2025年度AAMT/Japio特許翻訳研究会
報 告 書

機械翻訳及び機械翻訳評価に関する研究
並びに
国際ワークショップPSLT2025開催報告

2026 年 3 月

一般財団法人　日本特許情報機構

目 次

AAMT/Japio 特許翻訳研究会委員名簿

（敬称略・五十音順）

委　員　長	辻井　潤一	国立研究開発法人産業技術総合研究所 フェロー
		東京大学大学院 名誉教授
		マンチェスター大学 教授
副 委 員 長	須藤　克仁	奈良女子大学 教授
	綱川　隆司	静岡大学学術院 情報学領域 准教授
委　　員	荒瀬　由紀	東京科学大学 教授
	今村　賢治	国立研究開発法人 情報通信研究機構
		ユニバーサルコミュニケーション研究所
		先進的音声翻訳研究開発推進センター 先進的翻訳技術研究室
	越前谷　博	北海学園大学 教授
	岡崎　直観	東京科学大学 教授
	河野　誠也	京都工芸繊維大学 情報工学・人間科学系 助教
	菊井玄一郎	国立研究開発法人科学技術振興機構(JST)
	黒橋　禎夫	国立情報学研究所 所長
	後藤　功雄	愛媛大学 教授
	小町　　守	一橋大学大学院
		ソーシャル・データサイエンス研究科 教授
	鈴木　　潤	東北大学 言語 AI 研究センター センター長／教授
	田村　晃裕	同志社大学 大学院理工学研究科 准教授
	中澤　敏明	東京大学大学院 特任研究員
	二宮　　崇	愛媛大学 教授
	渡辺　太郎	奈良先端科学技術大学院 教授
オブザーバー	江原　暉将	元・山梨英和大学 教授
	高　　京徹	株式会社高電社
	園尾　　聡	東芝デジタルソリューションズ株式会社
	王　　向莉	株式会社ディープランゲージ

オブザーバー（(一般）日本特許情報機構）：

	大塩　只明	特許情報研究所 調査研究部 研究企画課
	笠田　和宏	特許情報研究所 調査研究部 研究企画課 課長
	木下　　聡	特許情報研究所 調査研究部 研究企画課
	小林　　明	専務理事/特許情報研究所 所長
	関口　明紀	特許情報研究所 調査研究部 部長
	塙　　金治	特許情報研究所 研究管理部 研究管理課
	舩戸さやか	特許情報研究所 調査研究部 研究企画課 係長
	三橋　朋晴	特許情報研究所 研究管理部 主幹
事　務　局		（一般）日本特許情報機構

2025 年度 AAMT/Japio 特許翻訳研究会・活動履歴

2025 年 5 月 7 日
第 1 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2025 年 6 月 24 日
第 11 回特許・技術文献翻訳ワークショップ（PSLT2025）
（於スイス・ジュネーブ開催）

2025 年 7 月 2 日
第 2 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2025 年 9 月 11 日
第 3 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2025 年 10 月 9 日
第 4 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2025 年 12 月 4 日
第 5 回 AAMT/Japio 特許翻訳研究会
（於ハイブリット開催）

2026 年 1 月 22 日
第 6 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

1. はじめに

AAMT/JAPIO　特許翻訳委員会委員長
産業技術総合研究所　フェロー
マンチェスター大学　教授
辻井　潤一

大規模言語モデル（LLM）がごく身近な存在となり、さまざまな業務に活用されている。私にとっても、原稿執筆時の関連文献検索などに欠かすことのできない道具となった。インターネット検索においても、検索結果のサイトを一つひとつ読むことはまれで、多くの場合、AI が提示する回答を読むことで必要な情報を得ている。実に便利な時代である。

翻訳においても同様である。自分で書いた英語を校正してもらうこともあれば、日本語で粗い概要を書いて英文にしてもらうことも多い。出来上がった英文は、自分で書くよりはるかに流ちょうで、時に気恥ずかしさを覚えるほどである。気がつくと、これまでやり取りをしていた、英語が母語ではない海外の友人からのメールも、見事な英語になっている。おそらく彼らも LLM を使っているのだろう。

しかし、かつて多少規範から外れた英語表現を読んだとき、その背後にある人柄を想像してほほえましく感じたことを思い出すと、どこか寂しさも覚える。

ビジネス調だけでなく、「カジュアルな英語に」と依頼すれば、それらしい英文を生成してくれる。ただ、その文章もまた母語話者的であり、自分では思いつかない言い回しであることが多い。そこに自分という人格が薄れてしまったような、不思議な感覚を抱くこともある。

一方で、翻訳された英文を読んでみると、自分の意図とは異なる情報が伝わっているのではないかと感じ、修正することもある。意図しない意味が付け加わっていたり、文脈から誤って補われていたりすることもある。外交文書では、二つの言語で可能な限り同じ情報が伝わるよう、慎重な校正が行われるという。異なる言語で同じ情報や効果を伝えることは、決して容易ではない。全く知らない言語への翻訳を LLM に任せ、そのまま相手に送ることができるかと問われれば、やはり躊躇せざるを得ない。

自分の原稿を LLM に校正させたり、部分的な修正を依頼したりすると、明らかに自分の書いた文章ではないという違和感を覚えることがある。思考の流れが微妙に変わり、論理が自分の意図からずれてしまうこともある。その結果、かえって修正に時間を要する場合も少なくない。単語の選択、文体、論理の構成など、文章には個人の思考や感性が深く反映されている。その個性が、LLM という一般化の中で均質化されていくことが、果たして望ましいことなのかと考えることもある。

さて、翻訳の話に戻ろう。LLM の登場によって、翻訳のコストが大きく低減されたことは確かである。しかし、外交文書ほどではないにせよ、高品質な技術翻訳を担う翻訳者は、文書の背景を調査し、ときには書き手と直接やり取りを行いながら作業を進めているという。現在の LLM に、そのような営みがどこまで可能であろうか。書き手の意図、さらには書き手と読み手の関係性まで含めて他言語へ写し取ることは、流ちょうな文章を生成する以上に困難な作業である。しかも、その流ちょうな文章の中に誤りが紛れ込むこともある。

翻訳において技術をどのように活用し、コストを低減しつつ質を向上させるか。そのためには、翻訳者の役割や求められる能力も含め、翻訳のプロセス全体を再設計する必要があるだろう。

やや抽象的な議論となったが、本委員会の活動が、将来の翻訳システムと人間の役割という課題に貢献していくことを期待している。

2. 年間報告書

2.1 PCT 出願に基づく多言語・多法域特許クレーム起案データセット

京都工芸繊維大学　情報工学・人間科学系
河野　誠也

＊本稿の内容は言語処理学会第 32 回年次大会(NLP2026)に投稿された論文に基づく

概要

特許クレームの起案は知的財産実務における重要なタスクであり、技術的専門知識と法的知識の両方を必要とする。従来の研究はクレームの自動修正を試みてきたが、これらの取り組みは単一言語・単一法域に限定されており、特許協力条約（PCT）制度における国際的な特許実務の本質を見落としている。本研究では、日本特許庁（JPO）と米国特許商標庁（USPTO）における PCT 出願から構築された多言語・多法域並列データセットを提示する。PCT 出願番号を介して、日本語と英語にまたがる出願公開段階と登録段階のクレームを対応付ける。本データセットにより、ある言語で作成された特許クレーム草案を、対象特許庁の要件に適合した別言語のクレームに変換する多言語特許起案タスクが可能となる。実験の結果、機械翻訳のみでは必要な変換を十分に行えないこと、また異なるアライメントカテゴリ（例：出願公開同士の対応と出願公開から登録への対応）が法域固有の適応、審査経過、およびクレーム起案戦略を反映する異なるパターンを示すことが明らかとなった。

2.1.1　はじめに

特許クレームは発明に対する法的保護の範囲を定め、知的財産戦略において重要な役割を果たす（Marco et al., 2019）。より強い権利を取得するため、あるいは審査要件を満たすためにクレームを修正するプロセスは、特許審査の成功に不可欠である（Faber, 1990; Reilly, 2018）。イノベーションのグローバル化が進む中、発明者は PCT 制度を通じて複数の法域で特許保護を求めることが多く、同制度は 150 以上の国で特許出願を行うための統一手続きを提供している。

特許検索や分類などの基本的な特許情報処理タスク（Lupu et al., 2017; Fujii et al., 2007）や特許クレームの理解支援（Shinmori et al., 2003）に関する研究は行われてきたが、特許クレームの自動修正についてはごく最近になって探索が始まったところである。JPO（Kawano et al., 2024）、USPTO（Suzgan, 2023）、EPO（Jian et al., 2025）を含む主要特許庁を対象に、大規模言語モデルを用いた特許クレーム修正のためのデータセットとモデルがいくつか提案されている。しかし、これらの取り組みは単一法域内での単言語修正に焦点を当てており、特許実務の国際的側面を無視している。

