MIA
Magíster en
Inteligencia Artificial

Evaluación y ajuste de modelos preentrenados para la transcripción de audio y corrección de disfluencias en personas con tartamudez en español

MAGÍSTER EN INTELIGENCIA ARTIFICIAL

Marcos Irving Mera Sánchez
Candidato a Magíster en
Inteligencia Artificial
Escuela de Ingeniería PUC

Agenda de hoy

- Introducción.
- Objetivos y contribuciones del proyecto.
- Descripción del problema.
- Definiciones teóricas.
- Aplicación de la teoría.
- Propuesta de solución.
- Resultados obtenidos.
- Conclusiones.
- Trabajos Futuros.

Introducción

MIA
Magíster en Inteligencia Artificial

Motivación

- ✓ Inclusión de personas con tartamudez en las tecnologías de reconocimiento de voz.
- ✓ Mejora de la calidad en las transcripciones, para garantizar la legibilidad y comprensión.
- ✓ Las disfluencias como repeticiones y prolongaciones suelen ser mal interpretadas por los ASR actuales, generando transcripciones imprecisas.

¿ Qué se espera del proyecto

- ➢ Adaptar el dataset base para realizar la evaluación y mejora de modelos preentrenados de reconocimiento automático del habla (ASR) enfocados en la transcripción de audios de personas con tartamudez.
- ➢ Implementar modelo para el manejo eficaz de las disfluencias, mejorando la precisión de las transcripciones y promoviendo la inclusión en aplicaciones tecnológicas, como terapias, educación y asistencia diaria.

Objetivos y contribuciones del proyecto

MIA
Magíster en Inteligencia Artificial

Objetivos

- Mejorar la precisión en la transcripción de audios de personas con tartamudez, evaluando la salida de texto generada por los modelos pre entrenados donde se obtenga un modelo robusto que maneje eficazmente las disfluencias.

Contribuciones del proyecto

OE1. Adaptación del dataset especializado

Incluye audios de personas con tartamudez con anotaciones precisas de disfluencias.

OE2. Evaluación de modelos ASR preentrenados

Comparación de modelos como wav2vec 2.0, Whisper utilizando métricas como Word Error Rate (WER), precisión, recall y F1-score.

OE3. Mejora de la salida textual

Implementación de técnicas de postprocesamiento para corregir repeticiones y pausas prolongadas.

OE4. Validar el modelo mejorado

En pruebas con datasets de audios de personas con tartamudez, garantizando su robustez y precisión.

Descripción del problema

MIA
Magíster en Inteligencia Artificial

Manejo incorrecto de disfluencias

Las aplicaciones que usan ASR no reconocen correctamente adecuadamente las disfluencias características de la tartamudez, como repeticiones, prolongaciones y bloqueos.

Transcripciones erróneas

Esto genera transcripciones erróneas que obstaculizan la comunicación eficaz para las personas que dependen de estas tecnologías.

Obstáculos para la comunicación

La no adecuada transmisión de voz en dichas tecnologías, obstaculiza la interacción efectiva en situaciones cotidianas.

Evaluación de los tipos de disfluencias para la aplicación mediante el el uso de las transcripciones de sistemas ASR actuales.

Definiciones teóricas

MIA
Magíster en Inteligencia Artificial

Dataset: Segmentación (oraciones)

- Similitud de oraciones base.
- Construcción de dataset de audio en relación a lectura base.
- Alineamiento de lectura a nivel de oraciones.
- Segmentación de audio a nivel de oraciones.
- Bitácora de dataset generado CSV.

- Whisperx
- **Timething**
- Sequence Matcher
- Calidad de datos.

Evaluación de modelos ASR

- Evaluación de modelos preentrenados, modo entrenamiento y evaluación de ASR

- Whisper (Open AI)
- Wav2vec
- Cloud Speech-to-Text (Google)

Corrección de texto basado en disfluencias

- NLP
- Red GAN
- Corrección de textos con disfluencias
- Análisis de disfluencias

- Redes GAN
- Modelos BERT

Evaluación de métricas ASR y semánticas

- Evaluación de métricas ASR

- WER
- CER

- Evaluación de métricas semánticas de texto

- BLEU Score
- ROUGE
- BERT

Aplicación de la Teoría - Estudios previos

Trabajo	Contribuciones Principales	Resultados Clave	Limitaciones/Desafíos
Construcción de recursos para la detección y clasificación automática de disfluencias producidas por tartamudez en Español (Cabrera Díaz, 2024).	- *Adaptación de dataset de disfluencias en español.* - Propuesta de DisfluencyNet preentrenado en wav2vec 2.0 XLSR53.	- *Avances en la clasificación de disfluencias en español.* - Necesidad de mayor volumen de datos para mejorar precisión.	- *Rendimiento limitado en español.* - Escasez de recursos específicos.
Adversarial Training for Low-Resource Disfluency Correction (Bhat et al.,2023).	- *Uso de técnicas adversariales para corregir disfluencias.* - Evaluación en lenguas de bajos recursos como bengalí e hindi.	- *Mejora significativa en F1-score (87.68%) al integrar datos sintéticos y no etiquetados.* - Resultados sobresalientes.	- *Enfocado en inglés y lenguas de bajos recursos.* - Necesidad de adaptación para español.
A Comprehensive Evaluation of Incremental Speech Recognition and Diarization for Conversational AI (Addlesee Angusen et al.,2020).	- *Evaluación de sistemas ASR y diarización para entornos conversacionales.* - Comparación entre Microsoft e IBM.	- *Microsoft: preserva mejor las disfluencias.* - IBM: mejor manejo de superposiciones de voces.	- *Complejidad en entornos con múltiples interlocutores y flujos de diálogo.*

Propuesta de solución

MIA
Magíster en Inteligencia Artificial

Dataset base - Disfluencias

•5 personas tartamudas.
•100 Lecturas base.

