BERT を用いた誤訳検出と LLM を用いた誤訳訂正による特許翻訳の自動後編集

武馬光星[1] 西村柾人[2] 宇津呂武仁[2] 永田昌明[3]
[1]筑波大学 理工学群 工学システム学類 [2]筑波大学大学院 システム情報工学研究群
[3]NTT コミュニケーション科学基礎研究所
{s2313594, s2320779}_@_u.tsukuba.ac.jp utsuro_@_iit.tsukuba.ac.jp
masaaki.nagata_@_ntt.com

概要

特許翻訳は，その専門性と厳密さから高い翻訳精度が求められる分野である．しかし，従来の Transformer ベースのニューラル機械翻訳 (NMT) モデルでは，特許文特有の長文構造や複雑な書式に起因する訳抜けや繰り返しといった誤訳が発生しやすいという課題がある．さらに，近年急速に発展している大規模言語モデルの翻訳訂正能力については，特許翻訳における有効性が十分に検討されていない．本論文では，mBERT を用いたトークンレベルの誤訳検出と大規模言語モデル (LLM) を用いた訂正手法を組み合わせた手法を提案する．BLEU および COMET のスコアで評価を行い，提案手法が最も高い翻訳精度を達成し，誤訳や繰り返しの改善に寄与することが示された．一方で，訳抜け文の訂正においては，課題も確認された．

1 はじめに

近年，機械翻訳は翻訳モデルの発展により飛躍的に向上している．しかし，特許文のように専門性が高く，厳密な表現や構造を必要とする文に対しては，依然として誤訳や訳抜け，繰り返し誤りといった課題が残されている．特許文は法的効力を持つ文書であり，翻訳の精度と一貫性は極めて重要である．したがって，高精度な誤訳検出および訂正技術の開発は，特許翻訳の実務や国際的な特許申請において不可欠である．

本論文では，特許翻訳における誤訳検出および訂正手法について評価を行う．特に，多言語対応された BERT [3](mBERT) を用いたトークンレベルの誤訳検出と，LLM を活用した翻訳訂正の組み合わせた手法を提案する．この手法により，誤訳検出の精度を向上させるとともに，検出結果を利用したより精度の高い翻訳訂正を実現する．具体的には，図 1 に示すように，特許文で学習を行なった mBERT を用いて，トークンレベルのタグ付けを行う手法，LLM を用いて翻訳文を訂正する手法の 2 段階の手法となる．

1 段階目の手法では，特許文で学習した mBERT を用いる．mBERT の学習を行うために，対訳特許文のターゲット文に人工的に誤りを作成する．人工的な誤りが発生した特許文で mBERT の学習を行い，翻訳文中の誤訳をトークンレベルで検出し，タグ付けを行う．2 段階目の手法では，検出された誤訳に基づいて，LLM(GPT-4o) に誤りを述べさせて，翻訳文の訂正を行う．

提案手法の評価を行うために，作成した人工誤り特許文，繰り返し誤り特許文，訳抜け特許文で評価を行なった．その結果，人工誤り特許文，繰り返し誤り特許文において，提案手法が他の手法と比較して，BLEU において統計的に有意な改善を確認した．

2 関連研究

2.1 翻訳における単語品質評価

Wei ら [13] は，単語品質推定のために mBERT を用いた教師あり学習手法を提案している．具体的には，翻訳文とソース文を連結した入力に対して回帰モデルを構築し，各トークンが BAD タグに該当する確率を出力するよう学習を行うものである．

