Die Zukunft der linguistischen Validierung:
Compliance des Abstimmungsprozesses mit generativer KI verbessern

Präsentiert von Lionbridge: Elisabet Sas Olesa, Karolina Elizondo, Kathryn Nolte, Nathalie Azuaje, Melinda Johnson

EINLEITUNG

Diese Studie untersucht, wie generative KI den Prozess der linguistischen Validierung (LV) unterstützen kann. Schwerpunkt ist hierbei die Steigerung der Effizienz der Dual Translation Reconciliation, bei der zwei oder mehr unabhängige Vorwärtsübersetzungen in einer finalen Übersetzung zusammengefasst werden (Koller et al., 2012).

Unsere Untersuchung sollte in erster Linie die Fähigkeit generativer KI evaluieren, Complianceprobleme in den mit herkömmlichen Methoden erstellten Abstimmungen (Reconcilations) zu erkennen.

METHODEN

Wir haben eine praxisnahe Analyse anhand einer PerfO-Stichprobe (1000 – 2000 Wörter) mit einer Mischung aus APAC-Sprachen sowie Nischensprachen und regionalen Sprachvarianten durchgeführt. Dazu wurden verschiedene Dateien aus einer zuvor abgeschlossenen Lokalisierung der Clinical Evaluation of Language Fundamentals®-Fifth Edition (CELF-5), einem flexiblen System individuell administrierter Tests, ausgewählt. Diese Tests werden von Klinikern häufig zur exakten Diagnose von Sprachstörungen bei Kindern und Jugendlichen herangezogen (NCS Pearson, Inc., 2013).
Die für die Analyse ausgewählten Zielsprachen waren Spanisch (Argentinien), Spanisch (Spanien), Französisch (Frankreich), Armenisch (Armenien), Japanisch (Japan) und Traditionelles Chinesisch (Taiwan).

EVOLUTION DES PROMPT-AUFBAUS

Der ursprüngliche Prompt sollte die Ausgabe des herkömmlichen Abstimmungsprozesses durch Vergleich von Übersetzung A mit Übersetzung B verarbeiten und ein binäres Ergebnis (Übereinstimmung/ Abweichung) als Beurteilung der abschließenden Entscheidung der für die Abstimmung zuständigen Person (Reconciliator) generieren.

Diese Herangehensweise brachte jedoch erhebliche Einschränkungen mit sich, insbesondere beim Einsatz für andere Sprachen als Englisch und in speziellen Kontexten – beispielsweise definiert durch bestimmte Therapiebereiche, Krankheiten oder Gesundheitszustände, die typischerweise mit Verfahren zur Bewertung der Lebensqualität untersucht werden.
Diese Feststellungen veranlassten uns, generative KI zur Evaluierung der Reconcilationergebnisse einzusetzen, um den QS-Workflow zu verbessern.

Der überarbeitete Prompt wurde dann in die Lionbridge-Plattform *Aurora Clinical Outcomes* integriert, um eine durchgängige linguistische Validierung zu ermöglichen. Der Einsatz generativer KI zur Unterstützung der internen Entscheidungsfindung liefert zugleich Erkenntnisse zum Entscheidungsprozess des Reconcilators und kann sicherstellen, dass der Vergleich unter Berücksichtigung der einschlägigen Standards und Branchenanforderungen ausgeführt wurde. Diese Anforderungen sind in den GOALS OF RECONCILIATION aufgeführt und in den Leitlinien enthalten, die der Reconciliator mit den aufgabenbezogenen Anweisungen erhält:

1. **Konzeptionelle Äquivalenz zur ursprünglichen Messung**
2. **Kulturelle Anpassung**
3. **Zugänglichkeit für die vorgesehene Studienpopulation/Zielgruppe**
4. **Erkennung von Biastrends**

ERGEBNISSE

Positive Ergebnisse

Durch Analyse der Ausgaben der generativen KI konnten Sprachexperten drei zentrale Bereiche bestimmen, in denen sich KI als nützlich erweist:

- **Ungültige oder unvollständige Begründungen:** Generative KI identifiziert zuverlässig Kommentare, die keine ausreichende sprachliche oder konzeptionelle Begründung enthalten (z. B. „Übersetzung A ist besser“ oder „Übersetzung B ist zu bevorzugen“). Außerdem erkannte sie ungelöste Probleme und nicht beantwortete Fragen. Daraus resultierten Optimierungsempfehlungen für den Prompt, damit solche Befunde zur weiteren Bearbeitung eskaliert werden.
- **Fehlende Begründungen:** Alle Experten stellten fest, dass generative KI fehlende, unvollständige oder unklare Begründungen zuverlässig identifizierte und dadurch die Maßnahmen zur Qualitätssicherung erheblich beschleunigte.
- **Effizienz:** Generative KI zeichnet sich durch hohe Verarbeitungsgeschwindigkeit aus und analysiert bis zu 300 Textsegmente in Sekundenschnelle.

Dank dieser Vorabprüfung können Sprachexperten Ergebnisse schnell filtern und entscheiden, ob eine Datei für den nächsten Schritt der linguistischen Validierung (LV) bereit ist.

Verbesserungsmöglichkeiten

Ungeachtet des Potenzials zeigten sich bei Ausführung des Prompts gewisse Einschränkungen, die aus der inhärenten Unvorhersehbarkeit von Antworten generativer KI resultieren:

- **Inkonsistentes Verhalten:** Die Ausgaben der generativen KI sind für identische Kommentare in verschiedenen Segmenten inkonsistent. Das gilt sogar innerhalb einer Datei. Diese Inkonsistenz verhindert die Nutzung der KI-Ausgabe als eigenständige Lösung.
- **Kontextbezogene Herausforderungen und Probleme mit Metareferenzen:** Die häufige Verwendung vager Kommentare und Verweise wie „Siehe Kommentar oben“ stellt für generative KI derzeit eine große Herausforderung dar, da sie den Kommentar, auf den Bezug genommen wird, nicht identifizieren kann. Experten empfehlen, die Verwendung prägnanter und expliziter Verweise sowie die segmentübergreifende Wiederholung von Erklärungen in das Pflichtenheft des Reconciliators aufzunehmen.
- **Fehlende Standardisierung der Viable Reconciliation Comments:** Unsere Experten sind sich einig, dass es einer präziseren, expliziten Definition sowie einer Vereinbarung aller Parteien über akzeptable und detaillierte Begründungen und Kommentare bedarf.

FAZIT

Diese Studie unterstreicht das erhebliche Potenzial generativer KI als Tool zur Qualitätssicherung bei der Identifizierung von Lücken, Complianceproblemen und fehlenden Begründungen.

Generative KI beschleunigt die Feststellung von Nachbesserungsbedarf und versetzt das Lokalisierungsteam in die Lage, Stakeholder gezielter anzuleiten und sie mit Erkenntnissen auszustatten, die zur künftigen Leistungssteigerung beitragen können. Menschliche Expertise bleibt jedoch unverzichtbar, um Ergebnisse zu interpretieren, nuancierte Entscheidungen zu validieren und die Erfüllung der projektspezifischen und rechtlichen Anforderungen sicherzustellen. Eine Human-in-the-Loop-Herangehensweise, die Analysen der generativen KI mit Expertenprüfungen kombiniert, verbessert die Qualität insgesamt, steigert die Leistung des Reconciliators dank fundierter Rückmeldungen und unterstützt die kontinuierliche Prozessoptimierung.

Letztlich führt diese Herangehensweise mit kontinuierlicher Prompt-Optimierung und gezieltem Training für COA-Lokalisierungsprojekte zu hochwertigeren Ergebnissen und einer besseren Kosteneffizienz.

Quellen:
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.