多言語特許処理は主に機械翻訳に焦点を当てており、対応文書が忠実な翻訳であるという仮定のもとで並列コーパスが構築されてきた（Utiyama & Isahara, 2007; Nagata et al., 2024）。しかし、この仮定は、特許クレームがオフィス固有の起案規則や独立した審査結果を満たすように修正される国際特許実務における法域固有の適応を考慮していない。実際には、PCT 制度を通じて出願する出願人は、法域の起案規則、法的用語、および審査実務を反映して、翻訳を超えたクレ

ームの適応を行う必要がある。

本研究では、PCT 出願から構築された多言語・多法域特許起案並列データセットを提示する。我々のデータセットは JPO および USPTO の特許文書を出願公開段階と登録段階で対応付け、単なる逐語訳ではなく多言語起案（例：ja_pre → en_granted）を可能にする。PCT 出願番号をアライメントキーとして活用し、データセットを構築し、ベースライン実験を通じて、機械翻訳が多言語特許起案には不十分であること、アライメントカテゴリごとに異なるパターンが観察されることを実証する。

2.1.2　関連研究

多言語特許処理は、特許ファミリーと文レベルのアライメントから導出された並列コーパスに支えられ、主に機械翻訳に焦点を当ててきた（Utiyama & Isahara, 2007; Nagata et al., 2024）。このようなコーパスは通常、意味保持の仮定のもとでキュレーションされ、内容の乖離をノイズとしてフィルタリングすることが多い。並行して、特許 NLP は検索および関連タスクをカバーする評価ベンチマーク（例：NTCIR、TREC、CLEF-IP）を通じて研究されてきた（Fujii et al., 2007; TREC, 2010; CLEF, 2011）。他方、クレーム修正と起案に関する最近の研究は、審査関連のシグナルとオフィス固有の実務を用いることを探索してきたが、主に単一法域内の単言語設定で研究されてきた（Kawano et al., 2024; Suzgan et al., 2023; Jiang et al., 2025）。我々の研究は、PCT 由来の対応における法域間の乖離を保持し、タスクを翻訳のみではなく多言語起案としてフレーミングする点で異なる。

2.1.3　デートセット構築

JPO および USPTO の特許文書を同一の PCT 出願番号を用いて出願公開段階と登録段階で対応付け、クリーニングし、文書レベル（特許クレームセット）のアライメントペアを構築した。ここで、ソース側の特許クレームの公開日がターゲット側のクレームよりも時間的に過去である場合のみをフィルタリングした。

2.1.3.1　PCT 出願制度

特許協力条約（PCT）は、単一の国際出願を通じて複数の国で特許保護を求めることを可能にする。出願後、出願人は選択した国で「国内段階」に入り、各特許庁が独立して出願を審査する。公報は種別コードで分類される：A 種別は出願公開、B 種別は登録特許である。PCT 出願番号と種別コードを組み合わせることで、法域をまたいで対応する段階の文書を対応付けることができ、これが本データセット構築の基盤となる。また、特許の審査の性質上、同一の出願であったとしても B 種別は A 種別と比較して、特許査定を受けるために必要なオフィスアクションへの対応や各種補正の結果が反映されている点で内容が異なる。

2.1.3.2　データセットの分析

本データセットは 2004 年から 2022 年の間に公開/登録された PCT 出願を対象とし、全カテゴ

リで 1,138,204 件のアライメント済みの特許クレームのセットのペアを含む。

表 1: アライメントカテゴリ別のデータセット統計

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

表 1 にアライメントカテゴリ別のデータセット統計を示す。ここで、ja は JPO 側、en は USPTO 側、pre は公開段階、granted は登録段階の公報である。表より、第一に、en_pre→ja_pre のような同一段階の翻訳アライメントカテゴリでは、平均クレーム数がほぼ同一（25.5 対 25.6）であり、クレーム列挙における構造的乖離が限定的であることを示している。第二に、ja_pre→en_granted では平均クレーム数が大きく異なり（26.2 対 17.9）、出願段階のクレームセットと登録段階のクレームセットの間の起案ギャップを定量的に反映している。第三に、登録段階の日本語テキストは出願公開段階の対応物よりも平均して長く（例：ja_granted→en_granted の 4,589 文字対 ja_pre →en_pre の 3,447 文字）、登録段階でのテキスト密度の増加を示唆している。

従来の特許翻訳コーパスとは異なり、法域間の出願公開段階の対応であっても、法域固有の実務により構造と内容が大幅に異なることがある。したがって、本データセットはこれらの自然に発生する差異をノイズとしてフィルタリングするのではなく、保持する。

2.1.4 実験設定

機械翻訳を用いたベースライン実験を実施し、単純な特許翻訳と法域をまたぐ特許起案の間のギャップを明らかにする。ソースクレーム（文書レベル）を翻訳し、対象法域の参照と比較することで、法域固有の起案規則と独立した審査プロセスから生じる乖離を定量化する。Google Cloud Translation API (Advanced) v3 をベースラインとして使用する。長いクレームテキストについては、必要に応じて入力を複数のチャンクに分割し、複数の API 呼び出しで翻訳する。評価には、BLEU と chrF は sacrebleu で、SARI は Hugging Face の evaluate ライブラリで、COMET は Unbabel/wmt22-comet-da チェックポイントを使用して comet ライブラリで計算する。

表 2: 各アライメントカテゴリのデータセット分割

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

2.1.4.1 ベースライン性能

表 3 に、全文書評価設定における 5 つのアライメントカテゴリにわたる Google Cloud Translation API (Advanced) v3 のベースライン性能を報告する。

表 3: Google Translate (v3) ベースラインメトリクス（全データセット評価）

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	38.17	54.38	0.90	46.46
en_pre→ja_granted	40.15	54.24	0.91	48.01

知見 1：同一段階の対応は忠実な翻訳ではない。特許翻訳研究で並列データとして扱われることが多い同一段階のカテゴリであっても、BLEU は飽和からは程遠い（例：ja_pre→en_pre で 35.81、ja_granted→en_granted で 29.57）。これは、PCT 由来の文書における法域間対応が必ずしも逐語訳ではなく、構造的・内容的差異を含みうることを示している。

知見 2：異なる段階の対応は同一段階よりも大幅に困難である。対象が登録クレームセットの場合、性能はさらに低下する（例：ja_pre→en_granted で BLEU 26.68）。これは表 1 のデータセット分析と一致しており、審査中にクレームが統合、追加、取消、または限定されうることを示唆している。

翻訳重視の特許コーパスとの比較：JaParaPat は数億の整列された文ペアで日英 MT モデルを訓練し、SacreBLEU 55.6-56.5 を報告しているのに対し、商用ベースラインは全文書評価で ja_pre→en_pre に対して 35.81 の BLEU を達成している。このギャップは、我々の整列されたクレームセットが逐語的な文レベルの翻訳としてキュレーションされていないために予想されるものである。

2.1.5 まとめ

本研究では、JPO と USPTO に出願された PCT 出願から構築された、多言語・多法域特許起案データセットを構築した。本データセットは出願公開と登録特許を日本語と英語にわたって対応付け、多言語特許起案タスクの研究を可能にする。

評価実験では、PCT 出願から導出された多言語・他法域の対応関係が単純な翻訳のみでは十分にモデル化できないことを示した。同一段階（出願公開→出願公開，登録→登録）の変換でさえ非自明な乖離を示し、異なる段階の変換設定（出願公開→登録）は法域固有の審査と修正の影響

により大幅に困難であった。

本研究の今後の展開としては：(1) EPO およびその他の主要特許庁を含む追加法域へのデータセットの拡張、(2) 法域固有の適応パターンを学習できる多言語特許起案の専用モデルの開発、(3) 言語間の起案プロセスを導くための審査経過情報の組み込み、に取り組む

参考文献

[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 multipurpose 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

2.2 大規模言語モデルを用いた特許文書からの固有表現および発明構成要素の抽出

奈良女子大学　須藤　克仁
岡崎 ひかり
米田　碧弥

2.2.1 はじめに

大規模言語モデル（LLM）の技術的進展により、従来であれば専用のモデルを学習、またはファインチューニングする必要があったさまざまな自然言語処理タスクがプロンプトによる指示や文脈内学習（In-Context Lea rning）により解くことができるようになりつつある。特許情報処理では発明を構成する要素の正しい認識が重要であり、特許翻訳においても発明の構成要素が正しく、また訳語の揺れなく訳出される必要がある。そのため、我々は特許文書から発明の構成要素を自動抽出し、それが機械翻訳を通じて正しく、揺れなく訳出されているかを評価・検証できる方式の確立を目指す。本稿ではその目標に向けた第一歩として、大規模言語モデルを用いた特許文書から固有表現と発明構成要素の自動抽出を試みた。本試行実験においては、GPT-5.2 と GPT-4.1 を同一のプロンプトを用いて動作させ、その結果を比較した。

