PROPUESTA DE TEMA DE TESIS DE DOCTORADO

DOCTORADO EN CIENCIAS DE LA INGENIERÍA

TÍTULO DEL TEMA DE TESIS PROPUESTO

Interfaz de Lenguaje Natural para la Consulta de Información en un Lago de Datos Hospitalario

PROPONENTE

Jonathan Zavala Díaz

DIRECTOR DE LA TESIS

Dr. Juan Carlos Olivares Rojas

Palabras clave: Lenguaje natural, NPL, Interfaz, español, Hospitalario.

Índice

1. Resumen

El área de la salud es una de las áreas más beneficiadas con el uso de las TICs en los procesos de cuidado de la salud y diagnóstico de los pacientes. Al igual que muchas otras áreas, los datos de salud crece de forma rápida y en grandes volúmenes en poco tiempo. Esto presenta enormes ventajas, pero también muchos retos como el detalle de poder realizar consultas y predicciones en poco tiempo. Los lenguajes de consultas generalmente no son amigables ni dinámicos para los usuarios finales. En los últimos años el avance en el procesamiento del lenguaje natural y de la inteligencia artificial a permitiendo la masificación de aplicaciones como los chatbots y asistentes digitales que permiten una comunicación de forma natural en diversos idiomas como el inglés. Desafortunadamente, la información médica es de un contexto muy específico, así como los desarrollos existentes están muy enfocado en otros idiomas como el inglés, dejando a los especialistas de la salud hispanoparlantes muy rezagado. Por tal motivo proponemos desarrollar una interfaz de lenguaje natural para consulta de información en un lago de datos hospitalario, con lo cual no solo los usuarios expertos en tecnología puedan tener acceso a este tipo de información, sino que un usuario no experto en estas tecnologías se pueda comunicar por medio de lenguaje natural a la interfaz y hacer consultas de información médica. Dicho lo anterior proponemos realizar esta interfaz enfocada en el dominio biomédico, ya que a diferencia de una interfaz genérica pretendemos que sea capaz de detectar vocabulario y abreviaciones médicas. Además ya que la mayoría de información valiosa se esta desperdiciando por no tener acceso a ella debido a que se encuentra en datos no estructurados como lo son las narrativas médicas, dicho lo anterior, proponemos implementar un lago de datos que contenga toda esta información para posteriormente integrarla a nuestra interfaz, implementar técnicas de NLP para extraer información relevante en estas narrativas medicas en texto libre y posteriormente aplicar técnicas de machine learning para buscar patrones que puedan ayudar a predecir complicaciones en pacientes con algún tipo de enfermedad.

2. Introducción

Con los avances y crecimiento en las Tecnologías de la Información y la Comunicación (TICs), grandes volúmenes de datos se siguen acumulando velozmente en los bancos de datos de cada organización. Mas sin embargo la mayor parte de esta información nunca se puede utilizar para generar un beneficio real para la organización, para ello se debe ser capaz de convertir la información en conocimiento. El conocimiento se puede utilizar para aliviar el proceso de toma de decisiones por parte de los expertos en el área dominio [1].

En los últimos 20 años, la recopilación y el almacenamiento de datos hospitalarios se ha incrementado enormemente con el uso generalizado de los sistemas de información clínica, los cuales contienen grandes cantidades de datos sobre la salud y atención medica de los pacientes [2]. Una amplia gama de estos datos se encuentra comúnmente dentro de narrativas clínicas las cuales son del tipo de datos no estructurados. Debido a la naturaleza de estos tipos de datos, un Data Lake es el tipo de almacenamiento que mejor se adapta a ello, ya que esta diseñado para almacenar datos sin procesar (Estructurados, no estructurados, semiestructurados y binarios). Los informes narrativos permiten la flexibilidad de expresión como dudas, negaciones o hipótesis diagnósticas y la representación compleja de enfermedades, examen clínico, historial del paciente y antecedentes médicos familiares [3]. La mayoría de los registros médicos actuales conservan un gran elemento de texto libre. Si bien esto es atractivo para la mayoría de los usuarios finales debido a la flexibilidad de expresión, crea desafíos para el uso continuo de la información contenida en las notas [4]. El procesamiento del lenguaje natural (NLP, por sus siglas en inglés) es una forma de aprendizaje automático que se puede utilizar en este contexto para procesar y analizar elementos de texto libre, por lo cual puede ayudar en la predicción de los resultados de los pacientes, aumentar los sistemas de clasificación de hospitales y generar modelos de diagnóstico que detectan enfermedades crónicas en etapa temprana [4].

