

大規模言語モデルによる 特許情報機械翻訳の試み

—令和 7 年度特許庁事業「特許情報の機械翻訳における
大規模言語モデルの利用可能性に関する調査」の紹介—

An attempt at machine translation of patent information using Large Language Models

特許庁 総務部総務課 特許情報利用推進班長

長岡 真

平成 24 年特許庁入庁。化学分野の特許審査に従事のほか、調整課審査基準室、総務課法規班、審査第三部審査調査室を経て令和 7 年度 10 月より現職。

1 はじめに

特許庁では、公報、審査書類をはじめとする様々な特許情報を提供している。特許情報は、知財に係る権利の確認や新技術情報の入手等を可能とし、出願人を含むユーザーの知財活動促進の一助となる重要な情報である。特許庁が提供する情報には、日本語によるもの、外国語によるものがあり、また、提供先には、日本語ユーザーも外国語ユーザーも含まれる。そのため、翻訳と特許情報の関係は以下のように整理することができる。①日本語の公報や審査書類を外国語に翻訳する。これにより、海外庁の審査官による日本の審査結果の参照を促進することで、日本出願人の海外での円滑な権利化を後押しすることが期待される。また、日本での権利範囲や審査状況を把握しやすくすることで、海外企業等の日本への出願を活性化させる狙いもある。②外国語公報を日本語に翻訳する。外国特許文献の増加、特に、中国語の特許文献の急増が耳目を集めてすでに久しい。中国語などの外国語特許文献を日本語で容易にアクセス可能とすることは、企業活動におけるリスク回避や国内ユーザーによる出願活動の助けとなる。

さて、特許庁では、特許情報プラットフォーム（J-PlatPat）やワンポータル・ドシエ（OPD）において、リアルタイム日英機械翻訳を提供している。機械翻訳は機械翻訳プラットフォーム（MTP）が担う。また、外国語公報の日本語への機械翻訳を行う「外国特許文献の日本語機械翻訳データ作成及び機械翻訳精度向上に関する調査事業」を実施し、J-PlatPat での当該翻訳文を

対象とした外国公報等の検索及び照会や、翻訳文のダウンロードを可能としている。これらの機械翻訳の様式は、いずれもニューラル機械翻訳（NMT）である。NMT はその翻訳エンジン構築時に対訳コーパスを大量に学習させることで高精度な翻訳結果を得ることを可能とし、実際、MTP や前出の調査事業の翻訳精度は十分に高い。一方で、近年技術発展がめざましい大規模言語モデル（LLM）は、機械翻訳についても適用が進められている。LLM は事前学習により大量のデータを学習しているため、NMT とは異なり大量の対訳コーパスによる学習を行わなくとも翻訳用途に利用可能であるとされる。一般に特許文書は技術用語を多く含み、係り受けの複雑な長文も多く含まれるなどの特徴を有するが、LLM の中には様々な言語で公知になっている特許文書をすでに学習しているものもあると考えられ、特段の手当をせずとも特許文書の翻訳を行える可能性がある。

そこで、将来的な特許庁の機械翻訳システム検討を視野に入れ、LLM による特許文書の機械翻訳の状況を把握するため、令和 7 年度に「特許情報の機械翻訳における大規模言語モデルの利用可能性に関する調査」事業を実施したので紹介する¹。

なお、本稿は筆者が個人の資格で執筆したものであり、特許庁ほか所属組織の公式見解を示すものではない。

1 https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/document/kikai_honyaku/2025_01.pdf

2 調査の概要

本事業では、国内外の民間事業者、又はアカデミアが提供する LLM について、翻訳タスクへの適性、パラメータ数、学習データの量及び範囲、学習言語、ユーザーによる追加学習機能の有無、オンプレミス環境での動作可否、提供開始時期、料金の観点やその他の観点から、その内容を調査した。LLM を用いて特許文書を翻訳した際の精度評価を主な目的とすることから、公開情報調査に加えて翻訳試行により精度の確認を行い、claude-sonnet-4、gemini-2.5-pro、gemma3:27b の3つのモデルを調査対象として選定して、翻訳精度を評価した。

