El futuro de la validación lingüística está aquí:
Optimización del cumplimiento del proceso de conciliación mediante la IA generativa

LIONBRIDGE | Pearson

Presentado por Lionbridge: Elisabet Sas Olesa; Karolina Elizondo; Kathryn Nolte; Nathalie Azuaje; Melinda Johnson

INTRODUCCIÓN

Este estudio analiza cómo la inteligencia artificial generativa (IA generativa) podría respaldar el proceso de validación lingüística (VL), con especial atención en su capacidad para optimizar la eficiencia de la fase de conciliación de traducciones dobles, la fase en la que dos o más traducciones directas independientes se fusionan en una sola (Koller et al., 2012).

El objetivo principal de nuestra investigación fue evaluar la capacidad de la IA generativa para detectar incumplimientos en las conciliaciones realizadas mediante los métodos convencionales.

MÉTODOS

Realizamos un análisis práctico utilizando una muestra de PerfO (entre 1000 y 2000 palabras), que incluía una combinación de variantes lingüísticas de la región Asia-Pacífico, así como idiomas minoritarios y regionales. Basándonos en estos parámetros, seleccionamos varios archivos de un proyecto de localización previamente completado de la «Clinical Evaluation of Language Fundamentals® - Fifth Edition» (CELF-5), un sistema flexible de pruebas administradas individualmente. Estas pruebas se emplean habitualmente para ayudar a los profesionales sanitarios a diagnosticar con precisión los trastornos del lenguaje en niños y adolescentes (NCS Pearson, Inc., 2013). Los idiomas seleccionados para el análisis fueron: español (Argentina), español (España), francés (Francia), armenio (Armenia), japonés (Japón) y chino tradicional (Taiwán).

EVOLUCIÓN DEL DISEÑO DE PROMPTS

El prompt inicial se diseñó para funcionar junto con el resultado tradicional de la conciliación, comparando la traducción A con la traducción B con el fin de generar un resultado binario de aprobado o no aprobado según la decisión final del conciliador.

Sin embargo, este enfoque presentaba importantes limitaciones, especialmente al aplicarlo a idiomas distintos del inglés y a contextos altamente especializados, como los definidos por un área terapéutica concreta o por una enfermedad o afección habitualmente evaluada en los instrumentos de calidad de vida. A la luz de estos hallazgos, aplicamos la IA generativa directamente a la evaluación del resultado de la conciliación, lo que reforzó nuestro flujo de trabajo de control de calidad. A continuación, el prompt actualizado se integró en *Aurora Clinical Outcomes*, la plataforma propiedad de Lionbridge para la validación lingüística integral. El uso de la IA generativa para respaldar la toma de decisiones internas también aporta visibilidad sobre el proceso de razonamiento del conciliador. Permite confirmar que la tarea se haya realizado conforme a los estándares requeridos y los requisitos del sector. Dichos requisitos se reflejan en los OBJETIVOS DE LA CONCILIACIÓN y se destacan en las orientaciones proporcionadas al conciliador en las instrucciones de la tarea:

1. **Equivalencia conceptual con la medida original**
2. **Adaptación cultural**
3. **Accesibilidad para la población o el público objetivo del estudio**
4. **Detección de tendencias de sesgo**

RESULTADOS

Resultados positivos

El análisis de los resultados de la IA generativa realizado por expertos lingüistas reveló tres áreas clave a las que esta contribuye:

- **Justificaciones inválidas o incompletas:** la IA generativa señaló sistemáticamente los comentarios que carecían de un fundamento lingüístico o conceptual suficiente (p. ej., «La traducción A es mejor» o «La traducción B es preferible»). También detectó consultas no resueltas o preguntas sin respuesta. Esto dio lugar a recomendaciones para perfeccionar el prompt, garantizando así que estos casos se remitiesen para su seguimiento.
- **Ausencia de justificaciones:** todos los expertos señalaron que la IA generativa identificaba de forma fiable los casos en que faltaba una justificación, estaba incompleta o era poco clara, lo que aceleraba significativamente la revisión de control de calidad.
- **Eficiencia:** la IA generativa demuestra una alta velocidad de procesamiento, con capacidad para analizar hasta 300 segmentos de texto en cuestión de segundos.

Esta preselección permite a los expertos lingüísticos filtrar rápidamente los resultados y decidir si un archivo está listo para pasar a la siguiente fase del proceso de traducción.