本論文においても，Wei ら [13] の手法を基に，特許文を対象とした教師あり学習を適用することで，特許文における単語品質評価の精度向上を目指す．

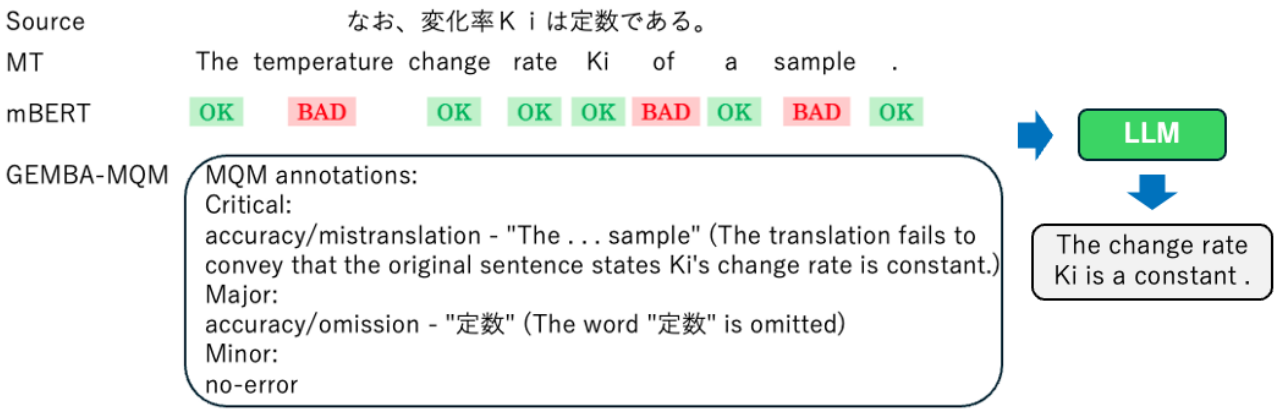

図 1: 誤り検出→ LLM での誤り訂正

2.2 翻訳における後編集

近年，機械翻訳の品質向上に向けて，後編集において LLM を活用する研究が進展している．

Xu ら [14] は，LLMRefine と呼ばれる手法を提案した．LLM での翻訳結果に対して，エラーカテゴリとエラースパンのリストを生成するように学習した LLM で誤りの検出を行う．出力された詳細なフィードバックを用いて誤りを特定し，LLM が反復的に修正を行うことで，英独，中英翻訳タスクにおいて翻訳スコアの向上を実現した．

また，Ki ら [8] は，外部からのフィードバックを用いて，LLM での機械翻訳の後編集を行う手法を提案した．彼らの研究では，Multidimensional Quality Metric (MQM) [1] によるエラーアノテーションを活用し，LLM に外部からのフィードバックを与えることで，翻訳品質の改善を図った．中英，英独，英露のデータセットを用いた実験において，TER，BLEU，COMET スコアの向上が確認された．

3 誤訳検出

3.1 mBERT を用いた誤訳検出

本論文では，多言語学習済み言語モデルである mBERT を活用し，トークンレベルでの翻訳品質評価を行う．具体的には，翻訳文中の誤りを検出するために mBERT の事前学習された知識を利用し，誤りのラベル付けを行う．mBERT の訓練においては，Deguchi ら [2] の手法に従い，NTCIR-7 [6], NTCIR-8 [7] の対訳特許文に対して以下の操作を行うことによって，人工誤りデータを作成する．

- 削除: トークンを 5%の確率で削除
- 挿入: トークンを 10%の確率で挿入
- 置換: トークンを 30%の確率で置換

挿入および置換の処理では，mBERT を用いて MASK トークンを補完する手法を採用する．具体的には，挿入箇所および置換箇所に MASK トークンを配置し，その位置に適合するトークンを mBERT により予測する．この際に予測するトークンは，元のトークンとの類似度を下げるために，Transformers の Pipelines [4] で出力できる score の最も低いトークンとする．作成した擬似データの誤りトークンに BAD タグを付与し，残りのトークンに OK タグを付与することで教師データを作成する．以上の手法を用いて作成された教師データ 8000 文を用いて，mBERT の学習を行う．

作成した教師データの有効性を示すために，mBERT のタグ付け精度の比較を行う．結果を付録 A に示す．

3.2 LLM を用いた誤訳検出

LLM を用いた誤訳検出手法として，Kocmi ら [9] が提案した GPT ベースの評価手法である GEMBA-MQM を利用する．この手法は，GPT-4 を基盤とし，言語に依存しない固定の 3-shot プロンプトを用いて翻訳エラー範囲・種類を出力するものである．GEMBA-MQM の手法に基づき，以下の二種類の設定で誤り検出を行う．

- 0-shot: 事前の例示なしで翻訳誤り検出を行う．
- 3-shot: 言語に依存しない 3 つの具体例を提供し，翻訳誤り検出を行う．

特に，3-shot の手法は，GPT-4 を用いた最も高い誤り検出精度を達成する手法として報告されている [9]．mBERT，LLM を用いた誤訳検出は，後続す

表 1: 人工誤りを対象とした誤り検出評価

(a) 日英翻訳

モデル	項目	Precision	Recall	F1
GPT-4o(0shot)	OK	0.843	0.0774	0.142
	BAD	0.389	0.976	0.556
	TOTAL	F1: 0.298, MCC: 0.111		
GPT-4o(3shot)	OK	0.755	0.268	0.395
	BAD	0.409	0.853	0.553
	TOTAL	F1: 0.454, MCC: 0.141		
mBERT	OK	0.831	0.895	**0.862**
	BAD	0.797	0.696	**0.743**
	TOTAL	F1: **0.817**, MCC: **0.609**		

