大規模言語モデルによる特許クレーム起案支援

―単一法域におけるクレーム自動修正から多言語・多法域特許起案へ―

Toward LLM-based Support for Patent Claim Drafting: From Monolingual Claim Refinement to Multilingual and Multi-jurisdictional Claim Drafting

京都工芸繊維大学　情報工学・人間科学系　助教

河野　誠也

2021 年奈良先端科学技術大学院大学にて博士（工学）取得。JSPS DC2、理化学研究所 GRP リサーチアソシエイト・特別研究員・研究員等を経て、2025 年より京都工芸繊維大学助教。自然言語処理・音声言語処理の研究に従事。

kawano@kit.ac.jp

1 はじめに

特許出願書類の中核は「特許請求の範囲」、すなわちクレームである。明細書が発明の技術的開示を担うのに対し、クレームは発明が受ける法的保護の範囲そのものを定義する「権利書」としての性格を持ち、その一語一句が権利範囲の広狭や特許の有効性に直接影響する[1]。特許クレームの起案（ドラフティング）とは、発明の技術的本質を的確に捉え、新規性･進歩性･記載要件といった審査要件を満たしつつ、可能な限り広く強い権利として言語化する作業を指す[2]。

重要なのは、起案が出願時の一度きりの作業ではない点である。多くの出願は審査過程で拒絶理由通知（オフィスアクション）を受け、出願人は意見書・補正書の提出によりクレームを修正して応答する。この補正は、新規事項追加の禁止（特許法第 17 条の 2）をはじめとする法的制約の下で行われ、権利範囲の広さと登録可能性のトレードオフを見極める高度な判断を要する[3]。すなわち、特許起案とは出願から登録に至るまで継続する反復的な「書き換え」のプロセスである。

さらに、国際出願ではもう一つの次元が加わる。特許協力条約（PCT: Patent Cooperation Treaty）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するが、国内段階に移行した後は各国特許庁が独立に審査を行う。このため出願人は、各法域の起案規則、法的用語、および審査実務を反映して、クレームを単なる翻訳を超えて「適応」させる必要がある。このように特許起案は、技術・法律・言語の三領域にまたがる専門的作業であり、実務上は多大な時間と労力を要する。

近年の大規模言語モデル（LLM: Large Language Model）の急速な発展は、この特許起案の過程を計算機によって支援する研究に新たな可能性を開きつつある。本稿では、特許起案をめぐる関連研究を概観したうえで、筆者の取り組みとして、単一法域内における特許クレームの自動修正フレームワーク ClaimBrush と、その発想を多言語・多法域へと拡張した PCT 出願に基づく特許起案データセットの構築を紹介し、今後の展望を述べる。

2 関連研究

2.1　特許情報処理の基盤タスク

特許文書を対象とする自然言語処理（特許 NLP）は、NTCIR、TREC-CHEM、CLEF-IP といった評価型ワークショップを中心に、特許検索や分類などの基盤タスクとして長く研究されてきた[4][5][6][7]。また、独特の長文･入れ子構造を持つクレームの理解を支援するため、構造解析や用語説明による可読性向上の研究も早くから行われている[8]。クレームの書き換えに関しても、マルチマルチクレーム（2 以上の他のクレームを択一的に引用するクレームをさらに択一的に引用するクレーム）を検出して複数のクレームに展開するルールベース手法が提案されているが[9]、特定の目的に限定された書き換え支援にとどまっていた。

表 1　ClaimBrush データセットの統計（平均文字数・クレーム数は B 基準、括弧内は A からの平均変化率）

種別	件数	平均文字数	平均クレーム数
Type 1：A のみ（対応する B なし）	2,610,893	1402.16	7.84
Type 2：A と B が同一（拒絶理由なし）	369,807	1259.83（+0.0%）	6.16（+0.0%）
Type 3：A と B が同一（拒絶理由あり）	35,953	822.14（+0.0%）	4.64（+0.0%）
Type 4：A と B が相違（拒絶理由なし）	628,882	1438.76（−13.2%）	6.79（−25.9%）
Type 5：A と B が相違（拒絶理由あり）	1,210,998	1401.28（−14.8%）	6.60（−28.5%）

2.2　特許機械翻訳

多言語の観点では、特許ファミリーや文レベルのアライメントから構築された日英並列コーパスに基づく機械翻訳研究が中心であった[10][11]。特に JaParaPat[11] は約数億の整列済み文対を提供し、高精度な特許翻訳を実現している。ただし、これらのコーパスは対応文書が忠実な翻訳であるという仮定の下でキュレーションされており、法域間で生じる内容の乖離はノイズとして除去されることが多い。この仮定の限界については 4 章で改めて議論する。

