UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

Transcripción y extracción automáticas de información clave desde audios clínicos en español.

Maicol Alam Fernández Rodríguez

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN INFORMÁTICA MÉDICA.

Director de Tesis: Prof. Jocelyn Dunstan, PhD
Co-Director de Tesis: Fabián Villena, MSc

(2022)

UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

Transcripción y extracción automáticas de información clave desde audios clínicos en español.

Maicol Alam Fernández Rodríguez

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN INFORMÁTICA MÉDICA.

Director de Tesis: Prof. Jocelyn Dunstan, PhD

Co-Director de Tesis: Fabián Villena, MSc

(2022)

Agradecimientos

Agradezco el financiamiento recibido del Centro de Modelamiento Matemático (CMM), AFB170001, ACE 210010, así como también el proyecto FONDECYT 11201250 y por los proyectos Basal FB210005 y FB0008. Esta tesis fue parcialmente apoyada por la infraestructura de supercomputo del NLHPC (ECM-02). En particular me gustaría agradecer a María José Medina por su colaboración tanto en la creación de la muestra como en el presente escrito. Agradecer también a la Dra. Eugenia Díaz por su apoyo en la redacción de este trabajo. Dar las gracias a Soledad Rodríguez y Camila Fernández por su abnegada colaboración en la confección de la muestra estudiada.

Expreso mi mayor gratitud a la Dra. Jocelyn Dunstan, quien me invitó a ser parte de su gran equipo de trabajo durante estos últimos años, cuales se han visto plenamente afectados por la pandemia, pero siempre apoyándonos de la tecnología para continuar en función. Agradecer también a Fabián Villena quien, junto a la directora del presente trabajo, siempre estuvieron dispuestos a colaborar y entregar las más sinceras recomendaciones y consejos.

Por último, quiero agradecer a la familia López Machuca quienes me han acogido como uno más de su familia durante los últimos años, quienes siempre tuvieron un mensaje de apoyo y una mesa dispuesta para compartir.

Índice

Resumen

Los sistemas actuales de documentación clínica son ineficientes y costosos. Esta tesis propone un escriba digital, un sistema unificado de grabación, transcripción y anotación automática de audios clínicos en español en un entorno simulado. El proceso comienza con la transcripción del audio a texto y la aplicación de un modelo de extracción de información clave.

Se grabaron noventa audios en español de cuatro sujetos: treinta de dominio médico, treinta de dominio dental y treinta de dominio general. Los audios se transcribieron utilizando el servicio *Google Cloud Speech-to-Text* y luego se anotaron con un modelo de Reconocimiento de Entidades Nombradas (NER por sus siglas en inglés) desarrollado por el grupo de investigación. Evaluamos el rendimiento de las transcripciones mediante la tasa de error de palabras (WER por sus siglas en inglés) y la concordancia en la detección de entidades mediante la métrica *F1-score*.

Nuestra hipótesis plantea que el servicio *Google Speech-to-Text* puede tener un WER inferior al 30% y que el modelo NER, entrenado en un corpus de referencias de listas de espera, puede alcanzar una puntuación F1 superior a 0,75 en comparación con las anotaciones manuales.

Los resultados muestran una WER media entre los cuatro hablantes del 10,44%, 9,98% y 9,06% para los dominios dental, medico y general respectivamente. Existe una diferencia de medias significativa entre los tres tipos de grabación (prueba de Kruskal Wallis, valor P 0,010). En la transcripción, los errores típicos son las

palabras cambiadas de plural a singular o viceversa, y un gran número de verbos cambiados de pronombre o de tiempo verbal.

Las palabras no reconocidas por el servicio de transcripción son específicas del ámbito médico-dental, como "nevo", "endogastrio", "mesiodens". Para el reconocimiento automático de entidades, la media para el *F1-score* es de 0,86 para textos del dominio médico y de 0,80 para el dominio dental. No se encontraron diferencias de medias significativas al comparar los distintos grupos de anotaciones.

Los resultados presentados pueden ser la base para el desarrollo de un escriba digital y su evaluación en el ámbito clínico chileno. La transcripción y el modelo NER se integraron en una plataforma. El trabajo futuro incluye la búsqueda de estrategias para mejorar la transcripción y la detección de entidades como, por ejemplo: utilizando sistemas para la eliminación de ruido ambiental o pre entrenar el reconocimiento de voz al usuario. Además, permitir que más profesionales de la salud utilicen la plataforma para obtener una retroalimentación sobre la usabilidad de esta plataforma.

Abstract

Current clinical documentation systems are inefficient and costly. This thesis proposes a digital scribe, a unified system for automatic recording, transcription and annotation of clinical audios in Spanish in a simulated environment. The process starts transcribing audio into text and then applying a model for extracting key information.

Ninety audios were recorded in Spanish by four subjects: thirty medical, thirty dental, and thirty general domain. The audios were transcribed using the Google Cloud Speech-to-Text service and then annotated with a Named Entity Recognition (NER) model developed by the research group. We evaluated the performance of the transcriptions using Word Error Rate (WER) and the agreement in the detection of entities using F1-score.

We hypothesize that the Google Speech-to-text service can have a WER lower than 30% and that the NER model, trained on a corpus of waiting list referrals, can achieve an F1-score higher than 0.75 compared to manual annotations.

The results show an average WER among the four speakers of 10.44% (dental), 9.98% (medical), and 9.06% (general domain). These differences between the three types of recording are significant (Kruskal Wallis test, P-value 0.010). In the transcription, typical errors are words changed from plural to singular or vice versa, and a large number of verbs changed their pronouns or verb tenses. Words not recognized

by the transcription service are clinical/dental-specific, such as "nevo", "endogastrio", "mesiodens". For the automatic recognition of entities, the average F1-score is 0.86 texts in the medical domain and 0.80 for the dental domain. No significant mean differences were found when comparing the different groups of annotations.

The presented results can be the foundation of a digital scribe developed and tested in the Chilean clinical domain. The transcription and the NER model were integrated into a platform. Future work includes finding ways to improve the transcription and detection of entities and having more health professionals using the platform.

1 Introducción

1.1 Antecedentes

El proceso de documentación clínica implica el registro de información, para su posterior análisis y para tener en consideración en futuras atenciones de salud. Los pacientes acuden a los centros de salud por distintas dolencias y enfermedades que son pesquisadas mediante diversas técnicas y procedimientos, las que comprenden desde el examen físico inicial hasta los exámenes complementarios requeridos para el diagnóstico y posterior tratamiento. Una documentación clínica adecuada es clave para la recopilación de datos. Esta abarca no sólo los documentos generados durante la prestación de salud propiamente tal, sino que también incorpora todos los documentos creados durante los procedimientos de gestión asociados a dicha prestación como, por ejemplo: interconsultas, derivaciones, informes radiológicos y de laboratorio, entre otros.

Durante una cita de atención médica u odontológica se desarrollan diferentes etapas, las que se describen brevemente a continuación. Los primeros minutos son dedicados al saludo entre médico-paciente y la inscripción de los datos de información personal en los sistemas de registros utilizados en salud. Luego se dedican minutos al registro del motivo de consulta e historia actual de la enfermedad. Se registra, además, la anamnesis personal y próxima del paciente, etapa en la cual se recopila información correspondiente a las enfermedades y alergias que le han sido diagnosticadas con

anterioridad y cómo estos datos se podrían relacionar con el motivo de consulta y su tratamiento.

A continuación, se lleva a cabo el examen físico del paciente, etapa clave para pesquisar signos y síntomas de alguna enfermedad que podría estar afectándolo y explicar de alguna forma su motivo de consulta. Después de eso tienen lugar los exámenes complementarios, los cuales en muchas ocasiones son fundamentales para el diagnóstico de alguna patología. Luego de llevar a cabo estas etapas, se da paso al diagnóstico y a su tratamiento correspondiente. Esta última etapa es la que más varía, ya que depende de las patologías encontradas, y su resultado puede ir desde una simple prescripción hasta una resolución quirúrgica.

Al finalizar la consulta se completa el registro de la información recopilada durante la cita, para lo cual se desarrolla el proceso de evolución y epicrisis, etapas que son fundamentales en la documentación clínica. El registro de los procedimientos realizados, es decir, la ficha clínica, es de gran utilidad para futuras atenciones, no solo para el paciente en sí, sino que además puede servir de referencia para la utilización de nuevas técnicas y procedimientos, los cuales pueden facilitar el abordaje y resolución de patologías en otros pacientes.

Con respecto a la ficha clínica, esta ha evolucionado muy rápidamente durante las últimas décadas, desde la anotación manual de la información en papeles que podían llegar a formar un gran legajo, hasta la completa digitalización de esta.

Actualmente, la mayor parte de la recopilación de información clínica se realiza de manera electrónica (1). Los datos digitalizados disponibles en los sistemas de registros clínicos electrónicos han aumentado exponencialmente durante la última década. No obstante, cabe mencionar que dichos datos generalmente se subutilizan, ya sea por el desconocimiento de su potencial, por falta de herramientas para su procesamiento o debido a las políticas de restricción de acceso a datos sensibles por razones éticas. Se ha descrito que los historiales médicos digitales o sistemas de ficha electrónica, contienen tanto información estructurada, como no estructurada en forma de texto y se calcula que más del 40% de los datos de los sistemas de ficha electrónica contienen texto libre (2).

La calidad del texto clínico puede verse afectada debido a diferentes razones, entre las que se incluyen: el prolongado tiempo destinado a su documentación y el limitado tiempo con el que cuenta el profesional de la salud para realizar el registro, al punto que en ocasiones deben destinar tiempo extra al desarrollo del registro de información clínica. Se profundizará sobre el tiempo destinado a la documentación en el punto 1.1.1.

Por otra parte, la relación médico-paciente se ve en cierto modo afectada por la interferencia de un computador entre los interlocutores, lo cual puede llegar a impedir una conversación fluida, desconcentrando al profesional de salud el cual podría no lograr capturar algunos signos importantes en la conversación, como gestos o posturas propios de una comunicación paraverbal. A lo anterior se debe sumar la mala

usabilidad de los sistemas de registro clínico electrónico, hecho que repercute directamente en la calidad de la documentación clínica. Todos los puntos antes mencionados terminan por desencadenar el fastidio y agotamiento de los profesionales de salud, lo que influye en niveles de satisfacción significativamente empeorados frente al uso de estos sistemas (4).

1.1.1 El tiempo

Se ha descrito que el tiempo destinado a las citas médicas ambulatorias varía entre los diez a los cuarenta y cinco minutos, los cuales deben distribuirse en todas las etapas mencionadas anteriormente (4). Por ejemplo, en Argentina el tiempo de atención varía entre los diez a quince minutos, en Perú es de doce minutos, en Estados Unidos va desde los dieciséis a los veinte minutos y en Canadá va entre los diez a cuarenta y cinco minutos. El mayor tiempo reportado en Canadá se podría explicar debido al modelo de salud de dicho país, el que funciona en relación con el tiempo destinado a la atención del paciente. El caso opuesto ha sido descrito para Japón, país que tiene un sistema universal de salud, con un tiempo de atención muy acotado que iría entre los tres a los seis minutos de atención por paciente, lo que se aleja de lo recomendado por organizaciones de salud, que en general recomiendan que la atención se extienda entre los dieciocho y veinte minutos (4). En el caso particular y excepcional de Rusia, el tiempo de consulta está reglamentado en diez minutos, los que deben distribuirse de la siguiente manera: un minuto para saludar y despedir al paciente, tres minutos para la anamnesis, dos minutos para el examen físico, dos

minutos para la prescripción y dos minutos para completar formalidades administrativas (4). Cabe señalar que, aparte de lo riguroso de los tiempos intermedios durante la atención médica, el sistema ruso confiere un tiempo extremadamente acotado para llevar a cabo un registro detallado y de calidad.

