UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

Graduado en Ingeniería de la Salud

Automatización del Resumen de Consultas Médicas: Conversión de Audio a Texto y Generación Automática de Informes a partir de Interacciones entre Médicos y Pacientes

Realizado por

Raúl Obrero Berlanga

Tutorizado por

José Manuel Jerez Aragonés
Francisco Javier Moreno Barea

Departamento

Lenguajes y Ciencias de la Computación

MÁLAGA, (Julio, 2024)

UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADO EN INGENIERÍA DE LA SALUD

Automatización del Resumen de Consultas Médicas: Conversión de Audio a Texto y Generación Automática de Informes a partir de Interacciones entre Médicos y Pacientes

Functional Physician Consultation Summary Automation: Audio-to-Text Conversion and Automatic Reporting from Doctor-Patient Interactions

Realizado por
Raúl Obrero Berlanga

Tutorizado por
José Manuel Jerez Aragonés
Francisco Javier Moreno Barea

Departamento
Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA

MÁLAGA, JUNIO DE 2024

Abstract

Keyboard slavery is a problem in medicine that must be abolished. Currently, healthcare professionals spend a significant amount of time during medical appointments typing data into a computer, rather than dedicating it to patients. This situation creates a disconnect between the doctor and the patient, decreasing the quality of healthcare. Additionally, it increases the administrative burden on healthcare professionals, which can lead to errors and lower satisfaction for both doctors and patients.

To address this issue, this work proposed the use and comparison of different transformer models in order to automate the documentation process. Transformer models, known for their ability to process and generate natural language, have the potential to transform the way clinical documentation is carried out. We can pass a medical audio dialogue into text and then generate a coherent and concise summary of relevant information. This not only saves valuable time, but also allows doctors to focus on what matters most: direct patient care and attention.

Furthermore, the implementation of this technology can significantly reduce digital fatigue and professional burnout, thereby improving the efficiency and job satisfaction of healthcare professionals. Ultimately, avoiding the need for doctors to perform this tedious process represents an improvement in the efficiency of the healthcare system and the quality of life of healthcare professionals.

Keywords: NLP, Transformers, Medical Report, Speech-to-text, summary

Resumen

Actualmente, los profesionales de la salud dedican una cantidad significativa de tiempo durante las citas médicas a escribir datos en un ordenador, en lugar de dedicárselo a los pacientes. Esta situación crea una desconexión entre el médico y el paciente, lo que disminuye la calidad de la atención médica. Además, aumenta la carga administrativa sobre los profesionales de la salud, lo que puede llevar a errores y a una menor satisfacción tanto para los médicos como para los pacientes.

Para abordar esta problemática, se propone utilizar y comparar distintos modelos transformer con el fin de automatizar el proceso de documentación. Los modelos transformer, conocidos por su capacidad para procesar y generar lenguaje natural, tienen el potencial de transformar la manera en que se lleva a cabo la documentación clínica. A partir de un audio grabado durante la consulta médica, estos modelos pueden realizar una transcripción precisa a texto y, posteriormente, generar un resumen coherente y conciso de la información relevante. Esto no solo ahorrará tiempo valioso, sino que también permitirá que los médicos se centren en lo más importante: la atención y el cuidado directo de los pacientes.

Además, la implementación de esta tecnología puede reducir significativamente la fatiga digital y el agotamiento profesional, mejorando así la eficiencia y la satisfacción laboral de los profesionales de la salud. En última instancia, evitar que el médico tenga que realizar este proceso tedioso representa una mejora en la eficiencia del sistema de salud y en la calidad de vida de los profesionales de la salud.

Palabras clave: PLN, Transformers, Informes médicos, Speech-to-text, Resumen

Índice

Índice de figuras

Índice de cuadros

1 Introducción

1.1. Motivación

Recientemente, las citas médicas han experimentado una gran transformación. Anteriormente, cuando las tecnologías no estaban tan avanzadas, había un trato más humano entre médico y paciente. Es decir, todo el tiempo de la consulta se dedicaba al paciente, priorizando la atención médica, el bienestar y la comunicación con el paciente.

Actualmente, esto ha cambiado considerablemente. Los médicos pasan más tiempo interactuando con sistemas informáticos y escribiendo en teclados que manteniendo una conversación directa con sus pacientes. Esto ha dado lugar a un nuevo término denominado "esclavitud del teclado", que se refiere al hecho de que los médicos pasan más tiempo escribiendo en un teclado los datos del paciente que prestando atención real a la persona.

Este cambio también lleva a una falta de optimización en las consultas médicas. En una época en la que los servicios médicos son altamente demandados, es crucial optimizar estas consultas para reducir su duración, minimizando el tiempo que los médicos dedican a la toma de datos y la escritura en sistemas informáticos. Alternativamente, se podría aprovechar mejor el tiempo de la consulta para enfocarse en el paciente, lo que podría reducir la necesidad de citas posteriores.

Además del problema que enfrentan los pacientes, los médicos también experimentan agotamiento debido a la gran cantidad de datos que deben ingresar en diferentes sistemas informáticos, lo cual consume mucho tiempo. Según una encuesta publicada por AthenaHealth[1], más del 90 por ciento de los médicos reportan sentirse regularmente agotados por la carga de trabajo administrativo.

A pesar de que los sistemas de registros electrónicos de salud han traído ventajas, como una mejor organización de los datos, también han generado preocupaciones sobre la pérdida

de la conexión humana en la atención médica.

Como solución a estos problemas, se propone la utilización de modelos Transformers para realizar las tareas cotidianas que los médicos deben llevar a cabo y que dificultan su trabajo, impidiéndoles realizarlo de la manera más eficiente posible. Desde la extracción de datos de informes médicos a través del procesamiento del lenguaje natural, la elaboración automática de informes, hasta la generación de informes sintéticos para su uso en el entrenamiento de otros modelos. Todas estas son actividades que un médico realiza durante y después de la cita. Al eliminar estas tareas, se podrían lograr consultas médicas más humanizadas y con un trato más cercano al paciente.

1.2. Objetivos

El objetivo fundamental que se persigue con la elaboración de este trabajo es la utilización de distintos modelos Transformers con usos específicos para que desde un simple audio que recoja la conversación de una cita médica entre un médico y un paciente se pueda pasar directamente a un resumen o informe médico de la cita, para solucionar los siguientes problemas que se han mencionado en el apartado anterior de motivación:

- Optimización de consultas médicas.
- Consultas médicas centradas en pacientes y no en datos.
- Problema de agotamiento en médicos.
- Investigación del estado del arte.
- Investigación y comparación de modelos Transformes tanto para pasar de audio a texto como resumen.

1.3. Estructura del documento

Este trabajo comienza introduciendo la motivación que ha fomentado su realización y los objetivos que se pretenden alcanzar con su desarrollo.

En la siguiente sección, se realizará un estudio del estado del arte en los ámbitos relacionados con este trabajo, incluyendo los modelos Transformers, los modelos de conversión de audio a texto y los modelos de resumen.

La sección siguiente expondrá la metodología empleada en el desarrollo del trabajo, los resultados obtenidos y la discusión de estos.

Finalmente, se concluirá con las conclusiones derivadas de la realización del trabajo y se presentarán posibles líneas futuras de investigación.

1.4. Tecnologías usadas

En esta sección se expondrán todas las tecnologías utilizadas en el desarrollo del trabajo.

- El código de programación ha sido totalmente escrito en el lenguaje de programación Python en el entorno online Google Colab.
- Los modelos transformer utilizados han sido empleados a través de pipelines que proporciona el repositorio Hugging Face.
- En cuanto a librerías del lenguaje de programación, se han utilizado las siguientes:
    - transformers
    - speechbox
    - pyannote.audio
    - speechbrain
    - kenlm
    - pyctcdecode
    - sacremoses
    - rouge-score
    - re

  Estas librerías se expondrán en siguientes secciones para detallar su uso.

2

Estado del Arte

La presente sección tiene como objetivo presentar la tarea de Conversión de conversaciones médicas a resúmenes médicos. Para ello, se realiza una revisión de los trabajos más relevantes en cuanto a modelos de conversión de audio a texto y la generación de resúmenes mono documento, analizando la evolución de las técnicas de estas disciplinas.

2.1. Modelos Transformers

Para comenzar, se dará una explicación de los modelos Transformers, ya que estos son los tipos de modelos que se utilizarán en este trabajo.