El NLP ahora se usa cada vez más en la medicina para mejorar la utilización de registros de salud electrónicos no estructurados y para proporcionar una forma de comunicación con los pacientes para responder preguntas y realizar consultas [5]. Usar el lenguaje natural es una forma más fácil para recuperar información de registros de salud. Las computadoras no pueden entender el lenguaje natural por lo que se necesita una interfaz; esa es la razón para desarrollar una interfaz de lenguaje natural de consulta de información. La interfaz de lenguaje natural es capaz de traducir la consulta de lenguaje natural dada por el usuario a una equivalente en lenguaje de consulta. Se han desarrollaron interfaces de lenguaje natural para bases de datos y así convertir el lenguaje natural a una consulta SQL y obtener el resultado correspondiente de la base de datos. Todavía hay mucho trabajo de investigación en el campo de la interfaz de lenguaje natural y se están desarrollando nuevas interfaces para las bases de datos que brindan respuestas más precisas.

Con el avance en el poder de procesamiento del hardware se han desarrollado herramientas basadas en lenguaje natural para consulta de información, más sin embargo dichas herramientas están diseñadas para un sistema de base de datos sin tomar en cuenta datos no estructurados, donde se necesitan otras técnicas para analizar la información contenida en texto libre [6]. Debido a que los desarrollos más recientes de interfaces de lenguaje natural se centran en entender la pregunta y trasladarla a un lenguaje de consulta como SQL con la finalidad de obtener la información de una base de datos, se necesita crear una herramienta que pueda interpretar la pregunta en lenguaje natural y realizar con la ayuda de NLP y Machine Learning el análisis en datos no estructurados como narrativas clínicas para encontrar información requerida por el usuario. En la literatura de los últimos años han propuesto herramientas basadas en lenguaje natural para consulta de información, más sin embargo dichas herramientas están desarrolladas para idiomas principalmente en inglés y algunos otros idiomas específicos como francés, afgano y cingalés [2], [7], [8]. Por lo cual nace la necesidad de crear una herramienta especial para el idioma español.

Lo expuesto anteriormente ha sido la principal motivación para llevar a cabo el trabajo de investigación que se propone en este protocolo. Cuyo objetivo principal es desarrollar soluciones basadas en tecnologías de procesamiento de lenguaje natural, tecnologías de la información y Machine Learning para convertir la información en conocimiento que pueda ayudar a encontrar patrones de alguna enfermedad en narrativas clínicas. Para cumplir con este objetivo, se ha seguido una metodología propuesta en este protocolo.

3. Antecedentes

Un gran número de investigadores han enfocado sus esfuerzos en desarrollar herramientas basadas en lenguaje natural para consulta de información como Interfaces en Lenguaje Natural (NLI, por sus siglas en inglés), dichas herramientas enfocadas a diversos dominios y aplicando diversas técnicas de procesamiento de lenguaje. A continuación, se presentan los trabajos más relevantes encontrados en esta área de NLP.

En el trabajo publicado por Rencis [9] se presenta un lenguaje natural controlado, más sin embargo solo es el primer paso para la creación de una interfaz basada en lenguaje natural para consulta de información. En los trabajos [10] y [11] interfaces de lenguaje natural para visualización de datos ambos en el idioma inglés.

Otra herramienta es Doc'EDS presentada en [2], una herramienta de búsqueda semántica francesa para consultar documentos de salud de un Data Warehouse clínico, si bien el sistema proporciona una interfaz fácil de usar, está diseñada únicamente para el idioma francés. En [7] , [8] y [12] presentan interfaces en leguaje natural en las cuales traducen las consultas de lenguaje natural a un lenguaje de consulta para eztraer información de bases de datos, recalcando de estos trabajos el enfoque a datos estructurados solamente.

La herramienta MedCAT presentada en **[13]** se trata de un el kit de herramientas de anotación de conceptos médicos de código abierto que proporciona un novedoso algoritmo de aprendizaje automático auto-supervisado para extraer conceptos utilizando cualquier vocabulario de conceptos y una interfaz .

En el trabajo publicado por Trivedi et al. [14] presentan NLPReViz: una herramienta interactiva para el procesamiento del lenguaje natural en texto clínico, utilizan la técnica Bag-of-words de NLP, también se encuentra desarrollado para el idioma inglés.

En el trabajo publicado por Rojas et al. [15] presentan Clinical Flair, un modelo de lenguaje de dominio específico entrenado en narrativas clínicas en español. Clinical Flair, un modelo de lenguaje a nivel de carácter para la PNL clínica en español.

Los autores en [16] utilizan técnicas de procesamiento de lenguaje natural como base de su algoritmo de desidentificación para eliminar datos personales de las historias clínicas y así proteger su identidad.

En la Tabla 1 se presentan los trabajos mencionados anteriormente destacando sus características como el idioma el ámbito, técnicas utilizadas y si desarrollan una interfaz de lenguaje natural.

Tabla 1.- Trabajos encontrados en la literatura sobre NLP y herramientas basadas en lenguaje natural.