Audio

Lectura Base

Dataset segmentado (texto y audio)

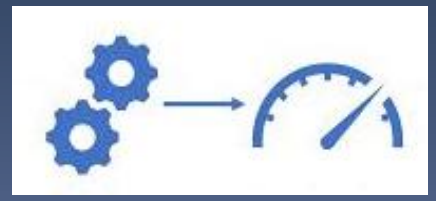
Segmento Audio

Segmento Lectura Transcrita

Gold Sentence

- Librerías de alineamiento:
- Whisperx
- Timething

Evaluación de modelos ASR

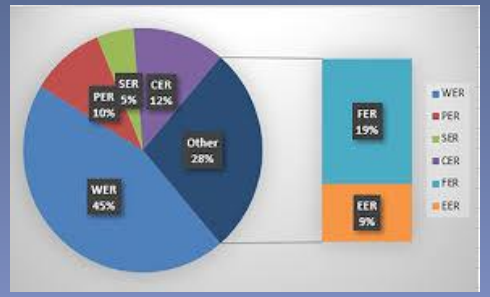


- Modelos ASR:
- Whisper
- Wav2vec Cloud Speech-to-Text

Corrección y análisis de texto basado en disfluencias

Corrección del segmento de texto

Evaluación con métricas ASR(WER, CER) y semánticas



Descripción de tareas del Proyecto a nivel funcional y técnico

Resultados obtenidos

MIA Magíster en Inteligencia Artificial

Evaluación de modelos ASR	*El modelo Wav2Vec2 ASR* es el más efectivo para evaluar disfluencias a nivel de palabras y caracteres, superando en precisión a modelos como Whisper ASR y Speech-to-Text ASR. El desempeño cuyo desempeño lo posiciona como ideal para tareas de transcripción en idioma español, mientras que los demás modelos requieren optimizaciones específicas.