A principios de los años 2000, los modelos basados en redes neuronales comenzaron a reemplazar a los modelos previamente utilizados, que se enfocaban en técnicas estadísticas y de aprendizaje automático, como las máquinas de vectores de soporte (Support Vector Machines) y los árboles de decisión. Durante la década de 2010, el surgimiento del aprendizaje profundo [2] impulsó un gran desarrollo en los modelos de procesamiento del lenguaje natural (NLP), permitiendo manejar grandes cantidades de datos y aprender patrones complejos en el lenguaje natural.

2.1.1. Redes neuronales

Las redes neuronales [3] son un conjunto de algoritmos de aprendizaje automático inspirados en el funcionamiento del cerebro humano. Están compuestas por unidades básicas llamadas neuronas artificiales, organizadas en capas interconectadas. Estas redes son capaces de aprender a partir de datos y realizar tareas como el reconocimiento de patrones o la clasificación.

El funcionamiento de una red neuronal, en específico con una estructura feed-forward o arquitectura lineal, consta de las siguientes fases, las cuales se van a referenciar en la figura 1:

1. **Entrada de datos:** La red neuronal recibe datos de entrada como imágenes o texto, en el caso de la figura 1 estos datos de entradas son los nodos de color naranja.

2. **Propagación hacia adelante:** Los datos se propagan en la red neuronal a través de las capas ocultas (nodos verdes).

3. **Cálculo de activación:** En cada capa oculta, se calcula la activación de cada neurona mediante una combinación lineal de las entradas ponderadas por los pesos de conexión, seguida de la aplicación de una función de activación no lineal.

4. **Propagación hacia atrás:** Una vez que la salida de la red se ha calculado, se compara con la salida deseada y se calcula el error. Luego, este error se propaga hacia atrás a través de la red utilizando el algoritmo de backpropagation para ajustar los pesos de conexión y minimizar el error.

5. **Ajuste de pesos:** Los pesos de conexión entre las neuronas se actualizan iterativamente utilizando algoritmos de optimización como el descenso del gradiente.

6. **Entrenamiento y ajuste de hiperparámetros:** El proceso de ajuste de pesos se repite iterativamente a lo largo de múltiples épocas hasta que el modelo converja a una solución aceptable.

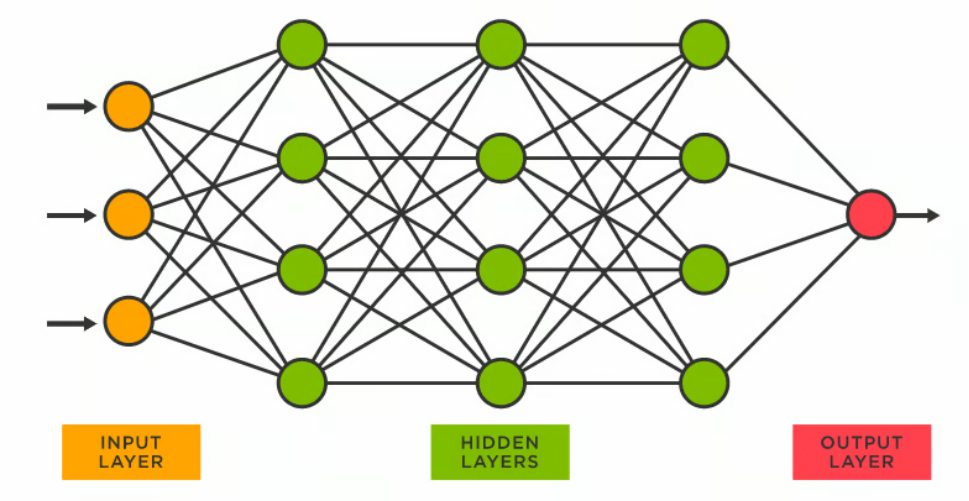

Figura 1: Ejemplo de red neuronal.

Hasta aquí, se ha explorado los conceptos fundamentales de las redes neuronales. Sin embargo, a pesar de los avances significativos que estas arquitecturas han permitido, presentan limitaciones en el procesamiento secuencial y en la captura de dependencias a largo plazo en los datos. Para abordar estas limitaciones y mejorar la eficiencia en tareas de procesamiento del lenguaje natural y otras aplicaciones, se desarrolló la arquitectura Transformer.

2.1.2. Arquitectura y funcionamiento

Estos modelos se basan en la estructura codificador-decodificador. El codificador se utiliza para capturar el contexto de la secuencia de entrada, mientras que el decodificador genera la secuencia de salida en función del contexto capturado. En tareas como la síntesis de texto, la secuencia de entrada y la de salida están estrechamente relacionadas.

Los componentes de la arquitectura de los modelos Transformers, como se ilustra en la figura 2, son los siguientes [4]:

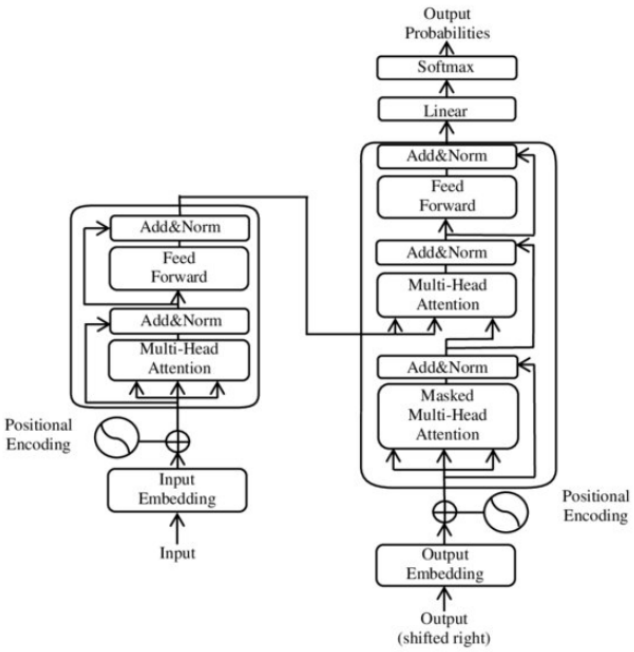

Figura 2: Arquitectura de modelo transformer.

- **Embeddings:** Transformación de palabras en vectores numéricos para operaciones matemáticas.

- **Codificaciones posicionales:** Añaden información sobre la posición de las palabras en la secuencia.
- **Autoatención:** Capacidad del modelo para comprender el contexto de una palabra en relación con otras palabras de la secuencia.
- **Atención por producto escalar:** Asignación de pesos de atención a cada palabra basados en la similitud entre la consulta y las claves.
- **Atención multi-cabeza:** Permite al modelo prestar atención a diferentes partes de la secuencia simultáneamente.
- **Capa Neuronal Feed-Forward (FFN):** Transformaciones lineales seguidas de funciones de activación no lineales para aprender representaciones más abstractas de la información de entrada.

El funcionamiento de los modelos Transformers se podría dividir en las siguientes etapas:

- **Entrenamiento de dos redes neuronales diferentes:** El modelo consta de un codificador (dos subcapas) y un decodificador (tres subcapas), entrenados para generar vectores distintos con las palabras de una secuencia de texto dada como input.
- **Codificación y autoatención en el codificador:** El codificador transforma las palabras en embeddings y añade codificación posicional sobre su posición en la secuencia. Luego, aplica el mecanismo de autoatención de múltiples cabezas para capturar patrones de dependencias a diferentes niveles de abstracción en la secuencia.
- **Transformaciones en la capa neuronal feed-forward:** Después de la autoatención, se aplican transformaciones oportunas en la capa neuronal feed-forward para la tarea en cuestión. La salida del codificador es un vector de representaciones continuas.
- **Funcionamiento del decodificador:** El decodificador recibe la salida del codificador y la información de todas las palabras procesadas anteriormente en la secuencia. Aplica el mecanismo de autoatención de múltiples cabezas, asegurando que cada posición preste

atención a todas las posiciones anteriores. Se utiliza enmascaramiento para evitar prestar atención a posiciones posteriores. Luego, se añade la salida del codificador para el momento actual y se repite el proceso de autoatención. Finalmente, se pasa por la capa feed-forward para obtener las probabilidades del siguiente token.

