構成要素の形式的な対応を考慮した強化学習による特許請求項翻訳

浅見遥斗[1] 宇津呂武仁[1] 永田昌明[2]
[1]筑波大学大学院 システム情報工学研究群 [2]NTTコミュニケーション科学基礎研究所
s2420710_@_u.tsukuba.ac.jp utsuro_@_iit.tsukuba.ac.jp
masaaki.nagata_@_ntt.com

概要

大規模言語モデル (LLM) の自然言語処理分野での活用が広がり，特許翻訳においても活用されているが，特許請求項は文長が極端に長く，構造も独特であるため，翻訳に特化させた LLM でも未だに訳抜けや構成要素対応の乱れが残る．本研究では，特許データを用いた継続事前訓練 (CPT) と教師ありファインチューニング (SFT) に加えて，Group Relative Policy Optimization(GRPO) を用いた強化学習を行うことで，8B 規模の LLM を特許請求項翻訳に特化させる．強化学習においては，行の長さを用いた構成要素の対応を考慮した報酬を導入する．日英，英日翻訳の評価では提案手法により GPT-5.2 に匹敵する性能を確認し，特に長文入力に対しては，本手法により大幅な性能向上を得られた．これらの結果から本手法が特許請求項翻訳に有効であることを示した．

1 はじめに

大規模言語モデル (LLM) は，その広範な事前訓練により，汎用性が優れており，要約や質問応答など多様な自然言語処理タスクで有用性が示されている．機械翻訳分野においても，GPT-4 のようなクローズドな LLM での翻訳が既存の翻訳モデルよりも人手評価が高いことが報告されている．特許分野においても，特許データを用いて訓練を行うことで従来の Transformer encoder-decoder ベースモデルや，GPT-4o に対して自動評価で有意に差をつける結果が得られている．本研究では，特許対訳データを用いた，継続事前訓練 (CPT) と教師ありファインチューニング (SFT)，Group Relative Policy Optimization(GRPO) を用いた強化学習を組み合わせることで特許請求項翻訳に特化した LLM を構築する．さらに，日本語および英語の双方で，改行を含む請求項の一行が個別の構成要素に対応するという性質に着目し，各行の長さの整合性を GRPO の報酬として利用する手法を新たに提案する．具体的には，特許文の中でも文の長さや独特の記述形式から翻訳難易度の高い特許請求項の日英，英日翻訳を行い，XCOMET や MetricX といった評価尺度を用いて翻訳精度を測定した．本研究の主な貢献は以下の二点である．

- 特許対訳データによる CPT と SFT に加えて GRPO を適用する三段階学習により，8B 規模のオープンモデルでありながら，クローズドモデルである GPT-5.2 と比較可能な水準の特許請求項翻訳性能を達成した点．
- 一部の請求項では改行が構成要素ごとの区切りと対応するという性質に着目し，日本語と英語の行長の整合性を GRPO の報酬として利用する手法を提案し，特に極端に長い請求項において翻訳精度を向上させた点．

2 関連研究

Xu ら [1] は，Llama-2[2] に対して 2 段階の fine-tuning を行う ALMA という手法を提案している．ALMA は，まず単言語データを用いた fine-tuning を行った後，少量の高品質な対訳データを用いた fine-tuning を行うことで，13B の LLM を使って GPT-3.5 に匹敵する翻訳精度を達成した．近藤ら [3] は，LLM の翻訳性能向上を目的として，「CPT」と「SFT」を組み合わせた二段階の訓練手法を提案している．この手法では，大規模な対訳データを用いて CPT を行い，その後高品質な対訳データを使用して SFT を実施している．このアプローチは，大量の対訳データを活用してモデルの翻訳性能を向上させ

るだけでなく，高品質なデータを用いたファインチューニングによって，翻訳精度をさらに高めることを目的としている．これらの手法を適用したモデルを，12 種のテストデータで評価を行い，対訳データで訓練された Transformer encoder-decoder ベースのモデルより統計的に有意に向上することを確認した．浅見ら [4] は，近藤らの手法を特許対訳データを用いて行うことで，LLM を特許翻訳に特化させ，特許請求項の翻訳において，特許対訳データで訓練された Transformer encoder-decoder ベースのモデルおよび，近藤らの用いた対訳データを使用して訓練された LLM に対し，翻訳精度が統計的に有意に向上することを確認した．また，MT-R1-Zero[5] では，強化学習の一つである GRPO(Group Relative Policy Optimization)[6] アルゴリズム内でルールベースによるフォーマットチェックと，複数の自動翻訳評価指標を組み合わせた報酬をすることで，SFT を行うことなく高い翻訳精度を達成した．