また、LLM は翻訳に特化した NMT とは異なり、翻訳そのものの以外への適用が期待される。そのため、翻訳の前後処理や、翻訳精度評価手法等、機械翻訳に適用可能な LLM の活用手法の調査を行った。さらに、NMT にはない特徴として LLM の処理はプロンプトに従うことから、翻訳精度向上のために有用なプロンプトの調査も行った。

加えて、選定した3つのモデルのうち、Gemma はオープンモデルであるため追加学習が可能である。そこで、少量の対訳コーパスによる追加学習が、翻訳に与える影響を調査した。

表1 LLM を用いた前後処理

	概要
①	翻訳対象文からLLMを用いて重要語を抽出、当該重要語を含むコーパスを既存のコーパス群から選択し、FewShotの例に含める前処理。
②	翻訳対象文を翻訳する際の参照用文対を、LLMを用いて作成する前処理。
③	翻訳対象文を翻訳する際の参照用文対を既存コーパス群から類似検索した上で選択し、FewShotの例に含める前処理。

表2 LLM に与えるプロンプト

	概要
①	翻訳対象の種類（「特許文書の請求項」等）と想定読者（「特許のサーチャー」）をプロンプトに含める。
②	翻訳対象の技術分野（特許分類の説明）をプロンプトに含める。
③	プロンプト①とプロンプト②の内容を含める。

表3 LLM を用いた翻訳精度評価手法

	概要
①	原文と翻訳文とを入力し、翻訳文に対して誤りを抽出。誤訳を Accuracy(文の誤訳)、Fluency(流暢さ)、Style(文体)、Terminology(語の誤訳)等の誤訳カテゴリ、Critical, Major, Miner の誤訳レベルで評価。
②	LLMに原文と翻訳文、人手翻訳文とを入力し、「正確性（文）」を5段階評価、「正確性（用語）」を4段階評価、「流暢さ」を5段階評価。
③	LLMに原文と翻訳文、人手翻訳文とを入力し0～100で評価。

公開情報調査の結果、調査対象とした前後処理、プロンプト、翻訳精度評価手法の一部をそれぞれ表1～3に記載した。

前後処理、プロンプト、翻訳精度評価手法を検討する際に翻訳を実行するモデルには Claude を選択した。前後処理及び翻訳精度評価には Gemini を用いた。

本調査では、翻訳精度に加え、各モデルが翻訳に要する時間から翻訳速度の比較も行った。

3 結果

3.1 LLM による特許文書の翻訳精度

翻訳精度は自動評価及び人手評価により確認した。

・自動評価

3種の LLM と、比較のための NMT の全4モデルについて自動評価のスコアを比較した（図1）。公報の「発明の名称」、「要約」、「特許請求の範囲」、「明細書」の4項目から500文ずつ、計2,000文の機械翻訳を行い、自動評価を行った。図1には言語方向に応じて「英日」、「中日」、「韓日」、「日英」と示す。また、日英審査書類として「拒絶理由通知書」、「意見書」、「補正書」からそれぞれ500文、計1,500文についても機械翻訳を行い、評価結果を「審査」に示した。

LLM の BLEU 値は、評価の高い順に NMT、Gemini、Claude、Gemma となり、一見すると NMT が優れるように見えるが、いずれも一般に機械翻訳で高品質といわれる BLEU 値 40 を大きく上回った。

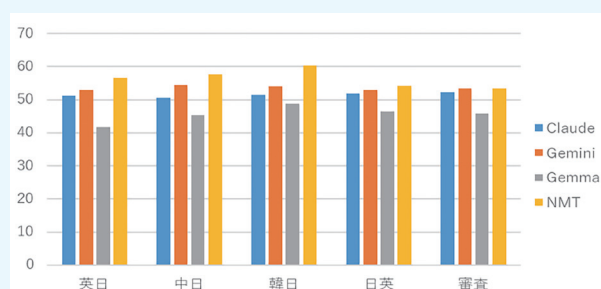


図1 BLEU 値

・人手評価

人手評価は、特許庁が公表している「特許文献機械翻訳の品質評価手順(ver1.0)²」の品質評価手順に則り「内

2 https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/document/tokkyohonyaku_hyouka/O1.pdf

