FRANCISCO JAVIER CRISTÓFOL RODRÍGUEZ
ALEXANDRA SANDULESCU BUDEA
ROBERTO MORENO LÓPEZ
CAROLINA PORRAS FLORIDO
ÁNGELA MARTÍN GUTIÉRREZ
(coords.)

INNOVACIÓN Y TRANSFORMACIÓN DIGITAL EN LA SOCIEDAD DEL SIGLO XXI

ARCO/LIBROS

© Arco/Libros-La Muralla, 2026
Edita: Arco/Libros-La Muralla (Barcelona)
www.arcomuralla.com
ISBN: 978-84-7133-990-4
Depósito legal: B-9.005-2026.
Diseño, realización y coordinación: Arco/Libros-La Muralla
Calle L'Equador, 39-45, Plt. Bj., Local 6, 08029 Barcelona (Imagine Content, S. L.)

© FÓRUM XXI, 2026
Editor: David Caldevilla Domínguez
Primera edición, 2026, Madrid (España)

Impreso en España-*Printed in Spain*

23

DESAFÍOS DE LA TRADUCCIÓN AUTOMÁTICA: LA IMPORTANCIA DE LA INTERVENCIÓN HUMANA EN LA ERA DE LA INTELIGENCIA ARTIFICIAL

Blanca Hernández Pardo
Universidad Pontificia Comillas (España)

1. Introducción y objetivo

La creciente demanda de la traducción automática ha dado lugar a una plétora de investigaciones y avances en el campo del procesamiento del lenguaje natural (PLN). Sin embargo, a pesar de las mejoras tecnológicas en la traducción automática (TA), sigue habiendo varios retos que deben abordarse, sobre todo en lo que se refiere a la calidad de las traducciones generadas por los sistemas automatizados.

El proceso de TA se enfrenta a varios retos, principalmente en cuanto a la capacidad de captar con precisión el tono, el estilo y la estructura del texto original. Esto puede ser especialmente difícil en la lengua materna de cada uno, donde la traducción automática tiende a fallar a la hora de transferir información más allá de la mera traducción de palabras. Por ello, la TA requiere una mayor intervención humana para garantizar la calidad y la precisión de las traducciones generadas (Romana García y Hernández Pardo, 2022).

En este artículo, pretendemos abordar estos retos y explorar tanto la calidad de varios motores de traducción automática de textos especializados como la importancia que posee en la actualidad la intervención humana para corregir y mejorar la calidad de estas traducciones. Para ello, se tomó un fragmento de texto científico en inglés que se pasó por una serie de motores de TA para comparar la calidad de los resultados. A continuación, se analizaron los textos generados con métricas de evaluación automática y se interpretaron los datos obtenidos.

El objetivo del presente estudio consiste en demostrar la importancia del trabajo humano en la traducción automática, así como destacar la necesidad de que exista colaboración entre investigadores y traductores profesionales para mejorar la calidad de este tipo de traducciones (Romana García y Hernández Pardo, 2022).

2. Panorama de la traducción automática

Los orígenes de la traducción automática (TA) se remontan a 1933, con el diccionario mecánico multilingüe de Georges Artsrouni, basado en tarjetas perforadas, y la propuesta de mecanización en tres fases de Petr Petrovich Trojanskij (Tertoolen, 2012). Más adelante, Warren Weaver y Andrew Booth informatizaron estas ideas, explorando el potencial de los ordenadores para analizar estructuras lingüísticas y establecer reglas de traducción a partir de grandes volúmenes de datos (Booth, 1958).

Durante los años posteriores, diversos centros de investigación estadounidenses se enfocaron en el desarrollo de la TA, y en 1954, la Universidad de Georgetown e IBM presentaron el primer prototipo público. Aunque era limitado, este hito despertó gran interés y supuso el inicio de una era de financiación masiva para proyectos de TA en EE. UU. (Hutchins, 2004).

Este periodo marcó también el desarrollo de los primeros modelos estadísticos de traducción automática (TAM), que trataban la traducción como un problema probabilístico. Mediante el uso de corpus bilingües, los modelos aprendían correspondencias entre palabras y frases (Casacuberta y Peris, 2017, p. 68). Sin embargo, estos sistemas eran poco precisos y no captaban la complejidad del lenguaje natural debido a la simplicidad de sus algoritmos y a la escasez de datos representativos. A pesar de ello, sentaron las bases de los enfoques actuales.

