言語的検証の未来の到来:
生成 AI で照合プロセスのコンプライアンスを強化

プレゼンター (ライオンブリッジ): エリサベト サス オレサ、カロリナ エリソンド、キャスリン ノルテ、ナタリー アスワヘ、メリンダ ジョンソン

はじめに

この調査では、生成人工知能 (生成 AI) がどのように言語的検証 (LV) プロセスをサポートできるかを、特に二重翻訳照合 (Dual Translation Reconciliation) の工程の効率性を高めるうえでの生成 AI の役割に焦点を当てて検証しました。二重翻訳照合とは、2 件以上の独立した順翻訳を 1 本にまとめる工程のことです (Koller et al., 2012)。

この調査の主な目的は、従来式の手法で生成された照合結果に含まれる規格違反や逸脱を検出する生成 AI の能力を評価することでした。

プロンプト設計の進化

初期プロンプトは従来式の照合結果と並行して機能し、翻訳 A と翻訳 B を比較して、照合担当者の最終判断を合格/不合格の二値で示すように設計されました。

しかしこのアプローチには大きな制約があり、特に英語以外の言語や、専門性の高い状況 (特定の治療領域やよく QOL 評価の対象となる疾患・状態など) については、十分な対応が難しいことがわかりました。こうした知見をふまえて、当社では 生成 AI を利用して照合結果の評価を行い、QA (品質評価) ワークフローをサポートすることにしました。そして改良したプロンプトを、ライオンブリッジ独自の包括的な言語的検証プラットフォームである Aurora Clinical Outcomes に統合しました。生成 AI を利用して内部判断をサポートすることで、照合担当者の思考プロセスに関するインサイトも得られ、必要な基準や業界要件に沿って作業が行われたことを確認できます。そのような要件は照合目標 (GOALS OF RECONCILIATION) に反映されており、照合担当者に渡される照合作業の指示書にも、照合担当者向けのガイダンスとして明示されています。主な要件は次のとおりです。

1. **元の評価票との概念的等価性**
2. **文化的ローカリゼーション**
3. **対象の被験者層や読者にとってのわかりやすさ**
4. **バイアス傾向の検出**

手法

APAC (アジア太平洋地域) 言語、ロングテール (話者人口の少ない) 言語、および地域別変種を組み合わせた、1,000 ～ 2,000 語の PerfO (パフォーマンス アウトカム) サンプルを用いて、実践的分析を実施しました。上記の条件に基づいて、個別に実施できる柔軟な言語能力テスト システムである Clinical Evaluation of Language Fundamentals®-Fifth Edition (CELF-5) の、過去に完了したローカリゼーション プロジェクトから、さまざまなファイルを選定しました。このテストは一般に、小児や青少年の言語障害の正確な診断をサポートする目的で使用されます (NCS Pearson, Inc., 2013)。分析対象の言語としては、スペイン語 (アルゼンチン)、スペイン語 (スペイン)、フランス語 (フランス)、アルメニア語 (アルメニア)、日本語 (日本)、繁体字中国語 (台湾) が選択されました。

結果

良好な成果

言語専門家による生成 AI 出力の分析の結果、AI によるサポートが有効と考えられる 3 つの領域が明らかになりました。

- **不適切または不十分な理由付け:** 生成 AI は、言語的または概念的な根拠が不十分なコメント (「翻訳 A のほうがよい」、「翻訳 B が推奨される」など) を一貫して指摘したほか、未解決のクエリや未回答の質問も検出しました。それによってプロンプト改良のための推奨事項がまとめ上げられ、同様のケースはフォローアップ対象としてエスカレーションされるようになりました。
- **理由付けの欠如:** 専門家全員が、生成 AI は理由付けが欠如している箇所や理由付けの不十分・不明瞭な箇所を確実に検出し、QA レビューを大幅に効率化できることを確認しました。
- **効率性:** 生成 AI は処理速度に優れており、数秒で最大 300 のテキスト セグメントを分析できます。

この AI による事前スクリーニングを行うことで、言語専門家がそのレビュー結果をすばやく取捨選択し、ファイルを言語検証プロセスの次の工程に進められるかを判断できるようになります。