3 提案手法

先行研究 [3] にならい，単言語データを用いた事前訓練を行った LLM である Qwen/Qwen3-8B[1](以下 Qwen3-8B と呼ぶ) に対し，特許対訳データを用いた CPT および SFT を行った後，GRPO を適用した．ここで Qwen3-8B は日本語を含めた多言語の事前訓練がなされているモデルである．

3.1 CPT

事前訓練済みのモデルに対して特許対訳データを用いた CPT を行う．用いるデータの形式としては，{原言語文}\n{目的文} の形式である．

3.2 SFT

事前訓練済みのモデルに対して特許対訳データを用いた CPT を行った後，高品質な特許請求項対訳データを用いた SFT を行った．なお，損失計算の際は翻訳部分のみを学習させるために，指示部分と原言語文部分については除外して計算を行っている．用いる指示は GRPO で用いたプロンプトを使用した．

3.3 強化学習アルゴリズム: GRPO

Feng ら [5] の先行研究より，Group Relative Policy Optimization (GRPO) アルゴリズム [6] を強化学習アルゴリズムとして使用した．ロールアウトに用いるプロンプトとしては，Feng ら [5] に従った．

また，以下の長さ報酬を与える GRPO の際は，構成要素の明示のため，生成される翻訳に改行が含まれるようにするため，プロンプトにおけるフォーマット指定部分に改行部分を<break>と出力させる命令を追加した．GRPO の報酬関数としては，MT-R1-Zero の手法と同様のルールベースとメトリックによる報酬関数に加え，特許請求項翻訳を強化するために，改行で区切られた構成要素の行の長さを用いた報酬関数を組み合わせた．以下にその詳細を記載する．

3.3.1 フォーマット報酬

プロンプトにて指定したフォーマットに従っているかによって報酬を与える．すなわち，<think></think>の間に思考部分，<translate></translate>に翻訳を出力しているかを判定する．もし従っていれば 1, 従っていなければ –1 とした．

3.3.2 メトリック報酬

MT-R1-Zero [5] の結果より，参照文との文字的類似度を得る BLEU[7] と，参照なしの翻訳尺度である Cometkiwi[8] の数値を統合させたものを報酬として用いる場合が最も翻訳精度が向上すると報告されているため，本研究でも用いた．ここで，BLEU は sacreBLEU[9] を用いて計測し，Cometkiwi では WMT23-cometkiwi-da-xl[2] を用いた．それぞれ，$[0, 1]$ のスコアで与えられるため，ここでは $[0, 2]$ の報酬が与えられる．

3.3.3 長さ報酬

上記 2 種の報酬に加え，入力文と翻訳文の構成要素ごとの長さ比に基づき，構成要素対応の整合性を評価する長さ報酬を導入した．入力文および翻訳文を同一の構成要素単位に分割し，各構成要素の文字数比から対応の妥当性を測定する．

各構成要素 i に対し，入力側と翻訳側の文字数をそれぞれ l_i^{src}，l_i^{tgt} とし，長さ比を

$$r_i = \frac{l_i^{\mathrm{tgt}}}{l_i^{\mathrm{src}}}$$

と定義する．翻訳方向ごとに定義された期待長さ比

1） https://huggingface.co/Qwen/Qwen3-8B

2） https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

μ_{dir} からの正規化誤差を

$$\delta_i = \frac{r_i - \mu_{\text{dir}}}{\mu_{\text{dir}}}$$

とし，各構成要素の長さ整合度スコアを

$$s_i = \exp\left(-\frac{\delta_i^2}{2\sigma^2}\right)$$

として算出する．最終的な長さ報酬は，全構成要素に対する s_i の平均とした．ただし，日本語と英語で文頭と文末が入れ替わる場合があるため，最初と最後の構成要素に関しては考慮しないこととした．本手法は，文長差が正規分布に従うと仮定する Gale and Church らの手法 [10] に着想を得たものであり，期待長さ比による正規化とガウス型スコア化によって，構成要素対応の確率的整合性を簡潔に近似している．

3.3.4 各報酬の統合方法

フォーマット報酬を S_{format}，メトリック報酬を S_{metric}，長さ報酬を S_{length} とする．先行研究 [5] に従い，最終的な報酬 r は式 (1) で与える．

$$r = \begin{cases} S_{format} - 3, & \text{if } S_{format} = -1 \\ S_{format} + S_{metric} + S_{length}, & \text{if } S_{format} = 1 \end{cases} \tag{1}$$

