ISSN 1870-4069

Determinación automática del grado de dominio de una lengua extranjera usando modelos de lenguaje

Carlos Andrés Martínez González, David Pinto, Darnes Vilariño Ayala, Beatriz Beltrán Martínez, Yolanda Moyao Martínez

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

mg223470496@alm.buap.mx,{david.pinto,
darnes.vilarino,beatriz.beltran,
yolanda.moyao}@correo.buap.mx

Resumen. El artículo propone un sistema automatizado para evaluar el dominio del inglés como lengua extranjera utilizando modelos de lenguaje avanzados y técnicas de Procesamiento de Lenguaje Natural (PLN). El sistema integra el modelo GPT-J para la aplicación de pruebas y combina herramientas como Whisper [11] y SpeechRecognition [28] para evaluar respuestas orales, utilizando corpus como LibriSpeech (audio) [20] y el Open American National Corpus (texto) [10]. Las métricas empleadas incluyen F1-score [15], Perplexity [12], ROUGE [16] , BLEU [21] , WER [26] y PER [9] para medir aspectos como comprensión, gramática, vocabulario y pronunciación, alineándose con el Marco Común Europeo de Referencia para las Lenguas (MCER) [6]. Los resultados muestran que GPT-J tiene un rendimiento óptimo en comparación con otros modelos como GPT-Neo y GPT-3, aunque se identifican áreas de mejora en la generación de texto y la precisión fonética. El sistema demuestra ser capaz de evaluar de manera autónoma, reduciendo la necesidad de intervención humana. Como trabajo futuro, se sugiere incorporar interfaces 3D para una interacción más natural.

Palabras Clave: Language models, automated english proficiency assessment, GPT-J, natural language processing, automated evaluation.

Automatic Determination of the Degree of Mastery of a Foreign Language Using Language Models

Abstract. This article proposes an automated system for assessing English as a foreign language (EL) proficiency using advanced language models and Natural Language Processing (NLP) techniques. The system integrates the GPT-J model for testing and combines tools such as Whisper [11] and SpeechRecognition [28] to evaluate oral responses, using corpora such as LibriSpeech (audio) [20] and the Open American National Corpus (text) [10]. Metrics used include F1-score [15], Perplexity [12], ROUGE [16], BLEU [21], WER [26], and PER [9] to measure comprehension, grammar, vocabulary, and pronunciation, aligning with the Common European Framework of Reference for Languages (CEFR) [6]. The results show that GPT-J performs optimally compared to other models such as GPT-Neo and GPT-3, although areas for improvement are identified in text

generation and phonetic accuracy. The system proves capable of evaluating autonomously, reducing the need for human intervention. As future work, we suggest incorporating 3D interfaces for more natural interaction.

Keywords: Language models, automated English proficiency assessment, GPT- J, natural language processing, automated evaluation.

1. Introducción

Con el avance diario de la tecnología aplicada al campo computacional, se ha revolucionado la calidad de vida del ser humano, especialmente en áreas como la educación y la comunicación. En este contexto, los modelos de lenguaje han demostrado ser herramientas fundamentales para evaluar y mejorar la competencia lingüística en diversos idiomas. Este trabajo de investigación se centra en generar un modelo de aprendizaje automático que permita determinar de manera precisa el grado de dominio lingüístico en el idioma inglés como lengua extranjera. Se propone utilizar modelos de lenguaje basados en Procesamiento de Lenguaje Natural (PLN), para evaluar aspectos clave del dominio del idioma inglés, tomando en cuenta que el hablante no es nativo. El sistema evalúa competencias en comprensión, expresión oral y escrita, gramática, vocabulario y otros factores relevantes que determinan el nivel de competencia en inglés. Para ello, se utiliza un modelo GPT-J encargado de aplicar las pruebas, en conjunto con un submodelo interno que determina el nivel de dominio del idioma conforme a los parámetros establecidos por el Marco Común Europeo de Referencia para las Lenguas (MCER) [6].

Por lo tanto, el sistema se implementa en una plataforma web, permitiendo la interacción fluida entre el usuario y el sistema de evaluación del grado de dominio del idioma inglés. Esta interfaz utiliza tanto texto como audio, permitiendo medir las competencias lingüísticas de los usuarios, lo cual es clave para el aprendizaje efectivo del idioma, obteniendo datos reales para refinar el algoritmo basado en los resultados obtenidos. Estas pruebas permiten ajustar el modelo de acuerdo con las necesidades de los usuarios y asegurar una tasa de precisión adecuada para su implementación a mayor escala.

A fin de garantizar una evaluación precisa y confiable, se emplean métricas como F1-score [15], Perplexity [12], ROUGE [16], BLEU [21], WER [26] y PER [9]; además de un corpus representativo de inglés estadounidense proveniente del Open National American Corpus [10] y LibriSpeech [20], con el propósito de seleccionar los datos más adecuados para el entrenamiento y validación del modelo propuesto.

Una de las principales aportaciones de este trabajo es la propuesta de utilizar modelos de lenguaje de gran escala, como GPT-J, no únicamente como generadores de texto, sino como aplicadores automatizados de evaluaciones del idioma inglés. Este enfoque permite explorar nuevas aplicaciones educativas de los modelos de lenguaje, diferenciándose de los usos convencionales centrados en la generación libre de contenidos.

Además, este estudio no solo busca desarrollar un sistema de evaluación automatizada, sino también contribuir al campo del procesamiento del lenguaje natural (PLN) mediante una metodología que permita evaluar de manera efectiva la adquisición

de una segunda lengua mediante modelos avanzados de Inteligencia Artificial (IA). A continuación, se presentan los trabajos relacionados con el tema de investigación.