改善すべき点

生成 AI は大きな可能性を秘めているものの、一方で本質的に予測不能な性質を持つため、プロンプト実行時にはいくつかの制約も生じました。

- **一貫性のない動作:** 生成 AI は同一ファイル内の異なるセグメントに付された同じコメントを評価する場合でも、一貫しない挙動を示すことがあるため、生成 AI の出力だけに依存することは信頼性の面で問題があります。
- **文脈や他コメントへの参照に伴う課題:** 「See comment above (上記のコメントを参照)」のような曖昧なコメントや参照を多用すると、現状の仕組みでは生成 AI にとって大きな課題となります。生成 AI は参照されている前述のコメントを特定できないため、専門家たちは照合担当者に期待される作業方法として、簡潔で明確な参照を使ったり、複数のセグメントに繰り返し同じ説明を入力したりすることを推奨しています。
- **有効な照合コメントの統一基準の欠如:** 当社の専門家はみな、妥当とされる根拠やその詳細な説明について、より正確かつ明確に定義し、すべての関係者間で合意を行う必要があると指摘しています。

まとめ

この調査では、情報の欠落や規定違反、理由付けの不足などを検出するための、品質保証サポート ツールとしての生成 AI の大きな可能性が示されました。

生成 AI を利用することで、やり直し作業の必要性をより早く検知できるようになり、ローカリゼーション チームがより的を絞ったフィードバックを関係者に返して、将来のパフォーマンスを改善するためのインサイトを得られるようになります。ただし、レビュー結果を解釈して、微妙な判断の妥当性を検証し、プロジェクトの要件や規制要件の順守を確認するには、人間の専門知識が依然として不可欠になります。「人間参加型 (ヒューマンインザループ)」のアプローチ (生成 AI を活用した分析と専門家によるレビューの組み合わせ) を採ることで、全体的な品質が向上し、情報に基づくフィードバックによって照合担当者のパフォーマンスも上がるほか、継続的なプロセスの改善もサポートできます。

このアプローチでは最終的に、継続的な最適化と的を絞ったトレーニングにより、成果物の質を高め、COA のローカリゼーション プロジェクト全体のコスト効率を上げることができます。

引用元:
Koller, M. et al. (2012). The process of reconciliation: evaluation of guidelines for translating quality-of-life questionnaires. Expert Review of Pharmacoeconomics and Outcomes Research 12(2), 189–197.

生成 AI の事例:
AI を活用した比較レビュー出力の有効性調査

LIONBRIDGE | Pearson

プレゼンター (ライオンブリッジ): ステファニー カサーレ

目的

言語的検証 (LV) とは、対象地域全体で正確かつ一貫したデータ収集を行うために、臨床アウトカム評価 (COA) をローカライズし、それをレビューするプロセスのことです。このプロセスはあえて時間のかかる複雑な設計になっており、そのアプローチによって非常に高い品質と細部にまでわたる正確な翻訳が確保されますが、その複雑さには代償が伴います。このプロセスの金銭的・時間的な負担を軽減するため、この調査では納期を短縮し、アウトソーシングのコストを削減しつつ、このプロセスに求められる高い基準を維持することを目指しました。

このポスターでは、生成 AI を利用した比較レビュー (CR) の実行可能性について検討しています。比較レビューは LV プロセスにおける重要な品質保証の工程であり、原文テキストと逆翻訳テキストを比較して、概念的等価性を判断します。これは中間的な工程であり、その前後の工程はトレーニングを受けた経験豊富な言語担当者が実施するため、CR は自動化の有力候補となります。このアプローチによって、プロセスの完了時まで未検出のエラーが残るリスクを最小限に抑えることができます。

この研究は、現行の人間のサプライヤーによる比較レビューの品質を最低限維持できるプロンプトを開発することを目的として実施されました。

手法

まず、比較レビュー結果と比較レビュー コメントという、期待される出力を生成するプロンプトを開発しました。比較レビュー コメントは、比較レビュー結果に関連する詳細情報を示すものです。比較レビュー結果は、次の 3 つのカテゴリに分類されます。

同一 – 原文テキストと逆翻訳テキストが、大文字/小文字の使用や句読点などを含むあらゆる意味で、完全に同一であることを意味します。

同等 – 表現や文構造その他の詳細については相違もあるものの、各セグメントの意味合いは概念的に同等であることを意味します。読者はそれらが同じ情報を伝えていることを理解できます。

