Test Andres Parra Rodríguez

Celular: +57 3202347640

Correo: p.r.and@hotmail.com

LinkedIn: https://www.linkedin.com/in/andresparrarod/

Cargo: Data Scientist

Resumen

Se desarrolló un pipeline completo para transcripción, corrección y análisis de un archivo de audio educativo de 5 minutos, siguiendo la metodología CRISP-DM. Primero, se generó una transcripción automática con Whisper y se almacenó en un archivo CSV. Luego, se seleccionaron cinco fragmentos representativos para crear un conjunto de referencia manual (gold set), incluyendo la identificación de hablantes. Se aplicó un modelo de lenguaje para corregir los errores en la transcripción automática y se evaluó su desempeño mediante métricas WER y CER antes y después de la corrección. Finalmente, se identificaron los errores gramaticales más frecuentes cometidos por los estudiantes y se propuso una regla de negocio para recomendar ejercicios personalizados. Además, se integró una funcionalidad extra de diarización para mejorar la segmentación por hablante.

1. Comprensión del negocio

El objetivo del proyecto fue evaluar y mejorar la calidad de las transcripciones automáticas de un archivo de audio educativo de 5 minutos, a través de métricas lingüísticas y modelos de corrección gramatical. Se buscaba además extraer información pedagógica útil para mejorar el aprendizaje de los estudiantes, con la opción de integrar funcionalidades adicionales como diarización o detección de errores comunes.

2. Comprensión de los datos

Se proporcionó un único archivo de audio: audio_full.m4a. Este contenía la grabación completa, de la cual se extrajeron segmentos relevantes para generar un conjunto de referencia (gold set) y realizar evaluaciones posteriores.

Se encontró que era un audio de 5 minutos exactos, con dos personas como parlantes. La conversación trata sobre cómo los hablantes planean aprovechar un feriado local en Madrid, comentando sus rutinas, cancelaciones de clases y formas de relajarse durante el día libre. Esta es llevada acabo en idioma inglés.

Está pendiente un análisis exploratorio como la frecuencia de palabras para identificar las más repetidas y detectar posibles temas centrales. Otro sería un análisis de categorías

gramaticales (POS tagging) para ver si predominan sustantivos, verbos o adjetivos y hasta inclusive una nube de palabras para visualizar gráficamente el texto

3. Preparación de los datos

Se realizaron los siguientes pasos:

Segmentación del audio: Se extrajeron 5 clips aleatorios de 30 segundos para construir manualmente el conjunto de referencia (transcript_gold.csv), donde se anotó la transcripción manual y el hablante.

Diarización automática: Se evaluaron los modelos con y sin diarización. En cuanto a esta, luego de la

Transcripción automática: Se utilizó WhisperX (basado en Whisper) como sistema ASR para generar transcripciones automáticas del audio completo (transcript_raw.csv) y de los fragmentos.

4. Modelado

Se implementó un pipeline de corrección lingüística para mejorar la calidad gramatical y semántica de las transcripciones. Debido al bajo número de muestras (5) y a la corta grabación entregada, no se realizó fine tuning. De haberse realizado se perdería la generaliación de los modelos y tendría u sesgo a los pocos ejemplos suministrados. Se escogió el modelo de whisper de OpenAi debido a que trabaja de forma local y no tiene límites de caracteres o costos asociados como los de Google o Azure.

Por otro lado no se realizó una búsqueda de hiperparámetros. Se limitó a evaluar del modelo whisper los distintos modelos existentes (tiny, base, small, médium, large).

Modelos empleados para la transcripción

Modelo	**Basado en**	**Tamaño (parámetros)**	**Arquitectura**	**Entrenado para**
Whisper	OpenAI Whisper	Depende del tamaño: - tiny: 39M - base: 74M - small: 244M - medium: 769M - large: 1550M	Encoder-Decoder (Transformer)	Transcripción automática de audio a texto multilingüe

Modelo	Basado en	Tamaño (parámetros)	Arquitectura	Entrenado para
WhisperX	OpenAI Whisper + Pyannote (diarización)	Mismo número de parámetros que Whisper + modelo adicional de diarización (~20M-100M aprox.)	Whisper (Transformer) + Pyannote (Speaker Embeddings)	Transcripción + Diarización (quién habla y qué dice)

Modelos LLM evaluados para la corrección gramatical

Modelo	Base	Parámetros	Arquitectura	Entrenado para
vennify/t5-base-grammar-correction	T5-base de Google	220M	Encoder-Decoder (Text-to-Text)	Corrección gramatical en inglés
prithivida/grammar_error_correcter_v1	BART-base de Facebook	~140M	Encoder-Decoder	Corrección gramatical de errores comunes
duongna/grammar-correction	T5-small (modificado)	~60M	Encoder-Decoder	Corrección gramatical enfocada en oraciones cortas

La salida corregida se guardó en transcript_corrected.csv.

5. Evaluación

Se calcularon métricas de error para los 5 clips del gold set entre transcript_raw vs transcript_gold y Entre transcript_corrected vs transcript_gold.

