the intellectual property domain. In **CLEF (notebook papers/labs/workshop)**, 2011.

Table 5 Impact of document length on BLEU (length-matched vs. all-documents).

Category	Match/All	Skipped	BLEU(Match)	BLEU(All)	Δ
ja_pre→en_pre	588/857	269 (31%)	45.51	35.81	-9.70
ja_pre→en_granted	717/1000	283 (28%)	38.32	26.68	-11.64
ja_granted→en_granted	424/442	18 (4%)	28.59	29.57	+0.98
en_pre→ja_pre	909/1000	91 (9%)	5.46	9.59	+4.13
en_pre→ja_granted	911/1000	89 (9%)	7.60	11.36	+3.76

A Appendix.

A.1 Related Work

Cross-lingual patent processing has largely focused on machine translation, supported by parallel corpora derived from patent families and sentence-level alignment [10, 11]. Such corpora are typically curated under a meaning-preserving assumption, often filtering out content divergences as noise. In parallel, patent NLP has been studied through evaluation campaigns and benchmarks (e.g., NTCIR, TREC, CLEF-IP) that cover retrieval and related tasks [5, 12, 13]. Recent work on claim refinement and drafting has explored using examination-related signals and office-specific practice, but has primarily been studied in monolingual settings within a single jurisdiction [7, 8, 9]. Our work differs by preserving cross-jurisdiction divergences within PCT-derived correspondences and framing the task as cross-lingual drafting rather than translation alone.

A.2 Procedure of Dataset Construction

We construct a cross-lingual, multi-jurisdiction dataset by linking JPO-side and USPTO-side documents that originate from the same PCT application. The key identifier for alignment is the PCT application number, which is shared across national-phase entries.

Step 1: Extract and normalize PCT identifiers. We extract PCT application information from JPO-side records and normalize identifiers to enable reliable joins across data sources.

Step 2: Aggregate JPO-side records by PCT application number. For each PCT application number, we aggregate available JPO-side metadata and claim text into a single structured record. When multiple related publications exist for the same PCT application, we consolidate them in a consistent representation (e.g., storing multiple values as arrays when needed).

Step 3: Match to English-side candidates and select the best counterpart. For each grouped JPO-side PCT application, we retrieve candidate English documents within the same patent family and select a single best English counterpart for each stage. When multiple candidates exist, we prioritize (i) the existence of English claims, then (ii) country preference (US > WO > EP), and then (iii) earlier publication dates. This yields English `pre` and (when available) English `granted` counterparts linked to the same PCT application.

Step 4: Clean and normalize claim text. We apply text normalization to both languages to reduce formatting noise. This includes normalizing whitespace and headers, removing invisible characters (e.g., BOM/ZWSP), and expanding canceled-claim ranges into explicit lines when present.

Step 5: Export and split. The final aligned table is exported to parquet and split into train/validation/test sets for model development and evaluation.

A.3 Impact of Document Length

Patent claims are long-context documents, and evaluation that filters out long examples can bias conclusions about real-world drafting performance. We compare a length-matched setting (excluding long documents) against an all-documents setting to quantify the sensitivity of baseline results to document length; the results are summarized in Table 5.