Áreas de mejora

A pesar de su potencial, la naturaleza intrínsecamente impredecible de la IA generativa también introdujo ciertas limitaciones durante la ejecución del prompt:

- **Comportamiento incoherente:** los resultados de la IA generativa presentan un comportamiento variable al evaluar comentarios idénticos en diferentes segmentos, incluso dentro del mismo archivo. Esto socava la fiabilidad de los resultados de la IA generativa como solución autónoma.
- **Desafíos contextuales o de metarreferencia:** el uso frecuente de comentarios vagos y referencias como «Véase el comentario anterior» supone un serio desafío para la IA generativa en la configuración actual, ya que no es capaz de localizar el comentario anterior al que se hace referencia. Los expertos sugieren utilizar referencias concisas y explícitas, así como explicaciones repetidas en todos los segmentos, como parte de las entregas previstas del conciliador.
- **Falta de estandarización de los comentarios de conciliación viables:** nuestros expertos coinciden en que es necesario establecer una definición más precisa y explícita, así como un consenso entre todas las partes sobre lo que constituye un razonamiento aceptable y una justificación detallada.

CONCLUSIÓN

Este estudio demuestra el importante potencial de la IA generativa como herramienta de apoyo al control de calidad para identificar deficiencias, incumplimientos y justificaciones ausentes.

La IA generativa permite identificar con mayor rapidez la necesidad de repetir trabajos y ayuda al equipo de localización a ofrecer comentarios más precisos a las partes interesadas, ofreciéndoles la información necesaria para mejorar su rendimiento en el futuro. No obstante, la experiencia humana sigue siendo indispensable para interpretar los resultados, validar decisiones repletas de matices y garantizar la conformidad con los requisitos del proyecto y la normativa vigente. Un enfoque basado en la intervención humana —que combine el análisis basado en IA generativa con la revisión de expertos— eleva la calidad general, optimiza el rendimiento del conciliador mediante comentarios informados y favorece la mejora continua del proceso.

En última instancia, con la optimización continua de los prompts y una formación específica, este enfoque puede ofrecer resultados de mayor calidad y una mayor rentabilidad en todos los proyectos de localización de COA.

Citas:
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.

La IA generativa en acción:
estudio sobre la eficacia de los resultados de la revisión comparativa por IA

Presentado por Lionbridge: Stephanie Casale

OBJETIVOS

La validación lingüística (VL) es el proceso mediante el cual se localizan y revisan las evaluaciones de resultados clínicos para garantizar una recopilación de datos precisa y coherente en todas las variedades regionales de destino. El proceso es, por su propio diseño, largo y complejo. Este enfoque garantiza traducciones de la más alta calidad y precisión, pero dicha complejidad tiene un coste. Con el objetivo de reducir la carga económica y temporal de este proceso, nuestro estudio pretende acortar los plazos de entrega y los costes de externalización, manteniendo al mismo tiempo los altos estándares que requiere el proceso.

Este póster analiza la viabilidad de utilizar la IA generativa para realizar revisiones comparativas (RC). La revisión comparativa es un paso clave para garantizar la calidad en el proceso de traducción, ya que compara el texto original con la traducción inversa para determinar la equivalencia conceptual. Al tratarse de un paso intermedio, los pasos previos y posteriores los llevan a cabo lingüistas capacitados y con experiencia, lo que convierte a la RC en una candidata idónea para la automatización. Este enfoque minimiza el riesgo de que se produzcan errores no detectados antes de la finalización.

Nuestra investigación tenía como objetivo desarrollar un prompt que mantuviera, como mínimo, la calidad existente de nuestros proveedores humanos para la revisión comparativa.

MÉTODOS

Primero desarrollamos un prompt que produjera el resultado esperado de un resultado de revisión comparativa y un comentario de revisión comparativa, que proporcionaba más detalles relacionados con el resultado. Los resultados de la revisión comparativa se dividieron en tres categorías:

Idénticos: este resultado indica que el texto original y la traducción inversa eran exactamente iguales en todos los aspectos, incluidas las mayúsculas y la puntuación.

Equivalentes: este resultado indica que, aunque pueda haber diferencias en la redacción, la estructura de las frases u otros detalles, el significado de los segmentos sigue siendo conceptualmente equivalente. El lector entendería que transmiten la misma información.

Necesita revisión: este resultado indica que algún elemento de los dos segmentos los hace conceptualmente no equivalentes. Un lector podría malinterpretar el texto traducido, creyendo que transmite algo distinto de lo que pretendía el texto original.