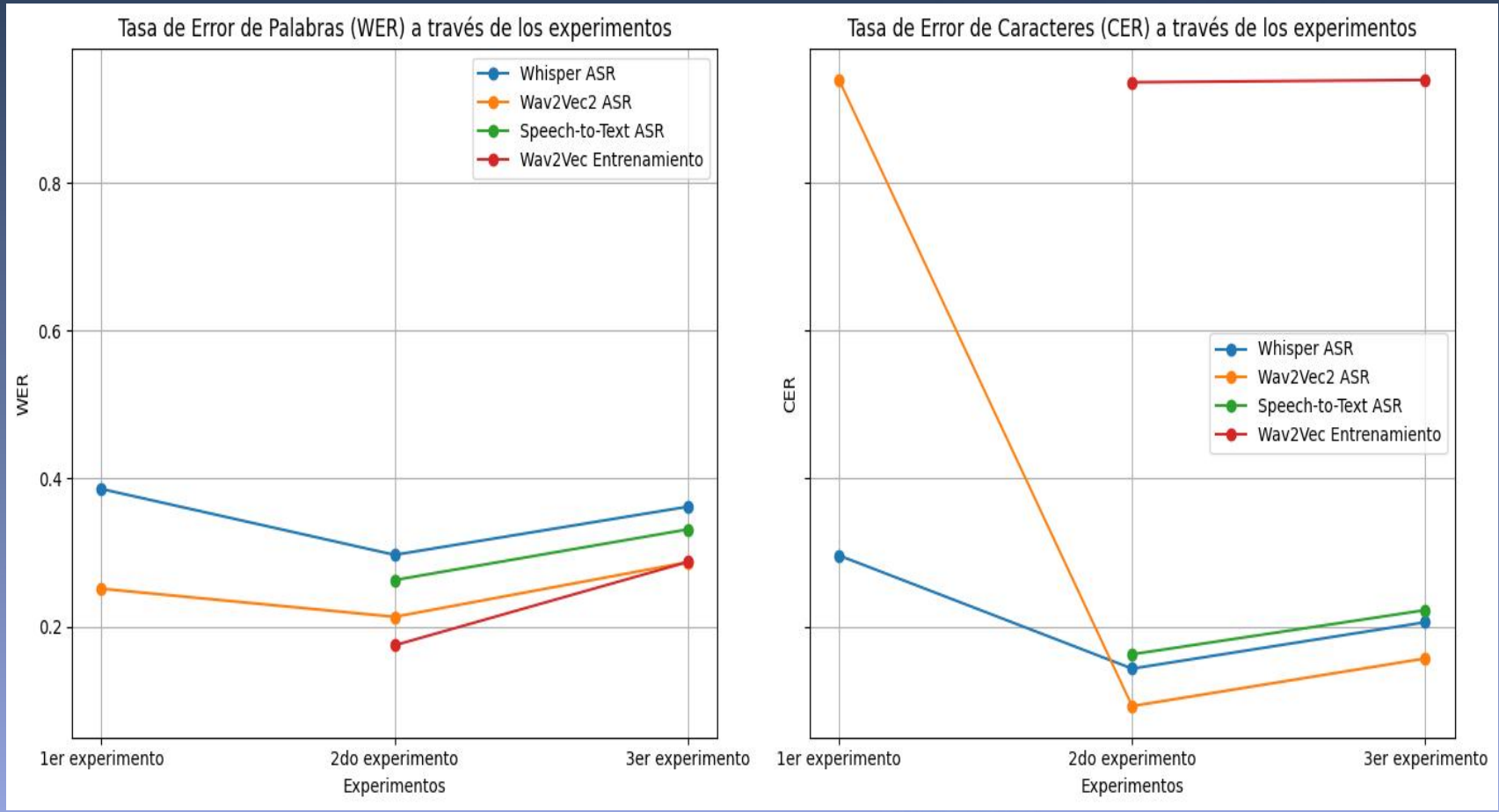
Información de la evaluación

Exp.	Librería	Nro. Lecturas | Segmentos
1	Whisperx	500 | 8348
2	Timething	327 | 3762
3	Timething	327 | 3960

Exp.	WER | CER	Modelo (Best)
1	0.2509 | 0.940	Wav2Vec2 Train
2	0.2129 | 0.0921	Wav2Vec2 ASR
3	0.2869 | 0.1566	Wav2Vec2 ASR

Mejor experimento en evaluación ASR

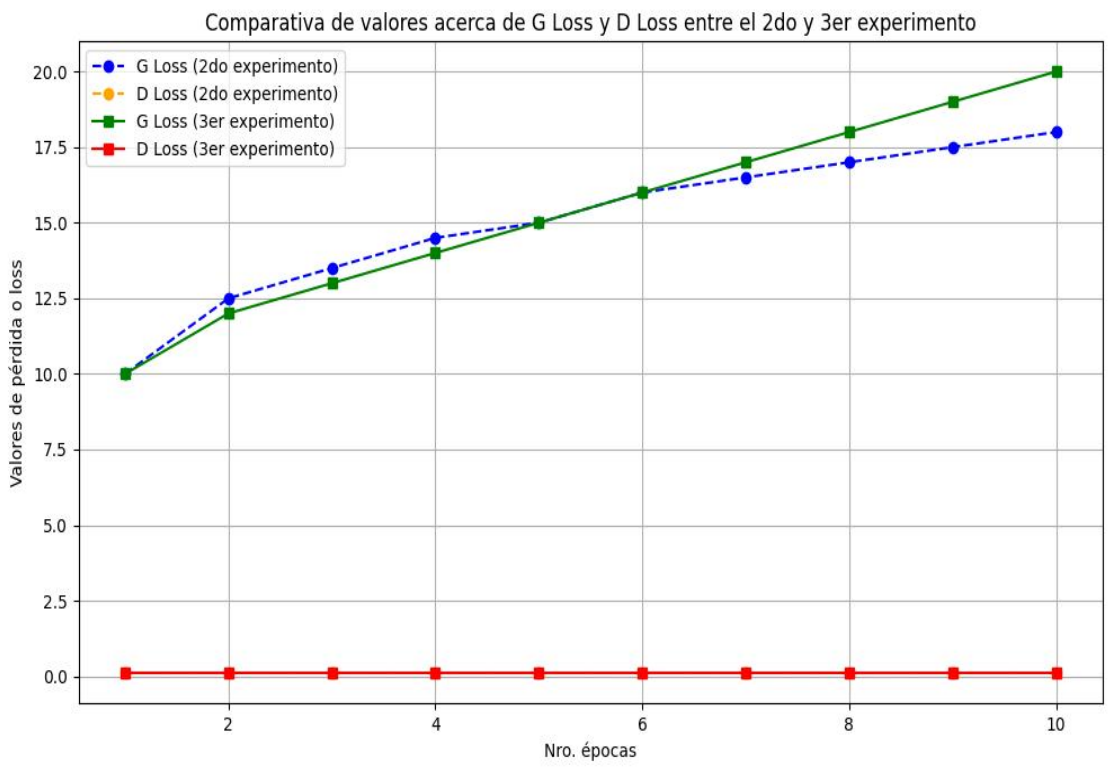
Mejor experimento en proceso de corrección de texto



Resultados obtenidos

MIA Magíster en Inteligencia Artificial

Evaluación de red GAN -Corrección de texto



La evaluación de la red GAN nos muestra que **Generator Loss** se incrementa con las épocas en ambos experimentos, mientras que los valores de **Discriminator Loss** permanece constante y cercana a cero en ambos casos, mostrando un desbalance en el entrenamiento ya que podría limitar la efectividad del modelo, cuyo objetivo es lograr que **Discriminador** aprenda de forma efectiva y que el **Generador** tenga mejor retroalimentación