Dentro de los modelos Transformers, podemos encontrar modelos de lenguajes que se ha hecho más populares, entre ellos tenemos BERT, modelo de lenguaje pre-entrenado desarrollado por Google; GPT, Modelos diseñados específicamente para la generación de lenguaje natural; o RoBERTa, modelo de lenguaje desarrollado por Facebook AI Research [5].

2.2. Modelos de voz a texto

Un motor de conversión de voz a texto es un sistema que transforma el contenido de un archivo de audio en palabras escritas que pueden ser leídas en una pantalla o procesadas por una computadora. Esto implica que el motor de reconocimiento de voz debe incluir programas informáticos capaces de tomar el contenido de un archivo de audio y convertirlo en texto editable en un procesador de texto u otro tipo de programa que permita su visualización. En resumen, convierte lo que se escucha en palabras que se pueden leer y manipular electrónicamente.

Para lograr esta tarea con los modelos mencionados anteriormente, se deben superar varios obstáculos [6], entre los que se encuentran:

- **Bidireccionalidad:** Hay un intercambio de habla entre dos personas o mas.
- **Incompletitud:** La información intercambiada siempre es mayor a la oral, debido a gestos o expresiones no verbales.
- **Variabilidad:** No es posible que una persona pronuncie dos veces de igual manera una palabra.
- **Multiinteractividad:** Existen varios niveles de comprensión, que interaccionan dinámicamente entre sí.

- **Continuidad:** Ni las sílabas, ni los fonemas, ni la palabras son fáciles de dividir automáticamente.
- **Transitoriedad:** Sólo las variaciones de una señal permiten transmitir información.
- **Incertidumbre:** La información que se recoge a través de un dispositivo es información ruidosa.

Entre las redes neuronales más empleadas para la conversión de habla a texto, se destacan las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN).

- **Redes Neuronales Recurrentes (RNN):** Las RNN integran conexiones retroalimentadas, lo que les otorga una capacidad de 'memoria' al considerar las secuencias temporales de los datos. Esta característica las hace especialmente adecuadas para tareas de procesamiento del lenguaje natural, donde el contexto y el orden de las palabras son cruciales para una comprensión precisa.
- **Redes Neuronales Convolucionales (CNN):** Las CNN ejecutan cálculos matriciales utilizando filtros predefinidos, lo que les permite identificar patrones espaciales y características en los datos. Aunque las CNN son más conocidas por su eficacia en tareas de reconocimiento y clasificación de imágenes, también han demostrado ser útiles en el procesamiento de señales de audio, ya que pueden captar características acústicas importantes en fragmentos de audio.

Ambos tipos de redes neuronales se utilizan en sistemas de conversión de habla a texto, aprovechando sus respectivas fortalezas para mejorar la precisión y la eficiencia del reconocimiento de voz [7].

2.2.1. Evaluación de modelos

Según un estudio de Belenko [8] en el que se comparan herramientas de código abierto speech to text, se analizaron factores, los cuáles son muy favorables para la elección de una herramienta u otra. Estos factores son:

- **Word Error Rate:** Sigue la ecuación mostrada en la figura 3, donde la S es el número de sustituciones de palabras, D es el número de veces que se omite una palabra, I las veces que se ha introducido una palabra que no ha sido pronunciada, N la cantidad de palabrasde referecia y C el número de palabras correctas.
- **Speed factor:** Indica el tiempo que tarda el modelo en reconocer un audio.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

Figura 3: Word Error Rate.

También están las métricas ROUGE [9]. Estas métricas son un conjunto de métricas de evaluación utilizadas principalmente en tareas de resumen automático y generación de texto. Comparan un resumen generado automáticamente con uno o más resúmenes de referencia para determinar la similitud y calidad del resumen generado.

Dentro de las métricas ROUGE se encuntras varias variaciones en las que encontramos:

- **ROUGE-1:** Esta métrica calcula la similitud entre los unigramas (palabras individuales) presentes en el resumen generado y los presentes en el resumen de referencia. Es decir, cuenta la cantidad de palabras en común entre el resumen generado y el de referencia. ROUGE-1 se denota como ROUGE-N, donde N representa el tamaño del n-grama, y en este caso, $N = 1$.
- **ROUGE-2:** Similar a ROUGE-1, pero en lugar de comparar unigramas, ROUGE-2 compara bigramas, es decir, secuencias de dos palabras consecutivas. Calcula la similitud entre los bigramas del resumen generado y los del resumen de referencia.
- **ROUGE-L:** Esta métrica tiene en cuenta la similitud entre las secuencias de palabras más largas, lo que la hace más robusta en algunos casos. ROUGE-L mide la similitud entre

las secuencias de palabras más largas (hasta cierta longitud) en el resumen generado y en el de referencia. Esta métrica considera la longitud común más larga entre los dos resúmenes, lo que puede reflejar mejor la coherencia global del resumen.

2.2.2. Whisper como estado del arte

Whisper es un modelo de transcripción de voz a texto desarrollado por OpenAI. Actualmente, es el modelo más eficiente y que mejor funciona en este ámbito. Este modelo ha sido entrenado con un amplio conjunto de diversos audios. Se caracteriza por ser un modelo multitarea, ya que puede realizar tareas como reconocimiento del habla multilingüe, traducción del habla e identificación del lenguaje.

Al igual que varios modelos transformers, Whisper ofrece varios modelos diferentes, que incluyen Tiny, Base, Small, Medium y Large (ver Cuadro 1). Estos modelos difieren en eficiencia, velocidad y número de parámetros. Mientras que los modelos Tiny o Base tienen menos parámetros y son menos precisos, tienen tiempos de ejecución mucho menores, mientras que en modelos como el Large es lo contrario. Dependiendo de la tarea que se esté abordando, puede ser más eficiente utilizar un modelo u otro [10].

Cuadro 1: Modelos whisper

Modelo	**Capas**	**Parámetros**
Tiny	4	39M
Base	6	74M
Small	12	244M
Medium	24	769M
Large	32	1550M

El artículo referenciado en esta subsección también aborda la evaluación de la velocidad y efectividad de la transcripción de audio mediante la realización de pruebas con dos grabaciones de audio. Durante esta evaluación, se investigará el tiempo necesario para transcribir cada audio y se asignará una puntuación al texto obtenido en comparación con el contenido original

del audio, identificando los principales problemas que pueden surgir.

El principal problema encontrado es que al utilizar archivos de audio con una longitud relativamente extensa, de aproximadamente 11 minutos, la calidad de la transcripción y la longitud de las partes transcritas disminuyen significativamente, lo que resulta en la incapacidad de transcribir todo el audio de manera efectiva.

2.2.3. Otros modelos conocidos

Teniendo en cuenta a whisper como el modelo por excelencia de audio a texto, existen modelos de otras empresas tambien mundialmente conocidas. El gran problema de estos modelos frente a whisper es que estos modelos requieren conectarse a una aplicación web externa, vulnerando la privacidad de los datos, mientras que los modelos Whisper pueden ser ejecutados en máquinas locales, aunque requieren una capacidad de cómputo elevada, a continuación se especificará más de cada modelo.

Primero encontramos a Microsoft, con su Interfaz de Programación de Aplicaciones de Voz (SAPI), que ofrece herramientas para programadores que deseen crear aplicaciones para el sistema operativo Windows. SAPI automatiza el reconocimiento y síntesis de voz desde una aplicación, siendo accesible desde varios lenguajes de programación. Sin embargo, su limitación principal es que solo funciona para software en equipos de escritorio, aunque es de distribución gratuita y eficiente en la conversión de discurso a texto.

Por otro lado, Apple ofrece una API de reconocimiento de voz para iOS, permitiendo la traducción de voz a texto y la interacción por voz con dispositivos Apple, como Siri. Aunque ofrece soporte para más de 50 idiomas, tiene limitaciones en el número de traducciones por día y en la duración máxima de entrada de voz.

Google Cloud Speech proporciona una API para la traducción de voz a texto, con soporte para más de 80 idiomas y modelos de redes neuronales para una alta precisión. Sin embargo, el mejor motor de voz para cada aplicación dependerá de sus necesidades específicas, y hay

investigaciones que comparan diferentes opciones, como Sphinx y la API de Google.

