



INSTITUTO TECNOLÓGICO DE MORELIA

MAESTRÍA EN CIENCIAS EN INGENIERÍA ELECTRÓNICA

DIRECTOR: JUAN CARLOS OLIVARES ROJAS

CODIRECTOR: GERARDO MARX CHÁVEZ CAMPOS

REVISOR: ADRIANA DEL CARMEN TÉLLEZ ANGUIANO

REVISOR: JOSÉ ANTONIO GUTIERREZ GNECCHI

REPORTE DE AVANCES DE INVESTIGACIÓN

**SISTEMA MULTIMODAL BASADO EN PROCESAMIENTO DE
LENGUAJE NATURAL PARA LA CAPTURA Y CONSULTA DE
INFORMACIÓN EN BASE DE DATOS MÉDICAS**

AUTOR: I. PEDRO MATA MARTÍNEZ

N.C.: M22121544

FECHA: 29/05/2024

Problemática

La captura de datos en el área de la salud, más precisamente hablando, de la diabetes; debido a que cuando se realiza el interrogatorio y la exploración física a un paciente, el medico no puede realizar tareas de manera simultánea, impidiendo dar una mejor atención al paciente. Dicho esto, el principal objetivo de esta investigación es la captura y recuperación de información para el llenado de campos de historiales médicos de pacientes con diabetes mediante el análisis de procesamiento de lenguaje natural a grabaciones de consultas médicas.

Objetivos

Principal

- Investigar métodos de procesamiento de lenguaje natural para la implementación de un sistema multimodal que permita agilizar la captura y búsqueda de información en base de datos médicas.

Específicos

- Obtener el listado de la información relevante para una consulta médica enfocada en diabetes por un médico especialista en el área.
- Separar en secciones los tipos de información para un historial médico de un paciente generado por una consulta médica enfocada en la diabetes.
- Crear una base de datos en MySQL que pueda almacenar todos los campos necesarios relacionados a un historial médico de un paciente con diabetes.
- Crear un código en Python que pueda realizar la distinción de voces entre dos distintas personas con tal de diferenciar entre los diálogos del paciente y del médico especialista en diabetes.
- Crear un código en Python que pueda almacenar una conversación de audio para posteriormente transcribirla a un archivo de texto.
- Crear un código en Python con la capacidad de analizar el texto obtenido con NLP con la finalidad de recuperar la información relevante para el historial médico.
- Almacenar la información relevante para el historial médico en la base de datos.

Hipótesis

El desarrollo de un sistema multimodal basado en procesamiento de lenguaje natural permitirá una captura y búsqueda más rápida de los datos en base de datos médicas.

Metodología

Para el proceso del proyecto de investigación es necesario iniciar con la obtención de la información relevante del sistema de los historiales médicos del Centro de Atención a la Diabetes del instituto Mexicano de Seguro Social (CADIMSS), para conocer las interfaces y los campos con los que son llenados los historiales.

Después de analizar la estructura de su base de datos, se procedió a crear un modelado de la conversación entre el médico y el paciente, con tal de conocer las preguntas y respuestas que son de utilidad para llenar los campos de un historial médico de un paciente con diabetes, en otras palabras, el contexto.

Una vez creado un primer modelo, se enlistaron palabras para crear un vocabulario con posibles preguntas y respuestas de la conversación, del médico y paciente respectivamente.

Por parte de la estructura de hardware (HW) se han creado las primeras pruebas en una computadora personal (P.C.) para aumentar la velocidad de pruebas y búsqueda de información.

Para la estructura del software (SW) se han realizado las actividades que se muestran en el siguiente apartado del documento.

Posteriormente se aplicarán métodos y algoritmos de I.A. para el análisis de las conversaciones y llenado de campos de la base de datos de los historiales médicos de pacientes con diabetes del CADIMSS.