2. Trabajos relacionados

Dentro del estudio de la IA, diversas investigaciones han hecho hincapié que su aplicación no únicamente se restrinja al desarrollo del dominio de una lengua extranjera; sino también en entornos virtuales donde los modelos de lenguaje puede ser una herramienta adicional al progreso humano, creando así diversas experiencias inmersivas donde la realidad da un paso más allá de la convivencia humana - computacional.

Con relación a la influencia de los modelos de lenguaje para determinar el grado de dominio de un idioma, existe una investigación que destaca cómo los modelos de lenguaje están revolucionando el aprendizaje del inglés al facilitar la creación de chatbots conversacionales que interactúan con los estudiantes mediante respuestas coherentes, mejorando así la práctica del idioma. Además, estos modelos permiten adaptar el contenido educativo a las necesidades individuales de cada estudiante, ofreciendo una experiencia de aprendizaje personalizada [3].

En un artículo de la Universidad de Florida [4] sobre aplicaciones de la IA, se exploran diversas formas en que la IA se integra en modelos virtuales para evaluar y mejorar la enseñanza hacia los estudiantes. Se concluye que estos modelos son herramientas de apoyo docente efectivas, basadas en más de 20 horas de pruebas de rendimiento y funcionalidad.

Este mismo tema fue abordado por Villaroel [7], quien señala cómo el aula tradicional ha evolucionado con la inclusión de tecnologías como la IA y técnicas de análisis de información como Big Data. Estas herramientas permiten personalizar las rutas de aprendizaje de los estudiantes y optimizar la gestión educativa. Además, la importancia de las plataformas colaborativas como GitHub agilizan tareas rutinarias y ofrecen propuestas predictivas, así como la integración de la gamificación con IA, la cual mejora las habilidades cognitivas y fomenta un aprendizaje significativo y actualizado con relación a las necesidades de la actualidad.

Por otra parte, en un artículo de la Universidad del Sur de Florida [24], se explora la combinación de la Realidad Aumentada (AR) con Voicebots y modelos de lenguaje como ChatGPT para la enseñanza de lenguas extranjeras a niños pequeños a través de un entorno agradable e interactivo. Además, el uso de la AR facilita una experiencia visual interactiva que atrae la atención de los niños, mientras que los Voicebots y modelos de lenguaje como ChatGPT permiten conversaciones personalizadas que refuerzan la inmersión lingüística. Esta combinación, ayuda a que el aprendizaje de una lengua extranjera sea más efectivo y entretenido para los niños, permitiendo una mayor retención de conocimiento a través de la interacción constante.

Otra forma de evaluar el dominio de un idioma se aborda en la investigación de Muñoz y Fuertes [19], que explora cómo la IA puede ser aplicada en el ámbito educativo; en particular, destacan que la interacción entre humanos y máquinas puede ser una herramienta valiosa en el aprendizaje de una segunda lengua. Esto incluye aspectos como el aprendizaje informal, la autonomía del estudiante y la auto-evaluación.

De acuerdo con el trabajo realizado en conjunto por Guano y otros investigadores [8] sobre el modelo de aprendizaje del idioma inglés utilizando algoritmos de machine learning, se analiza cómo estos algoritmos abarcan diversas áreas, como las ciencias psicológicas, sociales, lingüísticas y pedagógicas. Aunque estos algoritmos pueden ser herramientas valiosas para personalizar el aprendizaje, evaluar habilidades y optimizar recursos educativos, también presentan riesgos, especialmente en la evaluación del dominio de una lengua extranjera. Entre estos riesgos se incluyen problemas de precisión en la evaluación, sesgo en los modelos y dependencia excesiva de la tecnología, lo que puede afectar tanto la calidad del aprendizaje como la equidad en la evaluación.

Dentro de los parámetros de evaluación sobre modelos de lenguaje, se encuentra un artículo que examina en profundidad los aspectos clave de la evaluación, centrándose en las preguntas: ¿qué evaluar?, ¿cómo evaluar? y ¿dónde evaluar? [1]. El artículo utiliza diversos estudios de rendimiento, protocolos y tareas para analizar no solo los avances realizados en los modelos de lenguaje, sino también para identificar áreas clave de mejora, estos se centran en métricas clave como Perplexity y BLEU [21], señalando áreas específicas para mejorar la precisión y efectividad de estos modelos. Además, destaca la importancia de la evaluación en términos de aplicabilidad, ética y equidad, subrayando los desafíos que enfrentan estos modelos en cuanto a sesgo, generalización y manejo de tareas complejas.

En el trabajo desarrollado por Kasneci [13], los autores exploran cómo los modelos de lenguaje grandes pueden ser utilizados de manera positiva en el ámbito educativo; destacan que, además de determinar el grado de dominio de una lengua extranjera, estos modelos pueden apoyar diversos aspectos de la educación en distintas materias. Sin embargo, también abordan desafíos asociados con su inclusión, como la necesidad de proporcionar retroalimentación instantánea y facilitar el acceso a recursos educativos, así como problemas relacionados con la precisión de las respuestas, por lo cual el uso de IA en educación plantea preocupaciones sobre la dependencia excesiva y la privacidad.