2.2.2 関連研究

固有表現抽出（または固有表現認識; Named Entity Recognition、以下 NER）は自然言語からの情報抽出の基盤となる自然言語処理の課題であり、長年に渡って数多くの研究が行われてきた。NER の典型的な手法は系列ラベリングによるものであり、与えられた単語列の各単語に対して固有表現ラベルを予測することで固有表現を同定する。サポートベクタマシン（Isozaki & Kazawa, 2002）や条件付確率場（Finkel, Grenager, & Manning, 2005）、LSTM-CRF（Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016）等、系列ラベリングのさまざまなモデルを固有表現アノテーションの付されたテキストコーパスを用いて学習する試みが続けられてきた。深層学習に基づく大規模事前学習モデルである BERT（Devlin, Chang, Lee, & Toutanova, 2019）の登場以後は大量のテキストコーパスで事前学習された大規模なモデルを固有表現コーパスでファインチューニングする手法が精度面で有利であることから主流となった。さらに大規模言語モデル（Large Language Model; LLM）が急速に進化してきたことで、LLM を用いた NER の手法も提案されている（Wang, et al., 2025）。

一方で、NER では固有表現コーパスを利用して学習・ファインチューニングを行う手法が主流であったために、固有表現コーパスで利用されている固有表現種別、典型的には人名や地名、組織名に対象が限定されるという問題があった。そこで、さまざまなドメインにおける情報抽出の実現のため、特定の固有表現種別に限定したモデルを目指すのではなく、自然言語による指示に基づいて任意の種別の固有表現の抽出を図るオープンタイプ NER というタスクが提案されてい

る（Zaratiana, Tomeh, Holat, & Charnois, 2024）。このタスクに対して、様々な固有表現種別のデータを用いてLLMを指示チューニングし多様な固有表現種別へ適応させる手法（Zhou, Zhang, Gu, Chen Muhao, & Poon, 2024）や、LLMのプロンプトで抽出したい情報について詳細なガイドラインを与えることで固有表現抽出を含むさまざまな情報抽出をzero-shotで行う手法（Sainz, et al., 2024）が提案されている。本稿の試みは、このオープンタイプNERの考え方に基づいて特許文書における固有表現抽出や、一般的な意味での固有表現ではない発明構成要素の抽出を狙ったものである。

2.2.3 抽出対象

本稿では、以下の3種類の要素を抽出対象とする。

- 構成要素（Component）
- 化学物質名（Chemical）
- 数量（Quantity）

化学物質名や数量は一般的に固有表現として扱われるものだが、特許文書では発明を構成する要素として「インクタンク 100」や「負圧センサ 52」のように番号を付けて特定のものを表す表現や、「第1の領域」や「第1噴射燃料」のように「第〜」として他の同類のものと区別する表現が多用される。こうした表現は発明内容の正確な理解に不可欠であり、他言語への翻訳においては訳語の統一について十分な注意を払うべきものである。したがって、特許翻訳における訳語の統一やその評価において注目すべきものとして本稿ではこの3種類を扱うこととした。

2.2.4 実験設定

LLMを用い上記の要素を特許文書から自動抽出する試行実験を以下の設定で行った。

まず、対象とする特許文書として、特許日英対訳コーパスであるJaParaPat v1.0[1]（Nagata, Morishita, Chousa, & Yasuda, 2024）の日本語文および英語文を用いた。JaParaPat v1.0は1億行を超える日英対訳文で構成されるが、本試行実験ではそのうち日本で2016年に出願された特許を優先権主張の対象とする米国特許（jp-us）のデータの出願番号の若い順から200件分(JP2016000007-US20150362716からJP2016001780-US20150365110)、合計49,259行のデータのみを対象とした。

LLMとしてはGPT-5.2と、GPT-4.1を、OpenAIのChat Completion APIを通じて利用した。API呼び出しの実行にはPythonのopenaiモジュールを用い、温度パラメータやサンプリングのパラメータはデフォルト値（temperature: 1.0, top_p: 1.0）のままとした。また、GPT-5.2で有効化できる推論（reasoning）機能は今回利用しなかった。LlamaやGPT-OSS等のオープンウェイトLLMについても具体的な検討を行ったものの抽出の結果が芳しくなかったため、本試行実験ではAPIを利用することとした。

LLMのプロンプトには図1に示すシステムプロンプトおよびユーザープロンプトを用い、ユーザープロンプトの直後にJaParaPatのデータを1行付加してAPIに渡すようにした。文脈情報

[1] https://www.kecl.ntt.co.jp/icl/lirg/japarapat/

システムプロンプト：

You are an expert linguistic annotator.

ユーザープロンプト：

Please find named entities in the given {LANGUAGE} sentences and extract every instance in their exact order of appearance as Component, Chemical, or Quantity in a TSV format (Entity [TAB] Type). No title row is needed. Output ONLY the TSV data. Do not de-duplicate or skip identical entities; extract each occurrence separately. Use only the raw type names (Component, Chemical, or Quantity) without brackets. Extract all numerical expressions, including values with any units, ranges, and variables, as Quantity without omission. Importantly, do not extract abstract nouns that describe a state or property.

図 1:　本試行実験で用いた LLM へのプロンプト
（{LANGUAGE}には言語名 として Japanese もしくは English が入る）

を考慮するためには文書単位で処理を行うことが望ましいものの、文脈長が長くなることによる抽出漏れや出力フォーマットの乱れを懸念し、本試行実験では 1 行ずつの処理としている。プロンプトでは出力形式と抽出対象を指定しているが、抽出対象の定義については初期検討段階ということもあり数量（Quantity）に補足を加えた以外はほぼ種別名のみの記載に留まっている。

2.2.5　実験結果

GPT-4.1 と GPT-5.2 による構成要素および固有表現の抽出数を表 1 に示す。種別が「その他」となっているのはプロンプトで与えた 3 種別以外の種別ラベルを付した結果が返ってきたものであり、ここでは種別制約違反とみなした。抽出数の結果において日本語と英語で傾向に大きな違いはなく、GPT-5.2 では GPT-4.1 より数量の抽出数が増加し、その他では抽出数が減少している。特に指定外の種別ラベルを付した種別制約違反の数は GPT-5.2 では大きく削減される結果となった。出力された指定外の種別ラベルを見ると、GPT-5.2 では指定された種別ラベルに不要な文字や記号が連結されてしまったもの（例 : pChemical）が比率として多いのに対し、GPT-4.1 ではそれに加え別の種別ラベル（例：Material、Type）や一文字の種別ラベルが多く出現し、プロンプトで与えた出力の形式に違反しているものが多かったと言える。これより、GPT-5.2 は GPT-4.1 よりもプロンプトで与えた制約を守ろうとする傾向が強いことがうかがえる。

抽出結果を詳細に見てみると、GPT-4.1、GPT-5.2 ともそれぞれの種別に対応するとはみなせないものの過剰抽出が目立った。

発明の構成要素については、日本語では「本発明」「図 1」が GPT-4.1 と GPT-5.2 に共通して数多く抽出され、英語では GPT-4.1 は “surface” や “vehicle” 等の一般名詞、GPT-5.2 は “FIG.1” といった図番号が多く抽出する傾向が見られた。

表 1: 構成要素および固有表現の抽出数

言語	種別	GPT-4.1	GPT-5.2
日本語	構成要素（Component）	176,745	153,971
	化学物質（Chemical）	24,302	12,186
	数量（Quantity）	54,056	60,305
	その他 ※種別制約違反	54	11
英語	構成要素（Component）	178,460	161,970
	化学物質（Chemical）	28,134	16,880
	数量（Quantity）	60,292	82,675
	その他 ※種別制約違反	84	28

化学物質名については「アンモニア」「アルミニウム」等が抽出される一方で、「インク」「オイル」といった一般名詞の過剰抽出が目立った。英語についてもほぼ同じ傾向が見られた。

数量についてはモデルの差が大きく出ており、日英とも GPT-4.1 が「図 1」等を誤って数量として抽出した一方、GPT-5.2 は「第 1」等を誤って抽出した。

2.2.6 おわりに

本稿では特許翻訳における発明の構成要素の正確な翻訳とその評価の仕組みを確立することを最終目標にした、大規模言語モデルを用いた特許文書からの固有表現と発明構成要素の zero-shot 抽出の試みについて述べた。JaParaPat の日英対訳データに対して GPT-4.1 と GPT-5.2 を用いて行った試行実験では正確性の面で問題が多く、特に各種別に合致しない要素の過剰抽出が目立つ結果となった。

今回の試行実験の結果を踏まえた今後の課題は以下の通りである。将来的には対訳コーパスにおける固有表現や発明構成要素の言語間アラインメントの実現に繋げ、特許翻訳およびその評価のために利用することを目指す。

- 固有表現・発明構成要素抽出の検証用データセットの作成と、データセットを利用した再現可能かつ客観的な抽出性能評価の実現
- プロンプトにおける抽出すべき要素の定義の詳細化、また few-shot 学習のための事例利用等による、プロンプトの曖昧性に起因する過剰抽出の抑制
- オープンウェイトの軽量な LLM の利用
- LLM の推論機能を利用した出力形式への確実な追従と固有表現・発明構成要素抽出の精度向上
- 対訳コーパスの両言語の情報を利用することによる固有表現・発明構成要素のバイリンガル同時抽出
- より多くの特許文書データを用いた検証

参考文献

Isozaki, H., & Kazawa, H. (2002). Efficient Support Vector Classifiers for Named Entity Recognition. *COLING 2002: The 19th International Conference on Computational Linguistics*.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.*

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., . . . Guo, C. (2025). GPT-NER: Named Entity Recognition via Large Language Models. *Findings of the Association for Computational Linguistics.*

Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2024). GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Zhou, W., Zhang, S., Gu, Y., Chen Muhao, & Poon, H. (2024). UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *arXiv preprint arXiv:2308.03279*.