Evaluación de proceso Corrección de texto

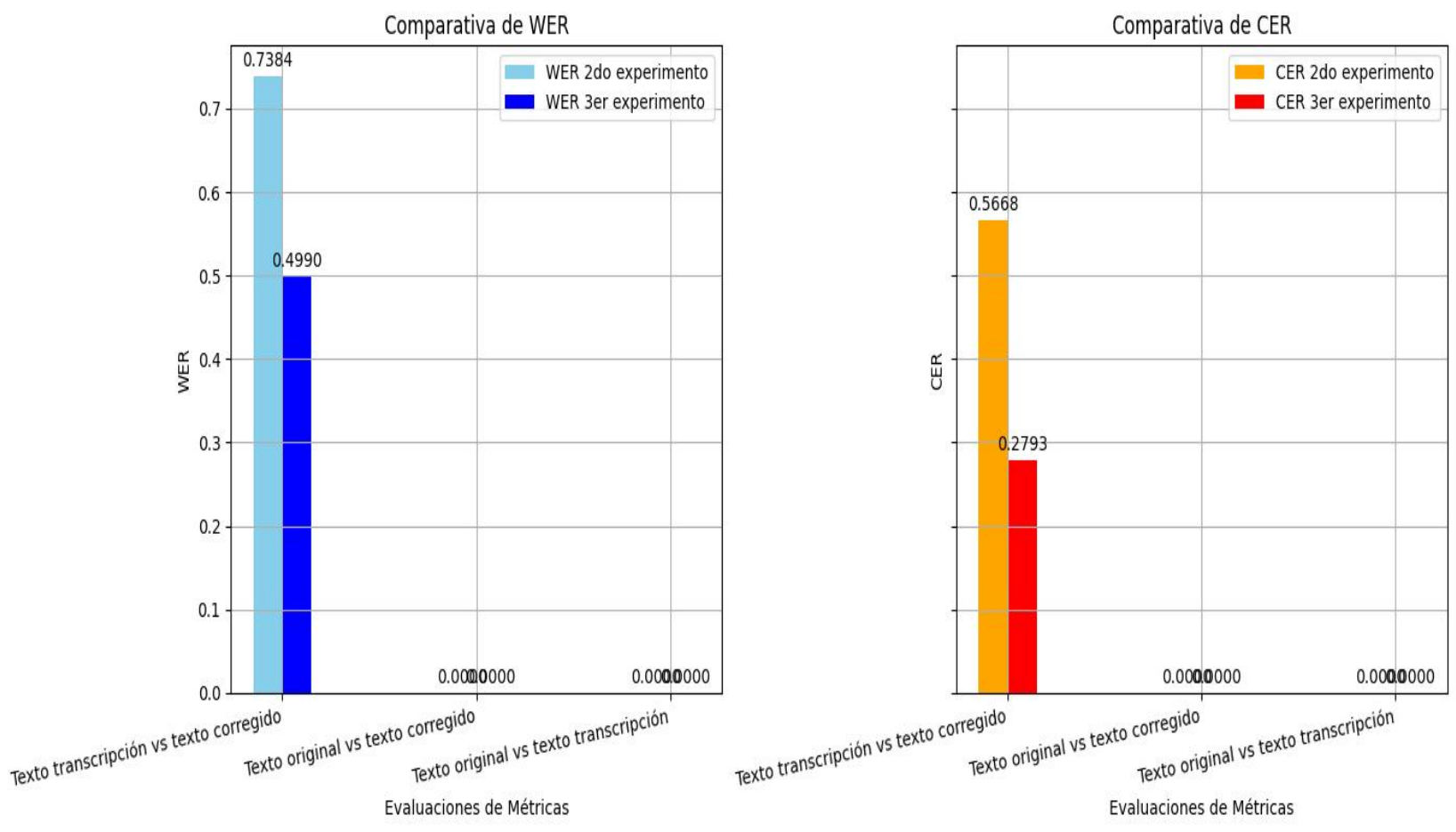
Tipos de texto	Descripción
Original_text	Texto de oración base de experimento.
Correct_text	Texto corregido por modelo prenentrenado textual.
Transcribed_text	Texto generado por ASR (Wav2Vec2)