Trabajo	**NLI**	**Idioma**	**Ámbito**	**Técnicas o herramientas utilizadas**
Rencis [9]	No	Letón	Biomédico	Lenguaje natural controlado
Álvarez et al. [16]	Sí	Español	Biomédico	Librería Flair
Setlur et al. [10]	Sí	Inglés	Demografía / Geografía	Gramática probabilística
Pressat-Laffouilhère et al. [2]	Sí	Francés	Biomédico	Algoritmos de procesamiento del lenguaje natural
Peduru-Hewa et al. [7]	Sí	Cingalés	Genérico	Método propuesto de conversión a SQL
S. Karimi, A. A. Rasel & M. S. Abdullah [8]	Sí	Afgano	Genérico	Técnica de mapeo de palabras
Yu & Silva [11]	Sí	Inglés	Genérico	Algoritmos de procesamiento del lenguaje natural
KraljevicYu et al. [13]	Sí	Inglés	Biomédico	Named Entity Recognition
Das & Balabantaray [12]	Sí	Inglés	Genérico	Técnica de mapeo de palabras
Trivedi et al. [14]	Sí	Inglés	Biomédico	Bag-of-words
Rojas et al. [15]	No	Español	Biomédico	Named Entity Recognition

4. Marco teórico

En lingüística, un lenguaje natural es cualquier lengua o idioma que ha sido generado en un grupo de hablantes con el propósito de comunicarse. Los lenguajes naturales pueden tomar diferentes formas tales como el habla, señas o la escritura. El NLP es un área de investigación de la Inteligencia Artificial (IA) que emplea un conjunto de tecnologías computacionales para analizar y generar de manera automática textos expresados en lenguaje natural. En la literatura existe un sinnúmero de definiciones de NLP. Por ejemplo, el autor en [17] define al NLP como: "un área de investigación que explora cómo las computadoras pueden utilizarse para entender y manipular texto escrito en lenguaje natural o del habla para hacer operaciones útiles". Con lo dicho anteriormente el lenguaje natural se refiere a la forma en que las personas se comunican y el procesamiento de lenguaje natural se encarga de analizar y procesar el lenguaje humano a través de herramientas y tecnologías de software. Esto involucra distintas áreas de la computación, tales como inteligencia artificial, lingüística computacional, etc. Mediante el procesamiento del lenguaje natural es posible procesar documentos de texto, mensajes SMS, email, páginas web, etc. y organizar el conocimiento para realizar tareas como análisis de sentimientos, análisis de contexto, generación de resúmenes, traducción automática, sistemas de diálogos, etc. [18]

El NLP generalmente se divide en un número de etapas enfocadas en los tres aspectos o dimensiones que constituyen la teoría lingüística o desde un punto más general, la teoría semiótica, las cuales son la sintaxis, la semántica y la pragmática [19]. En [19] los autores establecen un conjunto de cinco etapas de análisis en las cuales se descompone el NLP (ver Figura 1) siendo la entrada de este proceso un texto, y la salida el significado deseado del hablante.

Figura 1.- Etapas de análisis en el procesamiento de lenguaje natural.

El campo del NLP puede ser aplicado en aplicaciones tales como recuperación y extracción de información, traducción automática, minería de datos, generación de resúmenes, análisis de sentimientos y sistemas de búsqueda de respuestas, entre otras. A continuación, se explican brevemente estas aplicaciones de recuperación y extracción de información, así como sistemas de búsqueda de respuestas para propósitos del presente protocolo de tesis.

La recuperación de información es el proceso de encontrar material (usualmente documentos) de naturaleza no estructurada (usualmente texto) que satisfaga una necesidad de información, dentro de grandes colecciones (usualmente almacenada en computadoras) [20]. La extracción de información se define como una tecnología basada en el análisis del lenguaje natural para extraer fragmentos de información. El proceso toma como entrada textos y produce un formato fijo de datos inequívocos como salida [21]. Para ello, el proceso extrae fragmentos de texto con significado relevante ignorando los fragmentos irrelevantes que se emplean para estructurarlos. De esta manera, el ordenador es capaz de entender y almacenar la información extraída en un sistema de almacenamiento, como una base de datos, para su futura explotación [22]. La búsqueda de respuestas, llamado en inglés Question Answering (QA), puede ser definido como "un proceso capaz de entender preguntas formuladas en lenguaje natural como el inglés y responder exactamente con la información solicitada" [19].

Un concepto importante en el procesamiento del lenguaje natural son las **word embeddings** ya que se ha demostrado que si son previamente entrenadas son de gran utilidad para tareas posteriores de NLP, tanto por su capacidad para ayudar al aprendizaje y la generalización con información aprendida de datos no etiquetados, como por la relativa facilidad de incluirlos en cualquier proceso de aprendizaje [23].