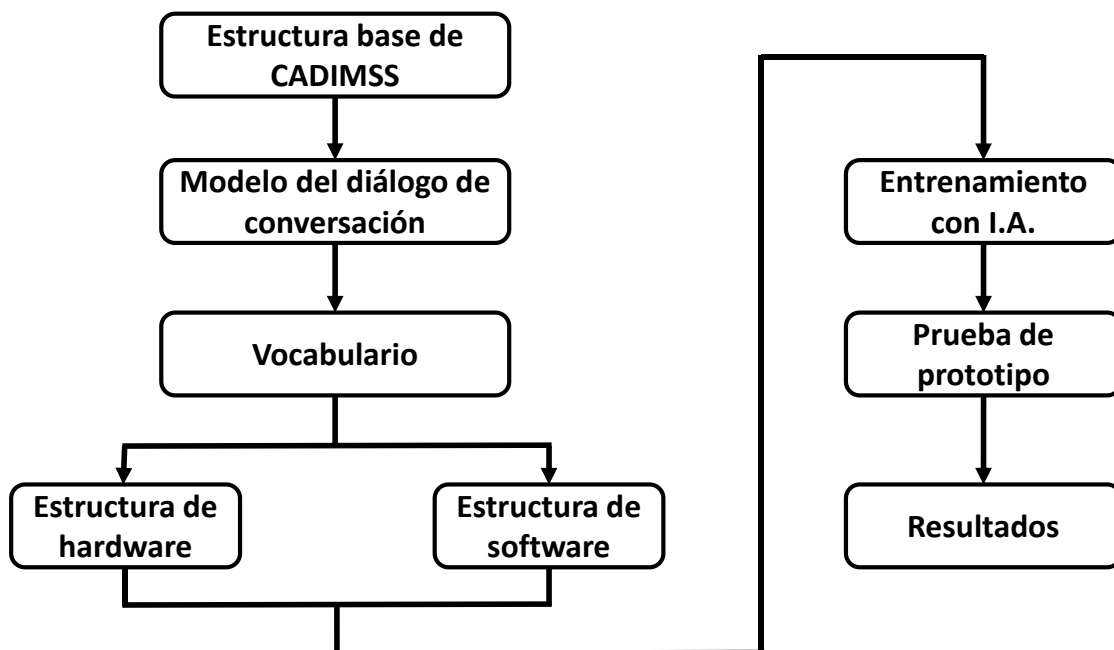


Ilustración 1 - Metodología para el desarrollo de la implementación de NLP en la captura y recuperación de información de consultas hospitalarias

Actividades

Para el periodo de Enero – Junio del año 2024 en el caso del proyecto de tesis “Sistema multimodal basado en procesamiento de lenguaje natural para la captura y consulta de información en bases de datos médicas” se dedicó el tiempo en las actividades optadas para el cronograma que se muestra a continuación con énfasis en el desarrollo del algoritmo para el Procesamiento de Lenguaje Natural (PLN), no obstante, para poder desarrollar dicho algoritmo es necesaria la realización de pruebas básicas como la ejecución de instrucciones por medio de voz, y de manera manual; así como el desarrollo base de PLN en Python; la generación del lexicón o diccionario especializado para el programa.

Tabla 1 – Actividades planeadas del cronograma para el periodo enero - julio 2024

Actividades	Periodo de enero a julio de 2024													
	Ene		Feb		Mar		Abr		May		Jun		Jul	
	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Redacción de tesis		X		X		X		X		X				
Creación del algoritmo NLP en la aplicación de Python	X	X	X	X	X	X	X							
Modificaciones a la DB	X								X	X				
Modificaciones a las interfaces web	X								X	X				
Prueba de programa NLP en consultas médicas para recuperación de datos														
Redacción del artículo científico			X	X		X		X	X					
Modificaciones finales														
Entrega final de documentación														

Como se puede apreciar en la tabla anterior (*Tabla 1*), se muestran las actividades planeadas para el periodo de enero a junio del año 2024, donde se resalta con una marca “X” en las que se lograron realizar, pero como se mencionó anteriormente, para continuar con el desarrollo del proyecto fue necesario implementar una serie extra de actividades para que el funcionamiento básico sea efectivo en pruebas futuras.

Las actividades extra que fueron necesarias de desarrollar fueron:

1. Ejecución de instrucciones por medio de comandos directos y grabaciones de voz.

Para que el proyecto tenga una base de PLN con el fin de ser usando en la captura y consulta de información en bases de datos médicas, fue requerido de primera instancia que por medio de comandos directos de voz se pudieran ejecutar instrucciones que modificaran la base de datos, por lo que se efectuaron pruebas en una base de datos alterna para

verificar su correcto comportamiento con las instrucciones existentes para la modificación de información en tablas previamente creadas, que son: "INSERT", "UPDATE" y "DELETE".