Word Error Rate (WER)

WER = N / (S+D+I)

- S: número de sustituciones de palabras
- D: número de eliminaciones de palabras (deletions)
- I: número de inserciones de palabras

- N: número total de palabras en la referencia de palabras

Character Error Rate (CER)

CER = N / (S + D + I)

- S: número de sustituciones de caracteres
- D: número de eliminaciones de caracteres
- I: número de inserciones de caracteres
- N: número total de caracteres en la referencia

Se presentó:

Una tabla comparativa de WER y CER antes y después.

audio_id	WER_raw	WER_corr	CER_raw	CER_corr	WER_mejora_%	CER_mejora_%
1.0	0.29	0.25	0.20	0.20	15.79	-1.39
3.0	0.46	0.43	0.29	0.30	8.00	-6.59
5.0	0.40	0.44	0.22	0.24	-9.52	-7.94
7.0	0.60	0.58	0.40	0.39	2.70	3.23
9.0	0.16	0.15	0.09	0.06	11.11	34.62

Un gráfico de barras mostrando la mejora porcentual en precisión tras la corrección.

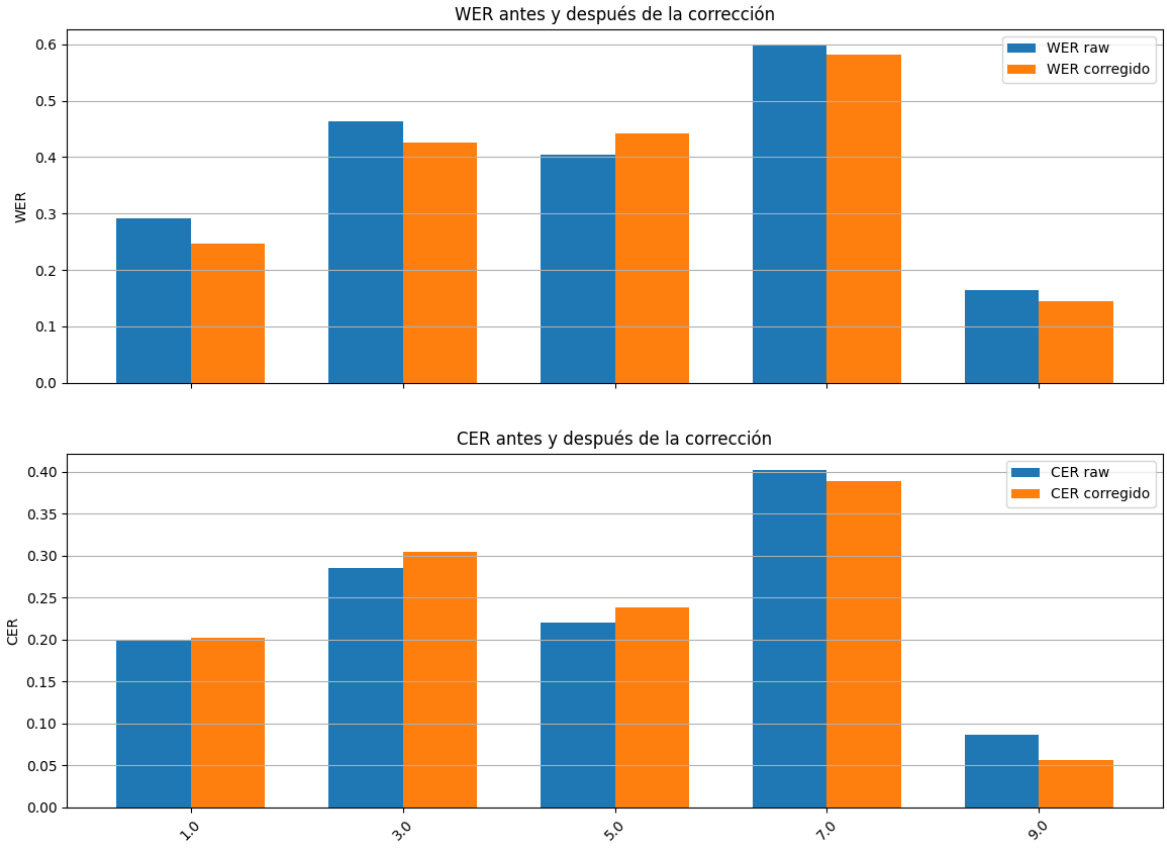

Tabla de 5 palabras más corregidas

palabra	# correcciones	# adiciones	# eliminaciones
a	3	3	0
I	2	1	1
go	2	0	2
Tomorrow	1	1	0
is	1	1	0

6. Insights

Se identificaron los 5 errores gramaticales y léxicos más frecuentes cometidos por los estudiantes en las transcripciones corregidas

Se propuso una regla de negocio:

“Si un estudiante comete el error X más de Y veces, se recomienda un ejercicio específico sobre Z”.

7. Funcionalidad adicional implementada

Se integró una funcionalidad opcional:

Diarización avanzada con WhisperX para mejorar la separación de turnos de habla.

Se exploró la posibilidad de retroalimentación pedagógica (corrección + sugerencia de ejercicios), que podría expandirse a un sistema automático de puntuación o recomendación.

Archivos entregados

notebook.ipynb: Contiene todo el flujo de trabajo (transcripción, diarización, corrección, evaluación).

requirements.txt: Lista de dependencias utilizadas.

transcript_raw.csv, transcript_gold.csv, transcript_corrected.csv: Archivos de resultados.

reporte.pdf: Descripción técnica, resultados, insights pedagógicos y propuestas de mejora.