En la última década, junto con el crecimiento del aprendizaje profundo, las representaciones (embeddings) basadas en redes neuronales han reemplazado casi por completo a los modelos convencionales basados en conteo y han dominado el campo. Dado que las **word embeddings** neuronales generalmente se entrenan con algún tipo de objetivo de modelado de lenguaje, como predecir una palabra faltante en un contexto, también se conocen como modelos predictivos [24]. Las incrustaciones de palabras fueron popularizadas por Word2vec, en [25] utilizan Word2Vec y BERT.

Existen herramientas avanzadas para el procesamiento del lenguaje natural como Apache OpenNLP [26], Stanford CoreNLP [27] y Stanza [28], nos proveen de un conjunto de herramientas para tareas comunes de NLP como la tokenización, la segmentación de oraciones, el etiquetado de partes del discurso, la extracción de entidades nombradas, la fragmentación, el análisis, la detección de idiomas y la resolución de correferencias.

Uno de los factores limitantes en la usabilidad de las computadoras corresponde a la usabilidad de las interfaces [29]. Un avance en la usabilidad de las interfaces son las interfaces gráficas de usuario (GUI, por sus siglas en inglés). A pesar de que las GUI han hecho que la interacción con la computadora sea más fácil para un gran número de personas, éstas requieren que el usuario tenga conocimiento de cada una de las opciones que le ofrece la interfaz, así como de la ubicación de cada una de estas funciones. Por el contrario, las interfaces de lenguaje natural (NLI) no requieren que el usuario cuente con un conocimiento especializado, ya que le permite usar todo el poder del lenguaje que ya posee en lugar de verse forzado a utilizar un modo de comunicación poco natural y limitante como lo son las GUI. Así, el objetivo de las NLI es superar la brecha existente entre el rendimiento lingüístico del usuario y la competencia lingüística del sistema computacional subyacente [30].

Smith en [29] propone una arquitectura genérica para una interfaz de lenguaje natural enfocada a algún tipo de aplicación funcional, como lo puede ser un gestor de bases de datos. Esta arquitectura se muestra en la Figura 2.

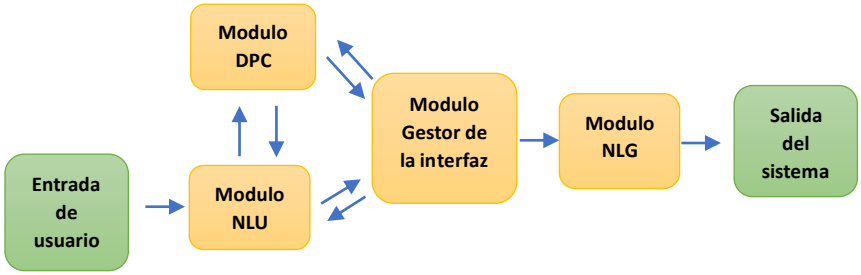

Figura 2.- Arquitectura genérica de una interfaz de lenguaje natural

Los elementos que componen la arquitectura mostrada en la figura anterior son:

- Módulo NLU (Natural Language Understanding): Este analiza la consulta provista por el usuario. El NLU es un campo de NLP que se ocupa de la comprensión de texto [31], es decir, se encarga de interpretar el fragmento de texto de entrada. Este proceso puede ser entendido como una traducción del texto expresado en lenguaje natural a una representación en un lenguaje formal no ambiguo.
- Módulo DPC (Domain Processing Components): Este componente obtiene la información específica del dominio, es decir, interactúa directamente con la fuente de información de la cual se espera extraer los datos que ayuden a dar respuesta a las solicitudes del usuario.
- Módulo gestor de la interfaz: Este módulo comprende todos aquellos componentes que mantienen información acerca de la interacción constante entre la NLI y el humano. En otras palabras, para que el usuario pueda aprobar, rechazar e incluso elegir de entre varias interpretaciones aquella que encaje con la petición realizada
- Módulo NLG (Natural Language Generation). En el contexto de IA y lingüística computacional, la NLG se encarga de generar texto o voz en lenguaje natural a partir de representaciones de datos estructurados y procesables por la máquina tales como las bases de conocimiento [32]. En este sentido, este módulo será el encargado de generar una respuesta ya sea solo en lenguaje natural o en combinación con otras modalidades tales como gráficos.

El exponencial crecimiento de los datos hospitalarios ha dado paso a la necesidad de contar con mecanismos capaces de procesar y comprender dicha información y con ello resolver necesidades específicas. Para ello, esta información debe convertirse en conocimiento la cual permita a los profesionales de la salud usar dicha información en la aplicación de su trabajo. Ante esta situación surge el NLP el cual nos provee de una serie de herramientas que se pueden implementar para dicha necesidad. Por lo cual este trabajo se propone una solución basada en procesamiento lenguaje natural y recuperación de información de bases de conocimiento. Se propone aplicar técnicas de NLP, cabe mencionar que la interfaz de lenguaje natural propuesta en este trabajo se desarrolla para el idioma español y para el dominio biomédico.