(b) 英日翻訳

モデル	項目	Precision	Recall	F1
GPT-4o(0shot)	OK	0.804	0.287	0.423
	BAD	0.415	0.879	0.563
	TOTAL	F1: 0.474, MCC: 0.190		
GPT-4o(3shot)	OK	0.709	0.503	0.588
	BAD	0.419	0.634	0.504
	TOTAL	F1: 0.558, MCC: 0.132		
mBERT	OK	0.824	0.925	**0.872**
	BAD	0.831	0.651	**0.730**
	TOTAL	F1: **0.821**, MCC: **0.615**		

る翻訳訂正の前処理として機能し，誤訳検出結果を用いて翻訳訂正の精度向上を目指す．

4 LLM を用いた誤訳訂正

原文と翻訳文を入力することで，LLM が翻訳文の誤りを分析し，適切な訂正文を生成する．具体的には，LLM は検出された誤訳部分について詳細に分析し，誤り箇所と誤りについての説明を行い，その内容に基づいて訂正文を生成する．LLM が誤りに対する説明を述べることで，訂正結果の透明性を高めるとともに，翻訳の改善過程を明確にする．

さらに，本論文では，前章で述べた誤訳検出結果を基に，LLM を用いて翻訳文の訂正を行う手法を提案する．LLM や mBERT の誤訳検出結果を入力として活用することで，訂正精度のさらなる向上を目指す．

5 評価

5.1 データセット

本論文では，以下の 3 種類のデータに対して評価を行う．

- 人工誤り特許データ
- 繰り返し誤り特許データ
- 訳抜け誤り特許データ

人工誤り特許データは，3.1 節で述べた手法で NTCIR7 と NTCIR8 の特許文の対訳データに人工的に誤りを発生させたデータである．この人工誤り特許データ 200 文で誤訳に対する訂正能力を評価する．

また，繰り返し誤りや訳抜け誤りに対する翻訳訂正精度を調べるために，2021 年の特許データの特許請求項を用いて評価する．この特許請求項を Transformer で翻訳したデータ (日英翻訳) から，以下の基準で抽出を行なった．

- 繰り返し誤り文: Transformer による翻訳文の文長が参照訳文の文長の 2 倍以上の文
- 訳抜け誤り文: Transformer による翻訳文の文長が参照訳文の文長の 0.5 倍以下の文

また，対訳データの質を高めるために，LaBSE [5] の埋め込みで原文と参照文の類似度を取り，類似度が 0.8 以上 0.98 以下の文を抽出する．抽出した繰り返し誤り特許データ 211 文，訳抜け誤り特許データ 200 文を評価に用いる．

5.2 評価手順

5.2.1 誤り検出

誤り検出の評価では，特許文に人工的に誤りを付与した 100 文に対して，トークン単位で翻訳誤り箇所に BAD タグを付与することで評価を行う．mBERT，LLM において図 1(左) と同様のタグのアノテーションを行う．以下の 3 種類のモデルで翻訳誤り検出を行い，評価を行う．

1. LLM(GPT-4o) - 0-shot
2. LLM(GPT-4o) - 3-shot
3. mBERT

F1 値と MCC を評価指標として用いる．

5.2.2 誤り訂正

誤り訂正の評価では，特許文に人工的に誤りを付与した文，繰り返し誤り特許文，訳抜け誤り特許文に対して，LLM で誤り訂正を行い評価する．LLM に，原文，翻訳文に加えて，mBERT，LLM での誤り検出の出力結果を与えて誤り訂正を行う．モデルの組み合わせを以下に示す．

表 2: 翻訳訂正評価 (BLEU/COMET)

	手法	人工誤り (日英)	人工誤り (英日)	繰り返し誤り (日英)	訳抜け誤り (日英)
1	修正なし	31.81/70.58	31.65/75.73	21.33/69.59	19.01/71.52
2	LLM - 検出 (誤り説明) + 修正	47.14/83.87	41.29/90.18	26.34/76.78	65.03/85.37
3	LLM - 検出 (GEMBA:0-shot)⇒ LLM - 検出 (誤り説明) + 修正	43.44/83.29	38.16/89.91	25.37/76.79	63.27/86.67
4	LLM - 検出 (GEMBA:3-shot)⇒ LLM - 検出 (誤り説明) + 修正	47.32/83.97	39.4/90.03	25.8/76.78	**66.08**/88.52
5	mBERT - 検出 (タグ) ⇒ LLM - 検出 (誤り説明) + 修正	**49.37**/84.08	**41.58**/90.11	**27.73**/76.65	56.65/83.71