En un proyecto desarrollado sobre la corrección de errores en sistemas de reconocimiento del habla (ASR por sus siglas en inglés) [17], se puede mejorar significativamente los resultados del reconocimiento de voz, especialmente cuando se utilizan modelos generativos como ChatGPT en configuraciones zero-shot o few-shot. Este enfoque, basado en las salidas N-best, ha mostrado ser eficaz incluso con arquitecturas avanzadas como transducers y modelos con atención. Modelos como T5 (Text-to-Text Transfer Transformer), que convierten cualquier tarea de procesamiento de lenguaje en un problema de generación de texto, se destacan por su versatilidad, haciendo que sean efectivos no solo para la corrección de errores ASR, sino también en diversas aplicaciones de PLN.

En el artículo de Millière [18] se exploran las implicaciones de los modelos de lenguaje modernos, como los de la familia GPT, para la lingüística teórica. A pesar de que estos modelos están diseñados con objetivos de ingeniería, su capacidad para adquirir conocimiento lingüístico complejo a partir de grandes cantidades de datos plantea la necesidad de reconsiderar su relevancia en el ámbito de la teoría lingüística. Esta información presenta evidencia empírica que sugiere que los modelos de lenguaje pueden aprender estructuras sintácticas jerárquicas y responder a fenómenos lingüísticos, aportando una colaboración más cercana entre lingüistas y científicos

computacionales podría ofrecer valiosas perspectivas, especialmente en debates sobre el nativismo lingüístico.

En otra investigación donde se propone determinar si los modelos de lenguaje extensos (LLMs por sus siglas en inglés) pueden reemplazar evaluaciones humanas [2], realizaron experimentos en tareas de clasificación de texto, calificación automática de ensayos y análisis de contenido. Como resultado, los LLMs demostraron ser consistentes y eficientes, sin embargo, aún necesitan supervisión humana para tareas complejas; lo que puede ser empleado en evaluaciones iniciales o como complemento en entornos educativos.

En un proyecto propuesto en el área de estudio [25], se desarrolló un marco de evaluación para modelos grandes de lenguaje aplicados al audio (LLM-A por sus siglas en inglés), a través de proponer un conjunto de tareas de evaluación en diferentes modalidades: clasificación, generación, transcripción y síntesis. Incluye datasets preexistentes y adaptados para estas tareas. Este proyecto identificó las fortalezas y limitaciones de varios LLM-A existentes, destacando la necesidad de mejorar la coherencia en la generación y la calidad de la transcripción. Lo que puede favorecer el entender cómo integrar modelos de lenguaje para dominios específicos como audio, ofreciendo ideas reveladoras sobre posibles extensiones para el modelo de lenguaje aplicado al inglés.

Recientemente, se creó el proyecto AudioPaLM [22], en el cual se desarrolló un modelo de lenguaje multimodal que entiende y genera texto y audio con alta calidad. Este modelo de lenguaje combina métodos de preentrenamiento textual y de audio en un único modelo, utilizando datasets de alta calidad en ambas modalidades, también incluye aprendizaje de transferencia para tareas específicas como síntesis y traducción multimodal. Sus resultados mostraron rendimiento competitivo en tareas de síntesis y traducción, que igualan o superan a modelos especializados en audio.

Otro proyecto relevante en el ámbito de la evaluación auditiva por parte de modelos de lenguaje es SpeechVerse [5], el cual acorde a los autores, su objetivo es diseñar un modelo escalable para tareas de lenguaje y audio con énfasis en generalización. Para ello, el modelo fue entrenado en un corpus masivo, que incluye múltiples lenguajes y modalidades, también se usaron técnicas como Fine-Tuning progresivo y enmascaramiento predictivo. Este modelo de lenguaje especializado en audio destacó en tareas de generación y comprensión en diversos lenguajes, superando modelos previos en métricas como BLEU y F1, lo que indica que su enfoque puede ser relevante para evaluar cómo los modelos de lenguaje podrían manejar idiomas y sus variantes, especialmente entre dialectos o regiones.

En la investigación desarrollada por Yang [27], se presentó AIR-Bench, una plataforma de evaluación diseñada específicamente para medir la comprensión generativa en modelos de audio, este sistema introduce tareas como la comprensión narrativa del audio y la generación de respuestas, utilizando datasets anotados manualmente. Los resultados evidenciaron que los modelos actuales son efectivos en tareas de comprensión básica, pero enfrentan desafíos en escenarios más complejos o especializados. Además, en futuras aplicaciones, AIR-Bench podría ofrecer métricas y tareas adaptables para evaluar componentes específicos de modelos de lenguaje aplicados al inglés, especialmente en contextos de evaluación avanzada. A continuación, se presenta la metodología propuesta en la presente investigación.

3. Diseño de la investigación

Para llevar a cabo la determinación del grado de dominio de una lengua extranjera usando modelos de lenguaje, se utilizó una metodología propuesta para ejecutar con éxito su desarrollo e implementación. Tal metodología se concentra en los puntos siguientes:

- Realizar un estudio de las metodologías propuestas hasta el momento, revisando que se ha hecho, cómo se ha hecho y qué resultados se tienen.
- Comparar los modelos de lenguaje para determinar cuál se utilizará en este proyecto de investigación.
- Investigar la tecnología necesaria para llevar a cabo el desarrollo del proyecto.
- Obtener un corpus para obtener el modelo de lenguaje a usar en el proceso de determinación del grado de dominio.
- Entrenar el modelo de lenguaje usando GPUs.
- Realizar las pruebas para determinar la eficiencia del modelo del lenguaje utilizando métricas adecuadas, por ejemplo, la medida F1 para la precisión del modelo o BLEU entre otros para la precisión de traducción de texto.
- Realizar los ajustes necesarios para mejora del modelo.
- Obtención del modelo ajustado.
- Evaluación del modelo generado.