Sainz, O., García-Ferrero, I., Agerri, R., de Lacalle, O. L., Rigau, G., & Agirre, E. (2024). GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction. *arXiv preprint arXiv:2310:03668*.

2.3 LLM による特許請求項翻訳のための後修正とプロンプト調整

愛媛大学　枝松　泰志
後藤　功雄
二宮　　崇

2.3.1　はじめに

特許文書の翻訳は発明の権利範囲を言語の壁を超えて正確に伝達し、国際的な審査・権利行使・訴訟において法的効力を確保するために不可欠である。そのため特許文書の翻訳が人手で行われているが、経済的・時間的なコストが大きく、コスト削減のために、翻訳の自動化を目的とした特許文書に特化した機械翻訳の手法[1]が提案されている。

特許文書のうち、特許請求項とは法的な保護範囲を定義する部分である。特許請求項は特許の中で法的な影響力を持ち、翻訳において非常に高い正確性が求められる。特許請求項翻訳の課題として、一文が 100 文字を超えるほど長く、構成が複雑であるため正確な法的解釈を維持したまま翻訳することが困難であるということがある。 さらに特許文書特有の表現や技術用語に対しての適切な訳語を選択することも困難である。近年、大規模言語モデル（LLM）の登場により、機械翻訳は既存の翻訳モデルを上回る性能を達成するようになった。特許請求項翻訳においても、LLM を用いて翻訳をした研究[2]が行われており、高い性能を示している。しかし、LLM を用いた特許請求項翻訳の課題は依然として明確になっていない。

本稿は、WAT2025 において我々が提案した商用 LLM を用いた三つの翻訳手法[3]について報告する。一つ目は Pinzhen Chen らの研究[4]の手法に基づき、LLM による翻訳後に LLM-as-a-Judge[5]で翻訳品質を評価し、その結果を用いて翻訳を修正する後修正翻訳手法を提案する。二つ目は特許請求項に特化したプロンプトを用いて翻訳するプロンプト修正翻訳手法である。三つ目は実際の特許請求項を用いた Few-shot 学習に基づく翻訳手法である。また、本研究で提案する手法を用いて、WAT2025 の日英特許請求項翻訳タスクに参加した。これは日英両方向の特許請求項翻訳の精度を競いながら、正確な特許請求項翻訳の自動評価指標の確立を目指すワークショップである。今回は日英方向のタスクに参加した。

LLM-as-a-Judge を用いた評価実験の結果、LLM を用いた後修正翻訳手法で有効性が確認された。一方で、プロンプト修正翻訳手法や Few-shot 翻訳手法では効果は見られなかった。これに対し、LLM-as-a-Judge を用いた後修正翻訳のみを対象に実施された人手評価では改善が確認されなかった。この結果、本研究における LLM-as-a-Judge の性能が十分でないことが確認された。また、自動評価指標では全ての手法において性能改善の効果が確認でき、特にプロンプト修正翻訳で高い性能改善の効果が確認できた。

2.3.2　LLM-as-a-Judge を用いた後修正

本手法では、特許請求項翻訳における課題である長文かつ複雑な構造を克服するため、LLM を

初回翻訳プロンプト

```
あなたは特許請求項の日本語→英語 翻訳の専門翻訳者です.
以下の完全なポリシーを厳格に守ってください.
与えられた単一の日本語行を，米国特許クレーム風の英語として「必ず 1 行」に翻訳してください.
返すのは翻訳結果のみ（注釈なし）. 内容の追加・省略・分割・結合を禁止します.
出力は最終的な英語クレーム行のみ. ラベル，注釈，Markdown は禁止.
入力 1 行→出力 1 行を厳守（余計な改行を入れない）.
必ず 1 文のみ. 末尾はピリオドで終える.
法的スコープを完全に保持. 内容の追加・省略・分割・結合・再解釈は禁止.
クレーム種別（装置/システム/方法/組成/用途/プログラム/媒体），番号，他クレーム参照（従属関係）を保持.
数値・単位・記号・式・不等号・ラベル・範囲は全て厳密に保持（ただし後述の ASCII 正規化は例外）.
ASCII のみで出力. 非 ASCII 文字は ASCII 等価に置換（ASCII ルール参照）. “^^e2^^80^^93 — ～ μ ° ” などは絶対に出さない.
従属クレームや wherein で新しい要素/限定を勝手に導入しない.
曖昧さを避ける."and/or" は絶対に使わない.
この単一の日本語クレーム行を英語に翻訳してください:
\{ここに実際の日本語文が入る.\}
```

図 1: 初回翻訳プロンプト

表 1: LLM-as-a-Judge の指標

指標
法的範囲の忠実性
米国特許クレームの様式，構造の適合
数値・単位・範囲・式の正確性
先行詞対応・参照整合性
用語の正確さ・統一性
表現の自然さ・読みやすさ

用いて、人間の翻訳作業で一般に行われる工程を再現する。具体的には初回翻訳、翻訳結果の評価（LLM-as-a-Judge）、および評価に基づく修正（後修正翻訳）という三つのステップを再現する。

まず、初回翻訳として、原文を LLM に入力して翻訳する。入力する原文は元の改行情報を保持したまま請求項単位で与える。これは請求項における改行情報が文構造や係り受け関係を反映することが多く、LLM が文構造を正しく理解するための情報として有用と考えたためである。図 1 に、モデルに入力したプロンプト例を示す。

初回翻訳の生成後、原文および翻訳文を入力として、参照文を用いずに LLM-as-a-Judge による翻訳品質評価を実施する。評価は表 1 の独自の六つの観点に基づいて行い、各項目についてエラー内容の記述と 100 点満点による評価点数を出力させる。図 2 に、モデルに入力したプロンプト例を示す。

LLM-as-a-Judge による評価の後、出力された評価結果、原文、および初回翻訳を入力として LLM に与え、最小限の修正のみを施した英文を生成させる。図 3 に、モデルに入力したプロンプト例を示す。

LLM-as-a-Judge プロンプト

```
あなたは非常に几帳面な特許請求項レビュアです\\
参照訳（正解訳）は一切使わず，原文日本語（JA）に対して英訳（PRED）を評価してください.\\
次の評価軸と出力形式に従ってください.\\

評価カテゴリ（各 0^^e2^^80^^93100，6 カテゴリは等加重）:\\
- fidelity\_legal\_scope（法的スコープの忠実性）\\
- us\_style\_structure（米国クレーム文体・構造）\\
- numbers\_units\_ranges（数値・単位・範囲）\\
- antecedent\_dependency（先行詞・従属関係）\\
- terminology（用語）\\
- naturalness（自然さ）\\

出力に以下のセクションを含めてください:

### Findings\\
- 具体的な問題点，または問題ない点（確認）を箇条書き.\\
  （法的文体，忠実性，数値/単位/式/範囲，用語一貫性，先行詞，従属関係，句読点/フォーマット，自然さ等）
- 各項目は [Fidelity]，[Numbers/Units] などのラベルで開始し，JA/PRED の該当箇所を正確に引用してください.

### Fix Suggestions\\
- Findings に対応づけた箇条書きで，最小限の修正で済む「修正案（英語断片）」を示してください.
```

図 2: LLM-as-a-Judge プロンプト

後修正翻訳プロンプト

```
あなたは特許請求項のプロ翻訳者で，第二稿（2nd pass）を作成します.\\
以下のポリシーを厳守してください. 参照訳は一切見ません.\\
与えられる「日本語行」「初回英訳」「評価者によって評価された評価」をもとに，\\
内容を追加・削除せずに英訳を修正してください.\\
出力は米国クレーム文体の英語 1 行のみ.\\


### Japanese（与えられたスコープを超えて翻訳しない）\\
\{実際の原文\}\\

### First-pass English\\
\{初回翻訳結果\}\\

### Issues to fix（抽象的. 参照文は使っていない／与えられていない）\\
\{LLM-as-a-Judge の結果\}\\

上記を踏まえて英語行を書き直してください（1 行のみ）.
```