index	original_text	corrected_text	transcribed_text
0	Feriado largo de agosto: ¿cuáles son los días no laborables de fin de mes?	Feriado largo de agosto: ¿cuáles son los días no laborables de fin de mes?	feriado largo de agosto cuales son los días no elaborables de fin de mes
1	Feriado largo 2019. Agosto también tendrá fin de semana largo.	Feriado largo 2019. Agosto también tendrá fin de semana largo.	feriado largo domilici nueve agosto también tendrá fin de semanas largo
2	El gobierno decretó días no laborables para celebrar el día de Santa Rosa de Lima.	El gobierno decretó días no laborables para celebrar el día de Santa Rosa de Lima.	el gobierno decretó días no laborables para celebrar el día de santa rosa el lima
3	Entérate más a continuación.	Entérate más a continuación.	enter a temás a continuación
4	Agosto de 2019 termina con un feriado en Perú.	Agosto de 2019 termina con un feriado en Perú.	agosto de dos milesinueve termina con un feriado en el perú
5	El día 30 del mes se celebra el día de Santa Rosa de Lima, la festividad que conmemora a Isabel Flores de Oliva o, Rosa de Santa María, como era conocida antes de su canonización, como patrona de la capital peruana.Y, desde 1989, como patrona de la Policía Nacional del Perú.	El día 30 del mes se celebra el día de Santa Rosa de Lima, la festividad que conmemora a Isabel Flores de Oliva o, Rosa de Santa María, como era conocida antes de su canonización, como patrona de la capital peruana.Y, desde 1989, como patrona de la Policía Nacional del Perú.	el día treinta del mes se celebra el día anta rosa lima la festividad que conmemora a isabel flores oliva o rosa de santa maría como reconocida antes de su canonización como patrona de la capital peruana y desde mil novecientos ochenta y nueve como patrona de la policía nacional
6	En ese sentido, y como ya es costumbre en los gobiernos de turno, se declaran días no laborables con el fin de promover el turismo al interior del país.	En ese sentido, y como ya es costumbre en los gobiernos de turno, se declaran días no laborables con el fin de promover el turismo al interior del país.	en este sentido y como y es costumbre en los gobiernos de turno se declaran días no laborables con el fin de promover el turismo el interior del país

Modelo preentrenado de texto usado: dccuchile/bert-base-spanish-wwm-cased.

Resultados obtenidos

Evaluación de métricas ASR - proceso Corrección de texto

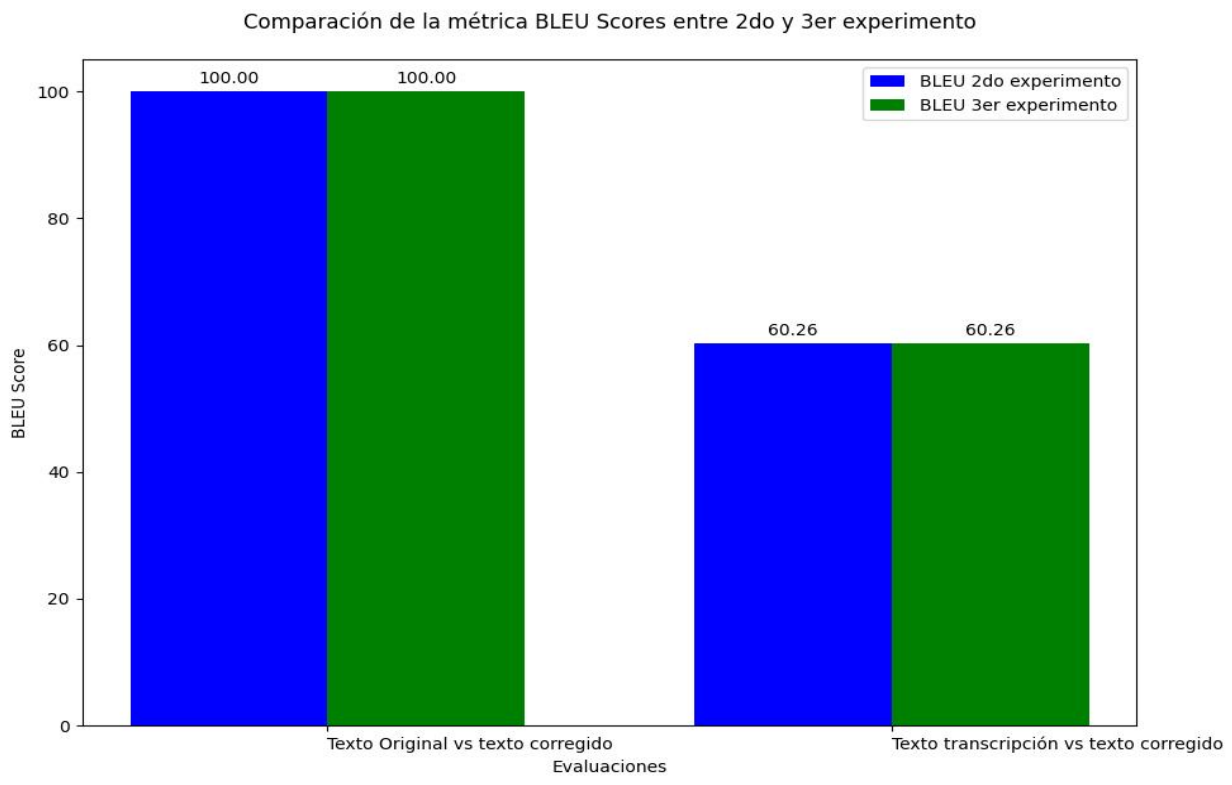


Al realizar mejoras acerca de calidad de datos durante el 3er experimento, durante la evaluaciones aplicando la corrección de texto acerca de "Texto Transcripción vs Texto corregido", se observan reducciones significativas en **WER** (de 0.74 a 0.50) y **CER** (de 0.57 a 0.28) lo que indica mejoras de calidad en el proceso de corrección de texto generados por el modelo preentrenado. Asimismo, las evaluaciones de "Texto Original vs Texto corregido" y "Texto Original vs Transcripción" permanecen con valores nulos, lo que sugiere sin diferencias significsticas u errores, lo que valida la coherencia del sistema.