2.3　LLM によるクレーム生成・修正

生成型言語モデルの台頭を背景に、クレームそのものの生成・修正を扱う研究が急速に立ち上がりつつある（この領域の全体像は Jiang らのサーベイ[12] に詳しい）。先駆的な研究として、Lee らは GPT-2 をファインチューニングした特許クレーム生成を試み[13]、その後、特許ドメインに特化した生成言語モデルの構築と評価も進められている[14][15]。米国特許商標庁（USPTO: United States Patent and Trademark Office）の出願を大規模に構造化したコーパス HUPD[16] は、この分野のデータ基盤として広く利用されている。近年では、明細書を入力としたクレーム生成における LLM の性能評価[17]、拒絶された出願クレームと登録クレームの対からなるクレーム修正データセット Patent-CR[18]、欧州特許庁（EPO: European Patent Office）のデータに基づくクレーム生成データセット EPD[19] など、生成・修正の両面で研究が展開されている。日本特許庁（JPO: Japan Patent Office）を対象とする筆者の ClaimBrush[20] もこの流れに位置づけられる。

これらの研究状況を整理すると、二つの特徴が浮かび上がる。第一に、クレーム生成・修正の研究はいずれも単一法域内・単一言語の設定で行われており、1 章で述べた国際特許実務における法域間適応の側面は扱われてこなかった。第二に、審査官による評価という特許制度に固有のフィードバック構造を、モデルの学習に直接組み込む試みは限られていた。筆者の研究は、審査過程の信号を活用した単一法域内のクレーム自動修正（3 章）から出発し、これを多言語・多法域の起案へと拡張する（4 章）ものである。

3　単一法域内の取り組み：ClaimBrush

3.1　審査過程を活用したデータセットの自動構築

クレーム自動修正モデルの訓練・評価には、修正前後のクレーム対を大量に収めた信頼性の高いパラレルコーパスが不可欠である。しかし、クレームは高度に技術的かつ法律的な文書であり、技術と知財実務の双方に精通した作業者による人手でのデータセット作成は、コストの観点から現実的でない。そこで ClaimBrush では、日本の同一出願に紐づく公開特許公報（A-kind）と特許公報（B-kind）のクレーム対を「望ましい書き換え」の事例とみなし、データセットを自動構築するアプローチを採用した。多くの出願は一度は拒絶され、必要な補正を経て登録に至るため、特許公報のクレームには拒絶理由への応答補正と自発補正の結果がすべて反映されており、審査の観点からは出願時のクレームよりも「洗練された」書き換え結果とみなすことができる。

データセットは、特許庁が公開する 2004 年から 2022 年までの公報データから同一出願番号を持つ公報対を抽出し、さらに審査経過情報を連結して、拒絶の

根拠法条（例：特許法第 29 条第 2 項）や審査官が引用した先行特許のクレームを各書き換え対に付与して構築した。構築したデータセットは総計 2,245,640 件の出願に対するクレーム対を含む（表 1）。このうち A-kind と B-kind の間でクレーム本文に差分がある事例は約 184 万件であり、差分のない約 41 万件と大別される。差分のある事例のうち、拒絶理由のない Type 4（628,882 件）は出願人の自発補正を、拒絶理由のある Type 5（1,210,998 件）はオフィスアクションへの応答補正を反映している。クレーム数の減少率は Type 4 の 25.9% に対し Type 5 では 28.5% と大きく、拒絶への対応がより大幅な書き換え（クレームの統合・削除・限定等）を伴うことが定量的に確認された。また、差分がないのに拒絶理由が存在する Type 3 は、補正によらず意見書の主張のみで拒絶を克服した事例と解釈できる。なお、公報と審査経過情報を組み合わせる本構築手法は日本に固有のものではなく、審査制度を共有する USPTO や EPO の公開データにも同様に適用可能である。

3.2　タスク設定と書き換えモデル

ClaimBrush では、特許クレーム修正を次のタスクとして定式化した。入力は、拒絶が予想される出願時のクレーム（クレームセット全体を一つのテキストとして扱う）と、予想される拒絶理由を連結したテキストであり、出力は、特許査定を受けられるよう書き換えられたクレームである。訓練には、構築したデータセットのうち一度拒絶された後に登録に至った書き換え対を用い、計算資源の制約から入出力の合計トークン長が一定以下の対に限定して、訓練 31 万 7 千件・検証 1,000 件・評価 1,000 件のデータを構成した。