3.1. Técnicas de PLN: Medidas empleadas

Divergencia de Kullback-Leibler

La divergencia de Kullback-Leibler [14] mide la diferencia entre dos distribuciones de probabilidad: distribución real o esperada y distribución aproximada o predicción.

Es fundamental en tareas que comparan distribuciones, como el análisis de texto; dentro del flujo del programa, esta divergencia evalúa la similitud del vocabulario del estudiante en relación con el de los hablantes nativos.

Similitud de coseno

Ésta medida se basa en la comparación de la dirección de los términos en un espacio matemático, sin importar su tamaño o frecuencia absoluta; en concreto, se calcula la similitud entre dos vectores basándose en el ángulo entre ellos, lo cual es ampliamente usado para comparar documentos o frases en términos de similitud semántica [23].

Distribución de Zipf

En la distribución de Zipf [29], se empleó la medida para conocer el flujo de palabras y su aparición en el texto comparado debido a que es un principio estadístico que describe cómo ciertas palabras en un idioma aparecen con mayor frecuencia que otras,

siguiendo un patrón predecible; lo que facilita comprender la distribución léxica en un corpus. Con relación a las técnicas especializadas para el desarrollo del proyecto, se comprende que existen métricas de evaluación del rendimiento del modelo de lenguaje, a continuación, se describen las métricas que fueron empleadas.

Métrica F1 Score

Esta métrica es adecuada para el proyecto [15], ya que representa el promedio armónico entre la precisión (proporción de predicciones correctas sobre todas las predicciones positivas) y la exhaustividad (proporción de aciertos sobre los datos positivos reales). Su uso es particularmente útil en escenarios donde existe un desbalance entre las clases, permitiendo evaluar el rendimiento del modelo de manera equilibrada.

Métrica Perplexity

Esta medida [12], es utilizada para cuantificar la incertidumbre asociada con una distribución de probabilidad, en los campos de la estadística, análisis de los datos y la ciencia de datos; la entropía cruzada calcula la incertidumbre promedio en las predicciones del modelo. Valores bajos de perplexity indican un mejor ajuste del modelo al texto, lo que permite medir la calidad de generación del modelo y la habilidad de construcción de frases en inglés.

Métrica ROUGE

Esta métrica permite evaluar textos no necesariamente grandes [16], comparado con la métrica BLEU, lo que permite una mayor flexibilidad en la calidad de evaluación si se visualiza desde el ámbito de la conversación y el análisis de frases cortas. En particular, se compara n-gramas generados por el modelo con los del texto de referencia, evaluando la similitud; lo cual es ampliamente usado en tareas de resumen y generación de texto, midiendo precisión, recall y F1 a través de diferentes niveles de análisis.

Métrica BLEU

La métrica BLEU (Bilingual Evaluation Understudy) [21] es un método cuantitativo ampliamente utilizado para evaluar la calidad de traducciones automáticas y sistemas de generación de lenguaje natural. Esta métrica combina dos componentes principales: la precisión modificada de n-gramas (que evita premiar repeticiones excesivas) y un factor de penalización por brevedad (para castigar traducciones demasiado cortas). El resultado, es un puntaje normalizado entre 0 y 1, donde valores cercanos a 1 indican mayor concordancia con las referencias humanas.

Métrica WER

El Word Error Rate (WER), o Tasa de Error de Palabras [26], es una métrica ampliamente utilizada en la evaluación de sistemas de reconocimiento de voz, esta

medida cuantifica la precisión de un sistema al comparar la transcripción generada automáticamente con una transcripción de referencia considerada como correcta.

Un WER más bajo indica un mayor nivel de precisión en el reconocimiento de palabras, siendo esta métrica fundamental para evaluar y comparar el rendimiento de sistemas de reconocimiento de voz.

Métrica PER

El Phoneme Error Rate (PER), o Tasa de Error de Fonemas, es una métrica que evalúa la precisión en el reconocimiento de unidades fonéticas [9]. A diferencia del WER, el PER se centra en la comparación entre la transcripción fonética generada por un sistema y una transcripción fonética de referencia. Esta métrica es particularmente útil en tareas relacionadas con el análisis de la pronunciación y la evaluación de sistemas de reconocimiento del habla a nivel fonético; por lo tanto, la métrica PER es una herramienta valiosa para analizar la capacidad de un sistema de reconocer y transcribir correctamente los sonidos del habla, lo que resulta esencial en aplicaciones como la corrección de la pronunciación o la síntesis de voz.

Toda esta información fue de utilidad para la construcción del modelo de lenguaje final. En el siguiente apartado se describen los resultados obtenidos utilizando todos los componentes antes descritos, así como los conjuntos de datos que fueron empleados para las pruebas de eficiencia.

4. Evaluación de resultados

Se utilizaron volúmenes de datos provenientes del Open National American Corpus y de LibriSpeech para hacer una correcta comparación de datos de referencia en contraste con la información obtenida del usuario en la prueba y verificar aspectos como la ortografía, la pronunciación, entre otros factores, con el fin de catalogarlos correctamente; a continuación, se describe a detalle cada corpus utilizado.