Con todo el aprendizaje anterior, hoy en día predomina la traducción neuronal (TAN), basada en redes neuronales profundas que representan las palabras como vectores numéricos (Bengio *et al.*, 2003; Casacuberta y Peris, 2017, p. 68) y modelos de atención como los *transformers*, lo que les permite captar de manera más satisfactoria las relaciones semánticas y gramaticales del texto, lo cual es muy eficaz para tareas de PLN (Hugging Face (s.f.a; Vaswani... *et al.*, 2017, p. 1). Estos modelos, especialmente desde la década de 2010, han superado a los anteriores en fluidez y precisión, gracias a su capacidad de aprendizaje contextual (Yang *et al.*, 2020, p. 1; Tan *et al.*, 2021).

Aunque la TAN ha demostrado ser una herramienta útil, sigue siendo imprescindible la intervención del traductor profesional, sobre todo en ámbitos especializados y según el uso final del texto (Casacuberta y Peris, 2017, pp. 71–72). A pesar de que algunos estudios, como el de Google (2016), apuntan a que Google Translate puede igualar al traductor humano en ciertos textos estructuralmente simples (Hasyim *et al.*, 2021, p. 186), la TA sigue mostrando deficiencias en la traducción de contenidos con carga cultural, expresiones idiomáticas o estilización literaria, donde la traducción humana continúa siendo insustituible.

En consecuencia, el panorama actual de la traducción automática está dominado por modelos de procesamiento del lenguaje natural (PLN), que hoy supone el eje central tanto en investigación como en aplicaciones prácticas. Estos modelos, basados en aprendizaje profundo y grandes volúmenes de datos etiquetados, entrenan redes neuronales que identifican patrones lingüísticos complejos y generan traducciones más precisas. Sin embargo, muchas de estas producciones se siguen considerando «traducciones tolerantes a errores» (TTE), es decir, textos

que no garantizan una calidad lingüística satisfactoria (Gil Sanromán *et al.*, 2021, p. 680), lo que limita su adecuación en contextos especializados o con fines comunicativos exigentes.

3. Metodología

El presente estudio pretende evaluar la calidad de las traducciones automáticas generadas por distintos motores de traducción y compararlas con el trabajo manual realizado por un traductor experto. Para ello, se seleccionó un pequeño fragmento de un texto científico en inglés y se pasó por dos motores de traducción automática gratuitos en línea: Google Translate y Reverso. A continuación, se analizaron los textos resultantes para determinar si habían pasado por alto errores importantes, que un traductor humano y experto tanto en el campo traductológico como en el de la propia especialidad del texto nunca habría cometido.

Una vez detectados los principales errores, se procedió a generar una versión correcta denominada «referencia», que es la que se emplearía para la segunda parte de la investigación, esto es, la extracción automática de métricas de evaluación. En esta segunda parte, se analizó y comparó la calidad y el rendimiento de varios traductores automáticos mediante métricas de evaluación y código de programación. Para ello, se empleó un cuaderno de Google Colaboratory y lenguaje Python.

La ventaja de este procedimiento es que supone una alternativa barata, reutilizable y replicable para automatizar el proceso de evaluación, si bien es importante tener en cuenta que no en todos los casos resultará fiable.

Asimismo, lo ideal de este proceso automatizado es que las métricas empleadas actúen con rapidez, que posean un tamaño que no requiera más memoria de la que podríamos disponer en el ordenador, que sean integrables en los flujos de trabajo y, por último, que sean personalizables, de modo que se puedan adaptar las necesidades puntuales del usuario.

Para llevar a cabo la evaluación automática de una traducción, siempre resulta esencial contar con un texto de origen y, por otro lado, con una o varias traducciones «humanas» de referencia (Romana García y Hernández Pardo, 2024). De este modo, será posible calcular la similitud entre las traducciones automáticas con respecto a la humana.