容伝達レベルの評価」、「重要技術用語の評価」、「流暢さの評価」、「誤訳のカテゴリ別カウント」及び「技術用語の訳ゆれの評価」を実施した。人手評価に用いた対象文は、自動評価に用いた各言語・各項目あたり 500 文のセットから 100 文ずつを選定した。

まず、内容伝達レベルの評価は、評価対象文に対し、機械翻訳結果が原文の実質的な内容をどの程度正確に伝達しているかを 5 段階の評価基準で主観的に評価するものであり、「総合評価」の位置づけとなる。高評価であるほど、外国語文献の内容把握という目的に対する実用性が高い機械翻訳文といえる。

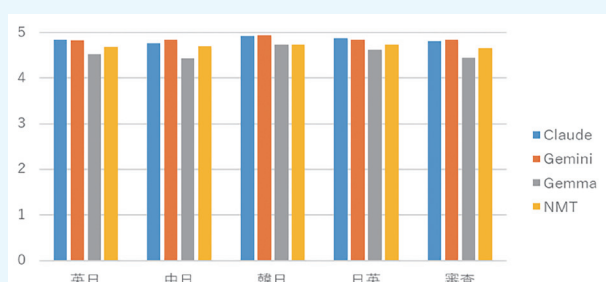


図2 内容伝達レベル

NMT の内容伝達レベルは全言語平均で 4.71 と機械翻訳として十分に良好であり、Claude は 4.84、Gemini は 4.86 と、これをさらに上回る高評価であった。Gemma は 4.56 と他の 2 モデルとはやや差があり、NMT よりも低値となったが、レベル 4 は大きく上回っており、良好な水準に達した。

次に、流暢さの評価では、機械翻訳文の「文としての読みやすさ、理解しやすさ」を 5 段階の基準で評価した。機械翻訳結果のみを独立した文として扱い、原文との整合性は考慮せずに和文として自然かどうかのみを評価する。

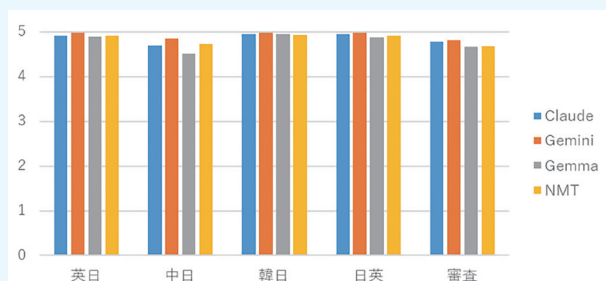


図3 流暢さ

モデル間の比較では全言語で Gemini が最上位となった。各モデル、各言語とも平均スコアは高い水準にあり、極端に低スコアのモデルや言語は見られない。各モデルとも十分に実用に足るといえ、さらにいえば、Gemini

と Claude については NMT と比してより流暢な、読みやすい翻訳を行うともいえる。

重要技術用語の訳質評価では、原則として全ての評価対象文について、文中に存在する技術用語 1 語を事前に指定し、この用語が機械翻訳結果で適切に翻訳されているかを A～D の 4 段階で評価した（A：適訳語、B：可訳語、C：誤訳語、D：不訳語）。

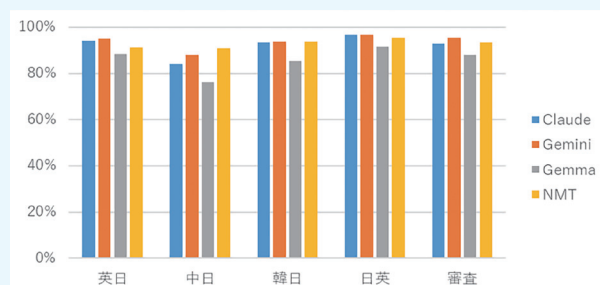


図4 重要技術用語 (A 判定率)