書き換えモデルには事前学習済み LLM（Qwen1.5 系列、0.5B ～ 7B パラメータ）を用い、全パラメータを更新する教師ありファインチューニング（SFT: Supervised Fine-Tuning）と、低ランク行列の追加学習により一部の層のみを適応する LoRA[21] の両方を検証した。また、比較対象として、無修正のコピー、クレームのランダム削除（RDC）、マルチマルチクレームの削除（DMMC）というヒューリスティックなベースラインに加え、GPT-4o-mini、GPT-3.5-turbo、Qwen1.5-72B といった LLM のゼロショット生成を評価した。

3.3　審査官選好の最適化

本フレームワークの最大の特徴は、審査官の選好に基づく選好最適化にある。人間のフィードバックによる強化学習（RLHF）は学習中に生成のサンプリングを繰り返すため計算コストが高く、直接選好最適化（DPO）[22] は「望ましい応答」と「望ましくない応答」の対データを必要とする。これに対し Kahneman-Tversky 最適化（KTO）[23] は、望ましさの二値ラベルさえあれば対でないデータからも学習できるため、審査結果という二値的な信号と相性がよい。

審査官の選好そのものは、次の選好モデルによって近似した。評価対象のクレーム、拒絶理由、および審査官が引用した先行特許のクレームを連結して入力とし、拒絶を克服できなかった出願時クレームを「望ましくない」例、拒絶を克服して登録に至ったクレームを「望ましい」例とする二値分類器を LLM ベースで学習した。この選好モデルは、両ラベルとも F1 値 72% を超える精度で審査官の判断を予測できることを確認した（表 2）。そのうえで、SFT 済みの書き換えモデルから各訓練サンプルにつき 12 個の書き換え候補をサンプリングし、選好モデルの判定に基づいて望ましさのラベルを自動付与して KTO の学習データとした。すなわち、人手アノテーションを介さずに、審査官選好に整合した選好最適化を実現している。

表 2　選好モデルによる審査官判断の予測性能

ラベル	Prec. [%]	Rec. [%]	F1 [%]	件数
望ましくない（拒絶）	71.81	77.70	74.64	1,000
望ましい（登録可能）	75.71	69.50	72.47	1,000

3.4　実験結果

評価には、入力・出力・参照の 3 者を考慮する必要があることから、文法誤り訂正分野の GLEU[24] とテキスト平易化分野の SARI[25] を単語・フレーズの 2 つの分割単位で用い、さらに選好モデルによる受理率（生成クレームが「登録可能」と自動判定される割合）を実務的指標として併用した。主要な結果を表 3 に示す。

結果からは複数の知見が得られた。第一に、指標の性質である。入力をそのまま出力する Copy は GLEU（単語）で最高値を示すが、これは入力からのコピーを過

表３　特許クレーム修正の性能

モデル	GLEU（word）	GLEU（phrase）	SARI（word）	SARI（phrase）	受理率
Reference（正解）	100.0	100.0	100.0	100.0	0.69
Copy（無修正）	63.23	23.08	28.38	20.21	0.22
RDC（ランダム削除）	47.22	18.15	28.38	20.21	0.30
DMMC（マルチマルチ削除）	57.85	21.89	28.38	20.21	0.23
GPT-4o-mini（ゼロショット）	47.04	5.17	32.00	24.99	0.70
GPT-3.5-turbo（ゼロショット）	40.19	11.75	29.62	27.33	0.17
Qwen1.5-72B（ゼロショット）	29.97	1.45	27.23	20.94	0.69
Qwen1.5-0.5B（LoRA）	54.16	22.17	34.55	37.83	0.53
Qwen1.5-1.8B（LoRA）	41.29	15.24	34.22	37.37	0.49
Qwen1.5-7B（LoRA）	55.67	24.10	35.42	39.12	0.57
Qwen1.5-0.5B（SFT）	55.06	24.65	36.66	40.54	0.63
Qwen1.5-0.5B（KTO）	58.45	25.76	38.95	43.92	0.96