En Chile la situación no es muy distinta a lo que describe la literatura. En nuestro país, el rendimiento de la actividad se ha determinado de acuerdo con estándares definidos o adecuados a nuestra población. El rendimiento se expresa en el tiempo que utiliza el profesional de la salud para realizar la actividad. Por ejemplo, en la consulta de morbilidad está establecido que el médico atienda a cuatro pacientes por hora, es decir, que destine quince minutos por usuario, no obstante, ese rendimiento debe tener relación con la posibilidad de ofrecer una solución efectiva al problema de salud que motiva la consulta. Las consultas programadas en la modalidad de telemedicina deberán regirse por un rendimiento sugerido de tres consultas por hora, en el caso de la modalidad sincrónica y seis consultas por hora en modalidad asincrónica, resguardando el proceso de referencia y contrarreferencia (5). Con respecto al tiempo destinado para las atenciones dentales, este varía entre los veinticinco a treinta minutos de duración, un tiempo muy acotado si se tiene en consideración que los cirujanos dentistas deben realizar no solo la prescripción, como en el caso del profesional médico, sino que además deben llevar a cabo un procedimiento que en muchas ocasiones se podría prolongar más allá del tiempo definido, o llegar a presentar complicaciones durante su realización, lo que podría incluso llegar a afectar la calidad de la prestación entregada. Es importante enfatizar

que son los mismos profesionales que atienden al paciente los encargados y responsables del proceso de documentación clínica, lo cual podría tomar un tiempo mayor del disponible por atención.

Se estima que los médicos dedican entre el 34% y el 78% de su jornada laboral a crear notas y revisar las historias clínicas en los sistemas de registro clínico electrónico, lo que supone un costo estimado de entre 90.000 y 140.000 millones de dólares anuales en tiempo médico (1). Esta considerable cantidad de tiempo destinado a la documentación clínica es considerada una de las principales causas de agotamiento profesional, ya que dedican casi la mitad del tiempo en teclear, hacer *click* y marcar casillas en los registros electrónicos (6). Es aquí donde las personas en rol de escribas médicos surgieron inicialmente como una solución alternativa a las problemáticas antes mencionadas. Tales personas están presentes en la sala de atención y elaboran notas clínicas coherentes en tiempo real, mientras la médico entrevista y realiza el examen físico, lo que permite una mayor participación con el paciente. Sin embargo, su presencia reduce de manera importante la privacidad de la interacción médico-paciente, además de aumentar los costes de personal (1). Estos factores convierten a los escribas médicos en una solución costosa y asociada a errores de documentación, por lo que la evidencia sugiere que el uso de un “escriba digital” puede facilitar una documentación más eficiente y menos costosa, solucionando así los problemas previamente descritos (7).

1.1.2 Escriba digital: una solución

Un “escriba digital” se define como un sistema de documentación sistematizado, capaz de capturar la conversación médico-paciente para luego generar la documentación, en forma similar a la función realizada por los escribas médicos humanos. Esto permitiría al profesional enfocarse en el paciente mejorando así la relación con estos, reduciendo también el tiempo invertido en documentación, aumentando la productividad y disminuyendo el agotamiento profesional (7).

En este sentido, una encuesta sobre el uso clínico de computadoras por dentistas de Estados Unidos, destacó el reconocimiento de voz como una forma de facilitar el registro directo, señalándola además como una de las mejoras más deseables en las aplicaciones actuales (8). Otro estudio basado en encuestas informó que más del 90% de los hospitales planean expandir su uso de sistemas de reconocimiento de habla al *front-end*, es decir, dictado directo en campos de texto libre de la historia clínica electrónica (EHR por sus siglas en inglés) (9). Una revisión sobre usabilidad y evaluación de la tecnología de reconocimiento del habla en entornos hospitalarios, mostró que estos sistemas se usaban comúnmente en el año 2008 para ayudar en la documentación clínica, aunque también se describía su uso en sistemas interactivos de respuesta de voz, control de equipos médicos y sistemas de traducción automática (9). La literatura del campo muestra que cada día toma mayor relevancia la explotación de las distintas capacidades que tiene la computación, con el

fin de lograr una mayor y mejor recopilación de la información, agilizando y optimizando los procesos de atención médica.

En relación con las mejoras esperadas en la utilización de sistemas de reconocimiento de voz y cómo esta tecnología contribuiría a la reducción del tiempo destinado a la documentación clínica, se ha relacionado el reconocimiento de voz con el *tiempo de respuesta* (TAT por sus siglas en inglés), definido como el tiempo necesario para todo el proceso de documentación clínica. Numerosos estudios han mostrado una disminución en el TAT cuando se utilizó reconocimiento de voz, reportando un descenso máximo de 81,16% (de 1.486 minutos hasta 280 minutos) y un descenso mínimo de 16,41% (de 329 minutos hasta 275 minutos). Estos estudios, en su gran mayoría radiológicos, involucraron predominantemente informes de texto libre. Cuando se analiza a lo largo del tiempo, parece haber una disminución en el TAT de 0.9% por año (figura 1). Las tasas de precisión de los sistemas de reconocimiento de habla van desde una precisión media de 88,9% a 96%. Las tasas de precisión generales en todos los informes mostraron una mejora mínima con el tiempo, al 0.03% por año (Figura 2) (10).

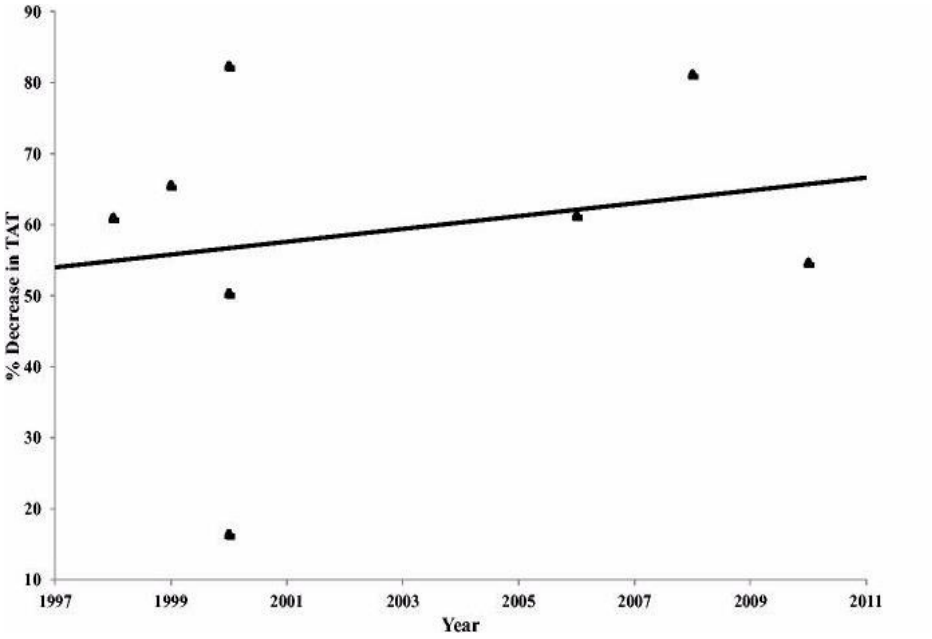

Figura 1: Aumento progresivo durante un lapso de 14 años, de la disminución del tiempo de respuesta (TAT) al utilizar sistemas de reconocimiento de voz en la práctica clínica (Tomado de Hodgson *et al*, 2016; Referencia 10).

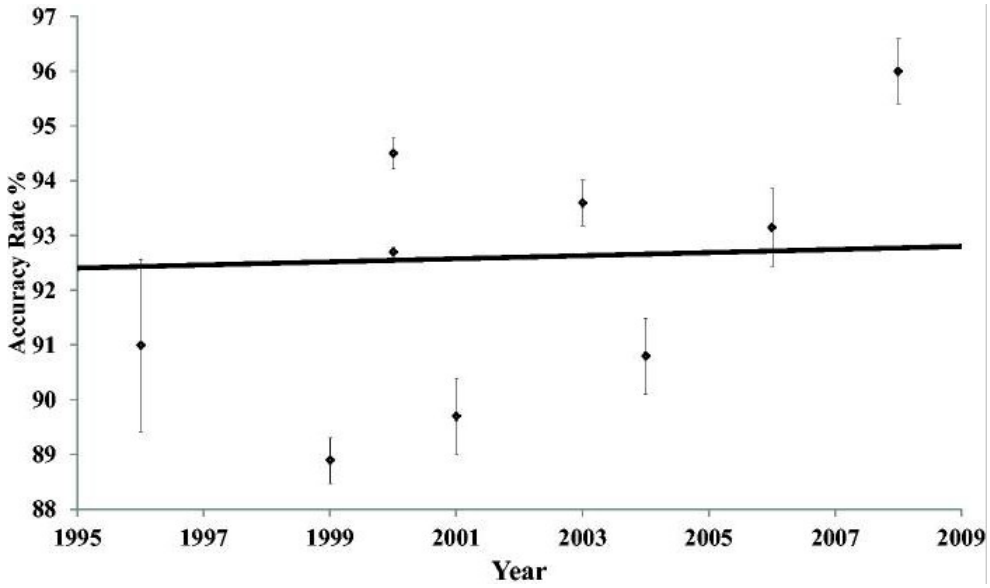

Figura 2: Precisión media de los sistemas de reconocimiento del habla, y su evolución entre los años 1995 al 2009 (Tomado de Hodgson *et al*, 2016; Referencia 10).

Múltiples iniciativas con diferentes nombres se han iniciado en varias compañías de desarrollo tecnológico, entre las que se encuentran Microsoft y Google. Se han identificado diferentes tareas y/o desafíos durante el desarrollo de un escriba digital (11) (Anexo 1). Un escriba digital o asistente médico automático debe ser capaz de desarrollar una secuencia de tareas no muy fáciles de cumplir. Estas incluyen: i) ser capaz de realizar la grabación de la conversación entre el profesional de la salud y el paciente, ii) generar la transcripción automática del audio y iii) extraer la información clave del texto para generar un resumen de la información recopilada durante la atención.

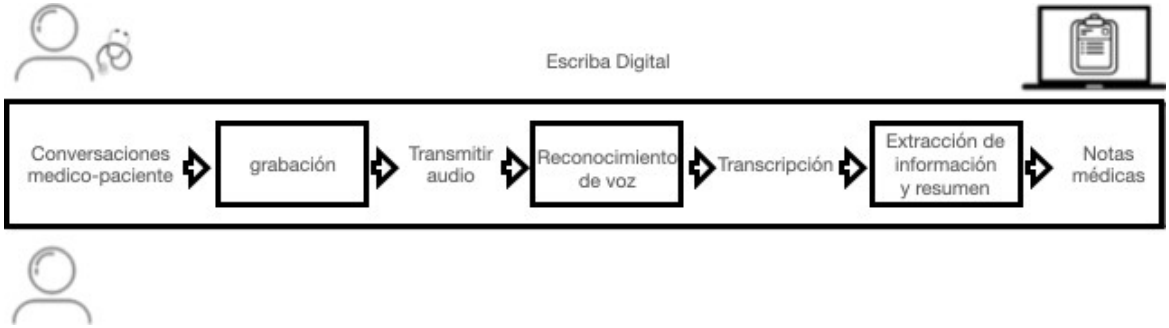

Figura 3: Descripción y secuencia temporal de las tareas asignadas al sistema de documentación Escriba Digital (Tomada y traducida de Quiroz *et al*, 2019; Referencia 7).

Desde la perspectiva del desarrollador, una herramienta de este tipo necesita una gran cantidad de información previa, que permita el entrenamiento de un modelo de *machine learning* para desarrollar las tareas requeridas. Se han identificado seis grandes tareas, cada una con desafíos específicos (Ver Anexo 1), que deben ser enfrentadas por los desarrolladores de sistemas para lograr una herramienta con las cualidades deseadas. En el presente trabajo se abordarán dos desafíos importantes:

1) el reconocimiento automático del habla desde audios clínicos y 2) la extracción de conceptos médicos usando procesamiento de lenguaje natural. A continuación, se profundiza en diferentes investigaciones que han abordado estos aspectos.

1.1.3 Trabajos relacionados

Un escriba digital es, en primer lugar, un sistema de conversión de voz a texto que necesita varios componentes para funcionar. Estos van desde un micrófono hasta un motor de reconocimiento de voz, de manera que el habla humana analógica es transformada en ondas digitales. Estas son relacionadas con las palabras mediante códigos de reconocimiento, identificando en primer lugar de dos a tres palabras, mediante la utilización de fonemas de vocales y consonantes. Mediante este proceso se crean vocabularios que facilitan el reconocimiento de patrones relacionados con la coherencia del texto (12).