5. Objetivos

5.1 Objetivo general

Desarrollar un interfaz de Lenguaje Natural para la Consulta de Información en un Lago de Datos Hospitalario

5.2 Objetivos particulares.

1. Investigar diversas técnicas y herramientas de lenguaje natural en español para el área de la salud.
2. Estudiar de Técnicas de conversión a lenguajes de consulta de datos.
3. Implementar un lago de datos hospitalarios provenientes de diversas fuentes y en diversos formatos.
4. Desarrollar interfaces en lenguaje natural (español) que permita la obtención de datos de sistemas hospitalarios en forma sencilla y funcional.
5. Realizar modelos de aprendizaje automático y de análisis de datos que permitan mostrar información de pronósticos.
6. Desarrollar una interfaz de usuario para visualizar la información proveniente de la interfaz de lenguaje natural

6. Metas.

1. 1 tabla con al menos 10 técnicas y herramientas de lenguaje natural en español para el área de la salud, resaltando sus características, evaluarlas, para elegir 3 de ellas con las cuales trabajar.
2. 1 tabla con diferentes Técnicas de conversión a lenguajes de consulta de datos y elegir la que tenga mayores prestaciones y se adapte a nuestros requerimientos.
3. 1 implementación lago de datos hospitalarios provenientes de diversas fuentes y en diversos formatos.
4. 1 interfaz en lenguaje natural en el idioma que permita la obtención de datos de sistemas hospitalarios en forma sencilla y funcional.
5. 1 implementación de modelos de aprendizaje automático y de análisis de datos que permitan mostrar información de pronósticos.
6. 1 interfaz de usuario para visualizar la información proveniente de la interfaz de lenguaje natural

7. Impacto.

Científico: Si hablamos de un impacto científico, este proyecto propuesto pretende aportar a la comunidad científica una nueva arquitectura y aportes en la aplicación del procesamiento de del lenguaje natural para un enfoque en el ámbito biomédico.

Social: La implementación de la interfaz propuesta pretende un impacto en la vida de los pacientes al ser una herramienta la cual sirva a los expertos ha un mejor diagnóstico y seguimiento oportuno de enfermedades en sus pacientes y con eso obtener una mejor de vida.

Económico: El tener datos almacenados y no convertirlos en conocimiento no sirve de nada, más sin embargo nosotros proponemos convertir en conocimiento todos los datos generados por los pacientes para así dar un mejor seguimiento a historiales clínicos con el objetivo de reducir costos por complicaciones en enfermedades.

Tecnológico: La interfaz de lenguaje natural propuesta para el mambito medico conlleva un cambio en la forma en la que hasta ahora se utiliza la información médica y como se consulta, por lo que una herramienta como la propuesta impactara en el uso de nuevas tecnologías en sistemas hospitalarios donde hoy en día no se están aprovechando el uso de nuevas tecnologías.

8. Metodología.

La metodología propuesta a seguir durante el desarrollo de este proyecto de tesis se divide en cuatro partes principales: en la primera se desarrolla un estudio del estado del arte que permita conocer los esfuerzos de investigación más relevantes dentro de las áreas de interés del proyecto; la segunda parte consiste en la formalización de los métodos propuestos en este trabajo para interfaces de lenguaje natural; en la tercera etapa, se lleva a cabo la implementación de los métodos propuestos; finalmente, en la cuarta parte se lleva a cabo la validación de la propuesta analizando narrativas clínicas de nuestro lago de datos.

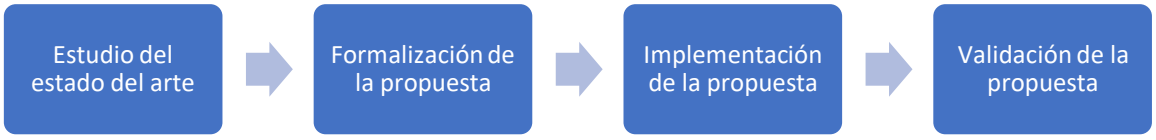

Figura 3.- Metodología propuesta

- Estudio del estado del arte.
  En esta parte de la metodología se llevó a cabo un análisis de todos aquellos desarrollos de última tecnología realizados en los contextos de PLN, interfaces de lenguaje natural y modelos de machine learning, principales tecnologías involucradas en este trabajo de tesis.
  - Procesamiento de lenguaje natural: Análisis de los diferentes niveles de PLN, así como las diversas aplicaciones de esta tecnología.
  - Interfaces de lenguaje natural: Análisis de las principales arquitecturas utilizadas en el desarrollo de este tipo de aplicaciones y de los esfuerzos de investigación más sobresalientes enfocados en proveer soluciones de este tipo.
  - Modelos de Machine Learning: Análisis de modelos y herramientas que permitan mostrar información de pronósticos.

