機械翻訳出力に対するLLMを用いた後編集

Post-editing of Machine Translation Output Using Large Language Models

株式会社アスカコーポレーション 早川威士
Takeshi Hayakawa, ASCA Corporation

Introduction

後編集 (post-editing: PE) は、機械翻訳出力の品質管理工程として行われる編集プロセスである。本研究では、大規模言語モデル (Large Language Models: LLM) による英語→日本語翻訳におけるPEのタスク精度を評価した。
LLMのタスク精度に影響を与える要因として、機械翻訳出力の誤り、LLMへの指示文、LLMのモデルの3つを取り上げ、それぞれのタスク精度への影響度を調べた。タスク精度は、タスクの正解率と、誤り訂正を超過した過剰な修正の程度 (修正率) によって評価した。

Materials and Methods

要因の定義

誤りタイプ：機械翻訳出力の誤りを5タイプに分類した。

- 数値誤り (numerical error: num)
- 湧出し (over generation: ogen)
- 訳抜け (under generation: ugen)
- 同じ単語の繰返し (repetition error: reperr)
- 意味的誤り (semantic error: semerr)

プロンプト：LLMへの指示文を8タイプに類型化した。

- Vanilla: 最低限の指示
- Persona: 役割の付与
- Context: タスクの背景説明
- Principles: 指示の範囲と制約の説明
- Oneshot: 例示 (1つ)
- Fewshot: 例示 (複数)
- Reason: 修正根拠の提示
- Score: 自己採点の実施

モデル：使用したLLMのモデル６種類

- GPT-4.1
- GPT-4.1-mini
- GPT-4.1-nano
- GPT-5
- GPT-5-mini
- GPT-5-nano

評価データセット

英語→日本語のパラレルコーパスに対し、日本語側に人為的に各タイプの誤りを挿入したデータセット200文を用いた。

評価指標（目的変数）

正解率：PEによって機械翻訳出力の誤りを正しく訂正できたかどうかを二値で判定し、その割合を要因ごとに算出した。

修正率：PEが機械翻訳出力を書き換えた文字の割合を要因ごとに算出した。修正率の高さは過剰な修正の程度が高いことを示すものとみなした。

評価方法

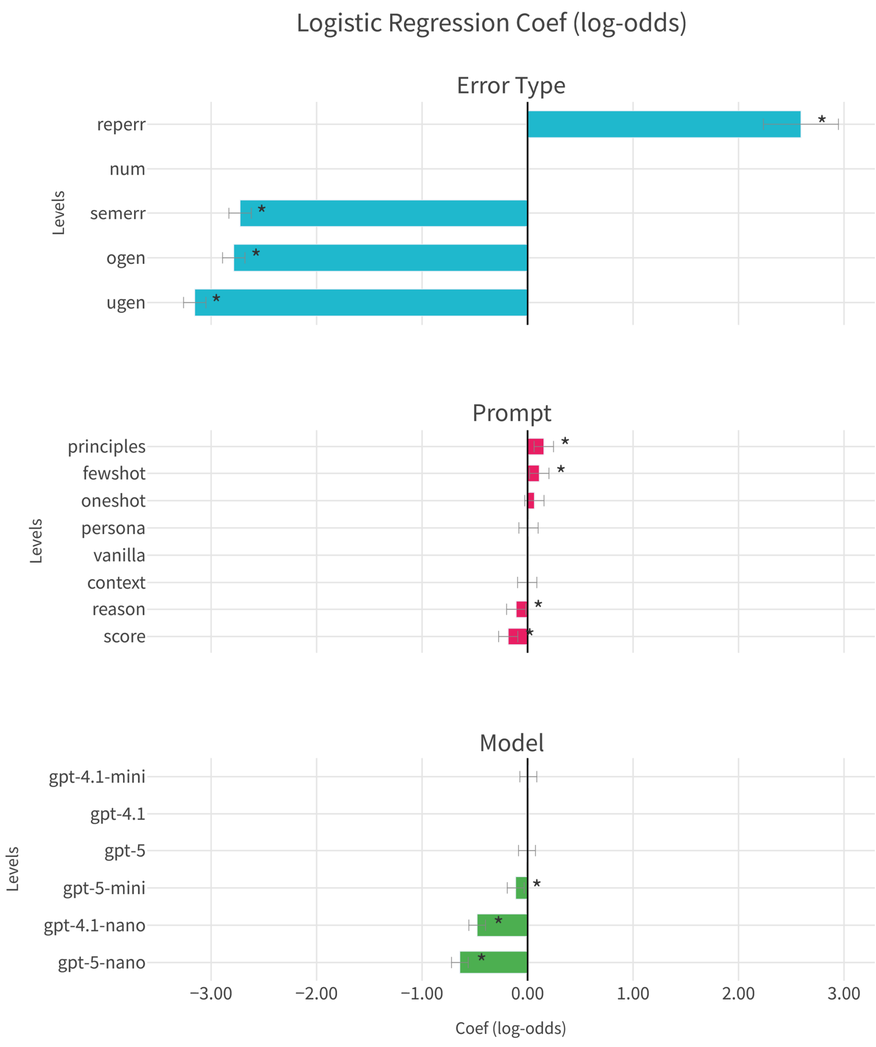
正解率：ロジスティック回帰分析

修正率：三元配置分散分析 (ANOVA)

Results

正解率の解析

PEの正解率に寄与する要因としては誤りタイプが最も強く、同じ単語の繰り返し、数値誤りといったエラーは修正が容易である一方、意味的誤り、湧出し、訳抜けは修正が困難であった。
プロンプトの影響は比較的少なかったが、指示の原則を説明することや複数の例示を行うことで正解率は改善した。しかし、自己評価を課するようなプロンプトでは正解率は悪化した。
モデルについては、世代ごとの差は無視できるものであったが、パラメータ数の少ないとされる軽量モデル (GPT-4.1-nano、GPT-5-nano) はパフォーマンスが有意に低かった。



修正率の解析

修正率の増加についてはプロンプトとモデルの主効果およびプロンプト x モデルの交互作用が極めて大きく、LLMによる過剰な修正はこれらの要因に強く依存することが示された。

Discussion

LLMによる機械翻訳出力のPEの正否は、出力に含まれる誤りのタイプに大きく左右されることが示された。ニューラル機械翻訳モデルに典型的に見られる数値の誤りや同じ単語の繰返し出力に対しては高精度で訂正できることが確認された一方で、自然文として破綻の生じにくい訳抜けや湧出しなどのエラーについては正解率が低くなる傾向が見られた。プロンプトによる改善効果は限定的であったが、指示の原則について説明することや複数の例示を与えることには一定の効果が認められ、これらのプロンプト文を洗練させる方向性は支持されたと言える。
翻訳に限らず、LLMによる自動的なPEでは意図を超えた過剰な編集がしばしば見られる。この問題に対しては、プロンプトの改善とモデル選択により対処できる可能性が示された。過剰な編集を行わないよう指示することと、その指示を理解できるモデルのペアリングを最適化する運用が望ましいと考えられる。