La instrucción "INSERT" en una base de datos tiene la función de agregar nuevos datos a los campos de la tabla existente seleccionada; para que dicha instrucción fuera ejecutada en las condicionales para el comando de voz se usaron palabras relacionadas o sinónimos además de la búsqueda de palabras específica relacionada al tema de investigación del proyecto que es la diabetes, así como glucosa, enfermedad, padecimiento, síntoma, etc.

La instrucción "UPDATE" tiene la función de modificar uno o varios campos existentes en donde se cumpla una condición de coincidencia, en otras palabras, se cambiará el o los datos en la tabla seleccionada donde la condicional que se esté buscando se cumpla. Por decir un ejemplo: modificar todos los campos donde se encuentre la palabra enfermedad y sustituirlas por padecimiento.

Por último, la instrucción de "DELETE" nos permite eliminar filas completas de la tabla designada en donde alguno de los campos seleccionados como condicionales se cumpla. Esta instrucción sería en teoría la menos usada debido a que anteriormente solo se deberán insertar datos que sean procesados de manera correcta por el algoritmo de PLN.

2. Generación del lexicón

Un lexicón es un listado de los elementos (palabras) que se forman a partir de una lengua sobre una rama o campo específico del conocimiento humano, en palabras más simples, es un inventario o diccionario donde se aglomeran todas las palabras relevantes al tema o campo en el que se desenvuelve.

Para ser analizados deben contar con una estructura creada a partir de la terminología y gramática de las palabras que son usadas al momento de hablar o interactuar sobre el campo en cuestión, esto con la finalidad de conocer cuáles palabras son las más referentes y relevantes del tema, así como su significado y similitud entre ellas. Para dicha actividad se realizó el proceso con el siguiente orden:

- Grabación de la interacción entre doctor y paciente de una consulta del CADIMSS
- Transcripción de la grabación de voz a texto
- Seccionamiento de los distintos campos de los tipos de antecedentes del paciente
- Tokenización de las palabras relevantes de cada uno de los diálogos entre el doctor y el paciente de la consulta del CADIMSS
- Recuento de palabras encontradas para analizar la similitud y diferencias entre cada uno de los diálogos de los distintos tipos de antecedentes del paciente

Una vez realizadas estas actividades queda de manera pendiente comparar por medio de un algoritmo el grado de similitud que se encuentra entre cada una de las palabras dentro de un mismo campo o tipo de antecedente, y, entre cada uno de los tipos.

3. Desarrollo de programa de Procesamiento de Lenguaje Natural en Python

Para el apoyo del desarrollo de algoritmo de PLN en el lenguaje de Python se ha usado como apoyo el libro “*Machine Learning con Python y Scikit-Learn*” de Raschka, Liu y Mirjalili. En el cual se muestra el desarrollo de programación para I.A. en Python desde casos más básicos como son el perceptrón hasta la aplicación de algoritmos para analizar el lenguaje, es decir, Programación del Lenguaje Natural.

Se han realizado ejemplos y ejercicios proporcionados por el libro y las ligas que éste muestra para comprender la sintaxis y los resultados que son otorgados por estos con la finalidad de analizarlos y darles un fin específico. Se planea que la base de este libro sirva como un apoyo extra que pueda permitir un mejor desarrollo de para el programa con PLN para el proyecto.

Dichos estos hechos, la tabla de actividades (*Tabla 2*) para el periodo de enero a julio del año 2024 quería de la siguiente manera, tomando en cuenta cada una de las actividades extras que se realizaron durante este periodo:

Tabla 2 – Actividades modificadas del cronograma para el periodo enero - julio 2024

Actividades	Periodo de enero a julio de 2024													
	Ene		Feb		Mar		Abr		May		Jun		Jul	
	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Redacción de tesis		X		X		X		X		X				
Generación del lexicon				X	X	X								
Ejecución de instrucciones en DB por voz							X	X	X	X				
Creación del algoritmo NLP en la aplicación de Python	X	X	X	X	X	X	X							
Modificaciones a la DB	X								X	X				
Modificaciones a las interfaces web	X								X	X				
Prueba de programa NLP en consultas médicas para recuperación de datos														
Redacción del artículo científico			X	X		X		X	X					
Modificaciones finales														
Entrega final de documentación														

Las celdas marcadas de color naranja representan las nuevas actividades que fueron realizadas durante el periodo de enero a junio de 2024, y las celdas marcadas de color amarillo representan aquellas actividades a las que se le dio un cambio de enfoque.