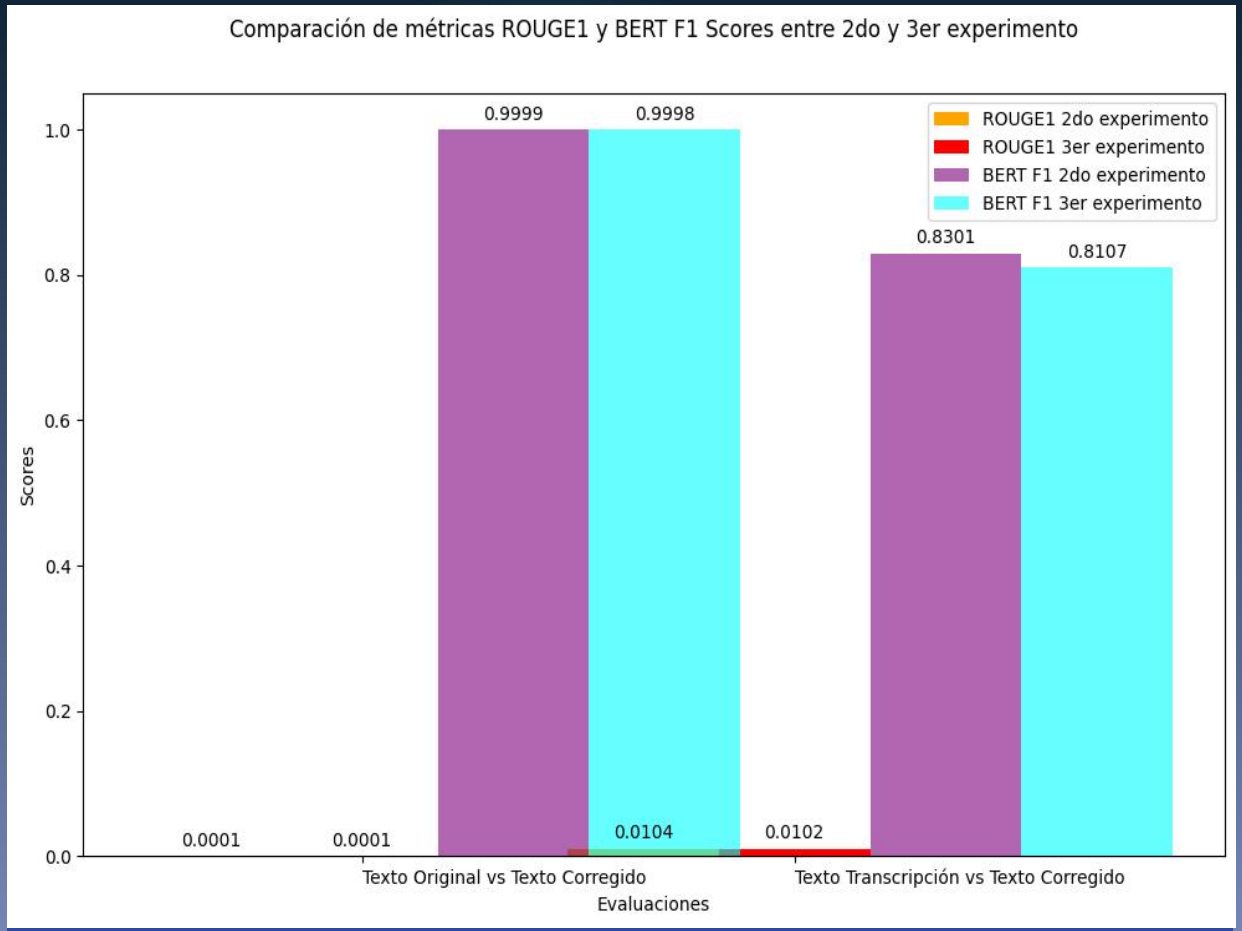
Resultados obtenidos

MIA
Magíster en
Inteligencia Artificial

Evaluación semánticas - proceso Corrección de texto



- La métrica **BLEU Score** mantiene valores perfectos ya que refleja alta similitud textual en ambos experimentos relacionado a "Texto Original vs Texto Corregido".
- En relación hacia "Texto Transcripción vs Texto Corregido", el valor se redujo a 60 aprox. debido a diferencias significativas entre las transcripciones y los textos corregidos.