図 3: 後修正翻訳プロンプト

2.3.3　特許請求項に特化したプロンプト修正

本手法では 2.1.2 節で導入した翻訳プロンプトを改良し、意味と法的効力を保持しつつ、米国特許請求項の慣習に、より忠実に従った翻訳を目指す。具体的には米国特許スタイルで翻訳すること、先行する他の請求項を引用しない独立請求項と引用する従属請求項とを明確にすること、句読点の統一、冠詞の用法などを正確に行うことを明記した。2.1.2 節のプロンプトに、図 4 のプロンプトを追加した。

2.3.4　特許請求項を用いた Few-shot 学習に基づく翻訳

本手法では特許請求項翻訳における専門用語や特許特有の表現を正確に翻訳するために、FAISS[6]および SentenceTransformer[7]により検索された翻訳例を用いた Few-shot 翻訳を行っ

プロンプト修正翻訳用プロンプト

```
原則として開放的遷移語は "comprising:"\\
並列要素の区切りはセミコロン.\\
2 個以上の並列列挙では最後の要素の前に "; and" を入れる.\\
要素名は明確にし，同じ要素は同じ語で一貫して参照する.\\
先行する他の請求項を引用しない独立請求項と引用する従属請求項とを明確にする.\\
  *既出要素への限定のみ追加. 新しい要素は導入しない.* \\
初出: "a/an/at least one [X]"（または "a first [X]", "a second [X]"）\\
- 再出: "the [X]"\\
- 定義後に "a/an" に戻さない. 単数/複数も一貫.\\
  - 各 wherein 節を，正しい先行要素に結び付ける.\\
- 並列条件は構造化して書く:\\
  "wherein: (i) ...; and (ii) ... ."\\
- wherein 節で新しい要素は導入しない.\\
```

図 4: プロンプト修正翻訳用プロンプト

検索用プロンプト

```
"あなたは日本語特許請求項の専門家です.
\\以下の日本語 1 行（1 クレーム）について，"\\
"FAISS で高精度に用例を拾うための『日本語クエリ』を**ちょうど 3 件**，JSON 配列で
出力してください."\\
"一般語（装置，処理，データ等）や曖昧語を避け，品詞は名詞句中心で**8〜24 文字程
度**に収めます."\\
"括弧・全角記号・機種依存文字は使用しません."\\
"説明や余計な文字は一切付けないでください."\\
```

図 5: 検索用プロンプト

た。FAISS は大規模なベクトルの集合から類似したベクトルを効率よく検索することを目的とした高速なベクトル類似度検索ライブラリである。FAISS インデックスの構築にあたっては、日英特許出願コーパスである JaParaPat[8]を用いた。本コーパスは日本語および英語の特許文書間でアライメントされた大規模な日英対訳コーパスである。2016 年から 2020 年の間に出願された特許文書ファミリーから自動抽出された、約 1 億文対から構成されており、出願種別ラベルや文書 ID などのメタデータを含んでいる。本研究では、出願種別ラベルのうち PCT ルートに分類されている日英対訳文対を使用する。PCT ルートでは単一の国際特許出願が翻訳を通じて複数の国の特許庁に提出されるため、その結果得られる多言語公開文書は実質的に対訳関係にある。これらの文書は同一出願に対する直接的な翻訳であることから、高い信頼性を有する対訳データとみなすことができる。また、本研究は、請求項の文のみを使用した。日本語の請求項文は SentenceTransformer に基づく、多言語埋め込みモデル（intfloat/multilingual-e5-base）[9]を用いてベクトルに埋め込み、文間の意味的類似度をコサイン類似度に基づいて評価できるようにする。これにより、入力された請求項に対して意味的に類似した日英文対を FAISS によって検索し、得られた用例を例文としてプロンプトに加えて Few-shot 翻訳を行うことが可能となる。

本研究では二種類の Few-shot 手法を用いた。一つ目の手法では特許請求項一文を FAISS インデックスに対して検索にかけ、類似度の高い上位三つの日英文対を取得し、それらを例文として翻訳プロンプトに挿入する。二つ目の手法では原文中に重要な用語を含む翻訳例を取得するため、まず LLM を用いて入力文から三つの用語を抽出する。図 5 にモデルに入力したプロンプトを示す。こうして抽出した用語を FAISS インデックスで検索にかけ、各用語との類似度が最も高い日

表 2: 各データの請求項数

	検証データ	評価用データ
特許数	13	26
特許請求項数	19	70

英文対を取得し、合計三つの日英文対を例文として翻訳プロンプトに挿入する。

2.3.5 実験

2.3.5.1 実験設定

本研究では基盤となる LLM として OpenAI の GPT-5[10]を使用する。

本研究では WAT2025 の「Patent Claims Translation/EvaluationTasks」において提供されている公式の検証、評価データを使用した。検証データは原言語の特許請求項とそれに対応する翻訳文から構成されているのに対し、評価データでは原言語の特許請求項のみが含まれている。これらのデータは特許ごとにファイルが用意されており、その中に特許請求項が存在している。各データセットに含まれている特許数および特許請求項数を表 2 に示す。

特許請求項は非常に長く、構造的に複雑な文を含むため、各請求項の構造を完全に理解することが難しい。さらに翻訳の正確性は意味だけでなく、法的範囲や技術用語の適切性といった多次元的観点から維持される必要がある。そこで本研究では、LLM を用いて翻訳品質を評価する LLM-as-a-Judge を採用し、長文全体にわたる翻訳の適切性を一貫して評価する。また WAT2025 における公式評価として、タスク主催者による人手評価が実施された。本研究では、これら二つの評価指標を用いて特許請求項翻訳の品質を評価すると共に、LLM-as-a-Judge による評価の妥当性および LLM を用いた特許請求項翻訳の課題について調査する。

また、本研究の主たる分析は LLM-as-a-Judge および人手評価に基づくが、特許請求項翻訳を従来の自動評価指標がどのように評価するかを確認することも有益であると考え、補足的な評価として既存の自動評価も実施した。

人手評価では各原文およびそれらに対応する翻訳文に対して、手動によるエラー注釈および評価スコアが付与された。エラー注釈は誤訳、訳抜け、湧き出しなどのエラーカテゴリに基づいて問題のある箇所に付与され、各エラーに重大度（Major または Minor）がラベル付けされた。 さらに各文に対して 100 点満点のスコアが割り当てられた。翻訳結果に評価優先度を付与し、LLM-as-a-Judge を用いた後修正翻訳のうち、初回翻訳と後修正翻訳の二つの翻訳結果がタスク主催者によって評価された。各翻訳結果に対して、70 文からなる評価文のうち 28 文が人手により評価された。

LLM-as-a-Judge による評価では、翻訳文を LLM が評価する LLM-as-a-Judge を用いる。評価基準は表 1 で定義したものと同一のものである。

表 3: 人手評価による誤りカテゴリ分類結果

	初回翻訳		後修正翻訳	
エラー種類	Major	Minor	Major	Minor
訳抜け	8	13	24	14
用語の一貫性	1	33	0	36
文法	1	8	2	1
誤訳	5	18	7	25
その他	0	2	0	5
文脈上不適切	3	11	2	14
湧き出し	0	10	5	34
原文エラー	0	2	1	2
句読点エラー	0	17	0	8
一貫性の欠如	0	0	0	4
ぎこちない	0	4	0	5
冠詞	0	0	1	2
エラー総数	18	119	42	150
エラー総数（Major+Minor）	137		192	

表 4: 人手評価による評価スコア

	初回翻訳	後修正翻訳
平均スコア	86.1	81.6

自動評価指標として COMET[11]および BLEU[12]を用いた。評価データは参照訳がなく、参照訳を使う自動評価手法は適用できないため、この自動評価では参照訳を含む検証データを用いて算出する。

2.3.5.2 人手評価の結果

人手評価における誤りカテゴリと評価スコアをそれぞれ表 3、表 4 に示す。初回翻訳と比較すると、後修正翻訳の出力は Major エラー数が 18 から 42 に増加し、Minor エラー数は 119 から 150 に増加した。その結果、総エラー数は 137 から 192 に増加し、平均評価スコアは 86.1 から 81.6 に低下した。

2.3.5.3 LLM-as-a-Judge の結果

LLM-as-a-Judge による評価結果を表 5 に示す。LLM による後修正翻訳では、初回翻訳よりも

表 5: LLM スコア

システム名	LLM-as-a-Judge スコア(%)
初回翻訳	91
後修正翻訳	92
プロンプト修正翻訳	80
Few-shot（全文検索）	84
Few-shot（用語検索）	78