Finalmente, Dragon NaturallySpeaking de Nuance ofrece una solución premium para el reconocimiento de voz, especialmente enfocada en aumentar la productividad en el sector empresarial. Aunque es potente y eficiente, su costo puede ser una limitación para algunos desarrolladores. Además, se está evaluando su uso en campos como la medicina, donde se compara con el dictado convencional en términos de eficacia y eficiencia [11].

2.3. Modelos de resumen

La generación automática de resúmenes de textos es una tarea del procesamiento de lenguaje natural que busca resumir el contenido de un documento manteniendo la información importante en un texto de una extensión menor. Una definición podría ser ”la creación de una breve pero exacta representación del contenido de un documento”.

Existen varias clasificaciones de la generación automática de resúmenes de documentos:

- **Técnica extractiva o técnica abstractiva**: La técnica extractiva selecciona y copia secuencias de palabras (frases, oraciones o párrafos) directamente del documento original, aunque puede presentar problemas de consistencia y coherencia. La técnica abstractiva genera secuencias de palabras que no necesariamente están en el documento original, y aunque presenta mayores desafíos computacionales, requiere técnicas avanzadas de generación de lenguaje.
- **Enfoque superficial o enfoque profundo**: Los enfoques superficiales representan documentos usando características básicas, como términos estadísticamente relevantes o posicionalmente destacados, y los extraen para el resumen. Los enfoques profundos utilizan técnicas avanzadas de procesamiento de lenguaje natural y análisis semántico, como tesauros y relaciones sintácticas, y pueden producir resúmenes tanto extractivos como abstractivos.
- **Propósito del resumen**: Los resúmenes pueden ser indicativos, informativos o críticos. Los indicativos proporcionan una visión abreviada de los principales temas del

documento, ayudando al usuario a decidir si leerlo (tamaño entre 5-10 % del texto original). Los informativos buscan sustituir al documento original, reteniendo información importante y detallada (tamaño entre 20-30 % del original). Los críticos, como revisiones y reseñas, capturan el punto de vista del autor del resumen, pero actualmente están fuera del alcance de la generación automática de resúmenes.

- **Audiencia del resumen**: Los resúmenes pueden ser genéricos, basados en consultas o enfocados en el usuario o tópicos. Los genéricos dan igual importancia a los tópicos principales del documento, dirigidos a una audiencia amplia. Los basados en consultas generan resúmenes en función de una pregunta específica. Los enfocados en el usuario o tópicos dan prioridad a las necesidades específicas del usuario o a ciertos tópicos, encuadrándose en un paradigma más ajustado a la recuperación de información.

Otras características importantes incluyen el número de documentos procesados, pueden ser monodocumentos o multidocumentos, el idioma del documento, monolenguaje o multilenguaje y el tipo de documentos resumidos, científicos, noticias, blogs, etc. [12]

2.3.1. Modelos extractivos

El enfoque extractivo selecciona y extrae frases o partes de ellas del texto original. Su principal ventaja es su robustez y facilidad de aplicación en contextos de propósito general, debido a su alta independencia del dominio y del género de los documentos. Normalmente, este proceso consiste en identificar las sentencias relevantes del texto original, reducir la redundancia y puntuar las sentencias basándose en varias características. Las sentencias con las puntuaciones más altas son extraídas y presentadas en el orden de aparición original. Para este propósito, se utilizan comúnmente modelos de puntuación o ranking basados en grafos, los cuales determinan la importancia de un vértice dentro del grafo usando información referencial global obtenida recursivamente.

Para el procesado semántico hay trabajos en los que los modelos de resumen se apoyan

en ontologías, como en el caso del artículo referenciado a continuación [13], el cual se apoyan en la UMLS, lo cual se hace con el objetico de desarrollar herramientas que ayuden al investigador en la representación del conocimiento e integración de información biomédica [14]. Se pueden dividir los algoritmos basados en extracción en 3 grupos, los cuales son algoritmos supervisados, semi-supervisados y no supervisados.

El principal inconveniente de las técnicas basadas en extracción es la posible inconsistencia en los resúmenes debido a la extracción de elementos aislados sin considerar las relaciones entre ellos. Esto puede provocar problemas como referencias anafóricas a elementos no incluidos o la falta de coherencia cuando las oraciones extraídas no aparecían de manera consecutiva en el texto original. Además, los resúmenes pueden resultar desequilibrados al no considerar todos los aspectos importantes del documento o su organización estructural. Este desequilibrio se refiere a la omisión de temas importantes o aspectos estructurales del documento original, como el propósito, los procedimientos y las conclusiones en un artículo de investigación [15].

Normalmente, se suelen llevar a cabo cuatros pasos en este tipo de técnicas: selección de términos, pesado de términos, pesado de oraciones y selección de oraciones. Sin embargo, esta selección de oraciones consiste en tomar las oraciones con mayor peso. Esto conlleva a que esta técnica funcione bien en las primeras oraciones, pero es probable que otras oraciones similares a las que ya han sido elegidas sean también seleccionadas, lo cual produciría redundancia [16].

2.3.2. Modelos abstractivos

Para comenzar, se puede definir a la abstracción como el proceso de reformular un contenido en otros términos. Fusión implica combinar porciones extraídas y compresión consiste en ignorar información irrelevante. Según estas definiciones, cualquier enfoque que no utilice extracción se clasifica dentro de la abstracción.

Los modelos de resumen abstractivo representan una de las técnicas más avanzadas en el campo del resumen automático. A diferencia de los métodos extractivos que simplemente seleccionan y combinan frases o segmentos del texto original, los modelos abstractivos generan

nuevas frases que pueden no estar presentes explícitamente en el documento fuente. Esto permite crear resúmenes más coherentes y naturales, capturando mejor la esencia del contenido original [17].

El funcionamiento de los modelos de resumen abstractivo se basa en dos componentes principales: el codificador (encoder) y el decodificador (decoder). Este enfoque es típico de las arquitecturas de secuencia a secuencia (sequence-to-sequence) utilizadas en diversas tareas de procesamiento del lenguaje natural. Estos encoder y decoder ya se han mencionado anteriormente en la explicación de los modelos Transformers.

- **Codificador:** El codificador toma el documento fuente y lo convierte en una representación interna (vectorial) a través de múltiples capas de redes neuronales. Esta representación captura la información semántica y sintáctica del texto original.
- **Decodificador:** El decodificador utiliza esta representación interna para generar el resumen palabra por palabra. En cada paso, el decodificador predice la siguiente palabra en la secuencia basándose en la representación interna y en las palabras generadas previamente.

Durante el entrenamiento, los parámetros del modelo se ajustan para maximizar la probabilidad condicional del resumen dado el texto original. Este proceso se realiza utilizando un corpus paralelo de documentos y sus correspondientes resúmenes.

- **Atención (Attention):** El mecanismo de atención es crucial en los modelos abstractivos. Permite al decodificador enfocarse en diferentes partes del documento fuente mientras genera cada palabra del resumen. Esto mejora la capacidad del modelo para capturar y recombinar la información relevante de manera más efectiva. Esta atención también se ha mencionado anteriormente en los modelos Transformers.
- **Modelos de Copia:** Los modelos de copia son una extensión importante que permite al decodificador copiar palabras directamente del texto fuente. Esto es especialmente útil

para manejar nombres propios, cifras y terminología específica que deben preservarse en el resumen.

- **Modelos de Cobertura:** Para evitar la repetición innecesaria de palabras, se utilizan modelos de cobertura que monitorean las partes del texto que ya han sido cubiertas durante el proceso de generación del resumen. Esto mejora la fluidez y coherencia del texto generado.

- **Señales de Guía:** Además del documento fuente, se pueden incorporar señales adicionales de guía (guidance signals) que proporcionan información complementaria al modelo. Estas señales pueden incluir palabras clave, plantillas, o información extraída de otros documentos relevantes, mejorando así la precisión y relevancia del resumen.

Como gran ventaja de estos modelos se puede mencionar que pueden generar resúmenes más coherentes y naturales en comparación con los métodos extractivos. Son capaces de parafrasear el contenido y combinar información de diferentes partes del texto, proporcionando una visión más integrada del contenido original.

A pesar de sus ventajas, los modelos abstractivos enfrentan varios desafíos:

- **Complejidad Computacional:** Los modelos abstractivos requieren una gran cantidad de datos y poder de cómputo para entrenarse efectivamente.

- **Fidelidad al Texto Original:** Es difícil asegurar que el resumen generado sea siempre fiel al contenido del documento fuente, lo cual puede resultar en la introducción de errores o interpretaciones incorrectas.