En la actualidad la tecnología del reconocimiento del habla se utiliza en diferentes ámbitos. Del mismo modo, también existen diferentes motores para dicho propósito, los que han evolucionado de tal manera que no solo han visto mejorado su rendimiento, sino que además ahora permiten un reconocimiento continuo del habla (*streaming*), así como la identificación de distintos hablantes, lo que en la actualidad ha permitido compararlos con el desempeño humano. Bajo esta perspectiva, en una investigación realizada por Kodish-Washs *et al*. (13) se evaluó el desempeño de ocho motores de reconocimiento automático del habla (*Bing Speech* API, *Google Cloud*

Speech API, IBM *Speech To Text*, *Azure Media Indexer*, *Azure Media Indexer* 2 *Preview*, *Nuance.Speech Anywhere*, *Amazon Transcribe* Preview y *Mozilla DeepSpeech*). Estos motores fueron evaluados en conversaciones entre un clínico y un paciente, a través de entrevistas sin un guion establecido. Luego se realizó la transcripción de los audios por humanos con el propósito de comparar los resultados entregados por los motores. Los resultados fueron deficientes, reportando valores de tasa de error por palabra (WER por sus siglas en inglés) que variaron entre un 34% y un 65%, lo que llevó a los autores a concluir que futuros trabajos se deben centrar en mejorar el rendimiento de los motores de reconocimiento automático del habla (ASR por sus siglas en inglés) antes de poder adoptar estas tecnologías para el habla conversacional en una amplia gama de casos de uso clínico (13).

Otro estudio desarrollado por Zuchowski *et al*. (14) evaluó un sistema para el reconocimiento del habla en documentación médica, concluyendo que dicha documentación disminuyó de 8.9 minutos en una documentación por mecanografía, a 5.1 minutos al utilizar un motor de reconocimiento de habla, lo que significa un ahorro significativo de tiempo. En esta misma investigación se concluyó que la tasa de error era menor al utilizar un sistema para el reconocimiento de voz lo que se contradice con la investigación realizada por Kodish-Washs *et al*. (13). Sin embargo, se debe aclarar que ambos estudios no son del todo comparables ya que la metodología utilizada en ambos es distinta: Zuchowski *et al*. (14) optaron por evaluar el sistema utilizando un texto estructurado por un guion, en tanto que la investigación realizada por Kodish-Washs (13) carece de un guion.

Así también, se realizó la evaluación sobre el funcionamiento de un *plugin* Jython de código abierto llamado *Speech-to-Text* para un *software* de utilización en la práctica forense llamado Autopsy, donde se logra un valor WER de 27.2% al transcribir mensajes de voz en inglés de hablantes no nativos y un valor WER de 7.8% para un conjunto de prueba del *corpus LibriSpeech* (15), afirmando también que la transcripción de un minuto de voz demora nueve segundos, reduciendo significativamente el tiempo necesario para encontrar información relevante contenida en archivos de audio (16).

El desempeño de los sistemas de reconocimiento del habla depende de cinco factores: el hablante, el habla, el vocabulario, las complejidades gramaticales del lenguaje y la entrada de la voz. El efecto causado por estos elementos puede reducirse mediante pruebas estándar y conjuntos de grabación continua. El hablante es el elemento más influyente; sin embargo, la precisión de las palabras no afecta su inteligibilidad por parte del sistema, ya que tanto el hablante como el sistema se ven afectados por la calidad del sonido, el ruido de fondo, el dominio del habla y la velocidad (12).

1.1.4 Reconocimiento automático del habla desde audios

Para lograr una buena transcripción de un audio en general, es necesario en primer lugar la obtención de un archivo de audio con la calidad suficiente para una disminución de potenciales errores en el reconocimiento de voz. Para esto es

necesario contar con un micrófono que permita una clara obtención de la voz de ambos interlocutores, lo cual se dificulta en un ambiente clínico. Bajo esas condiciones, un sistema de reconocimiento debe ser lo suficientemente robusto como para desarrollar la tarea de reconocimiento con la menor tasa de error a partir de un audio ya sea de alta o baja calidad. En la actualidad, se han desarrollado diferentes alternativas para sistemas de reconocimiento del habla, algunas de las cuales se describen brevemente a continuación:

- *Microsoft Speech API* (SAPI) es una biblioteca de código cerrado que admite ASR y *Text to Speech* (TTS según sus siglas en inglés) dentro de las aplicaciones de *Windows* (17).
- Google con *Speech-to-Text* utiliza tecnología de reconocimiento de voz de código cerrado, basada en aprendizaje profundo con una tasa de error del 8% (18).
- *Amazon Transcribe* es un servicio ASR que se utiliza en muchas aplicaciones típicas, incluidos los sistemas de atención al cliente por teléfono y subtitulado automático de audio y video. Este sistema está continuamente aprendiendo y mejorando para mantenerse al día con el desarrollo del lenguaje (19).
- IBM (*Speech to text*) permite a los usuarios agregar hasta noventa mil palabras fuera del vocabulario a un modelo de idioma personalizado. El servicio utiliza el aprendizaje automático para combinar el conocimiento de gramática y la estructura del lenguaje, así como la síntesis de señales de sonido y voz para transmitir la voz humana

con precisión. Además, es compatible con lenguajes de programación como Node, Java, Python, Ruby, Swift y Go SDK (20).

- Wit es un sistema gratuito, incluso para uso comercial. Es compatible con más de 130 lenguajes y aplica las leyes de protección de datos de la Unión Europea. Wit es compatible con los lenguajes de programación Node, Python y Ruby (21).

Tabla 1: Comparación del desempeño de 3 Sistemas de reconocimiento de habla.

Narrador A			Narrador B			Narrador C		
IBM	Google	Wit	IBM	Google	Wit	IBM	Google	Wit
30.10	10.60	26.87	47.73	20.45	23.28	36.51	24.85	58.87

Se evaluaron tres narradores (A, B y C) quienes realizaron la lectura de sesenta oraciones de texto en inglés. La Tabla muestra que el sistema de Google presenta la menor tasa de error por palabra (WER) en los tres participantes. Esto es: 16,60%, 20,45% y 24,85% para los Narradores A, B y C, respectivamente (Tabla tomada de Filippidou & Moussiades, 2020; Referencia 22).

Las conversaciones médico-paciente corresponden al tipo no estructurado, lo que dificulta aún más el reconocimiento correcto de la información grabada en audios. En la actualidad, existen diferentes herramientas disponibles para la transcripción automática de audios, las cuales incluso permiten diferenciar interlocutores en una conversación. Empresas como *Google* o *Mozilla* han creado algunas APIs para su uso en desarrollos independientes, herramientas disponibles para la comunidad y con amplia documentación. No obstante, se debe considerar que estas herramientas se han implementado para la transcripción de textos desde un ámbito no clínico, por lo

que las transcripciones podrían variar su calidad dependiendo de los temas a tratar en una conversación (22).

1.1.5 Extracción de conceptos médicos usando procesamiento de lenguaje natural

El segundo desafío que se establece para esta tesis recae en la extracción de información desde los audios generados a partir de dictados de interconsultas clínicas. La tarea de extracción se desarrolla con posterioridad a la transcripción de los audios, por lo que las técnicas utilizadas para la obtención de información clave desde el texto recaen principalmente en el procesamiento de lenguaje natural (PLN). Esta tarea presenta un alto grado de dificultad debido a que las técnicas de PLN tradicionales funcionan muy bien con oraciones gramaticalmente correctas y con límites establecidos. Estas características se pierden en los textos transcritos automáticamente, ya que la coherencia se ve perjudicada debido a la casi nula incorporación de signos de puntuación tales como puntos y comas. Es importante destacar que, con relación al desarrollo de herramientas para el análisis de texto, se ha llevado a cabo mucha investigación para el procesamiento de texto clínico en inglés, pero no tanto para otras lenguas. Sin embargo, se han desarrollado algunas herramientas para el idioma español como por ejemplo el corpus clínico español IXAMed del Hospital Galdakao- Usansolo, recopilado durante 2008-2012 y que contiene 141.800 registros de pacientes (23). También existe un corpus clínico argentino-español, que contiene 512 informes radiológicos anotados (2) y, por otra parte, nuestro grupo de trabajo se encuentra actualmente desarrollando un corpus de

texto anotado basado en la lista de espera del sistema de salud chileno (24).

Para desarrollar esta tarea durante 2018, mi directora de tesis Jocelyn Dunstan solicitó la lista de espera no GES a los veintinueve servicios de salud del país a través de la Ley de Transparencia. Como resultado de dicha gestión, el grupo tiene 5.157.902 derivaciones, la distribución entre las derivaciones médicas y dentales es 88% versus 12%, respectivamente. Durante los últimos tres años se ha llevado a cabo el etiquetado de entidades de este corpus de texto, en base a un sistema de preanotación automática y anotación manual, proyecto en el cual han trabajado estudiantes de medicina, médicos cirujanos y cirujanos dentistas. De esta manera, se ha logrado etiquetar e identificar las siguientes entidades: hallazgos clínicos, resultados de laboratorio, signos o síntomas, enfermedades, partes del cuerpo, medicamentos, abreviaciones, miembros de la familia, procedimientos de laboratorio, procedimientos diagnósticos y procedimientos terapéuticos (24).

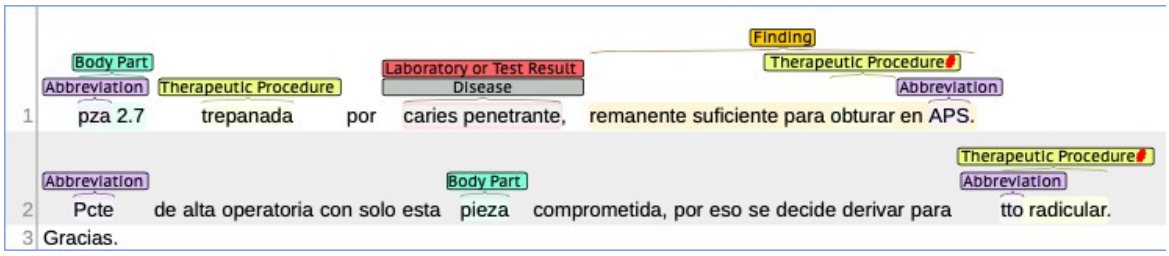

Figura 4: Ejemplo de interconsulta dental anotada en plataforma BRAT por nuestro grupo de investigación.

Para el desarrollo del modelo de reconocimiento de entidades nombradas (NER por sus siglas en inglés), es necesario la obtención de las anotaciones desde la plataforma BRAT (25), la cual genera un archivo en formato *standoff* (.ann) para cada derivación. Estos archivos siguen una estructura básica que consiste en tres columnas que contienen: un ID por anotación en orden consecutivo de aparición, el tipo de entidad a la cual corresponden con los índices del principio y final de la anotación, y por último la cadena de caracteres que constituyen esa entidad. Estos archivos fueron convertidos al formato CoNLL4 (26), ampliamente utilizado en la comunidad de PLN. Este formato no admite entidades anidadas, anotación realizada sobre otra anotación cuyas entidades son distintas, pero para solucionar este problema se entrenó un modelo para cada entidad. En la Figura 4 se observan entidades anidadas en cuatro ocasiones: un ejemplo de esto es la cadena de caracteres 'pza 2.7' (que se refiere a la pieza o diente 2.7), donde la entidad *Abbreviation* 'pza' está anidada dentro de la entidad *Body Part* 'pza 2.7'. También se observa que se repite una anidación para la cadena 'caries penetrante', la cual pertenece simultáneamente a un *Disease* y a un *Laboratory or Test Result*.

Estos modelos fueron fundamentales para el desarrollo de la presente investigación, su desempeño está descrito entre un *F1-score* de 0.77 para reconocimiento de enfermedades y de un 0.75 para el reconocimiento de medicamentos. En base a estos valores será posible identificar como el uso de esta herramienta se podría ver afectada por la utilización combinada con el sistema de transcripción *Speech-to-text* de *Google*. El uso en conjunto de estas herramientas

demostraran si es factible el desarrollo de un “escriba digital” que podría contribuir a solucionar los problemas más relevantes relacionados a la documentación, como lo son el agotamiento profesional, la interferencia médico paciente y el alto tiempo destinado a la documentación.