4 実験設定

4.1 データセット

CPT および SFT，GRPO には，日英特許対訳コーパスである JaParaPat[11, 12] を使用した．本研究で利用した内訳は表 1 に示す．

なお JaParaPat には文分割により，本来 1 つであった請求項が改行などにより過剰に分割されている場合があるため，それらを結合してから SFT のデータを構築した．なお，長さ報酬を加えた GRPO を行う際には，結合する際に<break>という記号を挿入した．また，SFT データの高品質化のため，埋め込みを用いた類似度フィルタリングによりデータを選定した．GRPO に対しては，SFT 用のデータセットより 10,000 件をランダムに抽出し使用した．テストセットとしては，WAT(Workshop on Asian Translation)2025 で公開された dev set を用いた．

4.2 ハイパーパラメータ

本論文における CPT および SFT のハイパーパラメータの設定は浅見ら [4] に従った．GRPO の実装には，verl [3] フレームワークに基づいて行った．訓練の際のバッチサイズは 16 に設定し，アルゴリズム内での 1 プロンプトあたりの 8 件のロールアウトを行った．学習率は 5×10^{-7}，ロールアウト時のサンプリング温度は 1.0．また，生成される回答の最大長は 4096 トークンに制限した．また，先行研究にならい，KL ペナルティを 0 とすることで，KL 制約を考慮しないこととした．

表 1 特許対訳データの使用用途とデータ内訳

使用用途	対象期間	データ種別	データ数
継続事前訓練	2016 年 ~ 2020 年	訓練データ	981,364,685
		開発データ	30,000
SFT	2020 年	訓練データ	30,000
		開発データ	3,000
GRPO	2020 年	—	10000

4.3 比較対象

比較対象として，浅見ら [4] により高い翻訳性能が報告されている rinna/llama-3-youko-8b[4] に CPT および SFT を適用したモデルと，Qwen3-8B に対して CPT，SFT，GRPO（長さ報酬の有無）を組み合わせた各モデルを比較した．また，クローズドな LLM として GPT-5.2 との比較も行った．

4.4 推論プロンプト

推論には GRPO におけるロールアウトに用いたプロンプトを使用した．

4.5 評価指標

評価指標として，参照ありの評価指標である XCOMET[13] と，参照なしの評価指標である，MetricX[14] を利用した．XCOMET のモデルは XCOMET-XL[5] を用いて計測した．また，MetricX では MetricX-24-Hybrid-QE-XL[6] を利用した．

5 評価結果

5.1 先行研究手法および訓練手法の組み合わせによる比較

表 2 に結果を記載した．先行研究の手法との比較として，本研究にて使用した Qwen3-8B は先行研究 [4] で高い翻訳精度を達成したモデルと比較し両

3) https://github.com/volcengine/verl
4) https://huggingface.co/rinna/llama-3-youko-8b
5) https://huggingface.co/Unbabel/XCOMET-XL
6) https://huggingface.co/google/metricx-24-hybrid-xl-v2p6

表 2 訓練方法ごとの XCOMET, MetricX スコア．GRPO 欄の ✓ はフォーマット報酬とメトリック報酬を，+length は加えて長さ報酬を用いたことを示す．* は Qwen3-8B と有意差あり ($p < 0.05$).

CPT	SFT	GRPO	英日翻訳		日英翻訳	
			XCOMET↑	MetricX↓	XCOMET↑	MetricX↓
youko-8b-cpt-sft			20.97	7.09	29.61	9.79
Qwen3-8B			29.96	6.35	39.54	4.96
✓	✓		17.12	8.49	37.78	6.91
		✓	24.41	5.64	49.53	4.23*
		✓(+ length)	24.94	6.77	39.54	4.96
✓	✓	✓	21.02	5.64	51.02*	4.02*
✓	✓	✓(+ length)	27.26	4.81*	51.46*	4.05*
GPT-5.2			**30.52**	**4.75***	**52.67***	**3.51***

翻訳方向に対して有意に翻訳精度が上回る結果となった．これは，Qwen3-8B が多言語事前訓練および Instruction Tuning により高い翻訳能力を備えているためと考えられる．

各訓練手法の比較では，日英方向においては CPT，SFT，GRPO をすべて組み合わせたモデルが両評価指標で最も高い性能を示したが長さ報酬の有無による有意差は確認されなかった．一方，英日方向では XCOMET では元モデルが，MetricX では提案手法を適用したモデルが最も高い性能を示した．