- **Conocimiento de Dominio:** La calidad del resumen puede verse afectada por la falta de conocimiento específico del dominio, lo que puede limitar la capacidad del modelo para generalizar correctamente.

2.3.3. Resumen de conversaciones

Modelos como BART y GPT-3, con su gran cantidad de parámetros, muestran un rendimiento excepcional en diversas tareas de propósito general. No obstante, su entrenamiento se basa principalmente en recursos de conocimiento como libros, documentos web y artículos académicos, y a menudo requieren datos adicionales específicos de dominios, como conversaciones o diálogos, para entender mejor los diálogos. La falta de conjuntos de datos apropiados disponibles públicamente crea un desafío para generar resúmenes abstractivos.

En cuanto a conversaciones médicas, los avances recientes en el resumen automático de diálogos médicos han impulsado significativamente el campo. Los modelos LSTM y de Transformers han demostrado ser capaces de generar resúmenes concisos de una sola frase a partir de conversaciones entre médicos y pacientes. Además, se han utilizado modelos de Transformers preentrenados para resumir directamente dichas conversaciones a partir de transcripciones.

El modelo jerárquico codificador-etiquetador ha surgido como un enfoque prometedor, produciendo resúmenes al identificar y extraer enunciados significativos, principalmente enfocados en declaraciones de problemas y recomendaciones de tratamiento. Sin embargo, es importante señalar que estos modelos se entrenan típicamente en conversaciones breves y generales entre médicos y pacientes, mientras que las conversaciones en el ámbito sanitario real tienden a ser más largas y detalladas. Entender las sutilezas de los patrones de comportamiento y pensamiento se vuelve crucial para la identificación precisa de enfermedades en tales contextos [18].

3
Desarrollo

En esta sección se desarrollará todos los aspectos relacionados con la metodología, exposición de resultados y su posterior discusión.

3.1. Metodología

Primero de todo, para la realización del trabajo es necesario audios como punto de partida. Para ello, se han obtenido varios audios de lectura de breves informes médicos, y un audio de una conversación entre médico y paciente en una consulta médica de una extensión más extensa, todos relacionados con el ámbito de la oncología. Se ha obtenido el consentimiento firmado del paciente para la experimentación realizada en el trabajo desarrollado de los audios en cuestión.

A continuación, se explican las etapan de las que ha constado el desarrollo del trabajo:

- **Pruebas Iniciales con Audios Breves:** Primero se realizaron pruebas con audios de una extensión breve para el paso de audio a texto. Se utilizaron modelos como Whisper [19] en sus modelos base y medium. Whisper es un modelo de transcripción de voz a texto desarrollado por OpenAI, caracterizado por su eficiencia y capacidad multitarea. También se utilizó otro modelo de audio a texto llamado jonatasgrosman/wav2vec2-large-xlsr-53-spanish [20], que ofrece una alternativa al modelo Whisper.

- **Pruebas con Audio de una Conversación Médica:** Una vez realizadas las pruebas con los primeros audios de la lectura de informes médicos y con el conocimiento de cómo funcionan los modelos, se utilizó el audio real de una cita médica y se probó estos modelos con él. Además de Whisper, se probó un pipeline que combina tanto Whisper como un modelo de diarización para distinguir los hablantes y pasar a texto una conversación en vez de texto plano.

- **Comparación y Evaluación de Modelos:** Posteriormente, se compararon los modelos de resumen a través de métricas Rouge1, Rouge2 y RougeL. Para ello, se generó un texto prototipo a mano del audio de la conversación y se eliminaron los nombres de los interlocutores para evitar sesgos. Se utilizó una función para eliminar palabras repetidas y se evaluaron los resúmenes generados por cada modelo.

- **Traducción del Texto Resultante:** Una vez seleccionado el mejor modelo de resumen, se tradujo su texto resultante al inglés utilizando el modelo Helsinki [21]. Debido a las restricciones de longitud de los modelos Transformers, se dividió el texto para evitar problemas y se tradujeron cada parte antes de unirlo nuevamente.

- **Resumen Automático:** Una vez traducido el texto, se aplicaron varios modelos de resumen automático. Entre ellos se encuentran:

    - **facebook/bart-large-cnn** [22]: Este modelo es un codificador-decodificador de transformador (seq2seq) con un codificador bidireccional similar a BERT y un decodificador autorregresivo similar a GPT. Es capaz de generar resúmenes coherentes y precisos.
    - **philschmid/bart-large-cnn-samsum**: Este modelo es muy similar al de Facebook (BART), también se basa en la arquitectura codificador-decodificador de transformador (seq2seq) y está especialmente diseñado para generar resúmenes de alta calidad.
    - **Saurabh91/medical-summarization-finetuned-starmpccAsclepius-Synthetic-Clinical-Notes**: Este modelo ha sido ajustado finamente en notas clínicas, lo que lo hace especialmente adecuado para la tarea de resumir textos médicos. Su entrenamiento específico en este dominio le permite capturar de manera más precisa la información relevante en los resúmenes.
    - **Falconsai/medical-summarization**: Este modelo T5 Large está optimizado específicamente para la tarea de resumir textos médicos. Utiliza una arquitectura de transformers para generar resúmenes precisos y contextualmente relevantes en el ámbito de la medicina.

- **google-t5/t5-large** [23]: Este modelo se propone para replantear todas las tareas de PNL en un formato unificado de texto a texto. A diferencia de los modelos estilo BERT, que solo pueden generar una etiqueta de clase o un tramo de la entrada, T5 Large puede generar texto completo como salida, lo que lo hace muy adecuado para la generación de resúmenes.

- **Proceso de Resumen:** Se generó una función summarizer para variar los modelos de resumen utilizados. Luego, se dividió el texto para evitar problemas de extensión de los modelos Transformers, resumiendo cada subconjunto del texto y finalmente uniendo los resúmenes generados.

- **Evaluación Final y Selección del Mejor Modelo:** Al carecer de un resumen prototipo y al variar los resúmenes en función de cada médico, se realizó una comparación personal de la fluidez del resumen y los aspectos médicos. Se creó un ranking con los mejores y peores modelos para el resumen. Una vez seleccionado el mejor modelo, se tradujo su resumen nuevamente al español con el modelo Helsinki.

3.2. Resultados

Primero, se van a presentar los resultados de la transcripción de audio a texto de los primeros audios de menor duración. Para facilitar la comparación, se ha seleccionado aleatoriamente uno de estos audios, y se han generado transcripciones utilizando diferentes modelos de reconocimiento de voz. A continuación, se muestran los resultados obtenidos con los modelos Whisper en sus versiones base y medium, así como con el modelo jonatasgrosman/wav2vec2-large-xlsr-53-spanish.

Los resultados se detallan en los cuadros 2, 3 y 4.

Cuadro 2: Modelo Whisper Base

Modelo Whisper Base:
Paciente de 68 años de anunistica del 1988 a los 25 años, de un carcinoma de ustal infiltrante de mama izquierda, de 2 de 2 centímetros, N1, un ganglio positivo de 6, tratada mediante más septomía izquierda a tipo Maden, en 1988, posteriormente, recibió que un bioterapia llevante con antraciclinas 6 ciclos, sin estudio de receptores hormonales ni de R2. de amnosticada en noviembre de 2023 a los 68 años, de un carcinoma total infiltrante de cuadrante superhistor no de mama derecha, G3, PT1B, 10 milímetros, gánilos negativos 0 de 3, M0. Estadío 1A, con invasión lifo mascular, receptores de estrógeno 10 % receptor de progesterone a 0 por ciento, ER2 negativo a 0, 15, 67, 60 por ciento, Tratada mediante tu morectomía y visóxia seletiva de gangliocentinelago. Se realizó un prosina que mostró un subtipo basal-leg con riesgo de residuiva intermedio y una probabilidad de residuiva a distancia de 10 años de ante por ciento. Dado que se trataba de un subtipo basal-leg se decide tratamiento ayubante con quimioterapia, es cremate C por cuatro ciclos. y

Cuadro 3: Modelo Whisper Medium

Modelo Whisper Medium:
Paciente de 68 años, diagnosticada en 1988 a los 25 años, de un carcinoma de octal infiltrante de mama izquierda T2 de 2 centímetros, N1, un ganglio positivo de 6, tratada mediante mastectomía izquierda tipo MADEN en 1988. Posteriormente recibió quimioterapia ayuvante con antraciclinas 6 ciclos, sin estudio de receptores hormonales ni DR2. diagnosticada en noviembre de 2023 a los 68 años de un carcinoma de utal infiltrante de cuadrante super externo de mama derecha G3-PT1B 10 mm ganglios negativos 0 de 3 M0, estadio 1A con invasión lifón vascular, receptores de estrógeno 10 % receptor de progesterona 0 %, ERDOS negativo 0 %, QI67 60 %, tratada mediante tumorectomía y visuóxida selectiva de ganglio-centinela. Se realizó un proxina que mostró un subtipo basal like con riesgo de recidiva intermedio y una probabilidad de recidiva a distancia 10 años de 11 %. Dado que se trata de un subtipo basal like se decide el tratamiento Ayuante con quimioterapia, Scrematoc por cuatro ciclos.