Son diversos los problemas que se han identificado durante el proceso de recopilación de información clínica. Entre ellos cabe destacar: i) El prolongado tiempo destinado al proceso de documentación, ii) Interferencia en la relación médico-paciente, iii) Degradación de la información y iv) Limitaciones en usabilidad de las interfaces de documentación actuales. Todos estos factores desencadenan el agotamiento del profesional médico encargado del proceso. Por dichas razones, en esta tesis se propone la creación de un “escriba digital”, es decir, un sistema que permite la transcripción de la voz para la posterior extracción de información con relevancia clínica. Este sistema favorecería una documentación más eficiente, menos costosa y demandante de tiempo, contribuyendo así a la solución de los problemas antes mencionados.

2 Problema

No existe información sobre el desempeño de un modelo para el reconocimiento de entidades nombradas sobre textos transcritos de manera automática a partir de audios clínicos en idioma español, en un ambiente simulado.

Responder este cuestionamiento permitiría aclarar algunos de los desafíos antes mencionados, como son la transcripción mediante un sistema de reconocimiento de voz y la extracción de conceptos médicos claves.

3 Solución

Desarrollar un sistema unificado de grabación, transcripción y anotación automática de audios clínicos en español, evaluando su desempeño en un ambiente simulado.

4 Hipótesis

Un sistema unificado entre el servicio *Google Cloud Speech-to-text* y el modelo NER entrenado sobre el corpus de la lista de espera, permite reconocer de forma automatizada entidades con un *F1-score* mayor de 0.75 en comparación a las anotaciones manuales realizadas sobre los textos originales con un promedio *word error rate* (WER) menor al 30%.

5 Objetivos

5.1 Objetivo General

Evaluar el desempeño de un sistema que logre la detección automática de información clave desde audios clínicos en español mediante la utilización de transcripción automática para el desarrollo de un sistema unificado de grabación, transcripción y anotación de relatos clínicos en español.

5.2 Objetivos Específicos

1. Construir un conjunto *Gold Standard p*ara la evaluación del sistema.
2. Evaluar el desempeño del servicio *Google Cloud Speech-to-Text* con respecto a los *Gold Standard*.
3. Evaluar el desempeño del modelo NER con respecto a las anotaciones manuales del conjunto *Gold Standard*.
4. Comparar el desempeño del modelo NER sobre los textos transcritos con respecto a las anotaciones automáticas del conjunto *Gold Standard*.
5. Desarrollar un sistema unificado de grabación, transcripción y anotación de audio clínico en español a partir de herramientas preexistentes.

6 Materiales y Método

6.1 Gold Standard

En el presente trabajo de tesis se llevaron a cabo dos evaluaciones: por una parte, se realizó la evaluación del servicio de *Google Cloud Speech-to-Text* y por otra, el desempeño del modelo NER sobre estas transcripciones. Para este propósito fueron necesarios tres conjuntos de textos, que constituyeron el conjunto *Gold Standard*, que fue leído por cuatro participantes. Estos corresponden a noventa textos, de los cuales se tomaron sesenta textos de interconsultas pertenecientes a la lista de espera del sistema de salud chileno. Estos a su vez se dividen en treinta textos pertenecientes a interconsultas dentales y treinta a derivaciones de dominio médico, todos pertenecientes al *Waiting List Corpus*. Con el objeto de realizar un mejor análisis y comparación, estas interconsultas fueron normalizadas, se expandieron sus palabras abreviadas para su correcta lectura y se eliminó el uso de mayúsculas. Asimismo, se recopilaron treinta textos de dominio general para evaluar el desempeño del servicio de transcripción automática en relación a los audios con lenguaje técnico, mediante la utilización de la métrica WER (*word error rate* / tasa de error por palabra). En resumen, el conjunto *Gold Standard* para la evaluación de la transcripción a noventa textos normalizados, comprendió treinta de dominio dental, treinta de dominio médico y treinta de dominio general.

6.2 Grabación

Los textos seleccionados en cada dominio fueron dictados por los participantes y grabados con el programa *Audacity* (27) disponible de forma gratuita en su página *web*. Los archivos fueron guardados en formato .wav no comprimido de dos canales, en una frecuencia de 44100 (Hz) y cuantificación lineal de 16 bits (PMC), utilizando un dispositivo del tipo manos libres (audífonos con micrófono incorporado) utilizando además mascarilla (cubreboca) de triple capa. Estas grabaciones se realizaron en un ambiente semi-silencioso, desechando los intentos fallidos, ya sea por la interferencia de algún ruido estruendoso o algún error de lectura. Además, solo se realizaron grabaciones con una duración menor a un minuto y de un tamaño menor a diez MB correspondiente al límite permitido para transcripción por el servicio.

Para la lectura y grabación de los textos se contó con la colaboración de tres personas externas a la investigación. Una mujer de veinte y ocho años, cirujana dentista; una mujer de cincuenta años, contadora y una, mujer de diecinueve años, estudiante de Derecho, además del autor de esta tesis, hombre de veintisiete años, cirujano dentista. Cada uno de los participantes realizó la lectura de los noventa textos pertenecientes al *Gold Standard*.

6.3.- Transcripción

Se identificó y evaluó la mejor alternativa para la transcripción de audios, tomando la decisión de utilizar el servicio *Google Cloud Speech-to-Text*, herramienta con buen desempeño registrado (22) y de bajo costo para su utilización, además de estar disponible en diferentes versiones y/o funciones. Se optó por la versión disponible para audios cortos, los cuales permiten procesar hasta un minuto de datos de audio de voz. Después del reconocimiento, los datos son enviados en una solicitud asincrónica o sincrónica. En este proceso se utilizó el código disponible para Python en el cual se hace envío del archivo de audio, se hace su procesamiento y se devuelve la respuesta. Se utilizó la transcripción a idioma español (Estados Unidos). El tiempo de respuesta varía dependiendo de la duración y calidad del archivo de audio. La API obtenida desde *Google Cloud* fue incluida e intervenida para que creara archivos en formato .txt los cuales fueron utilizados en la siguiente etapa.

6.4.- Anotación

Para la realización de esta etapa solo se utilizaron los textos de dominio médico y dental. Esto debido a que los textos de dominio general solo tenían la finalidad de comparar la calidad de la transcripción y verificar si esta se pudiera ver influenciada por la terminología específica médica o dental. Los resultados de esta comparación se mostrarán en la sección resultados.

Los sesenta textos de dominio médico-dental fueron anotados por el autor y consolidados en conjunto a dos profesionales de la salud, estableciéndose un consenso por las anotaciones realizadas y encontradas en los textos. El modelo computacional NER fue entrenado con las anotaciones realizadas sobre el corpus de la lista de espera y tiene la capacidad de reconocer varias entidades. En particular, en esta tesis se analizaron tres entidades: Enfermedades, Partes del cuerpo y Medicamentos, entidades con mayor relevancia clínica, utilizando la métrica *F1-Score* para la evaluación del desempeño. El resultado de la anotación fue un archivo *standoff* (.ann) para cada archivo .txt con las anotaciones encontradas por el modelo computacional.

6.5 Análisis

Posterior a esto, se inició la etapa de análisis de los resultados obtenidos, para lo cual fue necesario realizar 2 análisis independientes, los que se describen a continuación.

En el primer análisis se evaluó el desempeño del servicio de *Google Speech-to-text* mediante la tasa de error de palabra (WER). Dicha tasa expresa la distancia entre la secuencia de palabras que produce un API y la serie de referencia *(Gold Standard*), con la finalidad de obtener un valor comparativo sobre la calidad de transcripción (22). Para esto se utilizó una herramienta de código abierto llamada "*Asr_evaluation"* que permite la clasificación automática de errores de salida de traducción automática disponible en GitHub en la página

https://github.com/belambert/asr-evaluation, El cálculo WER se realiza obteniendo las inserciones, deleciones y sustituciones por cada oración, se realiza la sumatoria de errores dividido por el total de palabras por cien, lo que entrega un valor en porcentaje (ver Figura 5). Para conocer la cantidad de errores existentes, la herramienta compara la secuencia de caracteres en la oración de cada archivo transcrito, con su original del conjunto *Gold Standard.* Por lo tanto, en este trabajo se compararon noventa pares de archivos, cuyos resultados se mostrarán en la sección correspondiente.

$$WER = (S + D + I) / N1 = (S + D + I) / (H + S + D)$$

Figura 5: Definición del *Word Error Rate* (WER), donde I es el número total de inserciones, D el número total de eliminaciones, S el número total de sustituciones, H el número total de éxitos y N1 el número total de palabras de referencia.

Para la realización del segundo análisis, sólo se utilizaron los textos de dominio médico y dental, ya que su propósito fue evaluar la capacidad de reconocer términos médicos o dentales que no se observan en los textos de dominio general. En este análisis se compararon los resultados *F1-Score*, prueba que entrega una medida de rendimiento mediante una puntuación que valoriza la comparación entre las anotaciones realizadas en los conjuntos de datos. Se compararon las anotaciones manuales de los textos *Gold Standard* contra las anotaciones automáticas encontradas tanto en los textos G*old Standard* como en los textos transcritos (ver Figura 6). De esta forma, se evaluó si el servicio de transcripción de *Google Cloud* tuvo alguna interferencia en las anotaciones, con el propósito de validar su uso como al aplicarlo en herramientas conjuntas.

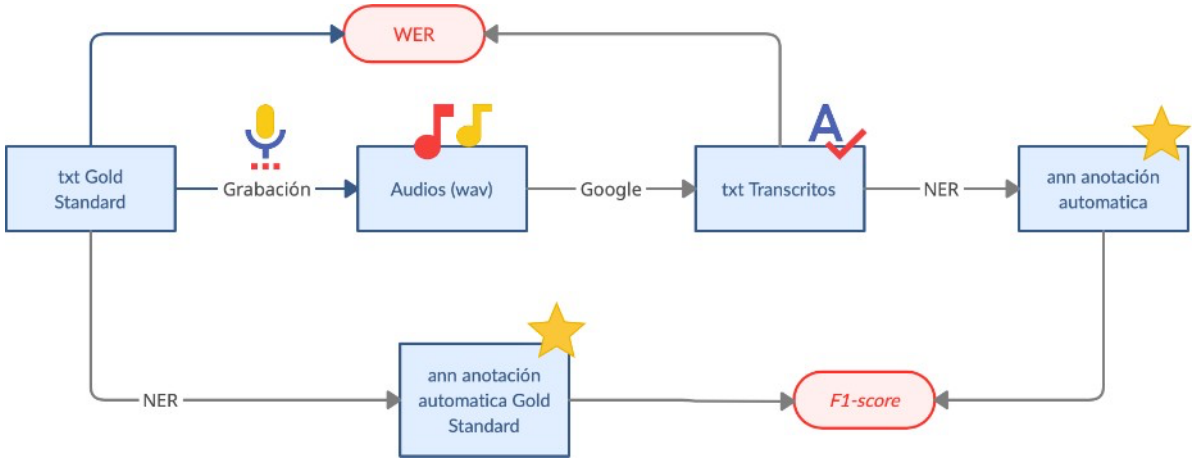

Figura 6: Descripción gráfica de la metodología, comenzando por los textos *Gold Standard* los cuales fueron grabados en audios para su transcripción por el servicio de *Google Cloud Speech-to-Text*, luego estos textos transcritos pasaron a una etapa de anotación automática y fueron comparados con la anotación automática de los textos *Gold Standard.*

6.6 Desarrollo de la plataforma

El desarrollo de la presente investigación contempló el desarrollo de una plataforma prototipo que unifique las herramientas antes mencionadas. Para el desarrollo de esta tarea se trabajó en conjunto con la empresa “Create Spa”, dedicada al desarrollo de interfaces gráficas y desarrollo *web*, a la que se le encargó el desarrollo de un sistema multiplataforma, que tenga la posibilidad de acceso remoto a un servidor.