要レビュー – 2 つのセグメントに含まれる何らかの要素によって、それらのセグメントの内容が概念的に同等ではなくなっていることを意味します。読者は翻訳されたテキストを誤解し、原文テキストの意図とは異なる意味に受け取ってしまう可能性があります。

次にそのプロンプトを調整し、同一でない結果について比較レビューコメントを生成させるようにしました。このコメントは 2 つのセグメント間の概念的な相違を説明するもので、専門家でない読者による誤解の可能性についての詳しい説明も含まれます。次に、句読点や大文字小文字の違いは、意味や理解に直接関係する場合を除いて無視するよう指示し、さらに原文テキストの意味に関係のない補足的テキスト (書式タグなど) も無視するよう設定しました。

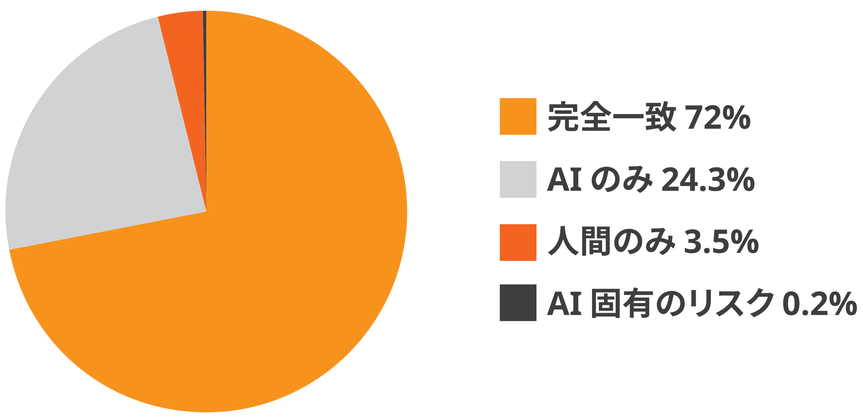
データを集めるにあたっては、以前にローカライズされた Pearson 社のアセスメント、Delis-Kaplan Executive Function System (D-KEFS、デリスカプラン実行機能検査システム) のトレイル メイキング テストを使用しました。これにはソース言語で約 1,000 単語が含まれており、逆翻訳は英語またはターゲット言語のネイティブ話者であるさまざまな翻訳者によって実施されました。そしてプロの比較レビュー担当者と、当社の COA ローカリゼーション チームのメンバーの両者が、複数回の比較レビューを実施しました。これらの担当者は比較レビューの経験レベルもネイティブ言語もさまざまであったため、幅広いレビュー結果を比較することができました。

結果

初期結果は有望なものであり、元の質問票と逆翻訳の相違が明確かつ簡潔に示され、全体的な暫定精度は 96.4% でした。

この結果の内訳を下の図に示します。人間による出力と AI の出力が完全に一致する部分は 72.09% であり、さらに 24.3% のセグメントが AI によって相違と判定されましたが、人間の比較レビュー担当者はそれを相違と判定していません。またこの結果には、人間のレビュー担当者が相違と判定したものの、AI は同等とみなした部分が 3.5% 含まれています。さらにまったく同一の判定のうち 0.17% については、順翻訳がラテン文字の言語ではなかったため、その文字種を考慮しないと見落とされる可能性が高いものとして、研究者が注意を促しています。

人間と AI による出力の一致率



注目すべき割合:
言語専門家による生成 AI 出力の分析の結果、生成 AI によるサポートが有効と考えられる 3 つの主な領域が明らかになりました。

AI 固有のリスク - 0.17%: この数値には、順翻訳にラテン文字以外の文字が使用されていたため、人間は検出できたものの、AI プロンプトでは見落とされた可能性のあるセグメントが含まれています。

一貫性のない判定 – 1.26%: レビューの過程で、AI は時折同一のセグメントに対して異なる出力を生成し、そうした出力が全データの 1% 強を占めました。

まとめ

生成 AI を利用することで、言語的検証プロセスにかかる時間と費用を大幅に節約できる可能性があります。

さらに調査を進めるにあたっては、データセットを拡張する必要があります。次のステップでは、概念実証 (POC) も実施して、比較レビュー照合の際にこの AI 出力を言語専門家の作業と組み合わせた場合の効果を検討することが求められます。またプロンプトをさらに改良することも、不一致やリスクの軽減に役立つ可能性があります。