表 6: 自動評価

システム名	COMET(%)	BLEU(%)
初回翻訳	84.6	53.8
後修正翻訳	84.9	48.9
プロンプト修正翻訳	85.4	56.6
Few-shot（全文検索）	85.1	53.9
Few-shot（用語検索）	85.0	53.5

高いスコアを達成した。一方で、プロンプト修正翻訳のスコアは初回翻訳と比較して低下し、Few-shot（全文検索）および Few-shot（用語検索）のいずれも初回翻訳より低いスコアを示した。したがって、LLM-as-a-Judge の評価では、これらの Few-shot およびプロンプト修正翻訳の手法の有効性は確認されなかった。

2.3.5.4　自動評価

実験結果を表 6 に示す。初回翻訳と比較して、提案手法全てにおいて COMET は上昇し、プロンプト修正翻訳と Few-shot 翻訳では BLEU も上昇した。

2.3.5.5　人手評価が実施された翻訳の分析

表 3 の人手評価の結果を元に特許請求項翻訳の分析を行う。文法誤りやカンマ・セミコロンの使用といった句読点に関する表層的な誤りは改善されたものの、他の種類の誤りについては改善が見られなかった。特に湧き出し、訳抜け、誤訳が大幅に増加しており、原文の特許請求項への忠実性に関わる誤りが増加したことが確認され、これらの数が多いほど翻訳スコアは減少した。具体的には「～手段」の「手段」や「～量」の「量」などの訳抜けや、「each of」や「configured to」などの湧き出しといった、細かい単語レベルのエラーがよく見られた。よく見られた誤訳のエラーは「a」と「the」の誤選択や脱落といった冠詞のエラーに起因していた。これらの冠詞のエラーを除くと、誤訳のエラーでは初回翻訳は 11 個、後修正翻訳は 13 個で差は小さかった。ま

た、湧き出しのエラーとして、原文では存在しなかったが要素を並列するときの「(i)」、「(ii)」の記号が湧き出しとしてカウントされた。一方で用語の一貫性や文脈上の不適切さに関する誤りも多数観察された。

今回、LLM-as-a-Judge の評価指標では法的な忠実性の他に米国特許クレームの様式や表現の自然さ、読みやすさといった翻訳の表層的な部分を同じ比重で評価している。その結果文法や句読点の修正、米国式請求項スタイルへの適合、特許翻訳で一般的な表現の導入といった表層的改善は見られたが、一方で範囲や比較方向の誤り、先行詞参照の誤り（誤訳)、原文に存在しない要素の追加（湧き出し)、および必須要素の脱落（訳抜け）などの原文忠実性の低下に関わる誤りが増加したと考えられる。特に、後修正翻訳では最小限の修正を行うのではなく、文全体を書き換える傾向が見られたため、原文への厳密な忠実性よりも整った英語表現の生成が優先された可能性がある。

人手評価と LLM-as-a-Judge 評価を比較すると、後修正翻訳では LLM-as-a-Judge の枠組みではスコアが向上した一方で、人手評価では翻訳品質が低下していた。したがって、本研究においては LLM-as-a-Judge フレームワークの性能が十分でなかったことが確認された。

2.3.5.6　人手評価が実施されていない翻訳の分析

プロンプト修正翻訳および Few-shot 翻訳について、LLM-as-a-Judge による評価結果を表 3 の基準を参考に分析した。プロンプト修正翻訳では米国特許請求項スタイルとの整合性、英語出力の自然さ、用語および単位表現の安定性を向上させた。一方で、請求項の権利範囲に関する不正確な修正や、誤った従属関係の挿入により、法的範囲の忠実性や先行詞対応・参照整合性のスコアは低下した。これはモデルが「自然な英語」の生成および「米国式請求項スタイル」への準拠に強く偏る傾向があり、その結果、構造保持や法的忠実性といった特許翻訳において本質的な側面が損なわれやすいためである。この忠実性の低下が、LLM-as-a-Judge の評価スコアの低下につながったと考えられる。

同様に、Few-Shot 翻訳では、句読点の配置、要素列挙、「configured to」の使用に代表される語彙的一貫性、および米国特許請求項特有の文構造の安定化など、表層的改善が見られた。しかし、Few-Shot 翻訳は文体的一貫性や用語の安定性を向上させる一方で、検索された例文の構造的バイアスの影響を強く受けるため、要素の再編成、節位置の移動、限定条件を導入するためによく使われる wherein 句の不要な挿入といった構造的歪みが翻訳出力に生じ、LLM-as-a-Judge の評価スコアの低下につながったと考えられる。以上の分析から、プロンプトによって LLM に制約を課した場合、本研究で用いたモデルでは、全体的な内容の忠実性を低下させることなく表層的な制約を満たすことが難しいことが示唆された。これは、今後の研究における重要な課題である。

2.3.6　まとめ

特許請求項の翻訳に焦点を当てて、LLM を用いた特許請求項翻訳の品質向上と課題分析を行った。本研究では LLM-as-a-Judge を用いた後修正翻訳、プロンプト修正翻訳、Few-shot 翻訳の三

つの異なるアプローチについて提案と比較を行った。その結果、LLM-as-a-Judge を用いた後修正翻訳では LLM-as-a-Judge による評価スコアは改善した。一方で、プロンプト修正翻訳と Few-shot 翻訳は、LLM-as-a-Judge 評価において改善を示さなかった。これに対して、公式の人手評価では、LLM-as-s-Judge を用いた後修正翻訳において、初回翻訳と後修正翻訳を比較した結果、総エラー数は増加し、平均スコアは初回翻訳の 86.1 から後修正翻訳の 81.6 に低下しており、人手評価に基づく品質が悪化した。この結果は、本研究において LLM-as-a-Judge フレームワークの性能が十分ではなかったことを示している。

分析の結果、提案手法の後修正翻訳は初回翻訳に比べて表層的な品質エラーの改善には寄与したものの、訳抜け、湧き出しといった日英文対での対応関係の誤りや係り受け関係の誤りといった元の特許請求項に対する忠実性に関してエラーが増加した。プロンプトによって LLM に制約を課した場合、本研究で使用したモデルでは、全体の忠実性を低下させることなくそれらの制約を満たすことが難しいという結果になった。

今後の課題として、まず日英文対における語単位での対応関係を厳密に保持し、元の特許請求項に対する原文忠実性の高い翻訳を行うことが挙げられる。また特許請求項翻訳に適した LLM-as-a-Judge を確立し、人手評価指標との相関を実証することも今後の課題である。

参考文献

[1] Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, Mikio Yamamoto. Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation, 2017.

[2] Haruto Azami, Minato Kondo, Takehito Utsuro, Masaaki Nagata. Patent Claim Translation via Continual Pre-training of Large Language Models with Parallel Data. In Proceedings of Machine Translation Summit XX: Volume 1, pp. 300–314, 2025.

[3] Taishi Edamatsu, Isao Goto, Takashi Ninomiya. Ehime-U System with Judge and Refinement, Specialized Prompting, and Few-shot for the Patent Claim Translation Task at WAT 2025. In Proceedings of The 12th Workshop on Asian Translation (WAT 2025), 2025.

[4] Pinzhen Chen, Zhicheng Guo, Barry Haddow, Kenneth Heafield. Iterative Translation Refinement with Large Language Models. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pp. 181–190, 2024.

[5] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, Jian Guo. A Survey on LLM-as-a-Judge, 2025.

[6] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, Hervé Jégou. The Faiss library. arXiv 2401.08281, 2024.

[7] Nils Reimers, Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019.

[8] Masaaki Nagata, Makoto Morishita, Katsuki Chousa, Norihito Yasuda. 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, pp. 9452–9462, 2024.

[9] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, Furu Wei. Multilingual E5 Text Embeddings: A Technical Report, 2024.

[10] OpenAI. OpenAI GPT-5 System Card, arXiv 2601.03267, 2025.

[11] Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie. COMET: A Neural Framework for MT Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 2685–2702, 2020.

[12] Matt Post. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 186–191, 2018.