- Formalización de la propuesta.
  Esta parte de la metodología contempla el desarrollo de un lago de datos hospitalarios provenientes de diversas fuentes y en diversos formatos, la base de conocimiento del dominio. Toda la información para almacenar será obtenida mediante el proceso de análisis de preguntas en lenguaje natural diseñado en este trabajo, el cual estará basado en técnicas tales como el análisis de dependencias, lematización y la búsqueda de sinónimos.

- Implementación de la propuesta: Esta etapa consiste en la implementación de la interfaz de lenguaje natural propuesta por medio de herramientas de PLN e integración a nuestro lago de datos con narrativas clínicas.

- Validación de la propuesta.
  Finalmente, esta parte de la metodología contempla la validación de la interfaz de lenguaje natural implementada en un lago de datos hospitalario. En concreto, la interfaz desarrollada se aplicará sobre el conjunto de datos de un data lake hospitalario que integren narrativas medicas

Arquitectura propuesta

La interfaz que se propone recibe una consulta formulada en lenguaje natural por los usuarios en idioma español, y mediante esta realizar una consulta a la base de conocimiento. Su arquitectura está dividida en tres módulos

principales: a) Base de conocimiento, b) El módulo NLP, y c) Procesamiento de la base de conocimiento. A continuación, se describen los módulos. En la Fig. 4 se presenta la arquitectura general de la interfaz propuesta.

La base de conocimiento donde se tomarán los datos en su mayoría se pretende que sean narrativas medicas en texto libre de donde la interfaz realizara las consultas y preprocesador analizara dicha información. Modulo NLP como se muestra a detalle en la Figura 5 son una serie de técnicas de NLP para poder entre la consulta hecha en lenguaje natural y así la interfaz pueda hacer la consulta a la base de conocimiento. Y por ultimo el preprocesador de la base de conocimiento se encargara en utilizar técnicas de NLP y machine learning para procesar la información de la base de conocimiento y poder encontrar patrones en narrativas clínicas.

