特許請求項機械翻訳およびその人手評価の課題と挑戦 ～WAT2025 特許請求項タスクからの知見～

中澤敏明[1] 綱川隆司[2] 後藤功雄[3] 笠田和宏[4]
須藤克仁[5] 奥山尚一[6] 家田堯[7] 永田昌明[8]
[1]東京大学 [2]静岡大学 [3]愛媛大学 [4]一般財団法人日本特許情報機構
[5]奈良女子大学 [6]日本知的財産翻訳協会 [7]発明推進協会 [8]NTT Inc.
nakazawa@nlab.ci.i.u-tokyo.ac.jp[1] tuna@inf.shizuoka.ac.jp[2]
goto.isao.fn@ehime-u.ac.jp[3] kasada.kazuhiro.cn@japio.or.jp[4]
sudoh@ics.nara-wu.ac.jp[5] okuyama@quon-ip.jp[6]
t-ieda@jiii.or.jp[7] masaaki.nagata@ntt.com[8]

概要

本稿では、特許請求項（Claims）の翻訳に関する初の共有タスクの結果と知見について報告する。我々は参加者に対して学習・開発・テストデータを提供し、提出された翻訳結果の人手評価を行った。今回は 2 チームからの提出があった。人手による誤り分析の結果、一般的な翻訳誤りだけでなく、特許翻訳に特有の誤りが明らかになった。また、人手によるアノテーション自体にも多くの課題があることが判明した。本稿ではこれらの知見について報告する[1)]。

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。

そこで我々は、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。

2 データセット

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。

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

1) 本稿は著者らによる国際ワークショップ WAT 2025 における報告内容 [1] (CC BY 4.0 https://creativecommons.org/licenses/by/4.0/) に基づくものである。

2) 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

表 1 学習データに含まれる文数

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。

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。

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 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

4 評価結果

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-XL[3] [6] および WMT23-CometKiwi-DA-XL[4] [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。

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。我々が用いたアノテーション基準 (エラーカテゴリー) を付録の表 6 に示す。また品質スコア評価の結果を表 4 に示す。

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。

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。

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。

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、

3） https://github.com/google-research/metricx
4） https://github.com/Unbabel/COMET

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	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 $_{\pm 0.078}$

表 3　文書レベルの自動評価結果

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 $_{\pm 0.098}$

表 4　人手評価の平均スコア

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 5　人手評価と自動評価の相関

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

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。

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。

6　結論と今後の展望

本論文では、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 for MT 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, Philipp Koehn, 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 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)**, 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, and Wolfgang 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.

A　人手評価の基準

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。我々が用いたアノテーション基準 (エラーカテゴリー) を表 6 に示す。

大カテゴリ	中カテゴリ	小カテゴリ	説明
正確性 (Accuracy)	湧き出し (Addition)		原文に存在しない不必要な情報が含まれる
	訳抜け (Omission)		原文にある情報が不足している
	未翻訳 (Untranslated text)		原文が翻訳されずに残っている
	誤訳 (Mistranslation)		原文を正確には表現していない
		数値・記号 (Numerals/Symbols)	誤訳の一部だが、特に数値や記号に関する誤訳
		冠詞 (Article)	冠詞の使用誤り
		係り先誤り (Incorrect dependency)	形容詞句や並列構造などで係り先が誤っている
		係り先が不明 (Unknown dependency)	原文の係り受け構造が保たれていない (例: 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)
		解釈の曖昧性 (Ambiguity)	訳文の解釈がただ一通りに定まらない
流暢性 (Fluency)	句読点 (Punctuation)		句読点が不正確 (ロケールやスタイルによる。特許請求項は 1 文で書く必要があるため、不適切な文分割も含む。)
	スペリング (Spelling)		スペルミスや大文字表記の誤り
	文法 (Grammar)		正書法以外の文法の問題
	言語使用域 (Register)		文法上の言語使用域の誤り (例: 不適切にくだけた代名詞)
	一貫性の欠如 (Inconsistency)		内部的な一貫性の欠如 (用語に関係しないもの)
	文字エンコーディング (Character encoding)		エンコード誤りによる文字化け
用語 (Terminology)	文脈上不適切 (訳語選択) (Inappropriate for context)		用語が標準的でないか、文脈にそぐわない
	用語の一貫性 (Inconsistent use)		用語が一貫性なく用いられている
スタイル (Style)	ぎこちない (Awkward)		翻訳に文体上の問題がある
ロケール規則 (Locale convention)	住所形式 (Address format)		住所の形式誤り
	通貨形式 (Currency format)		通貨の形式誤り
	日付形式 (Data format)		日付の形式誤り
	名前形式 (Name format)		名前の形式誤り
	電話形式 (Telephone format)		電話番号の形式誤り
	時間形式 (Time format)		時間表現の形式誤り
その他 (Others)			
原文エラー (Source error)			An error in the source.
非翻訳 (Non-translation)			分類不能な誤り

表 6　人手による誤りアノテーション基準