7. Resultados

7.1 Análisis WER

Se realizaron un total de trescientos sesenta grabaciones, repartidas entre los cuatro participantes (noventa por cada uno) las que fueron posteriormente transcritas, Se obtuvieron de esta forma trescientos sesenta archivos de texto plano los, cuales fueron comparados con su respectivo archivo *Gold Standard*. Los trescientos sesenta valores WER fueron graficados para ver su distribución y se les realizó una prueba de Shapiro-Wilk para evaluar su normalidad, obteniéndose un *P-value* de 0.017 de esta forma se rechaza la hipótesis nula, lo que demuestra que el conjunto de datos no se distribuye con normalidad, como se ilustra en la Figura 7.

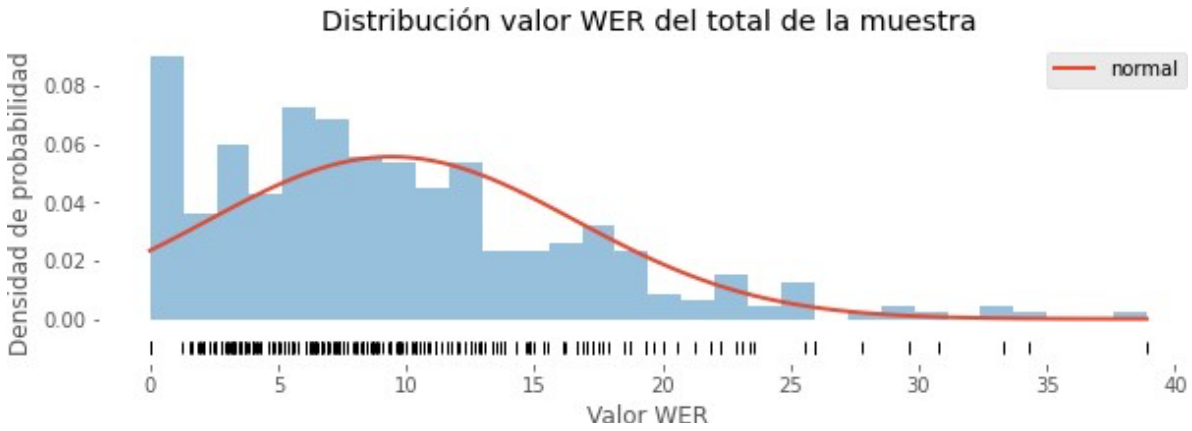

Figura 7: Se observa la distribución de la muestra, valores pertenecientes a los trescientas sesenta valores WER. Se observa también la no normalidad de la muestra, la cual se confirma con la obtención de un P-value de 0.017 en la prueba Shapiro -Wilk para evaluación de normalidad.

Los valores promedio de error para los tres dominios son de 10.44% para las transcripciones de dominio dental, 9.98% para las transcripciones médicas y 9.06% para las de dominio general. Se observa, por lo tanto, una mayor tasa de error en relación con los audios de dominio médico-dental en comparación a los de dominio general.

Se realizó además una prueba de Kruskal - Wallis con el propósito de comparar las medias de los tres grupos (dental, médico y general), debido a que corresponde a grupos independientes. Se obtuvo un *P-value* de 0.010 rechazando así la hipótesis nula y demostrando una diferencia de medias significativa de al menos uno de los tres grupos. Así mismo, se tomó la decisión de realizar una prueba de Mann - Whitney para comparar medias de dos grupos (Ver Tabla 2), donde se obtuvo una significancia entre los grupos de dominio médico y general, con un *P-value* de 0.006. También se observó un *P-value* de 0.0512 entre el grupo dental y el general, lo que podría tener una significancia discutible en la práctica. Del mismo modo, se compararon las medias entre los grupos dental y médico, donde se obtuvo un *P-value* de 0.199 mostrando que no existen diferencias estadísticamente significativas entre ambos.

Tabla 2: Comparación de medias de los grupos usando la prueba de Mann - Whitney

Grupos	*P-value*
Médico / Dental	0.199
Medico / General	0.006**
Dental / General	0.0512*

Se puede apreciar la clara separación del grupo de dominio general en comparación a los grupos medico y dental, de esta forma se podría identificar un mayor error en los textos de dominio médico-dental.

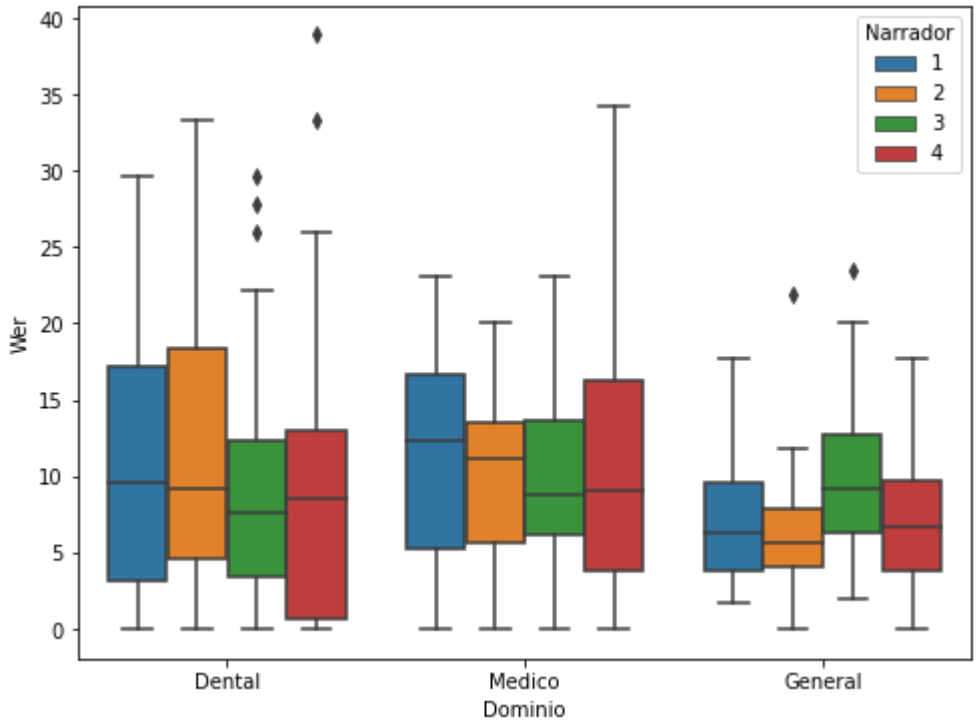

Figura 8: Se ilustran los valores WER de la muestra, separados por dominio y por narrador, con sus respectivas desviaciones estándares.

Por otra parte, se realizó una correlación de Spearman para evaluar una posible relación entre el largo del texto y el valor WER. La obtención de un *P-value* de 0.08 permite confirmar la hipótesis nula, concluyendo que no existe una correlación entre las variables antes mencionadas. Se realizó también una comparación de medias entre los diferentes participantes. Para esto, y debido a que esta vez las mediciones eran varias para cada participante, se tomó la decisión de realizar una prueba Wilcoxon para muestras relacionadas, la que arrojó un *P-value* de 0.088, es decir, no se observaron diferencias significativas entre los participantes.

7.2 Análisis por palabra

Se realizó un análisis exhaustivo para cada uno de los errores identificados en la transcripción, agrupándolos de manera arbitraria. De un total de 22.160 palabras analizadas por la herramienta "*Asr_evaluation"* (disponible en GitHub en la página https://github.com/belambert/asr-evaluation) se encontró un total de 2.087 errores, entre inserciones, deleciones y sustituciones. Entre dichos errores se identificaron varios grupos, siendo 2 los más relevantes y reiterados. En primer lugar, se observaron 139 errores que entraron al grupo denominado "Cambio de una palabra desde plural a singular o viceversa" (algunos ejemplos de este error son: diente/dientes, tiene/tienen, única/únicas). El segundo grupo con mayor cantidad de errores presentó un total de 99 errores, y corresponden al grupo denominado "Cambio de tiempo verbal o pronombre" (algunos ejemplos de este error son: estaba/está, acostumbran /acostumbran, soñó/soñé). En la tabla 3 se muestran los diferentes grupos identificados y el número de reiteraciones.

También se realizó un análisis de búsqueda de las palabras no identificadas por el servicio de transcripción de *Google Cloud*, constatando que se trató de palabras que fueron mal transcritas en los cuatro participantes, sin embargo, no es posible concluir que estas no sean reconocidas por el sistema de transcripción. El anexo 2 resume dichas palabras y muestra que entre las más relevantes se encuentran: urológico, endogastrio, nevo, distoclusión, trepanada, oftabiotico, entre otras. A su vez, el servicio de Google "inventó" las palabras "Inplantes" y "manecido"; la primera corresponde a un cambio de "m" por "n" y la segunda palabra no existe. Por otra parte, el sistema

corrigió cinco palabras mal escritas en los archivos *Gold Standard*, las cuales corresponden principalmente a tildes.

Tabla 3. Grupos de errores identificados y total de errores por grupo.

	Narrador 1	Narrador 2	Narrador 3	Narrador 4	Total	Ejemplo
Palabras / símbolos	1	0	1	1	3	más/+
Palabras / números	9	3	5	6	23	3/tres
Plural / Singular	41	23	37	38	139*	diente/ dientes
Cambio tiempo verbal/pronombre	30	20	27	22	99*	estaba/ está
eliminación de espacio	11	3	5	7	26	20 45/ 2045
aplicación o eliminación de tilde	12	11	5	1	29	citó/ cito
Espacio agregado	5	2	5	9	21	toda/ todo
Cambio de mayúsculas	12	3	8	6	29	hacer/ Hacer
Cambio género	5	3	2	3	13	cutáneo/ cutánea
Cambio de número	3	5	1	4	13	230816/ 230819
Total	129	73	96	97	395	

Se muestran los grupos de errores identificados y su nombre, además del total de errores por grupo y narrador, con sus respectivos ejemplos. Con un asterisco se denotan los valores mayores, en base al resultado del análisis con la herramienta "*Asr_evaluation".*

7.3 Análisis modelo de reconocimiento de entidades nombradas

El segundo análisis tuvo como propósito evaluar el desempeño del modelo NER sobre los textos transcritos. Como ya se mencionó anteriormente, solo se utilizaron para esta evaluación los textos de dominio médico y dental. En la Tabla 4 se identifican los dos dominios y las tres entidades estudiadas, con los respectivos resultados para cada una de las comparaciones realizadas, en base al conjunto *Gold Standard* anotado de manera manual y consolidado por tres profesionales del área de la salud. Este conjunto fue considerado el *Ground Truth* para la evaluación de los diferentes conjuntos, los que fueron anotados de manera automática por el modelo NER entrenado por Báez *et al*, 2020 (24), en base al corpus de texto de la lista de espera del sistema de salud chileno. Los resultados obtenidos se muestran en la Tabla 5, donde además se muestran los promedios obtenidos para los dominio médico y dental, con un valor *F1-score* de 0.80 y 0.86 respectivamente.

Tabla 4: Resultados F1- Score para las anotaciones automáticas realizadas sobre el conjunto Gold Standard y las anotaciones automáticas realizadas sobre los textos transcritos.

Dominio	Entidad	Gold Standard	Narrador 1	Narrador 2	Narrador 3	Narrador 4	Promedio
Dental	Enfermedad	0.8387	0.8511	0.8043	0.8298	0.7912	0.86
	Parte del cuerpo	0.973	0.973	0.8947	0.9636	0.9455	
	Medicamentos	0.7273	0.8485	0.4444	0.7273	0.6	
Médico	Enfermedad	0.8497	0.8188	0.8299	0.8188	0.8108	0.80
	Parte del cuerpo	0.9167	0.8116	0.8857	0.9014	0.9296	
	Medicamentos	0.963	0.88	0.88	0.88	0.9231	

Se consideró como referente de comparación, las anotaciones manuales y consolidadas realizada sobre los textos *gold standard*.

Para realizar una comparación entre los grupos y evaluar si existe diferencia entre la anotación automática realizada sobre un texto transcrito, se realizó una prueba estadística entre grupos. Estos se separaron en tres subgrupos y los textos de dominio dental y médico constituyeron tres subgrupos de diez archivos de texto cada uno. Se obtuvieron noventa valores *F1-score* para un correcto análisis en relación con las posibles diferencias encontradas entre grupos. Los valores correspondientes se presentan en la Tabla 5.