Generative KI in Aktion:
Studie zur Wirksamkeit des Comparative Review mit KI

Präsentiert von Lionbridge: Stephanie Casale

ZIELE

Als linguistische Validierung (LV) wird das Lokalisieren von COA und deren Prüfung auf akkurate und konsistente Datenzusammenstellung über alle Zielsprachen hinweg bezeichnet. Das Verfahren ist inhärent langwierig und komplex. Diese Herangehensweise sorgt für Übersetzungen höchster Qualität und Genauigkeit, die Komplexität hat aber ihren Preis. Um den finanziellen und zeitlichen Aufwand dieses Prozesses zu verringern, untersucht unsere Studie Möglichkeiten, die Bearbeitungszeiten zu verkürzen und die Kosten für das Outsourcing zu reduzieren, ohne die erforderlichen hohen Prozessstandards zu beeinträchtigen.

Dieses Dokument untersucht, ob generative KI für Comparative Reviews (CR) geeignet ist. Der Comparative Review ist ein wichtiger Schritt der Qualitätssicherung im LV-Prozess. In diesem Schritt werden der Ausgangstext und der rückübersetzte Text verglichen, um die konzeptionelle Äquivalenz festzustellen. Da es sich um einen Zwischenschritt handelt und der vorhergehende sowie der folgende Schritt von geschulten und erfahrenen Linguisten durchgeführt werden, bietet sich der Comparative Review als Kandidat für die Automatisierung an. Diese Herangehensweise minimiert das Risiko, dass Fehler vor Projektabschluss nicht erkannt werden.

Ziel unserer Untersuchung war die Entwicklung eines Prompts, der mindestens die Qualität unserer derzeit für Comparative Reviews eingesetzten Experten liefert.

METHODEN

Zunächst haben wir einen Prompt entwickelt, der zu einem Comparative Review Result und einem Comparative Review Comment mit weiteren Details zum Comparative Review Result die erwartete Ausgabe lieferte. Die Comparative Review Results wurden in drei Kategorien unterteilt:

Identisch: Ausgangstext und Rückübersetzung sind einschließlich Klein-/Großschreibung und Zeichensetzung in jeder Hinsicht identisch.

Äquivalent: Es bestehen Unterschiede bei Wortwahl, Satzstruktur oder anderen Details, die Bedeutung der Segmente ist jedoch gleichwertig. Der Leser würde verstehen, dass der Informationsgehalt identisch ist.

Überprüfung erforderlich: Die beiden Segmente weisen konzeptionelle Unterschiede auf. Ein Leser könnte den übersetzten Text missverstehen und ihm eine Bedeutung beimessen, die im Ausgangstext nicht beabsichtigt war.

Der Prompt wurde dann so gestaltet, dass für jedes nicht identische Ergebnis ein Comparative Review Comment erstellt wird. Diese Kommentare erläutern konzeptionelle Unterschiede zwischen den beiden Segmenten und stellen mögliche Fehlinterpretationen durch Laien dar. Der Prompt sollte alle Unterschiede in Zeichensetzung und Groß-/Kleinschreibung ignorieren, sofern diese nicht unmittelbar für Bedeutung und Verständnis relevant waren. Außerdem sollte der Prompt zusätzlichen Text ignorieren, der für die Bedeutung des Ausgangstextes irrelevant war (z. B. Formattags).