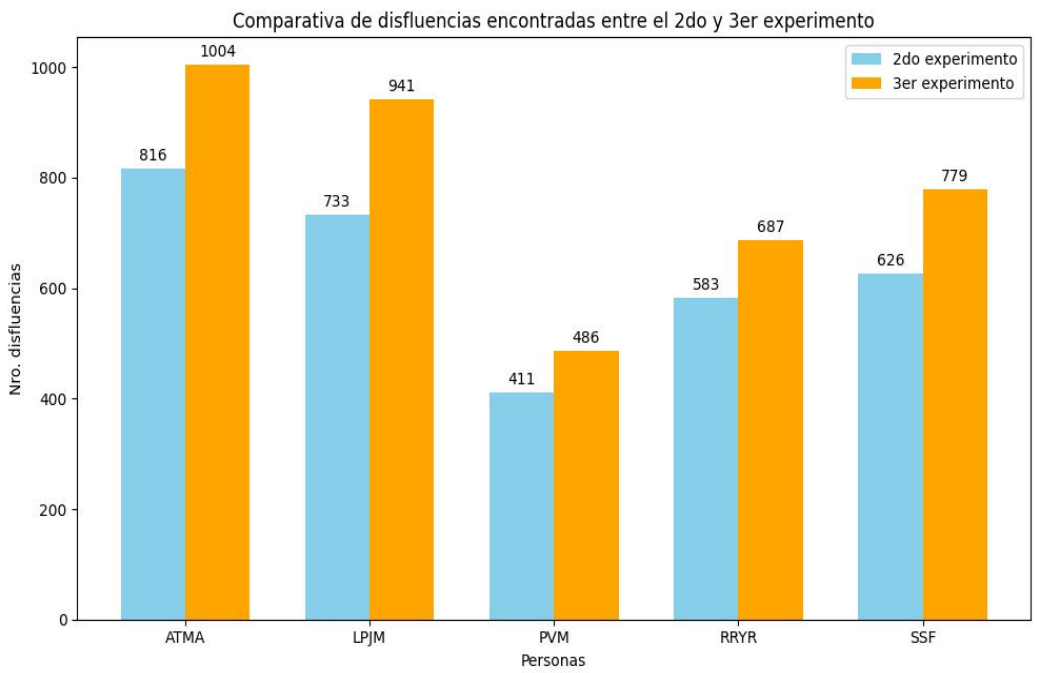


- La métrica **ROUGE1 Score** permanece cercano a cero en todas las evaluaciones ya que las similitudes basadas en fragmentos comunes son mínimas.
- En contraste, la métrica **BERT F1 Scores** fue cercano a 1, señala que el modelo captura bien la similitud y calidad semántica en ambos experimentos.