1. LLM(GPT-4o) - 検出 (誤り説明) + 修正
2. LLM (GPT-4o) - 検出 (GEMBA:0-shot)
   → LLM (GPT-4o) - 検出 (誤り説明) + 修正
3. LLM (GPT-4o) - 検出 (GEMBA:3-shot)
   → LLM (GPT-4o) - 検出 (誤り説明) + 修正
4. mbert - 検出 (タグ)
   → LLM (GPT-4o) - 検出 (誤り説明) + 修正

翻訳修正後の文の BLEU [10] と COMET [12] を評価する．BLEU は，sacreBLEU [11] を用いて評価した．COMET のモデルは，wmt22-comet-da を使用した．

5.3 評価結果

5.3.1 誤り検出評価

人工的に誤りを作成した特許データに対して誤り検出を行なった結果を表 1 に示す．表 1 の結果から，mBERT による翻訳誤り検出が F1 値および MCC の両指標で最も優れた性能を示したことがわかる．特に，Precision と Recall のバランスの良さが総合性能の高さに寄与している．一方，GPT-4o はプロンプト設定によって結果が異なり，3-shot 設定では 0-shot より改善が見られたものの，mBERT には及ばなかった．GPT-4o の出力を分析すると，ほとんどのトークンに BAD タグが付与されていることが確認された．その結果 BAD タグの Recall は高い値を示したものの，OK タグの Recall が著しく低下する傾向となった．

これらの結果は，特許文で学習を行なった mBERT が特許文に対してトークンレベルでの誤り検出とタグ付けを効果的に処理できる点に起因すると推測される．特に，mBERT によるタグ付けを活用した誤訳検出手法が有効であることを示唆している．

一方，GPT-4o は検出性能の低さやプロンプトの種類による性能のばらつきが課題として挙げられ，より高精度な誤訳検出・訂正のためには工夫が求められる．

これらの結果を踏まえ，特許文で学習を行なった mBERT の検出結果を利用して LLM での訂正を行う本論文の提案手法は，高精度な誤訳検出と訂正を一貫して処理できる手法として有望であると考えられる．

5.3.2 誤り訂正評価

表 2 に示す結果から，mBERT ベースの誤り検出と LLM を組み合わせた提案手法が，BLEU および COMET のスコアにおいて他の手法を上回り，翻訳訂正の精度が最も優れていることが確認された．

特に，日英翻訳の人工的に誤りを作成した特許文に対しては，他の手法と比較して BLEU が統計的に有意に向上した．この結果は，mBERT によるトークンレベルの誤訳検出が高い精度で機能し，その情報を活用して LLM が適切な翻訳訂正を行うことが可能であることを示している．

繰り返し誤り特許文においても，提案手法が最も良い精度を示し，BLEU が統計的に有意に向上した．この結果から，繰り返し誤りがある場合においても提案手法が有効であることが示された．

一方，訳抜け特許文においては，BLEU スコアが最も高かった手法は，LLM で検出および訂正を行う手法となった．提案手法が精度で劣った原因としては，ターゲット側のタグ情報が訳抜け誤りを表現することに限界があり，翻訳訂正に十分な情報を提供できなかった可能性が考えられる．

6 おわりに

本論文では，特許文で学習した mBERT で翻訳誤り検出を行い，LLM で翻訳誤り訂正を行うことで，BLEU が他の手法と比較して有意に向上し，誤訳や繰り返し誤りを改善することを確認した．特に，mBERT による高精度な誤訳検出が LLM による訂正を支え，翻訳精度全体の向上に寄与した．一方で，訳抜け誤りの訂正においては，LLM で検出および訂正を行なったモデルが提案手法を上回り，タグの使用において課題を残した．この結果から，誤りの種類に応じた手法の最適化が重要であることが示された．

参考文献

[1] A. Burchardt. Multidimensional quality metrics: a flexible system for assessing translation quality. In **Proc. TC**, pp. 1–7, 2013.

[2] H. Deguchi, M. Nagata, and T. Watanabe. Detector–corrector: Edit-based automatic post editing for human post editing. In **Proc. 25th EAMT**, pp. 191–206, 2024.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proc. NAACL**, pp. 4171–4186, 2019.