Cuadro 4: Modelo jonatasgrosman/wav2vec2-large-xlsr-53-spanish

Modelo jonatasgrosman/wav2vec2-large-xlsr-53-spanish:
Paciente de sesenta y ocho años diagnosticada mil novecientos ochenta y ocho a los veinticinco años de un carcinoma de tal infiltrante de mama izquierda de dos de dos centímetros en un ganglio positivo de seis tratada mediante más cemi izquierda tipo made en mil novecientos ochenta y ocho posteriormente recibió quimioterapia ayudante con antraciclina seis ciclos sin estudio de receptores hormonales ni dedos diagnosticada en noviembre de dos mil veintitres a los sesenta y ocho años de un carcinoma de tal infiltrante de cuadrante supersterno de mama derecha gtres pub diez milímetros ganglios negativos cero de tres mcro estadio una con invasión lifovascular receptores de estrógeno diez porciento receptor de profesorón a cero por ciento heridos negativo cero que sesenta y siete sesenta por ciento tratada mediante tumorectomía biseca selectiva ganglio centinela se realizó un prosina que mostró un subtipo vasalla con riesgo de recidiva intermedio y una probabilidad e recidiva distancia diez años de nceporciento dado que se trata de un subtipo basales decide tratamiento ayudante con quimioterapia extrema tec por cuatro ciclos.

Una vez dado esto, se mostrarán los resultados dados con el audio de la cita médica. Para ellos se mostrarán los resultados con el modelos whisper medium tal cuál, con el modelo whisper medium con el modelo diarization, y con el modelo whisper-large-v3. Debido a la extensión del audio, el texto será muy extenso, por lo que para la muestra se estos resultados se mostrará un fragmento del principio del texto. Estos resultados se muestra en los cuadros 5, 6, 7.

A continuación, se presentan los resultados de las métricas ROUGE en el cuadro 8, comparando con el texto prototipo. Se han evaluado tanto los modelos que utilizan diarización como

Cuadro 5: Modelo Whisper Medium

Modelo Whisper Medium:
¿Qué tal está? He visto que se cayó, se pegó un trompazo, que vino urgencias con un traumatismo que la encefálico que le hice en un talque, se pegó un guarazo. Bueno, pues el que me pegué ahora está de vacaciones y en la manga, que más o menos como está, me he pegado otro. ¿Y cómo se cae? Yo no lo sé, porque yo para mí es el torpídeo. La otra vez fue en un bordillo y esta vez ha sido en otro bordillo un poco más alto. El otro que he hecho, pues no lo he hecho bien. Yo no lo sé. ¿O no? Pues ya va. ¿Tota que acabó en el suelo? Que acabó en el suelo. La otra vez fue igual, con un bordillo. No sé. Yo tengo el oído, que lo tengo mal, que me ha dicho el médico. Sí, he visto ya que estuvo también en el lotorrino. que fue de... Por después del golpe. Pone que debió ir con su marido a revisión, a los torrinos, y le debió decir que le molestaba el oído y le vio en el momento, como fue. Sí, sí. Me mandó el hombre un... Mi tratamiento fue más amable, más... Y lo pone, pone paciente que viene acompañando. Ay, qué gracioso. Es que muy amable, fue muy amable. Me miró muy bien. Y me dijo que tenía moco detrás del tímpano. Sí. Me mandó un tratamiento, pero no me dijo seguridad de que le mandó un tratamiento a ver si la mejora y me mandó un poquito de cortisona para que no seme infectase.

los mismos modelos pero pasándolos por la función para eliminar la repetición de palabras. En la tabla, “P” significa precisión, “R” recall y “F” F-measure.

Una vez comparados estos modelos con las métricas ROUGE, el texto del modelo Whisper-v3-large, procesado por la función para eliminar la repetición de palabras consecutivas, se traduce. A continuación se muestra un fragmento en el cuadro 9:

Por último, se aplicaron los distintos modelos de resumen mencionados. Los resultados se muestran en los cuadros 10, 11, 12, 13 y 14.

Cuadro 6: Modelo Whisper Medium + diarization

Modelo Whisper Medium + diarization:
SPEAKER_03: ¿Qué tal está? He visto que se cayó, se pegó un trompazo, que vino urgencias con un traumatismo que la encefálico que le hice en un talque, se pegó un guarazo. Bueno, pues el que me pegué *SPEAKER_02: ahora está de vacaciones y en la manga, que más o menos como está, me he pegado otro. pero y cómo se ha vuelto y cómo se está? yo no lo sé porque yo para mí he hecho el pibia* *SPEAKER_03: la otra vez fue en un bordillo y esta vez ha sido en otro bordillo un poquito más alto* *SPEAKER_02: el otro que he hecho pues no lo he hecho bien, yo no lo sé o no lo he puesto bien* *SPEAKER_03: la otra vez fue igual con un bordillo* *SPEAKER_02: no sé, yo tengo el oído que lo tengo mal que me ha dicho el médico, perdón.* *SPEAKER_03: Sí, he visto ya que estuvo también en el lotorrino, que fue de...* *SPEAKER_02: Por después del golpe.* *SPEAKER_03: Pone que debió ir con su marido a revisión al lotorrino y le debió decir que le molestaba el oído y le vio en el momento.* *SPEAKER_02: Sí, sí, me mandó el hombre un... mi tratamiento fue más amable, más...* *SPEAKER_03: Sí, y lo pone, pone paciente que viene acompañándose.* *SPEAKER_02: Ay, qué gracioso, es que muy amable, fue muy amable, me miró muy bien. Y me dijo que tenía moco detrás del tímpano. Sí. Me mandó un tratamiento, pero no me dio seguridad de que... dice, le voy a mandar un tratamiento a ver si la mejora, y me mandó un poquito de cortisona para que no se me infecta se*

3.3. Discusión

En esta sección se discuten los resultados presentados en la subsección anterior.

Para empezar, se comparan los textos resultantes del primer audio. Se destaca que, para la extensión reducida de este audio, el modelo Whisper Medium proporciona una transcripción bastante precisa y coherente, con un buen uso de la puntuación (Cuadro 3). En contraste, los

Cuadro 7: Modelo Whisper-large-v3 + diarization

Modelo Whisper-large-v3 + diarization:
SPEAKER_03: ¿Qué tal está? He visto que se cayó, se pegó un trompazo, que vino urgencias con un traumatismo clarencefálico, que le hicieron un tal, que se pegó un guarrazo.
SPEAKER_02: Bueno, pues al que me pegué, ahora he estado de vacaciones, y en la manga, de más o menos que hemos estado, me he pegado otro.
SPEAKER_03: Se ha vuelto a caer, pero ¿y cómo se cae?
SPEAKER_02: Yo no lo sé Porque yo Para mí he hecho el pibio
SPEAKER_03: La otra vez fue en un bordillo
SPEAKER_02: Y esta vez ha sido en otro bordillo un poquillo más alto El otro que eché, pues no lo eché bien Yo no lo sé
SPEAKER_03: Total, que acabó en el suelo
SPEAKER_02: La otra vez fue igual, con un bordillo No sé Yo tengo el oído este, lo tengo mal Que me ha dicho el médico
SPEAKER_03: Sí, he visto ya que estuvo también en el otorrino, que fue de... Por después del golpe. Pone que debió ir con su marido a revisión al otorrino y le debió decir que le molestaba el oído y le vio en el momento.
SPEAKER_02: Sí, sí, me mandó el hombre un... mi tratamiento fue mamable.
SPEAKER_03: Sí, y lo pone, pone paciente que viene acompañándonos.
SPEAKER_02: Ay, gracias, es que es muy amable, fue muy amable, me miró muy bien. Y me dijo que tenía moco detrás del tímpano. Sí. Me ha mandado un tratamiento, pero no me dio seguridad de que... dice, le voy a mandar un tratamiento a ver si la mejora y me mandó un poquito de cortisona para que no se me infecta se

modelos Whisper Base y Jonatas muestran resultados menos eficientes, con frases sin sentido y errores ocasionales (Cuadros 2 y 4). Por lo tanto, en este caso, el modelo Whisper Medium parece ser la mejor opción.