Resultados obtenidos

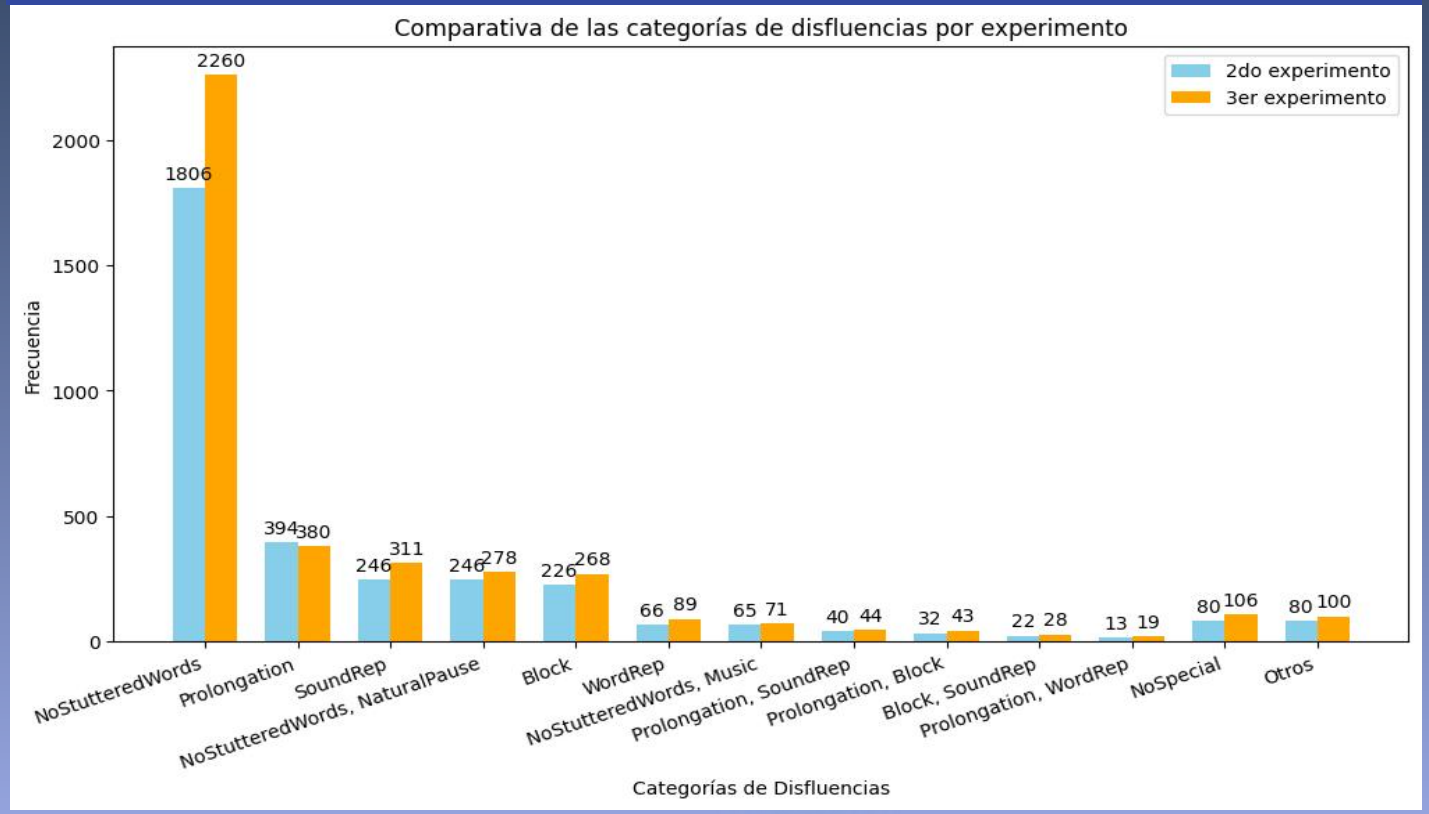
MIA
Magíster en
Inteligencia Artificial

Comparativa de resultados acerca de disfluencias detectadas



Al realizar mejoras en la segmentación de oraciones aplicando calidad de datos durante la 3er experimento, genera mayor identificación de disfluencias en relación hacia el 2do experimento, destacando el incremento en ATMA y LPJM al igual que en las otras personas lo que se muestra un crecimiento notable y consistencia de la tarea dada.

El análisis comparativo acerca de las categorías de disfluencias entre el 2do y 3er experimento, muestra que la categoría más frecuente en ambos casos es "NoStutteredWords", con un incremento significativo en el 3er experimento (2260 vs. 1806). Otras categorías como "Prolongation" y "SoundRep" registraron ligeros aumentos, reflejando mayor diversidad en las disfluencias capturadas durante el 3er experimento mejorando la detección durante el proyecto.



Conclusiones

MIA
Magíster en
Inteligencia Artificial

1. En base a las evaluaciones dadas de modelos ASR, el modelo preentrenado Wav2Vec2 en español es el más adecuado para evaluar sistemas ASR en las condiciones dadas.

2. El proyecto ha permitido la reducción de errores en transcripciones, alineamiento más preciso usando mejoras en los algoritmos lo que se traduce en mejoras en la evaluación de modelos ASR.

3. Señalar que las redes GAN no lograron un balance óptimo, el modelo alcanzó precisión en la detección de disfluencias, con avances en segmentación y alineamiento audio-texto usando Timething.

4. Para mejorar el modelo, se requiere un dataset más robusto y uso de técnicas de fine-tuning para la mejora de la corrección de texto, optimizando las aplicaciones prácticas hacia usuarios con tartamudez.

5. Para mejorar el modelo, se requiere un dataset más robusto y uso de técnicas de fine-tuning para la mejora de la corrección de texto, optimizando las aplicaciones prácticas hacia usuarios con tartamudez.

Trabajos futuros

MIA
Magíster en Inteligencia Artificial

Ampliación del dataset	✓ Incorporar una mayor cantidad de participantes con condición de tartamudez para generar aumento de datos e identificación de categorías de disfluencias, así como inclusión de textos de mayor complejidad lectora.
Mejoras en la segmentación	✓ Refinar las técnicas de alineamiento de oraciones para tratar los diversas categorías de disfluencias, adicional la identificación de pausas, palabras fuera de contexto y errores de pronunciación, generando un dataset robusto.
Gestión de calidad de datos	✓ Diseñar procedimientos sistemáticos para asegurar datasets consistentes y confiables desde las etapas iniciales del proyecto para obtener modelo más robusto y preciso en la medición de ASR.
Optimización de modelos	✓ Implementar técnicas de fine-tuning con mayor coste computacional para ajustar hiperparámetros en modelos preentrenados de ASR para alcanzar una precisión más alta.
Corrección de transcripción de disfluencias	✓ Desarrollar modelos más avanzados para la corrección de texto en español, considerando mejoras en el uso de hiperparámetros en redes GAN o adversariales.

MIA

Magíster en
Inteligencia Artificial

Preguntas de la Comisión

Pregunta 1

MIA
Magíster en
Inteligencia Artificial

¿ Cúales son las razones acerca de la técnica del alineamiento de oraciones como propuesta para la construcción del dataset orientado hacia la evaluación de los sistemas ASR?

Es muy interesante poder realizar evaluaciones de modelos preentrenados ASR haciendo uso de la transcripción por medio de oraciones separadas de forma establecida por el usuario mediante Gold sentences (oraciones base).

Este tipo de generación de dataset permite a nivel de oraciones, evaluar a detalle con mayor consistencia en lo relacionado a la ortografía y gramática presentada sino hacia un nivel sintáctico y semántico lo que permite profundizar el análisis de las transcripciones para encontrar disfluencias.

Pregunta 2

MIA
Magíster en
Inteligencia Artificial

¿Qué otras técnicas se pueden aplicar durante la corrección de textos?

Se pueden aplicar diversas técnicas como el uso de modelos de lenguajes pre-entrenados en idioma español, tomando en cuenta las clasificaciones de disfluencias identificadas que permitan resolver todos los casos posibles y no de forma prefijada o customizada.

Asimismo se pueden realizar evaluaciones usando:

- Algoritmos de gramática hacia la detección y ajuste de disfluencias.
- Uso de herramientas de corrección gramática.
- Normalización de texto.
- Correcciones basadas en contexto.
- Aplicación de sistema regulares y postprocesamiento semántico y otros.

MIA

Magíster en
Inteligencia Artificial

¡ Muchas gracias por su atención !