Tabla 5: Resultados F1- *Score* para anotaciones automáticas realizadas sobre el conjunto *Gold Standard* y anotaciones automáticas realizadas sobre los textos transcritos por narrador.

Entidad	Gold-standard	Narrador 1	Narrador 2	Narrador 3	Narrador 4	Subdivisión
Partes del cuerpo	0.9655	0.9286	0.9655	0.9286	0.8889	Dental sub1
	0.9744	0.9744	0.878	0.9744	0.95	Dental sub2
	0.9767	1.0	0.8636	0.9767	0.9767	Dental sub3
	0.303	0.6452	0.75	0.8125	0.8485	Med sub1
	0.7	0.9	1.0	1.0	1.0	Med sub2
	0.6667	1.0	1.0	0.9474	1.0	Med sub3
Enfermedades	0.8125	0.8485	0.8485	0.8485	0.8125	Dental sub1
	0.85	0.85	0.7368	0.85	0.7895	Dental sub2
	0.8571	0.8571	0.8571	0.7619	0.7619	Dental sub3
	0.5306	0.7917	0.7917	0.8163	0.7917	Med sub1
	0.9259	0.8235	0.8571	0.8	0.8627	Med sub2
	0.6909	0.84	0.84	0.8163	0.7755	Med sub3
Medicamentos	1.0	1.0	0	1.0	1.0	Dental sub1
	0.6667	0.6667	0.4	0.6667	0.6667	Dental sub2
	0.6667	0	0.6667	0.6667	0	Dental sub3
	0	0	0	1.0	1.0	Med sub1
	0.6667	0.8	0.8	0.8	0.8	Med sub2
	0.2	0.9474	0.9474	0.8889	0.9474	Med sub3

Se consideró como referente de comparación la anotaciones manuales y consolidadas realizada sobre los textos *Gold standard*, subdivididos en 3 grupos por dominio.

Se realizó una prueba de Friedman para comparar medias y evaluar si existen diferencias entre las anotaciones automáticas. El *P-value* fue de 0.50 por lo que se comprueba la hipótesis nula concluyendo que no existen diferencias entre los grupos comparados. También se realizó una comparación separándolos por entidad, realizando nuevamente una prueba de Friedman sin encontrar significancias.

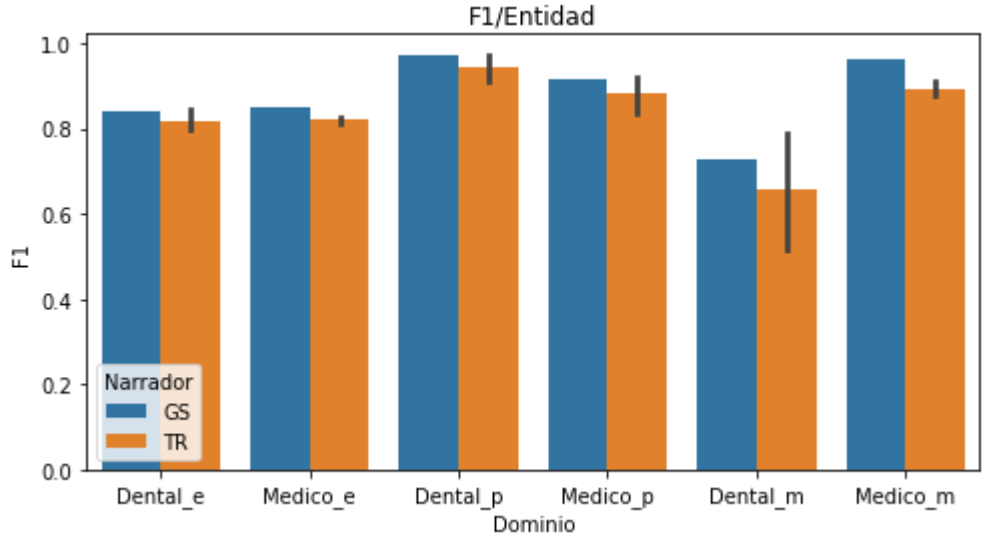

Figura 9: Resultados F1- Score de anotación automática para las entidades “Enfermedades”, “Partes del cuerpo”, “Medicamentos”, según dominio dental o médico comparado con su respectivo grupo *Gold Standard* anotado de manera automática.

7.4 Plataforma

La plataforma fue desarrollada en conjunto con ingenieros de la empresa “Create Spa”, la que en primer lugar solicita el inicio de sesión del usuario con la intención de realizar un registro de la información documentada por cada usuario. Para

este registro es necesario el ingreso de una dirección de correo electrónico y una contraseña que valida el ingreso (ver Figura 10). La plataforma está disponible tanto para dispositivos móviles como de escritorio.

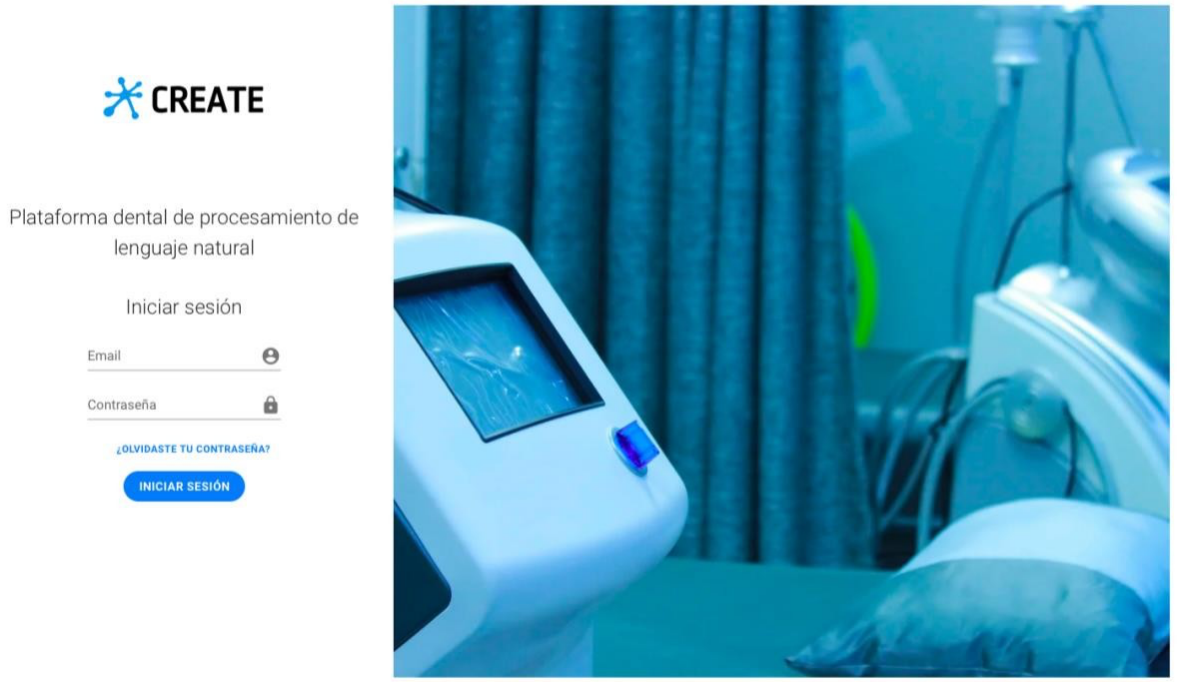

Figura 10: Imagen de inicio de la plataforma de transcripción y anotación, mostrando los campos de información requerida para el ingreso de los usuarios.

A continuación, se ingresa a la vista o pantalla de predicción en vivo (ver figura 11), donde una vez que se presiona el botón del micrófono se inicia la grabación del audio en modo *streaming*. Así, el sistema va autocompletando y corrigiendo la predicción de texto a partir del contexto, hasta detenerlo, ya sea de forma voluntaria desactivando el botón del micrófono o hasta que se completen los dos minutos (esto ha sido configurado por seguridad, para evitar la permanencia de micrófonos abiertos).

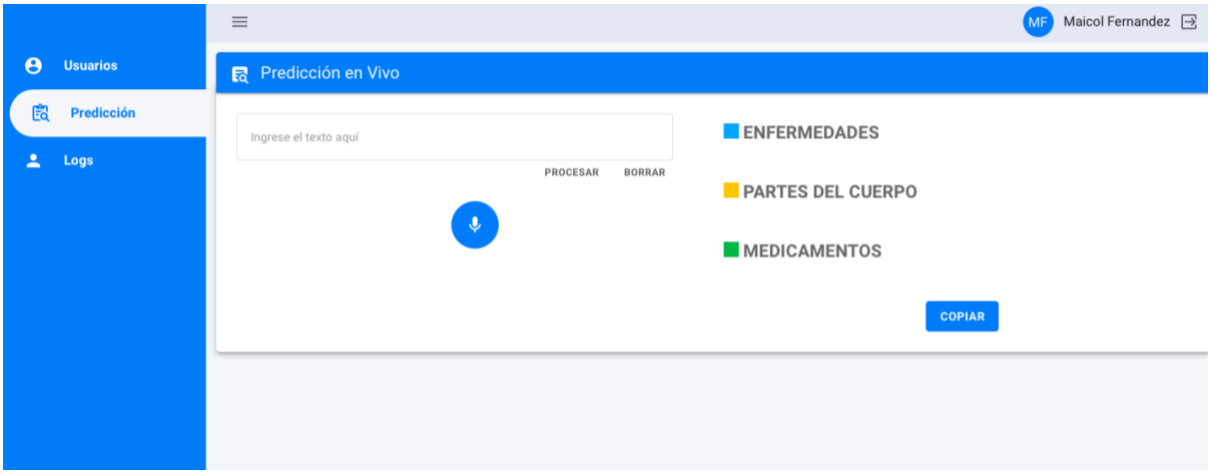

Figura 11: Imagen de la vista de predicción en vivo, con el área de texto al centro, el botón de grabación, representado por un micrófono, junto a los dos botones PROCESAR y BORRAR.

En forma posterior a la grabación y transcripción, se realiza la corrección sobre la predicción realizada por la plataforma, mediante el cuadro de texto de la vista de predicción en vivo (ver Figura 12) donde es posible corregir este tipo de errores. Para este propósito, existen dos botones bajo el área de texto, el botón BORRAR cuya función es borrar el texto dentro del cuadro y un segundo botón PROCESAR, que permite ejecutar el procesamiento del texto y la identificación de las tres entidades evaluadas en la presente tesis: Enfermedades, Partes del cuerpo y Medicamentos. La vista de predicción también presenta un botón COPIAR, en la zona inferior derecha (ver Figura 12), que permite copiar la información identificada y de esta forma transferirla a un sistema para el registro electrónico en salud.

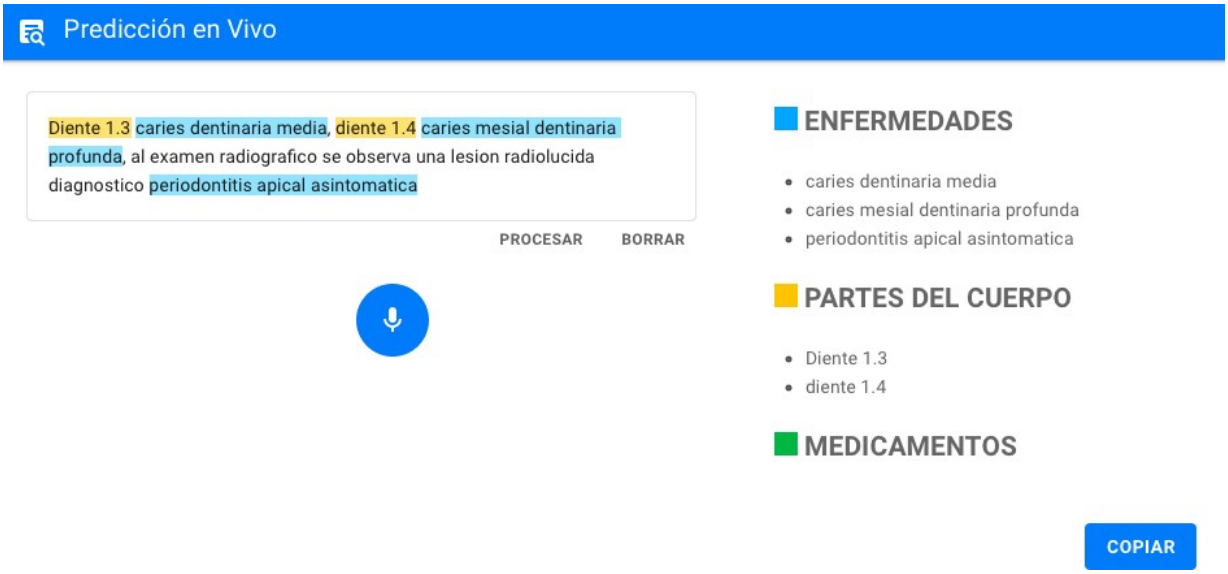

Figura 12: Imagen de la vista de predicción en vivo, de acceso posterior a la grabación de un audio y su transcripción correspondiente. A la izquierda se muestra el resumen de las entidades detectadas.