4.1. Conjunto de datos

El primer corpus de información proveniente de LibriSpeech se enfoca en los fonemas obtenidos al convertir audio a texto; para ello, se utilizaron herramientas avanzadas de captura de audio, las cuales permitieron comparar y verificar si el usuario pronunció correctamente en inglés, éste corpus se compuso de 100 horas de audio libre de ruido en temas de uso cotidiano.

El segundo corpus, por su parte, se centró en la comprensión de textos, así como en la ortografía, la gramática y el vocabulario. Para este análisis, se tomaron 7 GB de ejemplos del *Open National American Corpus* sobre diversos temas desde el aspecto técnico hasta de tipo cotidiano, con el objetivo de evaluar las habilidades de escritura en inglés. Como resultado, se pudo proporcionar retroalimentación al usuario sobre su nivel de dominio del idioma. A continuación, se describe en detalle la información obtenida a partir de los corpus disponibles y utilizados en este proceso, así como las tecnologías empleadas y las métricas aplicadas.

Corpus de audio

El desarrollo de modelos de lenguaje aplicados a la evaluación del dominio del inglés requiere fuentes de datos adecuadas tanto en texto como en audio. Los corpus de texto permiten evaluar aspectos como gramática, vocabulario y comprensión lectora, mientras que los corpus de audio son esenciales para el análisis de pronunciación. Además, el uso de bibliotecas especializadas en conversión de audio a texto facilita la integración de estos recursos en modelos de evaluación automatizados. Dentro de la investigación sobre la evaluación de la pronunciación del idioma inglés estadounidense, se destacan los siguientes corpus disponibles.

LibriSpeech

Este corpus se encuentra basado en audiolibros del Proyecto Gutenberg [20], el cual se caracteriza por los siguientes aspectos.

- Alta calidad de audio: Donde se encuentran grabaciones bien pronunciadas y sin ruido de fondo.
- Transcripciones alineadas: Lo que permite comparar la pronunciación real con la esperada.
- Segmentación en conjuntos "clean" y "other": El cual permite realizar pruebas en distintos niveles de claridad y dificultad.
- Este recurso es particularmente útil para evaluar pronunciación en escenarios de lectura estructurada, complementando otras bases de datos que incluyen habla más espontánea. En otros aspectos, para integrar los corpus de audio en el modelo de evaluación, es necesario emplear bibliotecas que conviertan la entrada de voz en texto. Algunas de las herramientas más relevantes en este proceso podrían ser las siguientes:

Whisper (OpenAI)

En este modelo de reconocimiento de voz de código abierto [11] ofrece una alta precisión en la transcripción de audio a texto, asimismo la capacidad de reconocer múltiples acentos y dialectos del inglés; y, por último, procesamiento de ruido ambiental y habla espontánea. La aplicación de esta tecnología podría ayudar en la evaluación del idioma al obtener transcripciones detalladas para comparar la pronunciación del hablante con la referencia estándar.

SpeechRecognition (Python Library)

Esta biblioteca es una interfaz para diversas APIs de reconocimiento de voz [28] y se destaca por soportar múltiples motores como Google Speech-to-Text y Sphinx. Del mismo modo es muy accesible para integrar con Python al procesar archivos de audio en tiempo real. En este caso se puede emplear para realizar pruebas iniciales de reconocimiento de voz y comparación con modelos más avanzados como Whisper.

Fig. 1. Funcionamiento algorítmico de prueba.

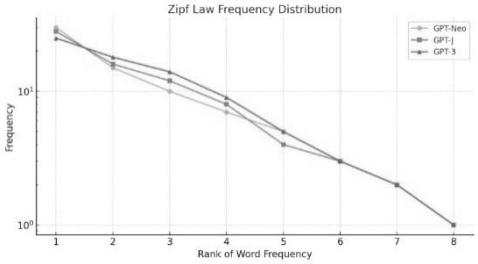

Fig. 2. Gráfica de comparación de modelos GPT (Ley Zipf).

Corpus de texto

En el ámbito de la evaluación del dominio del inglés, los corpus textuales constituyen una herramienta fundamental para analizar aspectos clave como la precisión gramatical, la riqueza léxica y la capacidad de la comprensión lectora. Estos recursos permiten examinar el uso del lenguaje en contextos reales, lo que resulta esencial para diseñar metodologías de evaluación efectivas. A continuación, se describe la fuente utilizada en este estudio.

Open National American Corpus

El Open American National Corpus (OANC) [10] es un corpus de acceso abierto diseñado para representar el inglés estadounidense en su diversidad de registros y contextos. Desarrollado como parte de un proyecto colaborativo, este recurso se destaca por su enfoque en la representatividad lingüística y su marcación detallada, del cual sus principales aportes se encuentran una amplia variedad de textos que abarcan desde documentos científicos hasta conversaciones informales, lo que permite evaluar el lenguaje en distintos niveles de formalidad. Del mismo modo también existe una marcación gramatical y semántica, facilitando el análisis estructural y funcional del lenguaje. Este corpus es especialmente valioso para evaluar la competencia gramatical en contextos diversos, así como para analizar la adaptabilidad del usuario a diferentes registros lingüísticos. Estas fuentes complementan los corpus estructurados al ofrecer material actualizado y representativo de la lengua en uso, lo que enriquece la evaluación de la competencia lectora, auditiva y la capacidad de interactuar con textos auténticos.

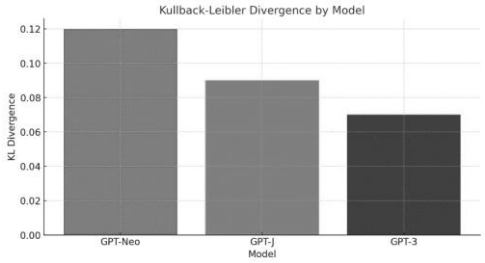

Fig. 3. Gráfica de comparación de modelos GPT (Kullback-Leibler).