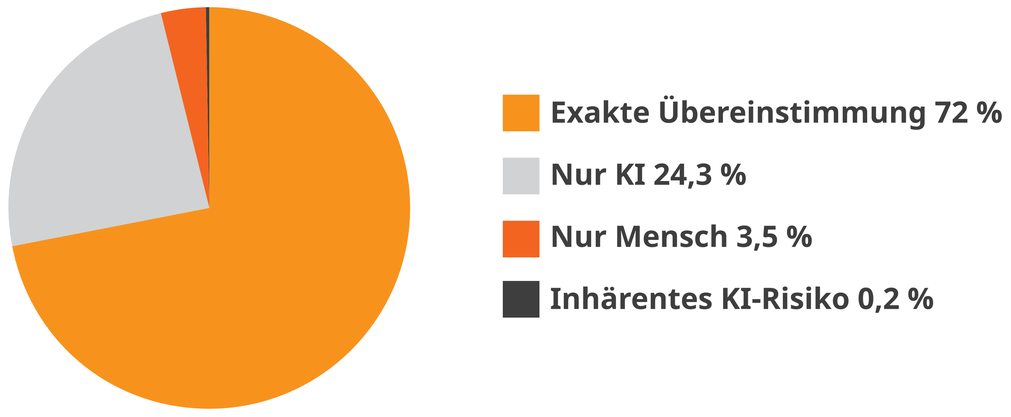
Als Daten haben wir eine zuvor lokalisierte Pearson-Bewertung (Delis-Kaplan Executive Function System (D-KEFS) – Trail Making Test) verwendet. Der Ausgangstext besteht aus etwa 1000 Wörtern. Die Rückübersetzungen wurden von verschiedenen Linguisten mit entweder Englisch oder der Zielsprache als Muttersprache angefertigt. Sowohl ein professioneller Comparative Reviewer als auch Mitglieder unseres COA-Lokalisierungsteams führten verschiedene Comparative Reviews durch. Der unterschiedliche Erfahrungsschatz in Bezug auf Comparative Reviews und die verschiedenen Muttersprachen der beteiligten Personen stellten ein breites Spektrum an Ausgaben für den Vergleich sicher.

ERGEBNISSE

Die ersten Ergebnisse sind vielversprechend, da generative KI klare und prägnante Beschreibungen der Abweichungen zwischen der ursprünglichen Bewertung und der Rückübersetzung mit einer vorläufigen Genauigkeitsrate von 96,4 % liefert.

Diese Aufschlüsselung ist in der folgenden Tabelle dargestellt: 72,09 % exakte Übereinstimmungen der Ergebnisse von Mensch und KI, weitere 24,3 % der Segmente wurden von der KI als Abweichungen gekennzeichnet, nicht jedoch vom Menschen. Weiterhin wurden 3,5 % der vom Menschen als Abweichungen identifizierten Fundstellen von der KI als äquivalent eingestuft. Schließlich gab es 0,17 % identische Antworten, was als riskant eingestuft wurde, weil die Vorwärtsübersetzungen nicht in lateinischen Buchstaben vorliegen und daher leicht übersehen werden können, sofern nicht speziell auf diesen Text geachtet wird.

PROZENTSATZ DER ÜBEREINSTIMMUNG VON KI UND MENSCH



Bemerkenswerte Prozentsätze:
Die Ausgaben der generativen KI wurden von unseren Sprachexperten analysiert, die drei Hauptbereiche für den effektiven Einsatz generativer KI ermitteln konnten:

Inhärentes KI-Risiko – 0,17 %: Diese Zahl umfasst Segmente, die von einem Menschen aufgrund des nicht-lateinischen Alphabets in der Vorwärtsübersetzung wahrscheinlich, vom KI-Prompt jedoch nicht identifiziert worden wären.

Inkonsistente Antworten – 1,26 %: Im Rahmen unserer Untersuchung gab die KI gelegentlich unterschiedliche Antworten für identische Segmente aus. Dies betraf jedoch nur etwas mehr als 1 % der Daten.

FAZIT

Generative KI kann potenziell erhebliche Zeit- und Kosteneinsparungen im Prozess der linguistischen Validierung erzielen.

Die Ergebnisse sollten mit weiteren Untersuchungen untermauert werden. So muss beispielsweise eine Machbarkeitsstudie durchgeführt werden, um die Auswirkungen der Nutzung dieser Ausgaben durch Linguisten im Rahmen der Comparative Review Reconciliation zu untersuchen. Einige Inkonsistenzen und Risiken lassen sich möglicherweise durch Optimierung des Prompts minimieren.