Resultados y evidencias

Para demostrar algunos ejemplos de los avances realizados durante el periodo de enero a junio del 2024, se han agregado una serie de imágenes como resultado de las actividades anteriormente mencionadas. Para una mejor resolución y una vista más completa de las evidencias se estarán constantemente agregando y/o modificando en el siguiente enlace:

https://github.com/Mata-ingmec/ITM_thesis_project

Para las evidencias de la generación del lexicón se hace uso de un programa en Python que puede convertir archivos de audio .wav a texto, guardados en la carpeta de archivos en donde se encuentra el programa, con la finalidad de tokenizar y contabilizar las palabras usadas en la conversación, además de encontrar aquellas que son relevantes para cada campo o tipo de antecedente para el historial médico de un paciente con diabetes, como se puede apreciar en la *Ilustración 2*.

```
{ 'transcript': 'usted tuvo algún traumatismo algún '
                        'esguince o fractura en su vida '
                        'pasada sí anteriormente me he '
                        'fracturado la costilla izquierda y '
                        'también tuve una un esguince en el '
                        'brazo derecho usted tuvo alguna '
                        'alguna otra factura en su vida sí de '
                        'niño me fracturé la pierna jugando a '
                        'football'}}],

'final': True}

Texto Completo:

usted tuvo algún traumatismo algún esguince o fractura en su vida pasada sí anteriormente me he fracturado la costilla izquier
da y también tuve una un esguince en el brazo derecho usted tuvo alguna alguna otra factura en su vida sí de niño me fracturé l
a pierna jugando fútbol

Tokens Totales: 48
['usted', 'algún', 'traumatismo', 'algún', 'esguince', 'fractura', 'vida', 'pasada', 'anteriormente', 'fracturado', 'costilla',
'izquierda', 'esguince', 'brazo', 'derecho', 'usted', 'alguna', 'alguna', 'factura', 'vida', 'niño', 'fracturé', 'pierna', 'ju
gando', 'fútbol']
['usted', 'algún', 'traumatismo', 'esguince', 'fractura', 'vida', 'pasada', 'anteriormente', 'fracturado', 'costilla', 'izquier
da', 'brazo', 'derecho', 'alguna', 'factura', 'niño', 'fracturé', 'pierna', 'jugando', 'fútbol']
[2, 2, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1]
```

Ilustración 2 – Resultados de la tokenización de grabaciones

En la siguiente evidencia, se muestra el programa usado para ejecutar instrucciones sobre una base de datos alternativa con el fin de verificar que se estén operando de manera correcta por medio de comandos de voz. En la de la *Ilustración 4* se guarda cada una de las instrucciones, así como las palabras clave o información con las que se desea interactuar, mientras que en otra (*Ilustración 3*) se efectúan de manera literal, modificando o eliminando filas en las que se cumplan las condicionales.

Opciones extra				
←T→				
			id	listado
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	2 enfermedad
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	3 diabetes
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	4 glucosa

Ilustración 3 – Tabla de inserción de instrucciones y palabras clave por voz


































←T→				id	instruccion	resultado
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	1	INSERT	enfermedad
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	2	INSERT	necrosis
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	3	DELETE	fatiga
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	4	UPDATE	enfermedad
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	5	INSERT	diabetes
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	6	INSERT	diabetes
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	7	DELETE	diabetes
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	8	DELETE	diabetes
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	9	INSERT	enfermedad
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	10	INSERT	necrosis
<input type="checkbox"/>	 Editar	 Copiar	 Borrar	11	UPDATE	diabetes
<input type="checkbox"/>	Editar	Copiar	Borrar	12	UPDATE	necrosis

Ilustración 4 – Tabla de prueba para ejecución de instrucciones de base de datos por voz