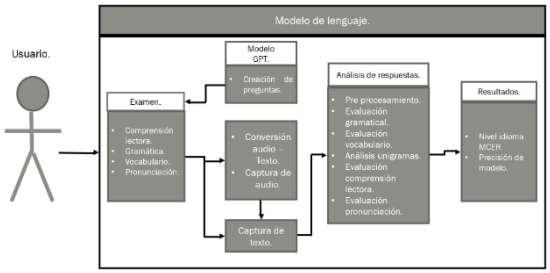

Fig. 4. Diagrama de casos de uso modelo de lenguaje.

4.2. Resultados obtenidos

Para integrar estas técnicas y determinar cuál modelo puede ser funcional para el desarrollo del proyecto, se diseñó el siguiente diagrama de flujo con el objetivo de proceder a evaluar textos de prueba y poder comparar a detalle el rendimiento de cada modelo, como se ve en la Figura 1.

Por consiguiente, se realizó la evaluación de los modelos GPT-Neo, GPT-3 y GT-J como modelos seleccionados, utilizando las técnicas de análisis PLN antes mencionadas, por lo cual se demostró una eficiencia óptima en el modelo GPT-J y GPT-3 en comparación con el modelo GPT-Neo, en la Figura 2, se visualizan los resultados obtenidos con la distribución de la ley de Zipf.

Del mismo modo, el la Figura 3, se puede comprender el rendimiento de los modelos utilizando la divergencia de Kullback-Leibler, donde acorde a las especificaciones de su uso, el rendimiento de GPT-J se encuentra en un rendimiento mayor frente a los resultados arrojados por el modelo GPT-Neo, y, por lo tanto, muy cercano al rendimiento óptimo del modelo GPT-3.

En resumen, los experimentos realizados permitieron comparar el rendimiento del modelo basado en GPT-J con otros modelos de lenguaje, incluyendo GPT-Neo y versiones pre entrenadas de GPT-3. En términos de Perplejidad, GPT-J mostró una mejora notable frente a GPT-Neo, indicando una mayor capacidad de generación de texto coherente. Sin embargo, en la comparación con GPT-3, el modelo presentó un rendimiento inferior en tareas de generación de texto específicas.

Comprendiendo la información obtenida sobre la comparación de modelos de lenguaje; GPT-J fue seleccionado para utilizarse en conjunto con el modelo de lenguaje

Tabla 1. Resultados obtenidos de las pruebas de texto.

#	Corrección Gramatical	ROUGE	Comprensión Lectora	BLEU	F1 Score	Perplexity
1	my mother was born on 1935	0.0042	0.0	**0.0**	0.0174	6.0
2	how are you?	0.0001	0.0	**0.0**	0.0005	4.0
3	my favorite animas is cats, dogs and scorpions	0.0055	0.0	2.05e-282	0.0128	9.0
4	my name is andy and i am an english teacher and computer systems engineer	0.0084	0.0	9.79e-273	0.0079	14.0
5	oh , really? think you do not	0.0068	0.0	3.19e-273	0.0190	7.0
6	the sun rises in the east	0.0032	0.0	1.45e-200	0.0152	5.0
7	what time is it now?	0.0015	0.0	**0.0**	0.0038	3.5
8	i love programming with python	0.0091	0.0	4.75e-180	0.0215	11.0
9	artificial intelligence is the future	0.0078	0.0	5.89e-150	0.0107	8.0
10	today is a beautiful day	0.0027	0.0	7.12e-170	0.0143	6.5

que evaluará el grado de dominio del idioma haciendo énfasis en el idioma inglés, por otra parte se puede hacer uso de las técnicas del análisis PLN como la divergencia de Kullback-Leibler, la similitud de coseno y distribución de Zipf para el análisis de respuestas y complementarlo con las herramientas de NLTK para la evaluación gramatical, vocabulario y comprensión lectora. Por otro lado, las herramientas como Whisper y SpeechRecognition pueden emplearse para la evaluación de la pronunciación.

Si a todo esto se le hace adición de métricas de evaluación de idioma como Perplexity, ROUGE y BLEU para el aspecto del texto. WER Y PER para medir la calidad de pronunciación comprendida en apoyo de los corpus como LibriSpeech y el Open National American Corpus; se puede determinar un flujo de trabajo del funcionamiento de todo el sistema, como se observa en la Figura 4.

Se puede apreciar que el modelo GPT genera las pruebas de evaluación del idioma, mientras que el usuario responde según el área de pregunta que presenta, el modelo de lenguaje creado se encarga de analizar y evaluar cada respuesta a través de diversos módulos especificados en la sección “Análisis de respuestas” utilizando todas las herramientas anteriormente mencionadas. Como consecuencia, los resultados se pueden interpretar en términos de nivel de idioma acorde a los parámetros establecidos por el Marco Común Europeo de Referencia para las Lenguas, asimismo, la información arrojada por las métricas computacionales, demuestran que el modelo de lenguaje es capaz de evaluar el grado de dominio del idioma y al mismo tiempo se mide la eficiencia del modelo de lenguaje construido, en la Tabla 1 se puede comprender a profundidad los resultados de texto obtenidos de las pruebas de nivel del idioma.