特許文書に用いられる技術用語の訳は、特許由来の大量の学習データを用いる NMT が得意とするところである。これと比して、Gemini、Claude は中日こそ NMT に及ばなかったものの、他については NMT と同等以上となり、事前学習によって、技術用語の訳においても十分に実用に足るものとなっているといえる。一方、Gemma では A 判定率の低さが目にとまる。技術用語のカバー範囲が他のモデルに比べて狭い可能性が示唆される。

「誤訳のカテゴリ別カウント」及び「技術用語の訳ゆれの評価」の詳細は省略するが、前者からは、いずれの LLM も技術用語の誤訳が多いことが分かり、後者からは、いずれの LLM も同じ文中、異なる文間での訳ゆれが存在することが分かった。これらは NMT と同じ傾向であった。

3.2 前後処理、プロンプトの工夫の影響

前後処理やプロンプトの工夫による影響は、いずれの言語や処理についても、翻訳精度の評価値の小さな変動にとどまり、向上と低下が混在する結果となった。翻訳に用いた Claude の翻訳精度が、特に工夫をせずとも高い水準に達していたため、誤差レベルの小変動に終始したものと推測される。しかしながら、詳細に確認すると、前後処理①により FewShot に特許由来の対訳コーパスが追加されたことで、中日翻訳における重要技術用語の評価の向上が認められたり、プロンプト①により翻訳対象の種類と想定読者が明示されたことにより、特許文書の

体裁に沿った翻訳が得られ、英日翻訳において請求項の用言止めが顕著に改善されたりといった効果は得られており、一定の翻訳精度向上効果を確認することができた。

3.3 追加学習の影響

Gemma への追加学習では、少量の対訳コーパスによる追加学習の効果を検討するため、1,500 文対程度の対訳コーパスでファインチューニングを行った。しかしながらその結果は、自動評価、人手評価ともに、言語毎に評価値の上昇する項目、低下する項目が混在し、変化幅も小さいものであった。残念ながら、1,500 文対程度の追加学習量では翻訳改善効果は乏しいと結論される。

3.4 LLM による翻訳精度評価

機械翻訳の評価は人手による評価と機械が評価を行う自動評価に分類される。LLM による評価は自動評価手法の1つとなることから、自動評価として有用といえるか調査を行った。調査は自動評価 BLEU、評価手法①の Accuracy（文の誤訳）に基づく評価、評価手法②の正確性（文）、評価手法③のスコアそれぞれについて人手評価の内容伝達レベルとの相関係数を求め、どの評価が人手評価との間で最も相関しているかを比較した。

表4 LLM による翻訳精度評価

	「内容伝達レベル」との相関係数
①	0.56
②	0.41
③	0.41
BLEU	0.07

人手評価と LLM による評価の間には、人手評価と BLEU 値よりは高い相関が得られた。これらの結果より、BLEU 値による評価と比べ、LLM による評価は、より人手評価に近いといえる。また、その評価を付した理由を説明できるのも LLM の特徴である。但し、LLM による評価は使用するモデルの性能に依存する。本調査では Gemini を用いて評価を行ったが、他のモデルでは人手評価との相関の度合いは異なる可能性がある。

3.5 翻訳速度

各モデルによる翻訳速度を比較するため、一文あ

たりの翻訳時間を調査した。オープンモデルである Gemma による翻訳速度はモデルを構築したシステム環境の影響を受け、Claude や Gemini の翻訳速度はモデル提供元やネットワークの混雑状況に影響を受ける点に留意する必要があるが、本調査においては以下の結果が得られた。