Las métricas de trabajo que se han empleado en la presente investigación, todas ellas ofrecidas por la librería *Evaluate* de HugginFace (s.f.b), incluyen las siguientes:

Métrica	Descripción
Word Error Rate (WER)	Mide el número mínimo de transformaciones que son necesarias para modificar la traducción generada en la de referencia («distancia de Levenshtein») (Haldar y Mukhopadhyay, 2011, p. 1). Se calcula dividiendo el número total de inserciones, eliminaciones y sustituciones requeridas para hacer coincidir la transcripción automática con la referencia, por el número total de palabras en la referencia (Koehn, 2010, pp. 224-225).
Bilingual Evaluation Understudy (BLEU)	Evalúa la precisión de la traducción automática mediante la comparación de n-gramas (secuencias de palabras contiguas) entre la traducción automática y las referencias humanas; además, se aplican penalizaciones por brevedad y no se analiza palabra por palabra, sino por corpus entero (Koehn, 2010, p. 226; Papineni *et al.*, 2002, p. 312).
Metric for Evaluation of Translation with Explicit Ordering (METEOR)	Utiliza una combinación de medidas de precisión (proporción de palabras en la traducción automática que coinciden con las palabras en la referencia humana), cobertura (proporción de palabras en la referencia humana que coinciden con las palabras en la traducción automática) y correspondencia (medida en la que las palabras en la traducción automática y las palabras en la referencia humana son similares entre sí) para evaluar la calidad de una traducción automática en comparación con una referencia humana (Koehn, 2010, p. 228; Lavie y Agarwal, 2007, p. 68).
Bidirectional Encoder Representations from Transformers (BERT) score	Entiende el significado de una palabra en función de su contexto en una oración, lo que ayuda a generar traducciones más precisas y coherentes. El rendimiento de esta métrica se evalúa mediante los valores *precision* (cuántas de las clasificaciones positivas del modelo son realmente positivas), *recall* (cuántas de las instancias positivas reales son identificadas correctamente por el modelo) y *f1-score* (la media de equilibrio entre las dos anteriores) para cada una de las frases de la traducción. A mayor puntuación F1, se supone una mayor calidad de la traducción (Zhang *et al.*, 2020, pp. 3-4).
Bilingual Evaluation Understudy with Representations from Transformers (BLEURT)	Muy parecida a la métrica BLEU, pero en este caso, en lugar de basarse en la coincidencia de n-gramas, BLEURT utiliza modelos de lenguaje avanzados para evaluar la calidad de las traducciones automáticas en un nivel más profundo, considerando la semántica y la coherencia del texto. Esto hace que BLEURT sea potencialmente más preciso y relevante para la evaluación de traducción automática (Sellam *et al.*, 2020, p. 2; Sellam y Parikh, 2020; Sravani y Mamidi, 2023, p. 214).

Tabla 1. Resumen de métricas empleadas en la evaluación de traducciones automáticas.
Fuente: Elaboración propia, 2024.

4. Análisis comparativo de traducciones automáticas

En primer lugar, el texto seleccionado con fines demostrativos del presente análisis comparativo, perteneciente al ámbito de especialidad médico-sanitaria (subcampo de los ensayos clínicos en el ámbito farmacéutico), se tradujo con dos motores de traducción automática, a disposición de todos los usuarios en línea de forma gratuita:

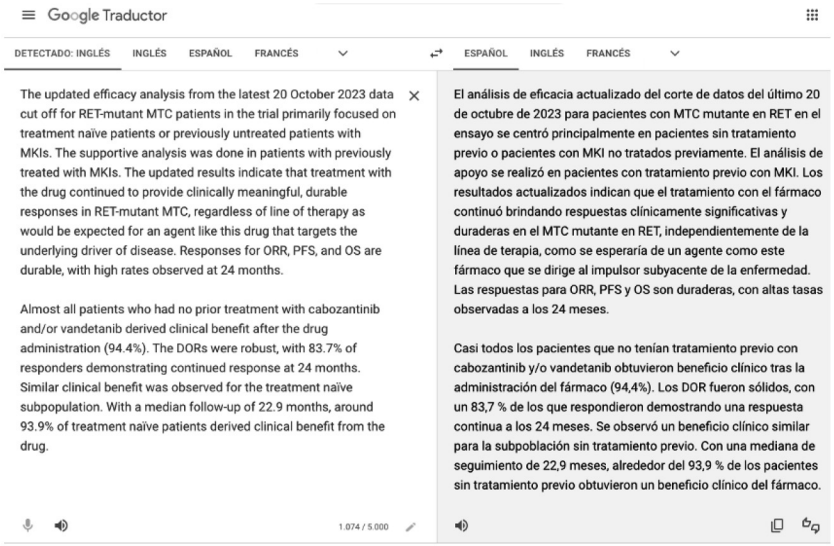

Figura 1. Fragmento de ejemplo original en inglés traducido en la plataforma Google Translate. Fuente: Elaboración propia, 2024.