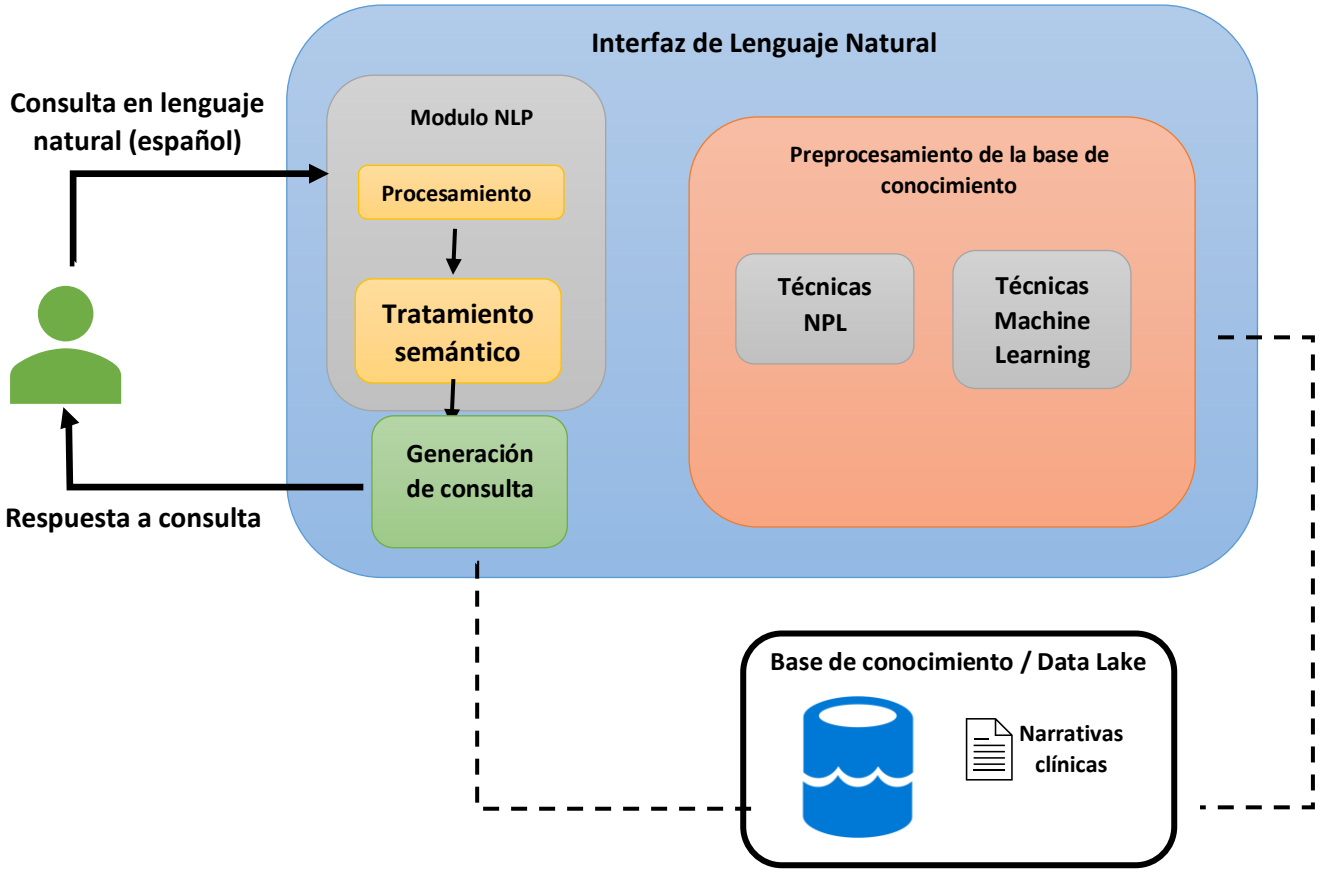

Figura 4.- Arquitectura básica propuesta para la interfaz de lenguaje natural

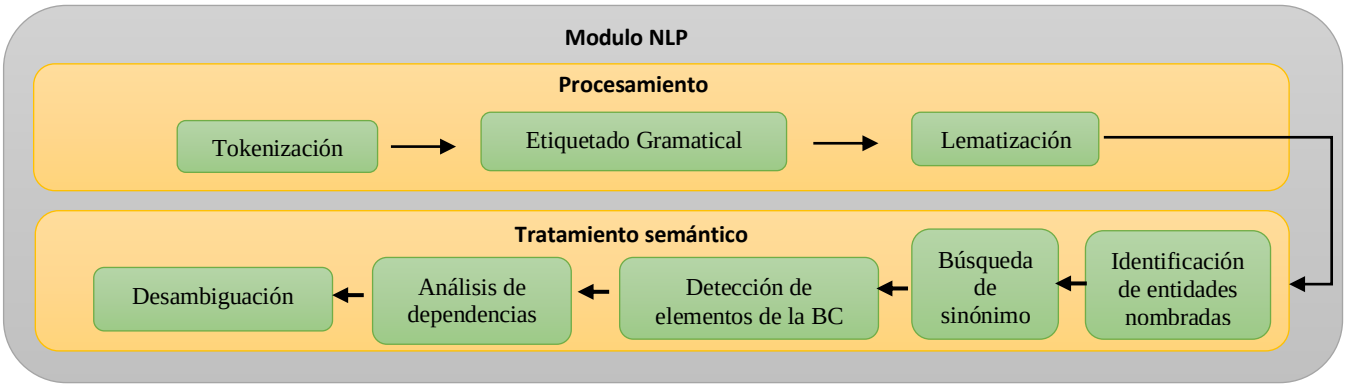

Figura 5.- Modulo NLP

9. Programa de actividades, calendarización.

10. Productos entregables.

- Software - Interfaces en lenguaje natural (español) que permita la obtención de datos de sistemas hospitalarios en forma sencilla y funcional.
- Método para modelos de aprendizaje automático y de análisis de datos que permitan mostrar información de pronósticos.
- Software - Interfaz de Lenguaje Natural para la Consulta de Información en un Lago de Datos Hospitalario
- Presentación en congreso internacional
- 1 publicación en el índice JCR
- Documento de tesis
- Certificado Toefl con 550 puntos

11. Vinculación con otras instituciones, empresas o sectores.

Pendiente…

12. Referencias.

[1] E. Rencis, "Application of a Configurable Keywords-Based Query Language to the Healthcare Domain," *J. Adv. Inf. Technol.*, vol. 12, no. 2, pp. 142–147, 2021, doi: 10.12720/jait.12.2.142-147.

[2] T. Pressat-Laffouilhère *et al.*, "Evaluation of Doc'EDS: a French semantic search tool to query health documents from a clinical data warehouse," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 34, Dec. 2022, doi: 10.1186/s12911-022-01762-4.

[3] N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, and A. Burgun, "Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse," *J. Am. Med. Informatics Assoc.*, vol. 24, no. 3, pp. 607–613, May 2017, doi: 10.1093/jamia/ocw144.

[4] S. Locke, A. Bashall, S. Al-Adely, J. Moore, A. Wilson, and G. B. Kitchen, "Natural language processing in medicine: A review," *Trends Anaesth. Crit. Care*, vol. 38, pp. 4–9, Jun. 2021, doi: 10.1016/j.tacc.2021.02.007.