Después de la grabación, corrección del texto y anotación está la opción de importar la información. Al presionar el botón COPIAR se entrega la siguiente línea de caracteres:

> “Diente 1.3 caries dentinaria media, diente 1.4 caries mesial dentinaria profunda, al examen radiografico se observa una lesion radiolucida diagnostico periodontitis apical asintomatica
>
> ENFERMEDADES:
>
> - caries dentinaria media

- caries mesial dentinaria profunda

- periodontitis apical asintomatica

PARTES DEL CUERPO:

- Diente 1.3

- diente 1.4”

La plataforma cuenta también con un historial de consultas, lo que facilita la observación de consultas anteriores para de esta forma utilizar distintos dispositivos durante la atención clínica. Por ejemplo, usar un *smartphone* durante la atención y al momento de terminar la atención pasar al procesamiento en un equipo de escritorio, realizando las correcciones pertinentes al texto para luego realizar la identificación de entidades y copiar directo a la ficha clínica del paciente. Este sistema permite la interoperabilidad de nuestra plataforma ya que es posible copiar y pegar estos elementos en cualquier área de texto de un sistema para el registro electrónico en salud. La información recopilada desde la plataforma será almacenada para su posterior análisis y así permitir el desarrollo de actualizaciones que se ajusten a la necesidad de los usuarios.

8. Discusión

La usabilidad de los sistemas de registro clínico representa desafíos para los profesionales de la salud que los utilizan. En la actualidad, el tiempo destinado al registro clínico ocupa gran parte del tiempo de una cita médica o dental. Es por lo tanto de suma importancia mejorar dichos sistemas, apuntando a que se conviertan en un aliado de los profesionales de salud y no un enemigo al momento de utilizarlos. En esta tarea, resultan fundamentales los medios de registro electrónico.

Son pocos los *softwares* que tienen la función de reconocimiento de habla y los que existen son, en su gran mayoría, productos de habla inglesa, sin muchos acercamientos al idioma español. Uno de estos sistemas que puede ser utilizado en español es “INVOX medical” (28) el cual se autodefine en su página *web* como “el *software* de reconocimiento de voz para el dictado automático de informes médicos. El programa de dictado por voz más intuitivo del mercado, que permite una transcripción de audio a texto cómoda e instantánea”. Además, su publicidad señala que es usado en dieciocho países. Dentro de sus características se define como un sistema fácil e intuitivo, ya que tiene un diseño claro y sencillo, lo que garantiza un manejo cómodo, rápido y preciso. Se define también como un sistema para múltiples especialidades, ya que posee un diccionario específico y adaptado a numerosas especialidades médicas, así como un sistema de alta precisión que reconoce con exactitud una amplísima variedad de terminología médica.

Una de características que llama la atención es que recomiendan la utilización de unos micrófonos en específicos, similares a grabadoras de voz, las cuales pueden ser un poco aparatosas. También este sistema requiere un *software* instalado en el computador y que al parecer carecería de una versión para dispositivos móviles. Quizás dentro de las características más deseable de estas herramientas es la utilización en dispositivos móviles como un teléfono celular o *tablet* y de esta manera entregar mayor versatilidad al momento de ingresar información sin la necesidad de la comprar micrófonos de alta gama. El sistema INVOX medical es un *software* disponible para Windows, y tiene un costo que va desde la versión gratuita (con funciones y horas de transcripción limitadas), hasta su versión PRO, que tiene un valor de 45 euros al mes (cercano a los $42.000 pesos chilenos).

Otra alternativa es el *software Dragon medical* perteneciente a la empresa Nuance, un sistema de similares características al INVOX medical, con un precio de US$ 1.360 (1.115.000.- pesos chilenos) que admite solo una licencia para la utilización del *software*. En su página *web* enuncia ser la mejor herramienta para el reconocimiento de habla en español, con una precisión del 99%. Dentro de los beneficios de utilizar esta herramienta se describe el ganar tiempo y ahorro de recursos, y que las soluciones de la empresa Nuance permiten el ahorro de un promedio de sesenta minutos al día en la molesta labor de la documentación clínica, según se indica en su página *web* (29). El *software* Dragon medical posee un sistema de navegación y corrección por voz que permite a los usuarios navegar y corregir dictados en su aplicación con sólo su voz. Los comandos de voz integrados simplifican el dictado, la navegación y las correcciones de texto. Por otra parte, se enuncia que los sistemas de documentación clínica que usan tecnologías de habla

capturan más y mejores datos que mediante la vía escrita habitual.

Una particularidad que presenta los sistemas evaluados en esta investigación, es que permitiría no sólo abaratar costos asociados a la compra de los *softwares* mencionados anteriormente, sino que también permitiría la recopilación de información clave, lo cual es una función que no está disponible en los sistemas ya comercializados y lo que le entregaría un valor agregado a nuestra herramienta.

En cuanto a los resultados obtenidos en el presente trabajo, el valor WER promedio para los tres dominios fue de un 9.82%, lo cual está por debajo de lo enunciado en la hipótesis, ya que se esperaba un 30%. Este valor de umbral fue tomado en base a la comparación realizada por Filippidou *et al*. del año 2020 (22) donde se obtuvo un valor WER promedio de 18.63% entre los tres participantes. Este estudio fue realizado para lengua inglesa y con temáticas distintas a las del ámbito de la salud. Es por estas razones que en el presente trabajo se consideró un valor conservador de 30% para la evaluación del desempeño de nuestro sistema. El bajo error promedio obtenido aquí (9.82%) podría deberse a que el servicio de Google Cloud pudo sufrir modificaciones y actualizaciones que permitieran la mejora en el desempeño.

Otro factor que se debe considerar así es el idioma utilizado, ya que en general, las herramientas funcionan mejor en el idioma inglés y al mismo tiempo existen más corpus de entrenamiento, lo que las torna más robustas. Cabe considerar que dada la estructuración más compleja del español y la gran conjugación de verbos podrían facilitar la

contextualización y predicción, de tal forma que podría presentar un menor grado de dificultad y por ende equivocarse menos mejorando su desempeño. Esto podría explicar en cierta forma los resultados de esta investigación en torno a la métrica WER. Existe un estudio del año 2006 en el cual se comparan 3 idiomas (español, inglés y mandarín) donde se observa una menor tasa de error WER en las transcripciones en español frente al inglés (30), pero debido a la antigüedad de la publicación este punto debería ser analizados en futuros estudios. Así también, en la investigación realizada por Filippidou *et al.,* 2020 (22) participaron tres narradores hablantes no nativos del inglés y su lengua materna es el griego, por ende, los sistemas evaluados fueron puestos a prueba y podría ser está la razón con mayor peso que podría explicar la diferencia de los valores encontrados en la presente tesis y los de la mencionada publicación.

Por otra parte, en este trabajo también se realizó la evaluación de las anotaciones realizadas por el modelo NER desarrollado por Báez *et al* (24) realizando una comparación entre tales resultados y los valores identificados en la presente investigación. Los valores *F1-score* publicados por el equipo de Báez van entre 0.75 a 0.92, por lo que pareció razonable considerar un promedio simple de 0.81, valor considerando la entidad de abreviaciones que tuvo el valor más alto de reconocimiento por el modelo. Es por esta razón, que se tomó la decisión de considerar un valor de 0.75 como un valor mínimo esperado, considerando los posibles errores de transcripción. Aun así, el sistema evaluado en esta investigación tuvo un valor *F1-score* que va entre 0.80 a 0.90, por lo que se podría considerar como resultado un valor promedio de 0.85, que es superior a lo esperado en la hipótesis. El mejor desempeño se identificó en relación con el reconocimiento de partes del

cuerpo en el dominio dental, con un *F1-score* de 0.944.

Por otra parte, se identificó el peor desempeño en relación con el reconocimiento de medicamentos en el dominio dental, con un valor *F1-score* de 0.655. También se realizó una correlación de Spearman para evaluar si las variables *F1-Score* y WER estaban relacionadas y se identificó que existe una correlación negativa y significativa entre las variables mencionadas, pero solo en la entidad Medicamentos, con un *P-Value* de 0.0037 lo que demostraría que al aumento del error por palabra (WER) existiría una disminución del *F1-Score* asociado al reconocimiento de medicamentos.

En relación con los errores más comunes identificados durante el proceso de análisis desarrollado con la herramienta de "*Asr_evaluation",* se identificó que existe un evidente error relacionado al cambio de palabras, desde el plural al singular o viceversa. Esto genera un efecto en cadena, ya que, al cambiarse una palabra, las que vienen a continuación también lo hacen. Un ejemplo de esto es: "las únicas piezas del mobiliario eran una cama una mesa y un banco" frase que cambió a "la única pieza de mobiliario era una cama una mesa y un banco". En esta oración se cambiaron las palabras: "las", "únicas", "piezas" y "eran", lo que significó el error de cuatro palabras y, consecuentemente, una penalización importante en el WER. De este error se podría responsabilizar al servicio de Google, pero también podría corresponder a un error de dictado por el narrador. A su vez, esto podría explicar la diferencia en el valor WER encontrado en esta investigación y los valores encontrados por Fillippidou *et al* (22) ya que, al ser textos de mayor longitud, un error como el mencionado anteriormente podría desencadenar un error en cadena que a la larga se ve reflejado en su valor WER. En esta

tesis se utilizaron textos cortos, lo que podría explicar que los valores WER fueran mucho menores a los esperados.

Otra observación realizada al servicio de transcripción recae en el segundo error con mayor número de reiteraciones, denominado “cambio de tiempo verbal o pronombre”, lo que ocurre en un gran número de ocasiones. Algunos ejemplos son: “toma” por “tomé”, “ponían” por “pongan”, “exilió”por “exilio”, “solicita” por “solicité”, entre otros. De esta forma, a pesar de que las palabras corresponden al verbo, su conjugación no corresponde a la del texto original, lo que genera su reconocimiento como un error por sustitución, aumentando así el valor WER.

Existen palabras que no fueron reconocidas por el servicio de Google como “nevo”, “endogástrio” y “mesiodens”. Estas palabras aparecen en forma acotada y puede deberse a factores tales como la dicción de los narradores o la calidad del micrófono. Sin embargo, estos términos son de gran interés clínico y podrían estar mostrando la punta del *iceberg* de un problema mayor. Es por esto que el servicio de transcripción de Google desarrolló modelos especiales para la transcripción médica, versiones que lamentablemente solo se encuentran disponibles para el idioma inglés.

Como se mencionó anteriormente, los valores esperados para la transcripción y reconocimientos de información clave planteados en la hipótesis, se vieron claramente superados. Es por esta razón que el próximo paso será la evaluación de nuestro sistema en comparación con otros sistemas de reconocimiento de voz para de esta manera evaluar

y comparar su desempeño, con el fin de demostrar cómo este beneficiaría el desarrollo de una conversación más fluida entre el profesional de salud y sus pacientes, y al mismo tiempo evaluar su efecto en disminuir el tiempo destinado a la documentación clínica.