大評価する GLEU の特性によるものであり、追加・保持・削除の書き換え操作を反映する SARI や受理率では Copy・RDC・DMMC といったベースラインはいずれも低い値にとどまる。第二に、ゼロショットの限界である。GPT-4o-mini や Qwen1.5-72B といった大規模モデルはファインチューニングした小規模モデルに及ばず、特にフレーズ単位の GLEU が極端に低い。これは、特許法第 17 条の補正要件に抵触しかねない破壊的な書き換えを生成する傾向を反映している。加えて、GPT-3.5-turbo や Qwen1.5-72B には英文特許の直訳調の出力が観察され、日本語かつ特許ドメインへの適応の重要性が浮き彫りとなった。実際、7B モデルの LoRA 学習が 0.5B モデルの SFT と拮抗する程度であったことも、ベースモデルの日本語特許生成能力の不足を LoRA では補いきれないことを示唆する。

第三に、選好最適化の効果である。KTO を適用した 0.5B モデルは受理率 0.96 を達成し、ベースの SFT モデル（0.63）から大きく向上したのみならず、GLEU・SARI も同時に改善した。すなわち、書き換え品質を維持しつつ審査官選好への整合を高められることが示された。ただし、この受理率は参照（実際に登録されたクレーム）の 0.69 をも上回っており、あくまで選好モデルという自動評価器に対する最適化の結果である点には留意が必要である。実審査での受理を保証するものではなく、選好モデル自体の精度向上と人手評価による検証が今後の課題となる。なお、拒絶理由を入力から除いた条件（w/o r）では LoRA 学習で一定の性能低下が見られたが SFT では限定的であり、拒絶理由のラベルだけでなく、その解消の手がかりとなる引用先行技術の内容等をより豊富に与える必要性が示唆された。

4　多言語・多法域横断への拡張

4.1　特許翻訳と特許起案のギャップ

ClaimBrush や Patent-CR などの自動特許ドラフティング研究が実現したのは、単一法域・単一言語内でのクレームの「磨き上げ」である。これを国際特許実務に広げようとするとき、まず直面するのが、従来の多言語特許処理が置いてきた前提との齟齬である。2.2 節で述べたとおり、特許翻訳研究の並列コーパスは、対応する文書が互いに忠実な翻訳であるという仮定の下に構築され、内容が食い違う対応はノイズとして除去されてきた[10][11]。この仮定は、翻訳品質の学習という目的に対しては合理的である。

表 4　アライメントカテゴリ別のデータセット統計

カテゴリ	ペア数	平均文字数（日）	平均語数（英）	平均クレーム数（日/英）
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

しかし、PCT 出願の国内段階で実際に起こっていることは、忠実な翻訳の生成ではない。クレームの記載形式や法的定型表現の慣行は法域ごとに異なり、クレーム数に応じた料金体系やマルチマルチクレームの制限といった制度的制約も一様でない。さらに各庁の審査は独立に進行するため、ある法域では認められたクレームが別の法域では拒絶され、法域ごとに異なる補正が施される。その結果、同一の PCT 出願に由来するクレームセットであっても、言語間・法域間で構造も内容も分岐していく。出願人が行っているのは翻訳ではなく、対象法域の要件に適合させる作業である。

ここに、特許翻訳と特許起案の本質的なギャップがある。翻訳コーパスがノイズとして捨ててきた法域間の乖離は、特許起案の観点からはむしろ学習すべき信号そのものである。そこで筆者は、この乖離を除去せずに保持したデータセットを構築し、タスクを「多言語特許起案」、すなわちある言語で作成された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へ変換するタスクとしてフレーミングし直すこととした。

4.2　PCT 出願番号に基づくアライメント

本データセットでは、JPO および USPTO の特許文書を、同一の PCT 出願番号をキーとして対応付けた。公報は種別コード（A 種別:出願公開、B 種別:登録特許）で分類されるため、PCT 出願番号と種別コードを組み合わせることで、法域をまたいで出願公開段階と登録段階の文書を段階横断的にアライメントできる。ここで、ソース側クレームの公開日がターゲット側よりも時間的に過去である対のみを採用するフィルタリングを行い、起案の方向性（草案から成果物へ）と時間的整合性を担保した。

4.3　データセットの統計と分析

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クレームセット対を含む（表 4）。表中、ja は JPO 側、en は USPTO 側、pre は公開段階、granted は登録段階を表す。同一段階の対応（en_pre → ja_pre）では平均クレーム数がほぼ同一（25.5 対 25.6）であり、クレーム列挙の構造的乖離は限定的である。一方、段階をまたぐ対応（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が定量的に現れている。また、登録段階の日本語テキストは公開段階よりも平均して長く（ja_granted → en_granted の 4,589 文字対 ja_pre → en_pre の 3,447 文字）、限定の付加等による登録段階でのテキスト密度の増加が示唆される。