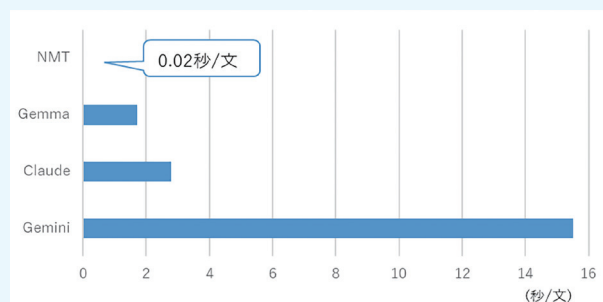


図5 翻訳に要した時間

翻訳速度は Gemini、Claude、Gemma の順に遅く、NMT と比して圧倒的に遅いことが分かった。

4 まとめ

本調査により、LLM を用いた特許文書の翻訳精度は、良好な水準に達していることが分かった。それは大量の特許文書を学習した高精度の NMT に匹敵し、人間が受ける印象により近い人手評価の結果に重きを置くならば、モデルによっては NMT を超えるといえる。そして、内容伝達のみならず、翻訳文の流暢さ即ち人間の読みやすさも、モデルによっては NMT を超えるものであることがわかった。一方で、誤訳の傾向の分析から、NMT と同様に技術用語の誤訳が多い点や同一用語に対する訳ゆれが一定数存在することが分かった。そして、翻訳速度については、いずれの LLM も NMT には及ばないことも確認された。

特許庁では、前述のとおり、MTP によって日英リアルタイム翻訳を提供し、外国特許文献の日本語機械翻訳データ作成及び機械翻訳精度向上に関する調査事業によって得た外国公報の機械翻訳文を蓄積して提供するという機械翻訳システムを提供している。本調査で判明した LLM の特徴を踏まえ、仮に LLM を用いた特許庁の機械翻訳システムを想定するに、まず、LLM を利用することで現状維持かそれ以上の翻訳品質を得ることが期待される。また、LLM のコンテキスト長が大きいこと

から、段落単位や文献単位の翻訳が可能となれば、異なる文の間で生じている訳ゆれへの耐性向上の可能性もある。さらに、前後処理やプロンプトの工夫等により、一定の訳質改善効果が見込まれる。翻訳処理速度については、現状では NMT が圧倒的に勝るため、リアルタイム翻訳や大量翻訳を行うためには翻訳速度向上の対策が必要となる。これに関し本調査では、パラメータ数が小さく処理速度の速いオープンモデルをベースにした追加学習モデルによる翻訳が、訳質の向上と翻訳速度の担保の観点から提案された。本調査では、オープンモデルである Gemma は 3 種の LLM のの中では最も翻訳精度の評価結果が低く、NMT を超えるものとはいえなかった。そして、その一因が技術用語のカバー範囲の違いによる可能性が示唆されながら、追加学習を行っても顕著な翻訳改善効果は見られなかった。しかしながら、本調査の追加学習データが比較的少量であったことが要因の 1 つとして考えられ、追加学習データの規模や追加学習手法を検討することで、翻訳精度と翻訳速度のバランスのとれた LLM を構築できる可能性は十分にあると考えられる。また、LLM の翻訳速度が NMT と同レベルまで改善しないにしても、リアルタイム翻訳は LLM、大量翻訳は NMT というようなそれぞれの特徴を活かした使い分けも可能かもしれない。

なお、本調査では、LLM を用いた機械翻訳の精度の評価に重きをおいたため、コストについて検討するための十分な資料を得られたとは言いがたい。もし、LLM を特許庁の提供する機械翻訳システムに組み込むのであれば、コスト面の評価も欠かせない。

LLM の進化のスピードはめざましいものがあり、判明した課題に対する状況は、あるいは本稿執筆現在にも変化しているのかもしれない。一方で NMT は大量の対訳コーパスを要するとはいえ、特許文書の翻訳に特化し、処理が速いという強みを有する。両者の強み、システム全体のコスト、そして、何より特許庁が提供する翻訳に何が求められるのかを見極め、利用者のニーズに沿った翻訳に係るサービスを模索していきたい。