Para demostrar el funcionamiento del código que ejecuta instrucciones por medio de comandos de voz se agregará una imagen de evidencia con las interpretaciones obtenidas por el programa y que tipo de instrucción se efectuó sobre la base de datos, tal y como se aprecia en la *Ilustración 5*.


```

Escuchando...
result2:
{ 'alternative': [ { 'confidence': 0.94797295,
                    'transcript': 'necesito agregar glucosa'},
                  {'transcript': 'necesito agregar glucoza'},
                  {'transcript': 'necesito agregar glukosa'}],
  'final': True}
Comando reconocido: necesito agregar glucosa
Se ejecutará INSERT
Se ejecutará INSERT
Escuchando...
result2:
[]
No se pudo entender el comando.
Escuchando...
result2:
{ 'alternative': [ { 'confidence': 0.94797289,
                    'transcript': 'ahora me gustaría'},
                  {'transcript': 'aora me gustaría'},
                  {'transcript': 'bien Ahora me gustaría'},
                  {'transcript': 'eh Ahora me gustaría'},
                  {'transcript': 'aora me gustaría'}],
  'final': True}
Comando reconocido: ahora me gustaría
Escuchando...
result2:
{ 'alternative': [ { 'confidence': 0.94797289,
                    'transcript': 'ahora me gustaría agregar necrosis'},
                  {'transcript': 'ahora me gustaría agregar nekrozis'},
                  {'transcript': 'ahora me gustaría agregar nekrosis'},
                  {'transcript': 'aora me gustaría agregar necrosis'},
                  {'transcript': 'aora me gustaría agregar nekrozis'}],
  'final': True}
Comando reconocido: ahora me gustaría agregar necrosis
Se ejecutará INSERT

```

Ilustración 5 – Resultados de la terminal del código para ejecución de instrucciones en base de datos por medio de comandos de voz

Conclusiones y planeación futura

Se reconoció que para dar el aporte científico al tema de investigación era necesario investigar sobre la generación del lexicón, debido a que para poder comprender cómo un programa de computadora pueda diferenciar entre sinónimos, antónimos, singulares, plurales, etc., es necesario tener una base o inventario de palabras con las cuales contar, dado que, para encontrar similitudes por vectorización es necesario primero tener ese orden y relevancia de dichas palabras.

Al ser un sistema multimodal es necesario que las instrucciones fueran ejecutadas por medio de comando de voz, en donde se encontraran las palabras clave necesarias para distinguir entre qué tipo de instrucción era requerido, es decir, si se necesitaba agregar nuevas palabras, cambiar alguna ya existente, o eliminar las que se seleccionaran.

Gracias a los avances realizados en dicho periodo, ahora se centrará el desarrollo de la investigación en la generación del código principal del proyecto; como es necesario para el PLN el uso de números en forma de vectores, a partir de la tokenización de los diálogos de las consultas médicas se pretende generar los valores y similitudes de cada palabra con el fin de que el programa funcione de mejor manera.

Para las actividades planeadas para el periodo de julio a diciembre del año 2024 se estipulan las marcadas en la siguiente tabla:

Tabla 3 - Actividades planeadas para el periodo de julio - diciembre 2024

Actividades	Periodo de junio de diciembre de 2024													
	Jun		Jul		Ago		Sep		Oct		Nov		Dic	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Redacción de tesis														
Generación de lexicón														
Creación del programa de PLN														
Modificaciones a DB y páginas Web														
Redacción de artículo científico														
Prueba de programa NLP en consultas médicas para recuperación de datos														
Modificaciones finales														
Entrega final de documentación														

Como se puede observar en la *Tabla 3*, las casillas rellenas de color azul muestran las semanas en las que se planea hacer enfoque a cada una de las actividades del cronograma, con la finalidad de realizar entrega de los avances de dicho proyecto para finales del mes de noviembre.

Es importante tomar en cuenta que debido a las modificaciones realizadas en la planeación anterior se omitirá del proceso la actividad de crear un “login” de usuario y contraseña para el acceso de la base de datos y su modificación, ya que de esa manera es posible evitar problemas por permisos de usuarios o redes en las que está conectado.

Debido a la necesidad de la entrega y presentación de un artículo científico relacionado al tema de investigación, se planea realizar los avances necesarios para la entrega en el mes indicado en dicha table.