4.4　機械翻訳ベースラインによる検証

前節までの分析が示す法域間の乖離を、翻訳ではどこまで埋められるのか。これを定量化するため、Google Cloud Translation API（Advanced）v3 で文書レベルの翻訳を行い、対象法域の参照クレームと比較するベースライン実験を実施した。評価には BLEU・chrF・SARI に加え、意味的な近さを捉えるニューラル評価指標 COMET を用いた。結果を表 5 に示す。

第一の知見は、同一段階の対応でさえ忠実な翻訳ではないことである。特許翻訳研究で並列データとして扱われることの多い同一段階カテゴリでも、BLEU は飽和には程遠い（ja_pre → en_pre で 35.81、ja_granted → en_granted で 29.57）。数億の文対で訓練

表５　Google Cloud Translation API (v3) ベースラインの性能（全文書評価）

カテゴリ	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

された JaParaPat[11] の文レベル翻訳が SacreBLEU で 55.6 ～ 56.5 を報告していることと対照すれば、この差は商用翻訳エンジンの品質不足では説明できず、本データセットの対応がそもそも逐語的な翻訳としてキュレーションされていないことの帰結である。

第二の知見は、段階をまたぐ対応はさらに困難なことである。日→英方向では、対象が登録クレームセットになると BLEU は 26.68 まで低下し、審査過程でクレームが統合・追加・取消・限定されるという表４の分析と整合する。

さらに指標間・方向間の比較からも示唆が得られる。BLEU が低い一方で COMET は全カテゴリで 0.84 以上と高く、意味内容は概ね保たれつつ表層・構造が大きく異なるという、まさに「適応」の性質が現れている。また、日→英方向では段階をまたぐほど性能が低下するのに対し、英→日方向では明確な低下が見られない（en_pre → ja_pre の 38.17 に対し en_pre → ja_granted は 40.15）など、言語方向・法域ごとに乖離の性質が異なる可能性も示唆される。総じて、法域間のクレーム対応は翻訳のみでは十分にモデル化できない「起案」の問題であり、法域固有の適応パターンを明示的に学習するアプローチの必要性が裏付けられた。

5 特許起案研究の展望

以上のように、筆者の取り組みは、単一法域内のクレーム自動修正から多言語・多法域の起案へと対象を広げてきた。今後の方向性として、第一に両者の統合が挙げられる。例えば、日本語の出願段階クレームから米国の登録段階クレームを直接生成するような多言語起案モデルに対し、ClaimBrush で確立した審査官選好最適化を適用することで、対象法域の審査実務に整合した起案支援が期待できる。第二に、審査経過情報の言語横断的な活用である。拒絶理由や引用先行特許といった審査シグナルは各法域で独立に生成されるため、これらを言語をまたいで対応付けることで、「ある法域でどのような拒絶を受け、どう補正したか」という法域固有の起案戦略をより直接的に学習できる可能性がある。第三に、検索拡張生成（RAG: Retrieval-Augmented Generation）[26] 等による先行技術や明細書本文の参照である。3.4 節で見たとおり、拒絶理由のラベルのみでは書き換えの手がかりとして不十分であり、その解消に必要な根拠情報をモデルに供給する仕組みが重要となる。加えて、EPO をはじめとする追加法域へのデータセット拡張にも取り組んでいく。

評価と実務適用の面でも課題は残る。表層的指標（GLEU・SARI・BLEU 等）と実務的指標（受理率等）を組み合わせてきたが、自動評価と人手評価の乖離はクレーム修正研究において広く指摘されており [18]、人手評価と整合する評価手法の確立は実用化に向けた重要な課題である。また、本技術はあくまで弁理士等の専門家を支援するツールであり、新規事項追加の禁止をはじめとする法的制約の遵守や、クレームの最終的な妥当性判断は専門家に委ねられるべきである。生成されたクレームの法的妥当性を機械的に検証する仕組みと、専門家のワークフローに自然に組み込まれるインタフェースの設計が、社会実装の鍵になると考えている。

6 おわりに

本稿では、特許起案という営みを技術・法律・言語にまたがる反復的な書き換えプロセスとして整理し、関

連研究の流れの中に、単一法域内のクレーム自動修正（ClaimBrush）から多言語・多法域特許起案データセットへと至る筆者の取り組みを位置づけて紹介した。単一法域内では、審査過程から自動構築した224万件規模のデータセットと審査官選好への最適化により、小規模なLLMでも実用的な水準の書き換えが可能になりつつあること、また法域間のクレーム対応は翻訳を超えた「適応」を含み、多言語特許起案という新たな研究課題を構成することを示した。これらの資源と知見が、知財情報の高度な活用と特許実務の効率化に資することを期待したい。

参考文献

[1] A. C. Marco, J. D. Sarnoff, and A. W. Charles: Patent claims and patent scope, Research Policy, 48(9), 103790, 2019.