[4] Hugging Face. transformers pipelines. https://huggingface.co/docs/transformers/en/main_classes/pipelines.

[5] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In **Proc. 60th ACL**, pp. 878–891, 2022.

[6] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In **Proc. 7th NTCIR**, pp. 389–400, 2008.

[7] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, T. Ehara, H. Echizen-ya, and S. Shimohata. Overview of the patent translation task at the NTCIR-8 workshop. In **Proc. 8th NTCIR**, pp. 371–376, 2010.

[8] D. Ki and M. Carpuat. Guiding large language models to post-edit machine translation with error annotations. In **Findings of NAACL**, pp. 4253–4273, 2024.

[9] T. Kocmi and C. Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In **Proc. WMT**, pp. 768–775, 2023.

[10] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

[11] M. Post. A call for clarity in reporting BLEU scores. In **Proc. WMT**, pp. 186–191, 2018.

[12] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **Proc. WMT**, pp. 578–585, 2022.

[13] Y. Wei, T. Utsuro, and M. Nagata. Extending word-level quality estimation for post-editing assistance, 2022. https://arxiv.org/abs/2209.11378.

[14] W. Xu, D. Deutsch, M. Finkelstein, J. Juraska, B. Zhang, Z. Liu, W. Y. Wang, L. Li, and M. Freitag. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In **Findings of NAACL**, pp. 1429–1445, 2024.

表 3: mBERT のタグ付け精度評価

(a) 日英翻訳

モデル	項目	Precision	Recall	F1
未学習 mBERT	OK	0.673	0.140	0.232
	BAD	0.372	0.882	0.523
	TOTAL	F1: 0.338, MCC: 0.031		
低品質教師データ	OK	0.707	0.913	0.797
	BAD	0.694	0.342	0.459
	TOTAL	F1: 0.673, MCC: 0.320		
提案手法	OK	0.855	0.883	**0.869**
	BAD	0.785	0.740	**0.762**
	TOTAL	F1: **0.830**, MCC: **0.631**		

(b) 英日翻訳

モデル	項目	Precision	Recall	F1
未学習 mBERT	OK	0.667	0.002	0.004
	BAD	0.362	0.998	0.531
	TOTAL	F1: 0.195, MCC: 0.002		
低品質教師データ	OK	0.747	0.873	0.805
	BAD	0.681	0.478	0.562
	TOTAL	F1: 0.717, MCC: 0.388		
提案手法	OK	0.870	0.880	**0.875**
	BAD	0.784	0.769	**0.776**
	TOTAL	F1: **0.839**, MCC: **0.651**		

表 4: 提案手法による人工誤りの訂正例

原文
ステップＳ１１において、プライマリプーリ１１への入力トルクを計算する。
参照訳文
In a step S11 , an input torque to the primary pulley 11 is calculated .
人工誤り文
In a processing stepd , The input torque to be primary pulley 11 is achieved :
提案手法
In step S11, the input torque to primary pulley 11 is calculated.

表 5: 提案手法による繰り返しの訂正

原文
a. 配電ハードウェアの構成部品として、少なくとも 1 つの受動電磁センサをインストールするステップと、b.
参照訳文
a.Installing at least one Passive Electromagnetic Sensor as a component of distribution hardware;b.
MT 文
a. installing at least one passive electromagnetic sensor as a component part of the electrical distribution hardware;b. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;c. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;d. ... controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;g
提案手法
a. installing at least one passive electromagnetic sensor as a component part of the electrical distribution hardware; b.

A　mBERT タグ付け精度比較

作成した教師データの有効性を示すために，未学習 mBERT，低品質教師データで学習した mBERT，提案手法の mBERT のタグ付け精度の比較を行う．低品質教師データは，作成時に mBERT を用いて最も確からしいトークンを挿入・置換することで生成した．この時，元のトークンと類似度が高いトークンが挿入・置換された場合にも誤りとして BAD タグが付与されるため，教師データの質が低下する．

表 3 の結果から，提案手法の誤り検出精度が最も高いことが確認できる．このことから，今回作成した教師データの有効性が確認できる．

B　人工誤り特許文，繰り返し特許文の修正例

表 4 より，提案手法が人工誤り文の誤訳を修正できていることが確認できる．表 5 より，提案手法が人工誤り文の誤訳を修正できていることが確認できる．特に先頭や末尾などの記号も維持して訂正を行うことができていることがわかる．