Otra herramienta, disponible solo en lengua inglesa, por lo que no fue incluida en esta tesis, pero que podría ser considerara en un futuro estudio es el reconocimiento de voz usando modelos médicos. Dentro del servicio de Google *Speech-to-text* existen 2 modelos para reconocimiento de voz en situaciones médicas (31). Estos dos modelos permiten el reconocimiento de palabras que son comunes en entornos médicos. El primer modelo cumple la función de reconocimiento de habla mediante el dictado de voz del profesional y el segundo modelo permite el reconocimiento de voz de una conversación entre un profesional y un paciente, entregando como resultado un texto con palabras como "*spk:provider*" y "*spk:patient*" ubicadas al comienzo de la oración con el propósito de identificar los interlocutores de la conversación.

9. Conclusión

Podremos identificar una muy buena evaluación del desempeño, ambas herramientas (*Speech-to-Text* y el modelo NER) funcionan muy bien, en particular al momento de reconocer entidades como partes del cuerpo y enfermedades cuyos valores *F1-score* son superiores a los esperados y validaría nuevos estudios en ambientes clínicos. Por otra parte, el desempeño del reconocimiento de entidades como Medicamentos se vio principalmente perjudicada especialmente en el dominio dental donde se identificó el peor valor en relación al desempeño de estas herramientas. Así también, se reconoció una correlación significativa negativa entre el valor WER y el *F1-score* por lo que se podría concluir que un fallo en la transcripción podría desencadenar un error al momento de reconocer información clave en los textos. Es importante destacar que, entre los valores del desempeño de las entidades, Partes del cuerpo y Enfermedades, no se encontró una correlación entre el valor WER y *F1-score*.

Existen errores agrupados y reiterados por el servicio de *Google Cloud* los cuales deberían ser profundizados por el proveedor para mejorar su rendimiento. Ejemplos son el cambio de palabras desde el plural al singular y el cambio de tiempo verbal o pronombres, errores que son reiterativos y aumentan el valor WER, yendo en desmedro del desempeño de una gran herramienta.

Así también, existen errores en la transcripción asociado a un posible desconocimiento de algunos términos, especialmente lenguaje técnico de utilización diaria en los establecimientos de salud. Una herramienta específica para la transcripción de

términos médicos o dentales beneficiarían el desarrollo de herramientas como la evaluada en esta investigación, disminuyendo la tasa de error.

En relación con el reconocimiento de entidades nombradas, no se identificaron diferencias significativas al momento de comparar a los narradores participantes, pero sí se podría señalar que un desempeño deficiente en relación al *F1-score* podría ser consecuencia de una transcripción deficiente.

El desarrollo de la plataforma expuesta en esta investigación sienta las bases para futuras investigaciones sobre su evaluación en ambientes clínicos y todo lo que esto conlleva: interrupciones por variados interlocutores, ruidos externos, problemas técnicos con el micrófono, etc. Al realizar una evaluación clínica del sistema será posible identificar errores que impidan su correcta utilización y como esta se ve afectada por factores externo. De esta forma se podrán realizar actualizaciones que permitan y faciliten su masificación.

10. Bibliografía

1.- Luh JY, Thompson RF, Lin S. (2019). Clinical documentation and patient care using artificial intelligence in radiation oncology. J Am Coll Radiol [Internet, citado el 13 de enero de 2022]; 16(9 Pt B): 1343–6. Disponible en: https://www.jacr.org/article/S1546-1440(19)30696-9/fulltext.

2.- Dalianis H. (2018). Clinical Text Mining. Cham: Springer International Publishing.

3.- Friedberg MW, Chen PG, Van Busum KR, Aunon F, Pham C, Caloyeras J, et al. (2014). Factors affecting physician professional satisfaction and their implications for patient care, Health systems, and Health policy. Rand Health Q. 3 (4): 1.

4.- Outomuro D, Actis AM. (2013). Analysis of ambulatory consultation length in medical clinics. Rev Med Chil. 141(3): 361–6.

5.- Minsal.cl. [citado el 13 de enero de 2022]. Disponible en: https://www.minsal.cl/wp-content/uploads/2021/09/2021.09.15_Orientaciones-para-la-planificación-y-programación-de-la-red-2022.pdf

6.- Harvard business review (2018). To combat physician burnout and improve care, fix the electronic health record. [Internet]. el 30 de marzo de 2018 [citado el 9 de julio de 2021]; Disponible en: https://hbr.org/2018/03/to-combat-physician-burnout-and-improve-care-fix-the-electronic-health-record.

7.- Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. (2019). Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ Digit Med. 2 (1): 114.

8.- Irwin JY, Harkema H, Christensen LM, Schleyer T, Haug PJ, Chapman WW. (2009).

Methodology to develop and evaluate a semantic representation for NLP. AMIA Annu Symp Proc. 2009: 271–5.

9.- Blackley SV, Huynh J, Wang L, Korach Z, Zhou L. (2019). Speech recognition for clinical documentation from 1990 to 2018: a systematic review. J Am Med Inform Assoc. 26 (4): 324–38.

10.- Hodgson T, Coiera E. Risks (2016). Benefits of speech recognition for clinical documentation: a systematic review. J Am Med Inform Assoc. 23 (e1): e169-79.

11.- Soltau H, Wang M, Shafran I, Shafey LE. (2021). Understanding medical conversations: Rich transcription, confidence scores & information extraction [Internet]. arXiv [cs.LG]. Disponible en: http://arxiv.org/abs/2104.02219.

12.-Ajami, Sima (2016). “Use of speech-to-text technology for documentation by healthcare providers.” The National Medical Journal of India vol. 29, 3: 148-152.

13.- Kodish-Wachs J, Agassi E, Kenny P 3rd, Overhage JM. (2018). A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. AMIA Annu Symp Proc [Internet, citado el 7 de marzo de 2022]. Disponible en: https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC6371385/.

14.- Zuchowski M, Göller A. (2022). Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. Br J Health Care Manag [Internet]. 28(1): 30–6. Disponible en: http://dx.doi.org/10.12968/bjhc.2021.0074.

15.- librispeech_asr · Datasets at Hugging Face [Internet]. Huggingface.co. [citado el 7 de marzo de 2022]. Disponible en: https://huggingface.co/datasets/librispeech_asr.

16.- Negrão M, Domingues P. (2021). SpeechToText: An open-source software for automatic detection and transcription of voice recordings in digital forensics. Forensic Science International: Digital Investigation [Internet]. 38 ;(301223). Disponible en: http://dx.doi.org/10.1016/j.fsidi.2021.301223.

17.- Sharma, F., Wasson, S.G. (2012). A Speech Recognition and Synthesis Tool : Assistive Technology for Physically Disabled Persons. International Journal of Computer Science and Telecommunications [Volume 3, Issue 4: 86-91.

18.- Këpuska, V.: Comparing speech recognition systems (Microsoft API, Google API And CMU Sphinx). Int. J. Eng. Res. Appl. 07(03), 20–24 (2017). https://doi.org/10.1007/s10772-014-9223-y.

19.- Amazon Transcribe – Automatic Speech Recognition - AWS, Amazon Web Services, Inc. https://aws.amazon.com/transcribe/. Accessed 26 Apr 2019.

20.- IBM Watson | IBM. https://www.ibm.com/watson. Accessed 05 May2019.

21.- Wit.ai. https://wit.ai/. Accessed 05 May 2019.

22.- Filippidou F, Moussiades L. (2020). A benchmarking of IBM, Google and wit automatic speech recognition systems. En: IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing; p. 73–82.

23.- Oronoz M, Gojenola K, Pérez A, de Ilarraza AD, Casillas A. (2015). On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. J Biomed Inform. ;56: 318–32.

24.- Báez P, Villena F, Rojas M, Durán M, Dunstan J. (2020). The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish. En: Proceedings of the 3rd Clinical Natural Language Processing Workshop.

Stroudsburg, PA, USA: Association for Computational Linguistics. DOI: 10.18653/v1/2020.clinicalnlp-1.32.

25.- Introduction - brat rapid annotation tool [Internet]. Nlplab.org. [citado el 26 de julio de 2021]. Disponible en: https://brat.nlplab.org/introduction.html

26.- Lenz Furrer, Joseph Cornelius, and Fabio Rinaldi. (2019). Uzh@ craft-st: a sequence-labeling approach to concept recognition. In Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, pages 185– 195.

27.- Crook J. About [Internet]. Audacityteam.org. [citado el 26 de julio de 2021]. Disponible en: https://www.audacityteam.org/about/

28.- INVOX Medical: la solución definitiva para dictar informes médicos [Internet]. INVOX Medical. 2017 [citado el 18 de enero de 2022]. Disponible en: https://invoxmedical.com

29.- Dragon Medical soluciones de voz y dictado para médicos I Nuance.es [Internet]. Nuance Communications. [citado el 18 de enero de 2022]. Disponible en: https://www.nuance.com/es-es/healthcare/physician-and-clinical-speech/dragon-medical.html

30.- Lamel, Lori & Gauvain, Jean-Luc & Adda, G & Barras, C & Bilinski, E. & Galibert, Olivier & Pujol, A & Schwenk, Holger & Zhu, X. (2006). The LIMSI 2006 Tc-Star transcription systems. Proc. TC-STAR Workshop on Speech-to-Speech Translation. 123-128.

31.- Recognize speech by using medical models [Internet]. Google Cloud. Disponible en: https://cloud-google-com.translate.goog/speech-to-text/docs/medical-models?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=sc

11. Anexos

Anexo 1: Tabla de descripción de las tareas encontradas en el desarrollo de un escriba digital y los desafíos que se deben enfrentar en cada una de ellas.

Tarea	Desafío
Grabación del audio	• Alto ruido ambiental.
	• Fidelidad del micrófono.
	• Múltiples hablantes.
	• Posición del micrófono relativa entre el profesional y el paciente.
Reconocimiento automático del habla	• Calidad de audio variable.
	• Alto ruido ambiental.
	• Múltiples hablantes.
	• Disfluencias, falsos comienzos, interrupciones, pausas no léxicas.
	• Complejidad del vocabulario médico.
	• Volumen variable del orador debido a la distancia al micrófono y a la posición relativa.

	• Diferenciación de varios oradores en el audio (diarización del orador).
Segmentación de tópicos	• Conversación no estructurada.
	• Progresión no lineal de los temas durante una conversación médica.
Extracción de conceptos médicos	• Resultados ruidosos de los programas de asignación de texto a UMLS.
	• Ajuste de los parámetros de las herramientas utilizadas para la asignación de texto al *Unified Medical Language System* (UMLS)
	• inferencia contextual (comprensión del significado apropiado de una palabra o frase dado el contexto).
	• Fenómenos en el habla espontánea como la anáfora cero, el pensamiento en voz alta, la deriva temática.
Resumir	• Resumir la comunicación no verbal no estructurada.
	• Integración de los conocimientos médicos para identificar la información pertinente.
	• Inferencia contextual.

	• Resolver la información conflictiva del paciente.
	• Actualización de las hipótesis a medida que el paciente revela más información.
	• Generación de resúmenes para entrenar un modelo de resumen mediante Machine Learning.
Recogida de datos	• Preocupación por la privacidad del médico y del paciente.
	• Recogida de datos y etiquetado costosos.
	• Consentimiento del paciente para ser grabado y utilizar los datos con fines de investigación.
	• Desidentificación y anonimización de datos.
	• Conjuntos de datos costosos.
	• Los datos se mantienen en privado como un activo de propiedad intelectual.
	• Reticencia de los médicos a ser registrados por temor a las responsabilidades legales y a la carga de trabajo adicional.

Tabla tomada y traducida de “Challenges of developing a digital scribe to reduce clinical documentation burden”. (Quiroz *et al*, 2019; Referencia 7).

Anexo 2: Observaciones realizadas al servicio de *Google Cloud Speech to text*, palabras no reconocidas por el sistema o palabras inventadas por este.

Observaciones	
Densitometría	
tscore	reconoce texcore en 2 casos
Urológico	reconoce urología en 1 caso
Endogastrio	
Nevo	
Oftabiotico	
Acenocumarol	Reconoce acenocumarina
Distoclusión	
Vestibularizado	
Irrigación	
Trepanada	
Defocación	
mesiodens	
Plancton	
Radiografía	
Ató	
Inplantes	mal escrito por Google
manecido	mal escrito por Google