A raíz de esta comparación, se selecciona el modelo Whisper Medium para la transcripción del audio de la cita médica. Además, para clasificar los hablantes, se compara este modelo con Whisper Large utilizando diarización. Entre los modelos Whisper Medium, se obtienen resultados similares (Cuadros 5 y 7), pero la diarización hace que la conversación sea más clara

	ROUGE								
Modelo	**ROUGE-1**			**ROUGE-2**			**ROUGE-L**		
Métrica	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Con repeticiones									
Medium	0.693	0.774	0.731	0.522	0.583	0.551	0.614	0.686	0.648
Large	**0.737**	**0.765**	**0.750**	**0.566**	**0.587**	**0.576**	**0.666**	**0.691**	**0.678**
Sin repeticiones									
Medium	0.694	0.771	0.731	0.523	0.581	0.550	0.615	0.683	0.648
Large	**0.831**	**0.764**	**0.796**	**0.637**	**0.586**	**0.610**	**0.750**	**0.690**	**0.719**

Cuadro 8: Resultados de las evaluaciones ROUGE

visualmente. En la comparación entre los modelos Whisper Large con diarización, se observa que Whisper Large produce resultados más precisos y coherentes (Cuadro 7). Sin embargo, se observa que estos modelos a veces fallan con repeticiones de palabras comunes, como ”si si”, lo que lleva a utilizar una función para eliminar estas repeticiones y comparar a través de las métricas Rouge.

Los resultados de las métricas ROUGE (Cuadro 8) muestran que la eliminación de repeticiones de palabras sucesivas aumenta las puntuaciones, especialmente para el modelo Whisper Large. Por lo tanto, se decide utilizar este último modelo con la eliminación de repeticiones para la generación de resúmenes.

A continuación, se comparan los resúmenes generados por cada modelo.

En el primer resumen (Cuadro 10), se destacan aspectos importantes como un traumatismo craneoencefálico y un problema en el oído, pero también incluye detalles irrelevantes como ir a Roquetas. El resumen tiene cohesión y se presenta desde una tercera persona.

En el segundo resumen (Cuadro 11), se recogen aspectos relevantes pero hay una alternancia entre la primera y tercera persona que puede afectar a la coherencia.

En el tercer resumen (Cuadro 12), se observa que es incompleto y carece de información relevante.

El cuarto resumen (Cuadro 13) tiene una extensión excesiva y parece ser una extracción de frases en lugar de un resumen.

En el quinto resumen (Cuadro 14) se observa una alternancia entre la primera y tercera

SPEAKER_03: How is he? I saw that he fell, hit a horn, that emergency came with a clarencephalic trauma, that they made him one, that he hit a slut.
SPEAKER_02: Well, to the one who hit me, now I have been on vacation, and in the sleeve, of more or less that we have been, I have hit another.
SPEAKER_03: He has fallen again, but how does he fall?
SPEAKER_02: I don't know because for me I have done the pipe
SPEAKER_03: The other time was on a curb
SPEAKER_02: And this time it was on The other one I threw out, because I didn't throw it right I don't know
SPEAKER_03: Total, which ended up on the floor
SPEAKER_02: The other time it was the same, with a curb I don't know I have this ear, I have it wrong That the doctor
SPEAKER_03: Yes, I have seen that it was also in the earpiece, that it was from... After the blow. It says that it must have gone with her husband to check the ear and must have told him that it bothered his ear and saw him at the time.
SPEAKER_02: Yes, yes, the man sent me a... my treatment was breast-feeding.
SPEAKER_03: Yes, and it puts him patient who comes with us.
SPEAKER_02: Oh, thank you, he's very kind, he was very kind, he looked at me very well. And he told me that he had mucus behind the eardrum

Cuadro 9: Conversation Transcript

persona.

En conclusión, el modelo philschmid/bart-large-cnn-samsum parece ser el más eficaz para resumir en este caso, considerando la coherencia y relevancia de la información proporcionada.

Cuadro 10: philschmid/bart-large-cnn-samsum

philschmid/bart-large-cnn-samsum:
Roquetas is going to Roquetas on the 10th of May. Her husband fell and hit a horn. He has a clarencephalic trauma. She has a problem with her ear. The doctor sent her a treatment, but he didn't give her any assurance that it would help. She is going back to the otorrino on the 16th. The tag has claimed it and she will claim it. Azumel's arm is very sore. Maria is going to give him treatment. Maria gives Azumel a phone number: 661 58 80 56 vale. Dani can't come in with her mother, because it's a very small place with many patients. Maria managed to park for her mother. Mary has a mammogram. She has an Ultrasound of the heart on May 22nd and a follow-up appointment on May 1st. She is asphyxiated and tired walking. She will bring Maria something after the trip.

Cuadro 11: facebook/bart-large-cnn

facebook/bart-large-cnn:
I have been on vacation, and in the sleeve, of more or less that we have been, I have hit anotherHe has fallen again, but how does he fall? I don't know because for me I have done the pipeI have instability because I also have my hands asleep. I sometimes drop the knife"Dani asks Maria if her husband is okay. Maria says he is fine, but his arm is very sore. Maria gives Dani a phone number for Maria. Dani asks if she can come in with her mother, but Maria says there is no room for family members. CNN.com spoke to a woman who had a mammogram at the hospital. The woman's mother asked if she could see the results of the mammogram. The doctor told her to claim the TAC instead of giving her for three treatments. After the trip, right? after the trip. I will bring you something. Thank you Maria.

Cuadro 12: Saurabh91/medical_summarization-finetuned-starmpccAsclepius-Synthetic-Clinical-Notes

Saurabh91/medical_summarization-finetuned-starmpccAsclepius-Synthetic-Clinical-Notes:
I have been seen by an otorrino in urgency, that he The answer to this question is to ask the doctor to give it treatment. The answer is to The thing that is tremendous is that the lady is tremendous. But tremendous. _02:

Cuadro 13: Falconsai/medical_summarization

Falconsai/medical_summarization:
_03: How is he? I saw that he fell, hit a horn, that emergency came with a clarencephalic trauma, that they made him one, he hit an slut. _02: Well, to the one who hit me, now I have been on vacation, and in the sleeve of more or less that we have been, I have hit another. this time it was on the other one I threw out, because I didn't throw it right. _02: Okay. _03: Come on, the otorrino who saw me in urgency, because he saw me with my husband later. it is that in urgency they were bad, to me they checked very well _01: and this man already checked me better than in the summer. this man said then go until the 16th day to see how he finds it, I believe that I have already heard _04: and the tag has claimed it. now what we are going to do, instead of giving it to, it's going to put on today's and another. _03: Okay. Is your husband from what hasn't come? Oh, help me. Help me, please. The arm, the arm that can't either up or back. I also have to go. Nothing happens. _02: Her husband is fine, what happens is that there is no parking lot. I've got that whole part very sore, but it's that one blow. _02: That's it. Cholesterol a high hair. _03: Yes? _01: Yes, but very little thing, don't worry. I didn't know, I hadn't given it. but as I said, I'm going to give it treatment. it is so dark all this of everything is state more like the view only of the party _01: can I come in with my mother? _03: No, you can't come in Dani because it's a very small place. then there are all the patients very... They're sitting, come on, they're in armchairs or they've been in beds but there's no room for family members to be there and then there are the nurses out there working so the family members can not come in that their thing would be that they could come in is that the friend's husband is coming. _02: My mother is tremendous the thing Thank you. _02: What do I have to do? Well, before, but from the knee I left, I've always done the exercise. _03: Well that would have to go back and see there for your house if there will be anything from the town hall or some story like that. here for the grown-ups we have a passive gymnastics but very good because we move everything. well that's the one that has to be pointed out, the two that sign up. it's okay _01: to make him decide for the Ultrasound of the heart we see on May 22 the next time I would touch it would be on May 1 but as it is party Yes I have put it on the 2 I put it in it.