3. 特許請求項翻訳タスクのワークショップ実施報告

3. 日英特許請求項翻訳タスクの実施結果と今後の展望

東京大学　中澤　敏明

3.1 はじめに

ニューラル機械翻訳(NMT) や大規模言語モデル(LLM) を用いた機械翻訳の性能は劇的に向上しており、言語やドメインによっては人間による翻訳を凌駕する場合もある。しかし、現在、機械翻訳の性能を正確に評価する万能な手法は存在しない。COMET [2] のような広く使われている指標であっても、学習データと異なるドメインに適用した場合、不安定または不正確な結果をもたらすことが報告されている[3]。

これは特許文書の翻訳にも当てはまる。平均的な翻訳品質は大幅に向上しているものの、適切な用語の使用や用語の一貫性などを正確に評価することは依然として困難である。特に特許請求項は、その長さと独特の文体により、正確な評価をさらに困難にしている。

そこで WAT2025[9]において、日英の特許請求項翻訳に焦点を当てた共有タスクを実施した。目的は翻訳品質を競うだけでなく、最終的に翻訳結果を正確に評価できる自動評価手法を開発することにある。第 1 回となる今回は、様々な手法による翻訳出力を収集し、人手で誤りをアノテーションすることで、将来的な自動評価モデル開発のための学習データを作成することを主目的とした。

3.2 データセット

3.2.1 学習データ

学習データとして、日英特許パラレルコーパス JaParaPat [4] の 2025 年 8 月版公開サブセット 2）を使用した。このサブセットは、2016 年から 2020 年までの期間をカバーし、1 億を超える対訳文対からなる。JaParaPat は日本国特許庁(JPO) と米国特許商標庁(USPTO) の公開特許公報から作成されたもので、パテントファミリー情報に基づいてアライメントされている。表 1 に提供した学習データの統計値を示す。

	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

表 1：学習データに含まれる文数

国際特許出願には主に 2 つのルートがあり、パリ条約ルートと特許協力条約（PCT）ルートである。JaParaPat には、これら両方のルートに基づくデータが含まれている。表 1 において、パ

リ条約ルートのうち、「jp-us」は日本で最初に出願され、その後アメリカ合衆国で出願された特許対を指す。「us-jp」はアメリカ合衆国で最初に出願され、その後日本で出願されたものを指す。「jp-x-us」は、日本およびアメリカ合衆国以外の国で最初に出願され、その後日本とアメリカ合衆国の両方で出願された特許を指す。公開版では、文書アラインメント、文分割、文アラインメントに異なる手法を採用しているため、元の JaParaPat 論文の表と比べて、文対の数が異なっている。

特許請求項翻訳の共有タスクにおける学習データとして見た場合、JaParaPat における最も重要な問題の一つは、特許請求項に対する文分割およびアラインメントである。JaParaPat では、長い請求項を改行によって複数のセグメントに分割し、セグメント単位でアラインメントを行うことが多いため、請求項レベルでのアラインメントを再構築することが困難である。我々は、この問題をどのように解決するかについて、JaParaPat の著者らと議論を行っている。

3.2.2 開発データ

今回は明細書本文ではなく請求項に焦点を当て、比較的難しい文構造、専門用語、非専門用語、曖昧な言語（複数の解釈が可能なフレーズなど）をエンジンがどのように処理するかを確認した。段落の長さ、用語の特殊性、構文、構造的/意味的曖昧性などを考慮して、既存の特許出願文書から日本語 13 件(19 請求項)、英語 11 件(11 請求項) を選択した。

予備的な人手評価として、JaParaPat で学習した NMT モデルとオープンウェイトの LLM を使用して開発データを翻訳し、アノテータに誤りの特定、全体的な品質スコアの付与、事後編集（ポストエディット）を依頼した。この事後編集された翻訳を参照訳として開発データを提供した。

3.2.3 テストデータ

テストデータのソーステキストは、既存の特許出願から選択された。選択にあたっては、以下の要素を考慮した。

・ 既存翻訳の有無: ファミリー出願が存在する場合、LLM が検索エンジンを通じて正解（公式翻訳）を見つけてしまう可能性がある。そのため、少なくともソーステキスト配布時点でターゲット言語での対応出願が存在しないものを選択した。これにより、公開データから参照訳を自動収集することができず、また予算の制約から参照訳を作成することもできなかったため、後述の通り参照なしの自動評価を行った。

・ 長さ/構成: 改行のない単一の長文テキストは、翻訳品質の低下を招く可能性があることが知られている[5]。本研究の主目的は、異なる翻訳エンジンが文の長さにどのように対処するかを検証することではなく、一般的な特許請求項の表現がどのように処理されるかを確認することであった。そのため、改行の有無にかかわらず、英語では概ね 220 語以内、日本語では 500 文字以内の原文テキストを選定した。

・ 分野: 原文テキストは、情報処理、通信、電気工学、化学など、さまざまな技術分野における出願から取得した。

・ 曖昧性と画像情報: 特許請求項は画像を参照しないと正しい解釈に到達できない場合がある。一部のマルチモーダル LLM などでは画像情報も利用した翻訳を行うことができるが、まだ一般的ではないため、今回は、画像などの追加情報なしで理解可能なテキストを選択した。例えば「区間」という用語は時間的概念(interval) か空間的概念(section) か曖昧になり得るが、文脈から意味が確定できる場合のみを含めた。

これらの要素を考慮し、日本語から英語(Ja-En) 方向に 26 文書（70 請求項)、英語から日本語(En-Ja)方向に 30 文書（81 請求項）を用意した。

3.3 参加システム

本タスクには 2 チーム(UTSK25, EHIME-U) が参加し、主催者が 3 つの商用システム(Commercial 1, 2, 3) の翻訳結果を収集した。

・ UTSK25: JaParaPat で継続事前学習を行ったオープンウェイト LLM

・ EHIME-U: クローズド/プロプライエタリな LLM に対するプロンプトチューニングと後修正

・ Commercial 1: オンラインサービス（標準プロンプト）

・ Commercial 2: クローズドシステム

・ Commercial 3: 翻訳用無料 LLM モデル

3.4 評価結果

3.4.1 自動評価

テストセットに対応する参照訳が存在しないため、MetricX-24-Hybrid-XL3）[6] および WMT23-CometKiwi-DA-XL4）[7] を用いた参照なし自動評価を行った。評価はセグメント（請求項）レベルと、文書全体を 1 つのセグメントとみなす文書レベルの 2 種類で実施した。その結果を表 2 および表 3 に示す。

System	ja-en		en-ja	
	MetricX ↓	CometKiwi ↑	MetricX ↓	CometKiwi ↑
UTSK25	3.761±1.654	0.544±0.122	3.623±1.474	0.641±0.111
EHIME-U 1	2.882±1.614	0.560±0.134	n/a	n/a
EHIME-U 2	2.987±1.607	0.568±0.131	n/a	n/a
Commercial 1	2.792±1.416	0.572±0.133	2.916±0.842	0.681±0.088
Commercial 2	3.879±2.454	0.567±0.139	3.126±1.031	0.676±0.093
Commercial 3	2.920±1.107	0.573±0.127	2.581±0.780	0.707±0.078

表 2：セグメントレベルの自動評価結果

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

表 3：文書レベルの自動評価結果

3.4.2 人手評価

予算の制約により、テストデータの一部（各方向 13 ファイル）に対して人手評価を実施した。アノテータには誤りの特定、および、全体的な品質スコアの付与の 2 種類の評価を依頼した。人手による誤りのアノテーション基準は Freitag ら[8] の指標を特許翻訳ドメイン向けに変更したものを使用した。品質スコア評価の結果を表 4 に示す。

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

表 4：人手評価の平均スコア

人手評価の平均スコアにおいて、両方向で最高精度を達成したシステムはなかったが、平均して Commercial 1 が最も高い精度を示した。また、人手評価と自動評価指標の相関係数を確認したところ(表 5)、どの指標も人手評価と実質的な相関を示さないという結果が得られた。この原因としては、参照なしの自動評価手法の限界、特許ドメインへの不適合、あるいは人手評価自体の不正確さが考えられる。

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

3.5 考察

翻訳出力と人手アノテーションの分析から、翻訳側とアノテーション側の双方に様々な問題が明らかになった。以下、主要な論点について議論する。

3.5.1 包括的用語と具体的用語の使用

文脈から正しい意味が導き出せるにもかかわらず、複数の解釈が可能なフレーズが存在する場合の翻訳について考察する。

例として、「前記信頼度情報が... 継続している区間を補正対象区間として検知して... 前記補正対象区間を走行している...」という原文がある。ここでは、他車両がその区間を「走行している(traveling)」ことから、「区間」が時間的な「期間(period/interval)」ではなく、道路の物理的な「区間(section)」を指すことは明らかである。

あるエンジンの出力は“... detect ... a section during which the reliability information...” となっていた。“Section” 自体は物理的・時間的双方の概念を取りうるが、前置詞“during” は時間的概念を決定づけてしまうため、この文脈では不正確である。一方で“in which” は曖昧（物理・時間の両方で解釈可能）だが、正解を包含している。概念が曖昧な場合、特定の用語(specific term) を選ぶよりも、包括的な用語(generic term) を選ぶ方が、正しく解釈される可能性が高まる場合がある。

また、特許特有の曖昧語として以下の例が挙げられる。

・ 挟まれる: 物理的または概念的に間に位置することを指す。“Sandwiched” と訳されることが多いが、文脈によっては不適切であり、単に“between” とする方が適切な場合がある。