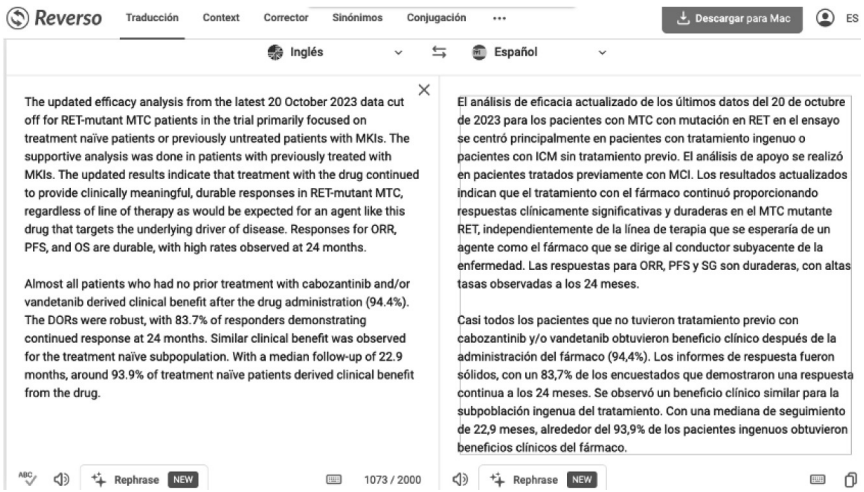

Figura 2. Fragmento de ejemplo original en inglés traducido en la plataforma Reverso. Fuente: Elaboración propia, 2024.

Para este estudio, se eligieron las herramientas de traducción automática («TA», de ahora en adelante) Google Translate («Traductor automático/Traducción automática 1» o «T1») y Reverso («Traductor automático 2» o «T2»). Si bien ambas traducciones ofrecen un resultado bastante mejor del esperado para

un TA, teniendo en cuenta que nos enfrentamos a una traducción especializada, en los dos casos nos encontramos con importantes y graves errores, claramente no permisibles en un contexto profesional.

Según lo comentado sobre el concepto de «traducción tolerante a errores» (Gil Sanromán *et al.*, 2021, p. 680), textos como los creados con ambas herramientas de TA no valdrían para su entrega y uso posterior por parte del destinatario. Ello se debe a que forman parte de documentos sumamente relevantes y delicados desde el punto de vista conceptual para el tratamiento de fármacos a nivel europeo, por lo que las traducciones tienen que gozar de una precisión semántica y una equivalencia funcional que, hasta ahora, solo el traductor humano es capaz de ofrecer.

A continuación, se comentan los principales errores detectados con ambas propuestas, recogidos en la Tabla 1:

#	Fragmento original	T1: Google Translate	T2: Reverso
1	*The updated efficacy analysis from the latest 20 October 2023 data cut off*	El análisis de eficacia actualizado del corte de datos del último 20 de octubre de 2023	El análisis de eficacia actualizado de los últimos datos del 20 de octubre de 2023
2	*RET-mutant MTC patients*	pacientes con MTC mutante en RET	pacientes con MTC con mutación en RET
3	*primarily focused on treatment naïve patients or previously untreated patients with MKIs*	se centró principalmente en pacientes sin tratamiento previo o pacientes con MKI no tratados previamente	se centró principalmente en pacientes con tratamiento ingenuo o pacientes con ICM sin tratamiento previo
4	*Almost all patients who had no prior treatment with cabozantinib*	Casi todos los pacientes que no tenían tratamiento previo con cabozantinib	Casi todos los pacientes que no tuvieron tratamiento previo con cabozantinib
5	*The DORs were robust, with 83.7% of responders demonstrating continued response at 24 months*	Los DOR fueron sólidos, con un 83,7 % de los que respondieron demostrando una respuesta continua a los 24 meses	Los informes de respuesta fueron sólidos, con un 83,7% de los encuestados que demostraron una respuesta continua a los 24 meses

Tabla 2. Comparación de fragmentos con errores de los traductores automáticos 1 y 2.
Fuente: Elaboración propia, 2024.

En el primer fragmento destacado, observamos que en ambas traducciones se genera un falso sentido, pues a lo que se refiere el original es a los datos de un ensayo clínico determinado recogidos hasta la fecha de corte específica: «El análisis de eficacia actualizado a partir de la última fecha de corte de los datos del 20 de octubre de 2023».