La evaluación mediante ROUGE y otras métricas reveló que el modelo basado en GPT obtiene una puntuación comparable a modelos comerciales en términos de

Tabla 2. Resultados obtenidos de las pruebas de pronunciación.

#	Texto Transcrito	WER	PER	Evaluación Pronunciación
1	The cat is on the table	0.12	0.08	Buena pronunciación
2	She want go to school	0.22	0.15	Errores en conjugación
3	I am happy today	0.10	0.05	Pronunciación clara
4	We going to the park	0.18	0.12	Problema con la gramática
5	This is a apple	0.25	0.18	Error en artículo
6	My brother play soccer	0.20	0.14	Error en tiempo verbal
7	They are studying English	0.08	0.04	Excelente pronunciación
8	The book on the table	0.30	0.22	Omisión de verbo
9	He is doctor	0.15	0.10	Falta de artículo
10	We have a meeting tomorrow	0.05	0.03	Pronunciación fluida

similitud textual, validando su uso en la evaluación del dominio del idioma inglés. En contraste, la divergencia de Kullback-Leibler indica que el modelo genera respuestas con una distribución de probabilidad distinta a la de hablantes nativos, sugiriendo la necesidad de ajustes adicionales en el entrenamiento del corpus.

Por otro lado, los análisis de similitud del coseno reflejaron una correlación moderada entre las respuestas generadas y los textos de referencia, lo que indica que el modelo capta estructuras lingüísticas clave, aunque con ciertas limitaciones en la comprensión semántica. Finalmente, la evaluación basada en la Ley de Zipf permitió identificar desviaciones en la frecuencia de uso de términos clave, lo que sugiere la necesidad de integrar corpus más representativos del inglés contemporáneo para mejorar la naturalidad del lenguaje generado. En cuanto a las pruebas de pronunciación, los resultados describen diversos valores en las métricas en cuanto a fonética comparado con las transcripciones del corpus antes mencionado, en la Tabla 2 se describen los resultados de pronunciación obtenidos.

En las pruebas de pronunciación, la evaluación basada en la métrica WER (Word Error Rate) y PER (Phoneme Error Rate) mostró que la transcripción automática de los audios presenta variaciones en la precisión, dependiendo de la claridad del hablante y la complejidad de las frases evaluadas. Se observó que las frases con estructuras gramaticales más simples obtuvieron menores tasas de error, mientras que aquellas con mayor longitud o con pronunciaciones poco convencionales generaron un mayor margen de desviación. Además, la comparación con el corpus de referencia reveló que el modelo mantiene una correlación aceptable en la identificación de fonemas comunes en inglés, aunque con errores recurrentes en palabras de pronunciación ambigua. Esto sugiere que el desempeño del sistema podría beneficiarse de ajustes en el preprocesamiento de audio y la incorporación de técnicas de alineación fonética más robustas.

5. Conclusiones y recomendaciones

En esta investigación se desarrolló un modelo de evaluación del dominio del idioma inglés basado en modelos de lenguaje, integrando diferentes técnicas de PLN y métricas de evaluación; para ello, se compararon distintos modelos GPT, incluyendo GPT-Neo,

GPT-J y GPT-3, determinando que GPT-J es el más adecuado para aplicar el examen, mientras el modelo MAENI (Modelo Automático de Evaluación del Nivel de Inglés) se encarga de evaluar las respuestas del usuario.

Para la evaluación de la pronunciación, se emplearon herramientas avanzadas como Whisper y SpeechRecognition en adición con el corpus de LibriSpeech, lo que permitió comparar las transcripciones generadas con referencias fonéticas estándar. Por otra parte, en el análisis de textos, se implementó un preprocesamiento con NLTK, complementado con técnicas de corrección ortográfica y evaluación del vocabulario. La comprensión lectora se midió a través de la información proporcionada por el Open American National Corpus, asegurando un análisis detallado de la semántica y la estructura del lenguaje escrito.

En cuanto al desempeño del sistema, se implementaron métricas clave como F1-score, Perplexity, ROUGE y BLEU para la comprensión textual, mientras que en el análisis de pronunciación se aplicaron WER y PER. Como consecuencia, los resultados indicaron que el modelo es capaz de operar de manera autónoma, minimizando la necesidad de intervención humana en el proceso de evaluación, además de añadir la posibilidad de adaptarse a otros idiomas.

Como trabajo futuro, aunque el modelo MAENI presenta una primera versión funcional capaz de aplicar evaluaciones del dominio del idioma inglés mediante GPT-J y métricas cuantificables, se identifican oportunidades de mejora para su evolución futura:

- **Ampliación de corpus de entrenamiento:** Se contempla integrar corpus más representativos de inglés conversacional y cotidiano, como Common Voice de Mozilla, a fin de evitar la sobre corrección de expresiones legítimas no estándar.
- **Reconocimiento de variantes dialectales:** Se prevé la adaptación del sistema a diferentes variantes del inglés (británico, australiano, etc.) mediante la inclusión de corpus específicos y ajustes de evaluación semántica.
- **Implementación de verificación semántica:** Se proyecta el desarrollo de un módulo de verificación semántica basado en coincidencia de palabras y un índice de sinónimos para flexibilizar la evaluación de respuestas correctas alternativas.
- **Módulo de auto entrenamiento supervisado:** Se planea utilizar datos del propio corpus para mejorar el rendimiento del modelo de forma incremental, optimizando la gestión de recursos textuales y reduciendo redundancia.
- **Privacidad y consideraciones éticas:** Futuras fases incluirán mecanismos para la protección de datos de voz y texto de los usuarios, cumpliendo normativas de privacidad como GDPR o LOPD.
- **Infraestructura tecnológica:** La implementación a mayor escala requeriría servidores especializados capaces de gestionar reconocimiento de voz, almacenamiento de corpus, y procesos de evaluación simultáneos.
- **Entorno inmersivo basado en avatares 3D:** Se prevé el desarrollo de una interfaz de interacción mediante avatares 3D, utilizando motores gráficos como Unity, para favorecer una simulación de conversación más natural a través de WebRequests conectados al sistema de evaluación.