・ 対象: 非常に曖昧で便利な用語。“Target”, “object”, “subject”, “... in question” など多義的である。

3.5.2 国や特許庁による慣習・法的制約の違い

翻訳先の国や地域の特許実務に合わせて、用語や概念を追加・省略する「調整(adjustment)」が必要な場合がある。

例として、「... プログラムであって、コンピュータを、... 記憶する記憶手段、... 判定する判定手段、...として機能させる、プログラム。」という原文がある。あるシステムは“means” （手段）という語を省略し、“causing a computer to: store ...; determine ...;” と訳出した。技術的には“means” があってもなくても同義である。しかし、米国特許実務の観点からは、means-plus-function (35 U.S.C. 112(f)) の解釈を避けるために“means” の使用を避ける実務家もいる。したがって、この省略は有益な場合がある。

しかし、アノテータはこの“means” の省略を「欠落：重大」なエラーとしてマークした。上記の理由から、これを重大なエラーとすべきではない。

3.5.3 アノテーションの問題

人手アノテーションには以下のような問題が見られた。

・エラーの見落とし: “thermostat”（サーモスタット）とあるのに「温度計」と訳された明白な誤訳の見落としなど。

・誤りではない箇所の指摘: 請求項の冒頭と末尾で主題（例：プログラム）を繰り返すのは日本特許の一般的な構造だが、これを「重大なエラー」とした。

・重大度の判定ミス: 文意を大きく歪める誤訳を「不自然：軽微」としたり、許容される構成を「重大なエラー」としたりするケースが見られた。

誤ったアノテーションを自動評価技術の開発データとして使用することは、悪影響を及ぼす。

3.6 結論と今後の展望

本稿では、第 1 回特許請求項翻訳タスクについてまとめた。本年は 2 チームから翻訳結果の提出があった。人手評価の結果に基づくと、いずれの翻訳方向においても一貫して高い性能を示したシステムは存在しなかった。しかし、自動評価結果との比較や人手アノテーションの分析を通じて、本論文で報告したように、さまざまな課題が明らかになった。

今後は、本研究で得られた知見を踏まえ、より安定的で高品質な人手評価の枠組みを定義するとともに、人手アノテーションを学習データとして活用し、特許翻訳に特化した高精度な自動評価手法の開発を目指す。

謝辞

本稿における人手評価は、アジア太平洋機械翻訳協会(AAMT) および AAMT/Japio 特許翻訳研究会の支援を受け実施したものである。

参考文献

[1] Toshiaki Nakazawa, Takashi Tsunakawa, Isao Goto, Kazuhiro Kasada, Katsuhito Sudoh, Shoichi Okuyama, Takashi Ieda, and Masaaki Nagata. Findings of the First Patent Claims Translation Task at WAT2025. In Proceedings of the Twelfth Workshop on Asian Translation, pp. 1–15, 2025.

[2] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework forMT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[3] 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, PhilippKoehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova,

Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, Proceedings of the Tenth Conference on Machine Translation, pp. 355–413, Suzhou, China,November 2025. Association for Computational Linguistics.
[4] Masaaki Nagata, Makoto Morishita, Katsuki Chousa, and NorihitoYasuda. JaParaPat: Alarge-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), pp. 9452–9462, Torino,
Italia, May 2024. ELRA and ICCL.
[5] Seiichiro Kondo, Kengo Hotate, Tosho Hirasawa, Masahiro Kaneko, and Mamoru Komachi. Sentence concatenation approach to data augmentation for neural machine translation. In Esin Durmus, Vivek Gupta, Nelson Liu, Nanyun Peng, and Yu Su, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 143–149, Online, June 2021. Association for Computational Linguistics.
[6] Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, Proceedings of the Ninth Conference on Machine Translation, pp. 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
[7] Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, Proceedings of the Eighth Conference on Machine Translation, pp. 841–848, Singapore, December 2023. Association for Computational Linguistics.
[8] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, andWolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, Vol. 9, pp. 1460–1474, 2021.
[9] Toshiaki Nakazawa and Isao Goto. 2025. Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025). Association for Computational Linguistics, Mumbai, India.

4. 国際ワークショップ開催報告

4. 国際ワークショップ開催報告：PSLT 2025

愛媛大学　後藤　功雄
奈良女子大学　須藤　克仁
静岡大学　綱川　隆司

4.1　開催概要

スイス・ジュネーブにて開催された機械翻訳サミット（Machine Translation Summit）2025 の併催ワークショップの一つとして、特許・技術文書翻訳ワークショップ（The 11th Workshop on Patent and Scientific Literature Translation; PSLT 2025）が 2025 年 6 月 24 日に開催された。本ワークショップは、アジア太平洋機械翻訳協会（AAMT）、および一般財団法人日本特許情報機構（Japio）による AAMT/Japio 特許翻訳研究会が中心となり 2005 年から隔年で開催しており今回で 11 回目を数える。ワークショップは現地開催で、現地に来られない発表者に対してはオンライン発表にも対応した形での開催となった。

本ワークショップでは日本国特許庁と WIPO からそれぞれ招待講演をいただき、2 件の一般講演があった。現地にて 20 名程度の参加があり、質疑も活発であり盛況であったと言える。

4.2　招待講演

招待講演 1 件目は日本特許庁の村上遼太氏から、特許庁の特許情報を用いた情報サービスに関するご講演をいただいた。特許庁では特許をデータベースに蓄積して、特許情報プラットフォーム（J-PlatPat）を通してユーザに特許情報を提供しており、J-PlatPat では機械翻訳を利用して日本語の特許を英語で提供したり、外国語の特許を日本語で提供していることが紹介された。さらにこのサービスの最近の追加機能や、日本語―インドネシア語の 500 万文対の対訳コーパスを機械翻訳システムで追加学習する効果などについて説明があった。

招待講演 2 件目はジュネーブに本部がある世界知的所有権機関（WIPO）の Bruno Pouilquen 氏から WIPO の機械翻訳に関してご講演をいただいた。WIPO の機械翻訳の歴史、WIPO の機械翻訳をドイツ、韓国、ユーラシア特許機構などの各国の特許庁が利用していることや、機械翻訳が機械学習により構築されていること、機械翻訳が IPC ドメインの情報を利用していること、訓練データの言語対毎のデータ量、他の翻訳システムとの BLEU スコア比較評価などについて解説いただいた。また IPC の自動分類や画像の自動分類についても紹介があった。

4.3　一般講演

一般講演 1 件目は Longhui Zou らによる英語から中国語への学術論文の翻訳での ChatGPT-4o と DeepSeek-V3 との比較評価についてであった。評価尺度として、参照訳を使わない自動評価（COMET-KIWI）、語彙多様性、構文の複雑さを用いた。結果は COMET-KIWI の平均スコアは DeepSeek-V3 の方が高く、語彙多様性は GPT-4o の方が高く、構文の複雑さも GPT-4o の方が高かった。LLM を学術翻訳に利用する実務者にとって、本研究の成果は、テキストの特性に基づく

モデルの選定に必要であることを示唆しているとのことでした。
一般講演 2 件目は Thomas Moerman らの発表で、彼らの手法は既存の 2 つの手法の組み合わせで、組み合わせた手法は、多分野にわたるデータから関連するデータを抽出するトピックフィルタリングと、利用可能なデータをより効率的に活用するためのファジーマッチ（FM）拡張である。3 つの科学分野における英語からフランス語への翻訳実験から、トピックフィルタリングと FM 拡張を組み合わせることにより、スクラッチから訓練したニューラル機械翻訳（NMT）モデルの性能が向上した。NMT システムは計算リソースが限られた状況下で科学文献を翻訳するための有力な選択肢となり得るが、翻訳性能と大規模言語モデル（LLM）のパラメータ数の増加に正の相関があることから、ファインチューニングなしでも大規模な LLM は、このような特化型 NMT モデルよりも優れた翻訳性能を発揮する可能性があることも示唆しているとのことでした。

4.4 所感

MT Summit 2025 には 5 つの併設ワークショップがあり、PSLT 2025 は同じ時間帯に他の 2 つのワークショップと同時開催になったものの比較的多くの方にご参加いただけ、また参加者は現地参加であったことから現地会場での質疑が活発であった。プログラム最後の総合討論の時間では、今年のアジア言語の翻訳ワークショップ（WAT 2025）で新たに始める特許請求項翻訳タスクについての紹介があり、参加者と情報共有できた。
今後、大規模言語モデルが特許や技術文書について翻訳に限らず作成支援や要約など様々な用途で活用されるようになると考えられ、大規模言語モデルの活用についても幅広く技術的課題の解決に向けた議論の場として本ワークショップをご活用いただくべく、募集分野やワークショップ名の再検討を行った上で引き続き実施していきたいと考えている。

2025年度AAMT/Japio特許翻訳研究会報告書

発 行 日　2026 年 3 月

発　　行　一般財団法人　日本特許情報機構（Japio）
〒135−0016　東京都江東区東陽町 4 丁目1番 7 号
佐藤ダイヤビルディング
TEL：(03) 3615−5511　FAX：(03) 3615−5521

編　　集　一般社団法人　アジア太平洋機械翻訳協会（AAMT）