En el segundo caso, el T2 ofrece una traducción válida del término, no así el T1, que genera un importante error semántico con su propuesta, puesto que lo que viene a decir dicha traducción es que el MTC es mutante, es decir, que produce mutaciones, lo cual es falso (en realidad, el MTC es mutante, puesto que

le ha afectado una mutación por un agente «mutante»). Además, en ambos casos, se observa que los TA no son capaces de identificar la forma desarrollada que les corresponde a las siglas del fragmento, problema que un traductor humano no habría tenido, pues por el propio término y la alusión a «mutación en RET», ya sabría que se refiere al *medullary thyroid cancer*, que en español se denomina «cáncer medular de tiroides» (TP del MedDRA). Si realizamos búsquedas en recursos especializados para lingüistas sanitarios, así como en la literatura médica, observaremos que en español no se habla tanto de «MTC» (siglas en inglés) como de «CMT», las siglas debidamente correspondientes al español.

El tercer fragmento, de cierta dificultad incluso para un traductor humano que no posea experiencia en la traducción de documentación relacionada con ensayos clínicos de medicamentos, ha generado bastantes problemas a los traductores automáticos. En la propuesta del T1, por la reformulación en español, lo que se entiende es que incluía a pacientes que no habían recibido previamente ningún tratamiento y a pacientes que recibían/habían recibido MKI y que no habían sido tratados previamente (error traductológico: sinsentido o bien falso sentido, si lo que se entiende al final es que no lo habían recibido de otro fármaco distinto que no se menciona en el fragmento –cosa que no es–). En la propuesta del T2 nos encontramos con el mismo problema, con la adición de dos errores graves más de tipo léxico: «con tratamiento ingenuo», que claramente se trata de un falso amigo, pues en inglés los *naïve patients* son aquellos que no han tenido experiencia previa con uno o varios fármacos. Por otro lado, observamos que, de nuevo, se genera un problema de siglas: traduce *MKI* por «ICM», y no se ha conseguido entender el procedimiento que la máquina ha llevado hasta llegar a dicha conclusión (se desconoce el significado de «ICM»). Sin embargo, por el contexto, cualquier traductor experimentado en esta área de especialidad entendería sin ningún problema que se refiere a los inhibidores multiquinasa. Además, tras una búsqueda documental de textos paralelos y literatura en español, se observa que no existe una sigla acuñada para el término en español (apenas se utiliza el equivalente «IMQ / IMK»), por lo que prevalece en los textos en castellano la sigla en inglés (MKI), tal y como ha dejado el T1. Una propuesta de traducción que subsanaría los problemas anteriormente comentados podría ser: «en pacientes que no habían recibido ningún tratamiento previo o en aquellos no tratados anteriormente con MKI».

En el cuarto caso no destacamos un error binario, sino más bien de estilo, por falta de fluidez y naturalidad tanto en español como en la jerga del lenguaje de especialidad que aquí nos compete. Un especialista quizá se habría referido a estos pacientes como aquellos que «no habían recibido anteriormente/previamente cabozantinib», en lugar de calcar la estructura del original, que no conlleva un problema sintáctico en español, pero sí resulta ciertamente artificial.

Por último, en el quinto fragmento, de nuevo nos encontramos con graves problemas de comprensión de la lengua de origen en el caso de las siglas por parte de los TA empleados. En estos contextos en los que se habla del beneficio clínico de un fármaco, existe multitud de criterios de valoración que actúan como parámetros para poder evaluar el nivel de dicha tasa. Para ello se emplean, entre

otros, los que en inglés se mencionan como DOR, ORR, PFS y OS (los últimos tres no se comentarán por limitaciones de espacio, pero las estrategias utilizadas por los motores difieren y, en algunos casos, son similares a las comentadas más arriba con respecto a otras siglas). El traductor especializado en este tipo de textos conoce sin problemas el concepto de «DOR» en lo que respecta al beneficio clínico de un fármaco y, con únicamente estos dos párrafos como contexto, sabe que se refiere a la *duration of response*, por lo que ninguno de los dos traductores automáticos ha acertado con este término: El T1 lo mantuvo en inglés (cosa que ocurre en gran parte de nuestra literatura para esta sigla, aunque en este caso también se ven algunas propuestas acuñadas en español, tales como «DdR»), pero considerándolo un término de género femenino, por lo que el lector difícilmente lo vinculará con el criterio de valoración de duración de la respuesta. Más grave es lo que ocurre con el T2, que lo ha traducido como «informes de respuesta», por lo que se genera un error semántico importante, pues genera ambigüedad frente al sentido original del TO.