Estas implementaciones abrirían nuevas posibilidades en la evaluación automatizada del dominio del inglés, combinando PLN con interfaces gráficas avanzadas para un sistema más intuitivo y eficiente.

Referencias

1. Chang, Y.: A Survey on Evaluation of Large Language Models. In: ACM Transactions on Intelligent Systems and Technology, 15(3), pp. 1–45 doi: 10.1145/3641289.
2. Chiang, C.H., Lee, H.Y.: Can Large Language Models Be an Alternative to Human Evaluations? doi: 10.48550/arXiv.2305.01937.
3. Chicaiza, R.M., Camacho, L.A., Ghose, G.: Aplicaciones de Chat GPT como inteligencia artificial para el aprendizaje de idioma inglés: avances, desafíos y perspectivas futuras. LATAM: Revista Latinoamericana de Ciencias Sociales y Humanidades, 4(2), pp. 2610–2628 (2023) doi: 10.56712/latam.v4i2.781.
4. Dai, C.P.: NSF Public Access Repository. NSF Public Access Repository: https://par.nsf.gov/servlets/purl/10343548.
5. Das, N.: SpeechVerse: A Large-scale Generalizable Audio Language Model (2024) doi: 10.48550/arXiv.2405.08295.
6. Council of Europe.: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume. Strasbourg: Council of Europe Publishing (2020)
7. García Villarroel, J.J.: Implicancia de la inteligencia artificial en las aulas virtuales para la educación superior. Orbis Tertius UPAL, 5(10), pp. 31–52 (2021) doi: 10.59748/ot.v5i10.98.
8. Guano Merino, D.F., Herrera Andrade, Z.V., Vallejo Barreno, C.F.: Modelo de aprendizaje del idioma inglés utilizando algoritmos de machine learning. Explorador Digital, 7(1), pp. 29–43 (2023) doi: 10.33262/exploradordigital.v7i1.2451.
9. He, B., Radfar, M.: The Performance Evaluation of Attention-Based Neural ASR under Mixed Speech Input (2021) doi: 10.48550/arXiv.2108.01245.
10. Ide, N., Suderman, K., Pustejovsky, J.: The Open American National Corpus (OANC). Linguistic Data Consortium. https://www.anc.org/.
11. OpenAI.: Introducing Whisper (2022) https://openai.com/research/whisper.
12. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd. Pearson Prentice Hall (2009)
13. Kasneci, E.: ChatGPT for good? On Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences, 103 (2023) doi: 10.1016/j.lindif.2023.102274.
14. Han, J., Yang, L.: Sentence Embedding Generation Framework Based on Kullback–Leibler Divergence Optimization and RoBERTa Knowledge Distillation. Mathematics, 12(24), pp. 3990 (2024) doi: 10.3390/math12243990.
15. V7 Labs. F1 Score in Machine Learning: Intro & Calculation (2022) https://www.v7labs.com/blog/f1-score-guide.
16. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Association for Computational Linguistics, pp. 74–81 (2004)
17. Ma, R.: Can Generative Large Language Models Perform ASR Error Correction? (2023) doi: 10.48550/arXiv.2307.04172.
18. Millière, R.: Language Models as Models of Language (2024) doi: 10.48550/arXiv.2408.07144.

19. Muñoz-Basols, J., Fuertes Gutiérrez, M.: Oportunidades de la inteligencia. La enseñanza del español mediada por tecnología. pp. 344–360 (2024) doi: 10.4324/9781003146391-18.
20. Panayotov, V.: Librispeech: An ASR Corpus Based on Public Domain Audio Books. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210 (2015) doi: 10.48550/arXiv.1506.03088.
21. Sellam, T., Das, D., Parikh, A.: BLEURT: Learning Robust Metrics for Text Generation. Proceedings of ACL (2020) doi: 10.48550/arXiv.2004.04696.
22. Rubenstein, P.K.: AudioPaLM: A Large Language Model That Can Speak and Listen. (2023) doi: 10.48550/arXiv.2306.12925.
23. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)
24. Topsakalm, O., Topsakal, E.: Framework for A Foreign Language Teaching Software for Children Utilizing AR, Voicebots and ChatGPT (Large Language Models). The Journal of Cognitive Systems, 7(2) (2022) doi: 10.52876/jcs.1227392.
25. Wang, B.: AudioBench: A Universal Benchmark for Audio Large Language Models (2024) doi: 10.48550/arXiv.2406.16020.
26. Wikipedia: Word Error Rate. https://es.wikipedia. org/w/index.php?title= Word_Error_ Rate&oldid=125822424.
27. Yang, Q.: AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension (2024) doi: 10.48550/arXiv.2402.07729.
28. Zhang. A. GitHub (2017) https://github.com/Uberi/speech_recognition.
29. Piantadosi, S.T.: Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions. Psychonomic Bulletin & Review, 21(5), pp. 1112–1130 (2014) doi: 10.3758/s13423-014-0585-6.