Cuadro 14: google-t5/t5-large

google-t5/t5-large:
otorrino sent me a treatment, but he didn't give me any assurance that he improves it. _02: a lot of sleep in my hands and feet. I sometimes drop the knife. otorrino who saw me in urgency, that I came, told me that suddenly he had mucus. _02: and this man, the day I went with my husband, he told him the same thing. _01: my husband is fine, what happens is that there is no parking lot. _02: I've got that whole part very sore, but it's that one blow. thanks to god. after the trip I will bring you something. thank you Maria.

4

Conclusiones y Líneas Futuras

En esta sección se van a exponer las conclusiones obtenidas del desarrollo del trabajo presentado en la sección anterior.

4.1. Problemas encontrados

En esta subsección se van a identificar los problemas encontrados en el transcurso del desarrollo del trabajo:

- **Informes médicos irreales:** Cuando se obtuvo el audio de la cita médica, el médico en cuestión también proporcionó el informe médico resultado de esa cita. Al compararlo con el audio de la cita médica, se puede observar claramente que el clínico incluye una gran cantidad de información clínica relevante en el informe final, sin que se haya hecho mención en el audio de la cita médica, ya sea por información previa a la cita médica o por conocimiento del experto en cuestión. Esto hace que sea mucho más complicado la comparación de informes sintéticos con informes reales.

- **Limitaciones modelos Transformers:** Como se ha podido observar en la sección de desarrollo del trabajo, se han tenido que generar varias funciones para la fragmentación del texto de entrada debido a las limitaciones en cuanto a la longitud. A priori, puede parecer que esto no influye mucho, en el caso de la traducción del texto al inglés no tiene ningún problema, pero en el caso de resumen si lo hay. Esto se debe a que se está produciendo un resumen de una parte del texto, por lo que el modelo puede tomar aspectos de ese fragmento del texto como importantes que en el texto entero no tendría,

además de que se puede perder bastante el contexto del texto dependiendo de donde se hagan los cortes en el texto.

- **Resúmenes médicos ineficientes:** Aunque se haya elegido un modelo como supuestamente el mejor, se puede observar claramente que ninguno se acerca un poco a un informe médico de un profesional ni a lo que espera un médico que se produzca.

- **Evaluación del informe médico:** Como se ha mencionado en la sección anterior no se tiene ningún resumen prototipo para realizar la evaluación. Esto se debe, que, al contrario de pasar el audio a texto, es algo muy subjetivo, ya que va a depender mucho del médico que haga este resumen y lo que él considere como aspectos importantes, por lo que es más complicado evaluar estos modelos.

4.2. Conclusión

Los resultados en cuanto a los modelos de transcripción de voz a texto han sido los esperados, ya que a la utilización del modelo Whisper, modelo que utiliza OpenAi, se esperaban que los resultados iban a ser bastante buenos, aunque sorprendentemente, mejora mucho con la eliminación de palabras sucesivas. En cuanto a los modelos de resumen sorprende un poco el bajo rendimiento en cuanto resúmenes médicos se refiere, ya que los modelos probados, supuestamente con un ajuste fino en el campo de la medicina, son los que peor resumen realiza, mientras que modelos más generales como el de Facebook si realizan un resumen del texto.

4.3. Líneas Futuras

Para continuar con el trabajo realizado, avanzarlo o mejorarlo se podrían cambiar la utilización de modelos Transformers para la realización de resúmenes por modelos LLM. Estos modelos son un tipo de inteligencia artificial diseñado para comprender y generar texto en lenguaje natural. Estos modelos se entrenan con grandes cantidades de datos textuales y utilizan técnicas de aprendizaje profundo, particularmente redes neuronales, para aprender patrones en el lenguaje. Con estos modelos no se encuentra la limitación que se tenía con los modelos Transformers, y además se le puede dar un contexto al modelo para que realice la tarea de la manera más exacta posible. Además permite realizar resúmenes estructurados, tipo informes

médicos, lo cual se puede acercar más a los informes médicos reales.

Referencias

[1] CNBC. *HIMSS 2024: Ambient clinical documentation steals the show*. Accedido el 9 de mayo de 2024. URL: https://www.cnbc.com/2024/03/16/himss-2024-ambient-clinical-documentation-steals-the-show.html.

[2] IBM. *Deep Learning*. URL: https://www.ibm.com/es-es/topics/deep-learning.

[3] IBM. *Redes neuronales*. Sin fecha. URL: https://www.ibm.com/es-es/topics/neural-networks.

[4] Yuening Jia. "Attention Mechanism in Machine Translation". En: vol. 1314. Institute of Physics Publishing, nov. de 2019. DOI: 10.1088/1742-6596/1314/1/012186.

[5] Mse A Fernando-Mercado Salinas Autor et al. *Número Especial de la Revista Aristas: Investigación Básica y Aplicada*. 2022.

[6] Instituto Tecnológico De Aguascalientes et al. *RECONOCIMIENTO DE VOZ*.

[7] Hernán Ordiales. *Comparativa de métodos de conversión de voz a texto Open Source*. URL: https://www.researchgate.net/publication/337367957.

[8] Mikhail Belenko et al. *Design, Implementation and Usage of Modern Voice Assistants*.

[9] Hugging Face. *ROUGE: Hugging Face Spaces*. https://huggingface.co/spaces/evaluate-metric/rouge. Accessed on May 15, 2024. Accessed 2024.

[10] Jaime Andrés Ruiz-Melendres y Jorge Eliecer Gómez-Gómez. "Transcription from audio to text in Municipal Sessions in Planeta Rica Transcripción de audio a texto en Sesiones Municipales de Planeta Rica". En: *Journal of Engineering Interfaces* 6 (2), págs. 1-14. ISSN: 2619-4465.

[11] Adriana Montoto et al. "El Reconocimiento de Voz como alternativa de inclusión para discapacitados auditivos en un entorno educativo". En: 9 (2022).

[12] Martha Eliana, Mendoza Becerra y Elizabeth Leon Guzmán. *Una Revisión de la Generación Automática de Resúmenes Extractivos A Review of the Extractive Text Summarization*. 2013.

[13] LAURA PLAZA MORALES. *GENERACIÓN AUTOMÁTICA DE RESÚMENES CON APOYO EN ONTOLOGÍAS APLICADA AL DOMINIO BIOMÉDICO LAURA PLAZA MORALES*. 2008.

[14] Manuel De La Villa y Manuel J Maña. "Setting a baseline for an automatic extractive concepts-based summarization on the biomedical domain". En: (2009). ISSN: 1135-5948. URL: http://www.uptodate.com/home/about/index.html.

[15] Alex Rosales Hechavarría Yordis Monteserin Matos. *Universidad de las Ciencias Informáticas Facultad 3 Algoritmos para la construcción automática de resúmenes de documentos de texto*. 2007.

[16] Romyna Montiel Soto et al. *Comparación de Tres Modelos de Texto para la Generación Automática de Resúmenes*. 2009.

[17] Iria da Cunha Fanego. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. URL: http://hdl.handle.net/10803/7508.

[18] Manjeet Yadav et al. "Fine-tuning Large Language Models for Automated Diagnostic Screening Summaries". En: (mar. de 2024). URL: http://arxiv.org/abs/2403.20145.

[19] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. DOI: 10.48550/ARXIV.2212.04356. URL: https://arxiv.org/abs/2212.04356.

[20] Jonatas Grosman. *Fine-tuned XLSR-53 large model for speech recognition in Spanish*. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-spanish. 2021.

[21] Jörg Tiedemann y Santhosh Thottingal. "OPUS-MT — Building open translation services for the World". En: *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*. Lisbon, Portugal, 2020.

[22] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". En: *CoRR* abs/1910.13461 (2019). arXiv: 1910.13461. URL: http://arxiv.org/abs/1910.13461.

[23] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". En: *Journal of Machine Learning Research* 21.140 (2020), págs. 1-67. URL: http://jmlr.org/papers/v21/20-074.html.

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

E.T.S. DE INGENIERÍA INFORMÁTICA