5. Análisis de métricas con programación

Mediante programación en Python, el uso de un cuaderno colaborativo en línea (Google Colaboratory) y la librería *Evaluate* de Hugging Face (s.f.b), procedimos a extraer datos sobre diversos parámetros de calidad a partir de cada una de las traducciones automáticas generadas, todas ellas con respecto a una humana de referencia.

Para conseguir una tabla que resultara más rica en datos comparativos, se decidió incluir en el análisis otras dos traducciones automáticas realizadas mediante programación: se trata de los modelos de traducción Helsinki EN-ES (Helsinki-NLP/opus-mt-en-es) y Facebook Multilingüe (facebook/m2m100_418M), ambos de Hugging Face. El primero es un modelo capaz de traducir únicamente de inglés a castellano (de ahora en adelante «Traduccion automática 3» o «T3»], mientras que el segundo es capaz de traducir entre múltiples idiomas (en adelante «Traduccion automática 4» o «T4»].

	T1: GoogleTrans	T2: Reverso	T3: Helsinki	T4: Facebook
WER	0.4205	0.4626	0,4579	0,5514
BLEU	0.3962	0.4107	0,4159	0,2670
METEOR	0.6520	0.6519	0,6560	0,5433
BERT Score	0.9176 (0.02)	0.9142 (0.03)	0,9136 (0,03)	0,8931 (0,03)
BLEURT	0.3194 (0.13)	0.2851 (0.13)	0,2906 (0,16)	0,2161 (0,15)

Figura 3. Resultados de métricas de evaluación sobre cuatro traducciones automáticas.
Fuente: Elaboración propia, 2024.

Como podemos observar, en lo que respecta a la métrica WER, el TA que mejor nota obtuvo con esta métrica fue Google Translate (0,4205). En contraposición, el modelo Facebook multilingüe obtuvo la nota más alta (0,5514), lo que significa una mayor tasa de error de palabras.

En segundo lugar, con la métrica BLEU, el que mejores resultados numéricos ofrece es el tercero, esto es, el modelo de TA Helsinki EN-ES, con una puntuación de 0,4159 (ligeramente superior a la de Reverso [0,4107] y ciertamente mayor a las puntuaciones de los otros dos traductores automáticos). Significa que esta traducción es la más precisa.

En tercer lugar, con METEOR, se obtuvieron los mejores resultados con Helsinki EN-ES (0,6560), aunque por una diferencia bastante pequeña con respecto a Google Translate (0,6520) y Reverso (0,6519), lo que demuestra, una vez más, que esta traducción automática es más precisa y fluida. Facebook multilingüe, por su parte, mostró la peor nota (0,5433).

Con BERT Score, la TA de Google Translate (0,9176) obtuvo la puntuación más alta seguido por muy poco del modelo Helsinki EN-ES (0,9136), lo que significa que es la traducción más similar a la de referencia humana. En contraposición, de nuevo es Facebook multilingüe el que peores resultados arrojó (0,8931).

Por último, BLEURT demostró que es Google Translate el que arroja un valor más alto (0,1394), lo que significa que es su traducción más precisa y fluida (como ya se demostró con otras métricas anteriores). Una vez más, en última posición estaría Facebook multilingüe, con un valor global para esta métrica de 0,2161.

Cabe destacar que las dos últimas métricas ofrecen datos adicionales entre paréntesis, que corresponden con la desviación estándar. En este caso, indica la variabilidad de las puntuaciones de BERT Score y BLEURT entre diferentes oraciones o segmentos de texto. Aquellos casos con desviación estándar alta implica que las puntuaciones son más variables. Por ello, observamos por ejemplo que con la métrica BLEURT, la desviación estándar de la TA de Google Translate es de 0,13 frente a la del modelo Helsinki EN-ES (0,16).

En cuanto a sus puntuaciones globales, Google Translate supera al modelo Helsinki EN-ES. Esto significa que, en general, las traducciones del primer TA son más precisas y fluidas. Además, Google Translate tiene una menor desviación estándar, así que las puntuaciones de BLEURT para este traductor automático demuestran ser más consistentes, lo que significa que muestra una calidad más uniforme de sus traducciones.