A continuación, se diseñó el prompt para generar un comentario de revisión comparativa para cualquier resultado no idéntico. Estos comentarios explican las diferencias conceptuales entre los dos segmentos, incluida una explicación detallada de las posibles interpretaciones erróneas por parte de un lector no experto. Se indicó al prompt que ignorara cualquier diferencia de puntuación y mayúsculas, a menos que estuvieran directamente relacionadas con el significado y la comprensión. El prompt también debía ignorar cualquier texto adicional que no estuviera relacionado con el significado del texto original (como, por ejemplo, etiquetas de formato).

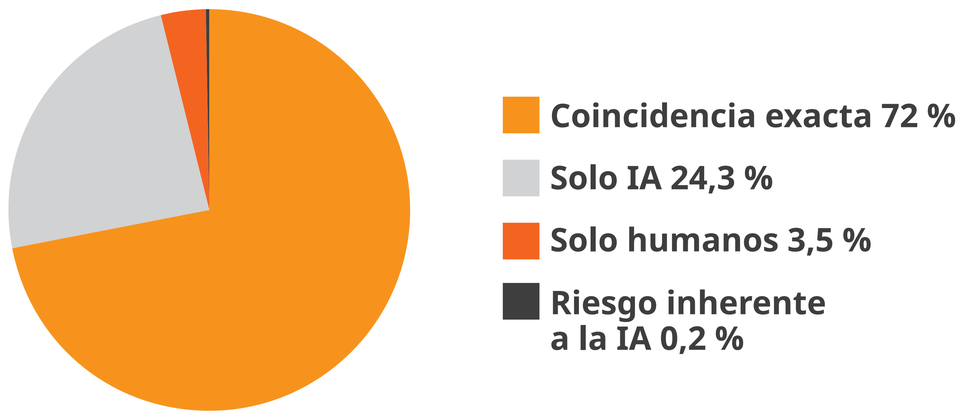
Para recopilar los datos, utilizamos una evaluación de Pearson previamente localizada, la prueba Trail Making Test (TMT) del Delis-Kaplan Executive Function System (D-KEFS), que contiene aproximadamente 1000 palabras de origen. Las traducciones inversas las realizaron diversos lingüistas nativos, tanto de inglés como del idioma de destino. Un revisor comparativo profesional y los miembros de nuestro equipo de localización de COA llevaron a cabo múltiples revisiones comparativas. Presentaban distintos niveles de experiencia en revisión comparativa y hablaban diferentes lenguas maternas, lo que garantizó una amplia variedad de resultados para comparar.

RESULTADOS

Los resultados iniciales son prometedores, con descripciones claras y concisas de las discrepancias entre la evaluación original y la traducción inversa, con una tasa preliminar de precisión global del 96,4 %.

Este desglose se puede ver en el siguiente gráfico. Incluye un 72,09 % de coincidencias exactas entre los resultados humanos y los de la IA, con un 24,3 % adicional de segmentos marcados como discrepancias por la IA, pero no por el revisor humano comparativo. Los resultados también incluyeron un 3,5 % de discrepancias detectadas por humanos que la IA consideró equivalentes. Los resultados incluyen asimismo un 0,17 % de respuestas idénticas, que los investigadores consideraron arriesgadas debido a que la traducción directa no estaba en caracteres latinos, por lo que podían pasarse por alto si no se tenía en cuenta ese texto.

PORCENTAJE DE COINCIDENCIA DE LA IA CON EL RESULTADO HUMANO



Porcentajes destacados:
Los resultados de la IA generativa analizados por nuestros expertos lingüistas revelaron tres áreas principales en las que la IA generativa puede ser de gran ayuda:

Riesgo inherente a la IA, 0,17 %: esta cifra incluía segmentos que podrían haber sido señalados por un revisor humano debido al uso de un alfabeto no latino en la traducción directa, pero que no habrían sido detectados por el prompt de la IA.

Respuestas incoherentes, 1,26 %: durante nuestra revisión, la IA generó ocasionalmente respuestas diferentes para el mismo conjunto de segmentos, lo que representó algo más del 1 % del total de los datos.

CONCLUSIÓN

La IA generativa tiene el potencial de ahorrar una cantidad significativa de tiempo y dinero en el proceso de validación lingüística.

Los estudios futuros deberían ampliar el conjunto de datos. Los próximos pasos deberán incluir una prueba de concepto para evaluar cómo afecta el uso de estos resultados al trabajo de los lingüistas durante la fase de revisión comparativa y conciliación. Una optimización adicional del prompt podría ayudar a reducir ciertas incoherencias y minimizar riesgos.