5.2 GPT-5.2 と提案手法の比較

GPT-5.2 と提案手法の結果について比較する．英日方向では両評価指標において統計的有意差はなく，提案モデルは 8B という大幅に小規模なモデルでありながら GPT-5.2 と同等の翻訳精度を達成した．日英方向でも XCOMET では有意差がなく，一部指標において GPT-5.2 に匹敵する性能を示した．

6 改行を含む請求項に対する評価

6.1 評価手順と結果

GRPO における長さ報酬の分析のため，WAT2025 dev set から原言語文に改行が含まれるもののみを対象に翻訳を行い評価を行った．なお，翻訳評価指標は 4.5 節と同様に XCOMET と MetricX を用いた．結果を付録 A に記載する．

6.2 長文入力に対する提案手法の効果

提案手法で導入した長さ報酬の効果を確認するため，特に英日方向で長文が多い WAT2025 dev set を用いて検証を行った．本セットには 400 トークンを超える文が含まれており，これらを対象として評価した．なお，トークナイズには Qwen3-8B を用いた．

結果を表 3 に示す．提案手法を適用したモデルは長文に対して安定した高い翻訳精度を示し，XCOMET では GPT-5.2 を上回る値を達成した．このことから，長さに基づく報酬設計が長文翻訳に寄与していることが示唆される．

表 3 英日方向における長文データを用いた訓練方法ごとの XCOMET, MetricX スコア．* は GPT-5.2 と有意差あり ($p < 0.05$).

CPT	SFT	GRPO	日英翻訳	
			XCOMET↑	MetricX↓
Qwen3-8B			17.3	8.12*
✓	✓		21.32	12.64*
		✓	18.83	7.83*
		✓(+ length)	17.3	8.12*
✓	✓	✓	13.92*	6.75*
✓	✓	✓(+ length)	**23.16**	6.47
GPT-5.2			20.94	**6.32**

7 おわりに

本研究では，特許請求項翻訳における構成要素対応や訳抜けの抑制を目的として，CPT および SFT に加え，GRPO を用いた強化学習を導入し，フォーマット報酬，メトリック報酬，さらに構成要素の長さ比に基づく長さ報酬を組み合わせた報酬設計を提案した．

これらの手法を統合した三段階学習モデルを用いた実験により，日英・英日の双方で翻訳性能の向上が確認された．特に長文入力に対しては，CPT，SFT，および長さ報酬を含む GRPO を組み合わせたモデルが高い翻訳精度を示し，一部指標では GPT-5.2 を上回る結果となった．これらの結果は，提案した報酬設計と強化学習の導入が特許請求項翻訳に有効であることを示している．

今後は，人手による定性的分析や，入力文の長さに依存しない報酬設計について検討を進める．

参考文献

[1] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In **Proc. 12th ICLR**, pp. 1339–1352, 2024.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv**, Vol. 2307.09288, pp. 1–77, 2023.

[3] Minato Kondo, Takehito Utsuro, and Masaaki Nagata. Enhancing translation accuracy of large language models through continual pre-training on parallel data. In **Proc. 21th IWSLT**, pp. 203–220, 2024.

[4] Haruto Azami, Minato Kondo, Takehito Utsuro, and Masaaki Nagata. Patent claim translation via continual pre-training of large language models with parallel data. In **Proc. 20th MTSummit**, pp. 300–314, 2025.

[5] Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. MT-R1-Zero: Advancing LLM-based machine translation via r1-zero-like reinforcement learning. In **Proc. EMNLP**, pp. 18685–18702, 2025.

[6] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. **arXiv**, Vol. 2402.03300, pp. 1–30, 2024.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

[8] Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, et al. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In **Proc. 7th WMT**, pp. 634–645, 2022.

[9] Matt Post. A call for clarity in reporting BLEU scores. In **Proc. 3rd WMT**, pp. 186–191, 2018.

[10] William A Gale and Kenneth Church. A program for aligning sentences in bilingual corpora. **Computational Linguistics**, Vol. 19, No. 1, pp. 75–102, 1993.

[11] Masaaki Nagata, Makoto Morishita, Katsuki Chousa, and Norihito Yasuda. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In **Proc. 14th LREC**, pp. 9452–9462, 2024.

[12] Masaaki Nagata, Katsuki Chousa, and Norihito Yasuda. JaParaPat: A large-scale Japanese-English parallel patent application corpus. **arXiv**, Vol. 2508.16303, pp. 1–12, 2025.