Tal y como observamos, no hay un único traductor automático que proporcione los mejores resultados para todas las métricas empleadas. Google Translate tiene el mejor rendimiento en tres de las métricas (WER, BERT Score y BLEURT). Lo que sí está claro es que Facebook multilingüe es el que peor puntuación ha obtenido en todas las métricas.

En consecuencia, la identificación del «mejor TA» dependerá de las necesidades específicas: Si lo que se requiere es una menor tasa de error de palabras o bien una traducción más consistente en términos de calidad, Google Translate demuestra ser una buena opción. Sin embargo, si lo que se espera es una mayor

fluidez y precisión general, entonces colocaríamos al modelo de traducción automática Helsinki EN-ES en cabeza.

Resultaría de gran interés añadir otras métricas de evaluación, e incluso llevar a cabo una comparación estadística de los datos arrojados por todas las métricas automáticas con respecto a la traducción de referencia realizada por un humano experto en el área traductológica y del ámbito de especialidad.

6. Conclusiones

Lo anteriores resultados indican que, aunque los motores de traducción automática pueden generar traducciones comprensibles y que transmiten hasta cierto punto el sentido general del texto original, sigue habiendo un gran número de errores e imprecisiones en la traducción automática de cierta importancia, lo que puede provocar errores en la interpretación del texto. Además, nuestra investigación ha demostrado que la intervención humana es esencial para mejorar la calidad de las traducciones automáticas.

Los errores inherentes a la traducción automática poseen una particular relevancia en el ámbito científico-sanitario, dada la trascendencia de las traducciones generadas en este contexto. La precisión y la fiabilidad de dichas producciones revisten una importancia crucial, puesto que sirven como base para generar publicaciones científicas e informes técnicos, cuyos contenidos fundamentan avances terapéuticos significativos. Además, la idoneidad y la exactitud de las traducciones adquieren una dimensión aún más crítica al considerar las implicaciones regulatorias y legales que conllevan. Una traducción defectuosa, imprecisa, errónea o sesgada podría conllevar graves consecuencias al comprometer la integridad de los procedimientos regulatorios y legales relacionados con la aprobación y el uso de medicamentos.

La terminología médico-sanitaria constituye un lenguaje especializado en constante evolución impulsado por los continuos avances en investigación y desarrollo. Esta dinámica obliga a que los motores de traducción automática, basados en modelos de aprendizaje profundo entrenados con grandes corpus, se actualicen y reentrenen casi a diario para mantener su eficacia.

En este sentido, las autoridades sanitarias dependen en gran medida de la exactitud y la fiabilidad de las traducciones a la hora de evaluar solicitudes regulatorias. La calidad de estas traducciones es un factor determinante para garantizar la seguridad y la eficacia de los productos farmacéuticos, así como para salvaguardar los derechos y la salud de los pacientes. En consecuencia, si los textos generados con motores de traducción automática en el ámbito científico-sanitario no alcanzan un estándar de calidad satisfactorio, sin duda se requiere la participación del profesional humano para corregir o trabajar directamente sobre dichos textos y así garantizar la integridad y la precisión de las traducciones.

Los sistemas actuales de evaluación automática de traducciones suponen una ayuda como punto de partida para determinar la calidad de los resultados obte-

nidos por modelos de aprendizaje profundos, pero requieren una interpretación cautelosa de los datos generados.

En conclusión, nuestro estudio apoya la importancia de la intervención humana en la traducción automática y subraya la necesidad de una mayor colaboración entre investigadores y traductores profesionales para mejorar la calidad de las traducciones automáticas (Romana García y Hernández Pardo, 2022). Además, nuestros resultados sugieren que la traducción automática no sustituye a la necesidad de revisión y corrección manual por parte de un traductor experto, pero puede servir como herramienta de apoyo útil para los traductores profesionales.

Referencias

Bengio, Y., Ducharme, R., Vincent, P. y Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137-1155.

Booth, A. D. (1958). The history and recent progress of machine translation. En A. D. Booth, Leonard Forster, D. J. Furley, R. Glement, Joseph Neddham, C. Rabin, L. W. Tancock (eds.), *Aspects of translation (Studies in Communication* 2*)* (pp. 88-104). Secker & Warburg.