[2] Faber, Robert C., and John L. Landis.: Faber on Mechanics of Patent Claim Drafting, Practising Law Institute, Practising Law Institute, 2015.

[3] G. Reilly: Amending patent claims, Harvard Journal of Law & Technology, 32, 1, 2018.

[4] A. Fujii, M. Iwayama, and N. Kando: Overview of the patent retrieval task at the NTCIR-6 workshop, Proc. of NTCIR, 2007.

[5] M. Lupu, A. Fujii, D. W. Oard, M. Iwayama, and N. Kando: Patent-related tasks at NTCIR, Current Challenges in Patent Information Retrieval, pp.77-111, 2017.

[6] M. Lupu, J. Tait, J. Huang, and J. Zhu: TREC-CHEM 2010: Notebook report, Proc. of TREC 2010, 2010.

[7] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz: CLEF-IP 2011: Retrieval in the intellectual property domain, Proc. of CLEF 2011, 2011.

[8] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama: Patent claim processing for readability - structure analysis and term explanation, Proc. of the ACL-2003 Workshop on Patent Corpus Processing, pp.56-65, 2003.

[9] 新森昭宏，大屋由香里，谷川英和：マルチマルチクレームの検出と書き換え，情報処理学会論文誌，49(7), pp.2692-2702, 2008.

[10] M. Utiyama and H. Isahara: A Japanese-English patent parallel corpus, Proc. of Machine Translation Summit XI, 2007.

[11] M. Nagata, M. Morishita, K. Chousa, and N. Yasuda: JaParaPat: A large-scale Japanese-English parallel patent application corpus, Proc. of LREC-COLING 2024, pp.9452-9462, 2024.

[12] L. Jiang and S. M. Goetz: Natural language processing in the patent domain: a survey, Artificial Intelligence Review, 58(7), 214, 2025.

[13] J.-S. Lee and J. Hsiang: Patent claim generation by fine-tuning OpenAI GPT-2, World Patent Information, 62, 101983, 2020.

[14] J.-S. Lee: Evaluating generative patent language models, World Patent Information, 72, 102173, 2023.

[15] Z. Bai, R. Zhang, L. Chen, et al.: PatentGPT: A large language model for intellectual property, arXiv preprint arXiv:2404.18255, 2024.

[16] M. Suzgun, L. Melas-Kyriazi, S. Sarkar, S. D. Kominers, and S. 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, pp.57908-57946, 2023.

[17] L. Jiang, C. Zhang, P. A. Scherz, and S. Goetz: Can large language models generate high-quality patent claims?, Findings of the Association for Computational Linguistics: NAACL 2025, pp.1272-1287, 2025.

[18] L. Jiang, P. A. Scherz, and S. Goetz: Patent-CR: A dataset for patent claim revision, Proc. of NAACL 2025 (Volume 1: Long

Papers), pp.2300-2314, 2025.

[19] L. Jiang, C. Li, and S. Goetz: Enriching patent claim generation with European patent dataset, arXiv preprint arXiv:2505.12568, 2025.

[20] S. Kawano, H. Nonaka, and K. Yoshino: ClaimBrush: A novel framework for automated patent claim refinement based on large language models, Proc. of 2024 IEEE International Conference on Big Data, pp.6594-6603, 2024.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen: LoRA: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685, 2021.

[22] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn: Direct preference optimization: Your language model is secretly a reward model, Advances in Neural Information Processing Systems, 36, 2023.

[23] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela: KTO: Model alignment as prospect theoretic optimization, arXiv preprint arXiv:2402.01306, 2024.

[24] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault: Ground truth for grammatical error correction metrics, Proc. of ACL-IJCNLP 2015 (Volume 2: Short Papers), pp.588-593, 2015.

[25] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch: Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics, 4, pp.401-415, 2016.

[26] P. Lewis, E. Perez, A. Piktus, et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks, Advances in Neural Information Processing Systems, 33, pp.9459-9474, 2020.