[13] Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 979–995, 2024.

[14] Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In **Proc. 9th WMT**, pp. 492–504, 2024.

表 4　改行付きデータを用いた訓練方法ごとの XCOMET, MetricX スコア．* は Qwen3-8B と有意差あり ($p < 0.05$).

CPT	SFT	GRPO	英日翻訳		日英翻訳	
			XCOMET↑	MetricX↓	XCOMET↑	MetricX↓
Qwen3-8B			23.80	7.12	31.55	6.24
✓	✓		17.99	12.23	29.95	8.88
		✓	**26.07**	6.74	35.51	4.42*
		✓(+ length)	23.80	7.12	31.55	6.24
✓	✓	✓	15.02	6.98	**36.02**	**3.92***
✓	✓	✓(+ length)	22.80	5.85*	32.48	4.71*
GPT-5.2			25.57	**5.52***	33.24	4.14*

A　改行を含む請求項に対する評価

A.1　評価手順と結果

GRPO における長さ報酬の分析のため，WAT2025 dev set から原言語文に改行が含まれるもののみを対象に翻訳を行い評価を行った．なお，翻訳評価指標は 4.5 節と同様に XCOMET と MetricX を用いた．結果を Table 4 に記載する．

結果としては，英日においては，XCOMET では Qwen3-8B に長さ報酬なしの GRPO を適用したモデル,MetricX においてはすべての手法を組み合わせたモデルが評価の高い結果となったが，GPT-5.2 には及ばぬ結果となった．それに対し，日英においては，長さ報酬を与えない GRPO を Qwen3-8B に CPT と SFT を行ったのちに適用したものが両翻訳指標において最も高い結果となり，GPT-5.2 にも勝る結果となった．また，単に Qwen3-8B に対して長さ報酬を与えない GRPO を適用した際の結果もその他組み合わせと比較し高い翻訳精度を達成した．なお，2 節の結果においても本評価の結果においても，Qwen3-8B に対して長さ報酬を加えた GRPO を適用した際の結果が元モデルと変わらない点があるが，これは訓練が進んでいく過程で全くフォーマットルールを守らなかったことにより，メトリック報酬，長さ報酬が全く関与していないことが原因と考えられる．

A.2　生成される翻訳の長さによる分析

提案手法にて与えた長さ報酬による GRPO を適用したことで，生成される翻訳の長さについての相関に向上があったかの調査を行った．なお，英日方向においては，テストデータとして用いた WAT2025 dev set での参照翻訳文に一切の改行が存在しないため，JaParaPat における改行を含む請求項との比較を行い，テストデータとの構成要素単位の長さについての分析は日英方向のみ行う．

表 5 に日英における原言語文，参照翻訳文，および Qwen3-8B に複数の手法を組み合わせたモデルによる翻訳結果における請求項内の構成要素単位のトークン数および文字数を記載し，表 6 に英日における構成要素単位のトークン数および文字数を記載した．なお，トークナイズには Qwen3-8B を用いた．

この結果より，日英においては，長さ報酬により，構成要素単位の区切り方，長さについては相関が上がったことが確認できる．なお，長さ報酬なしでの GRPO を適用した結果の構成要素単位の長さについては，出力された改行の個数が少ないために大きく参照から離れる結果となった．ただし英日においては，原言語文とのトークン比率が元モデルが最も JaParaPat 内の請求項と近い結果となった．

表 5　日英における，請求項単位の構成要素ごとの平均トークン数および平均文字数

データ種 / モデル	平均トークン数	平均文字数
src	75.47	103.71
ref	124.56	507.76
Qwen3-8B	226.86	457.06
Qwen3-8B-cpt-sft	118.27	390.21
Qwen3-8B-cpt-sft-grpo	260.60	1119.30
Qwen3-8B-cpt-sft-grpo(length)	121.16	290.58
GPT-5.2	52.10	217.10

表 6　英日における，請求項単位の構成要素ごとの平均トークン数および平均文字数

データ種 / モデル	平均トークン数	平均文字数
src(WAT dev)	105.13	295.46
ref(WAT dev)	460.86	791.35
JaParaPat En	58.74	245.81
JaParaPat Ja	62.00	90.57
Qwen3-8B	115.12	230.56
Qwen3-8B-cpt-sft	454.49	663.17
Qwen3-8B-cpt-sft-grpo	789.48	1185.15
Qwen3-8B-cpt-sft-grpo(length)	79.38	124.39
GPT-5.2	83.33	129.79