Haldar, R. y Mukhopadhyay, D. (2011). *Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach*. *ArXiv, abs/1101.1232*.

Hutchins, W. J. (2004). The Georgetown-IBM Experiment Demonstrated in January 1954. En: R.E. Frederking, K.B. Taylor (eds.), *Machine Translation: From Real Users to Research*. AMTA 2004. Lecture Notes in Computer Science, vol. 3265. Springer. https://doi.org/10.1007/978-3-540-30194-3_12

Koehn P. (2010). *Statistical machine translation*. Cambridge University Press.

Lavie, A. y Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgements. En *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 228-231). Association for Computational Linguistics. https://aclanthology.org/W07-0734.pdf

Gil Sanromán, Í., Hernández Pardo, B., Martín Matas, P. y Romana García, M. L. (2021). Innovación docente «Trujanews»: Cibercompetencias de gestión multilingüe. En S. Liberal Ormaechea y J. Sierra Sánchez (coords.). *Retos y desafíos de la innovación educativa en la era post COVID-19* (pp. 679-698). McGraw-Hill.

Hasyim, M., Saleh, F., Yusuf, R. y Abbas, A. (2021). Artificial Intelligence: Machine Translation Accuracy in Translating French-Indonesian Culinary Texts. *International Journal of Advanced Computer Science and Applications,* 12 (3), 186-191. https://doi.org/10.14569/IJACSA.2021.0120323

Hugging Face. (s.f.a). How Transformers solve tasks. *Transformers Documentation: Tasks explained*. https://huggingface.co/docs/transformers/tasks_explained

Hugging Face. (s.f.b). *Hugging Face Evaluate Library Documentation*. https://huggingface.co/docs/evaluate/index

MedDRA. (s.f.). *Medical Dictionary for Regulatory Activities*. https://www.meddra.org/

Papineni, K., Roukos, S., Ward, T. y Zhu, W. J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. En *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). https://doi.org/10.3115/1073083.1073135

ROMANA GARCÍA, M. L. y HERNÁNDEZ PARDO, B. (2024). Characterization of Institutional Texts for an Automated Golden Standard: Enhancing Machine Translation Quality Assessment between English and Spanish. En C. Orasan, T. Ranasinghe, G. Corpas Pastor, R. Mitkov, M. Kunilovskaya (eds.), *Proceedings of the Conference New Trends in Translation and Technology – NeTTT 2024* (pp. 138-155). Varna, Bulgaria.

ROMANA GARCÍA, M. L. y HERNÁNDEZ PARDO, B. (2022). Inteligencia artificial y gestión corporativa / Artificial Intelligence and Corporate Management. En R. C. López González (coords.), *Universalidad y multiversalidad en literatura, lengua y traducción* (pp. 241-252). Comares. ISBN: 978-84-1369-437-5.

SELLAM, T., DAS, D. y PARIKH, A. (2020). BLEURT: Learning Robust Metrics for Text Generation. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.704

SELLAM, T. y PARIKH, A. P. (26 de mayo de 2020). Evaluating Natural Language Generation with BLEURT. *Google Al Blog: Evaluating Natural Language Generation with BLEURT*. https://go.uv.es/pAZDm3a

SRAVANI, D. y MAMIDI, R. (2023). Enhancing Code-mixed Text Generation Using Synthetic Data Filtering in Neural Machine Translation. En *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 211-220).

TAN, Z., WANG, S., YANG, Z., CHEN, G., HUANG, X., SUN, M. y LIU, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open,* 1, 5-21. https://doi.org/10.1016/j.aiopen.2020.11.001

TERTOOLEN, R. C. (2012). *Desarrollos En La Traducción Automática: Esperando Aún Una Traducción de Alta Calidad,* 141-147. https://hdl.handle.net/10366/124880

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAIER, L. y POLOSUKHIN, I. (2017) Attention Is All You Need. Advances in Neural Information Processing Systems 30: *Annual Conference on Neural Information Processing Systems 2017, Long Beach, 4-9 December 2017*, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762

YANG, S., WANG, Y. y CHU, X. (2020). A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526.*

ZHANG, T., VARSHA, K., FELIX, W., KILIAN, Q.W. y YOAV A. (2020). BERT score: evaluating text generation with BERT. En *Proceedings of the International Conference on Learning Representations*. *arXiv preprint arXiv:1904.09675.*