[5] J. Wang *et al.*, "Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed," *J. Med. Internet Res.*, vol. 22, no. 1, p. e16816, Jan. 2020, doi: 10.2196/16816.

[6] E. U. Reshma and P. C. Remya, "A review of different approaches in natural language interfaces to databases," in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2017, pp. 801–804. doi: 10.1109/ISS1.2017.8389287.

[7] D. S. Peduru Hewa and C. Farook, "A Sinhala Natural Language Interface for Querying Databases Using Natural Language Processing," in *2021 21st International Conference on Advances in ICT for Emerging Regions (ICter)*, Dec. 2021, pp. 213–218. doi: 10.1109/ICter53630.2021.9774794.

[8] S. Karimi, A. A. Rasel, and M. S. Abdullah, "Natural Language Query and Control Interface for Database Using Afghan Language," in *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, Aug. 2022, pp. 1–8. doi: 10.1109/INISTA55318.2022.9894168.

[9] E. Rencis, "Towards a natural language-based interface for querying hospital data," in *Proceedings of 2018 International Conference on Big Data Technologies - ICBDT '18*, 2018, pp. 25–28. doi: 10.1145/3226116.3226133.

[10] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang, "Eviza: A natural language interface for visual analysis," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, Oct. 2016, pp. 365–377. doi: 10.1145/2984511.2984588.

[11] B. Yu and C. T. Silva, "FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 1–11, Jan. 2020, doi: 10.1109/TVCG.2019.2934668.

[12] A. Das and R. C. Balabantaray, "MyNLIDB: A Natural Language Interface to Database," in *2019 International Conference on Information Technology (ICIT)*, Dec. 2019, pp. 234–238. doi: 10.1109/ICIT48102.2019.00048.

[13] Z. Kraljevic *et al.*, "Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit," *Artif. Intell. Med.*, vol. 117, p. 102083, Jul. 2021, doi: 10.1016/j.artmed.2021.102083.

[14] G. Trivedi, P. Pham, W. W. Chapman, R. Hwa, J. Wiebe, and H. Hochheiser, "NLPReViz: an interactive tool for natural language processing on clinical text," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 1, pp. 81–87, Jan. 2018, doi: 10.1093/jamia/ocx070.

[15] M. Rojas, J. Dunstan, and F. Villena, "Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing," in *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 2022, pp. 87–92. doi: 10.18653/v1/2022.clinicalnlp-1.9.

[16] C. Álvarez *et al.*, "Estudio longitudinal para el desarrollo de modelos predictivos de complicaciones crónicas de la diabetes mellitus tipo 2," *Komput. Sapiens*, vol. 3, pp. 10–15, 2022, [Online]. Available: http://komputersapiens.smia.mx/publicaciones.php#KSXIV-III

[17] G. G. Chowdhury, "Natural language processing," *Annu. Rev. Inf. Sci. Technol.*, vol. 37, no. 1, pp. 51–89, Jan. 2005, doi: 10.1002/aris.1440370103.

[18] F. Pech-May, L. A. López-Gómez, and J. Magaña-Govea, "Procesamiento de lenguaje natural con aprendizaje profundo," *Komput. Sapiens*, vol. 2, pp. 56–61, 2019, [Online]. Available: http://komputersapiens.smia.mx/publicaciones.php#KSXI-II

[19] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*, 2nd ed. Chapman & Hall/CRC, 2010.

[20] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[21] H. Cunningham, "Information Extraction, Automatic," in *Encyclopedia of Language & Linguistics*, Elsevier, 2006, pp. 665–677. doi: 10.1016/B0-08-044854-2/00960-3.

[22] J. Cowie and W. Lehnert, "Information extraction," *Commun. ACM*, vol. 39, no. 1, pp. 80–91, Jan. 1996, doi: 10.1145/234173.234209.

[23] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 54–59. doi: 10.18653/v1/N19-4010.

[24] M. T. Pilehvar and J. Camacho-Collados, "Word Embeddings," 2021, pp. 25–40. doi: 10.1007/978-3-031-02177-0_3.

[25] J.-C. Hernández-Hernández, D. Juárez-Morales, J.-J. Guzmán-Landa, and G. J. Hoyos-Rivera, "Análisis de Sentimientos en Twitter," *Komput. Sapiens*, vol. 2, pp. 59–63, 2022, [Online]. Available: http://komputersapiens.smia.mx/publicaciones.php#KSXIV-II

[26] "Apache OpenNLP." https://opennlp.apache.org/

[27] "CoreNLP." https://stanfordnlp.github.io/CoreNLP/

[28] "Stanza." https://stanfordnlp.github.io/stanza/

[29] R. W. Smith, "Natural Language Interfaces," in *Encyclopedia of Language & Linguistics*, Elsevier, 2006, pp. 496–503. doi: 10.1016/B0-08-044854-2/00975-5.

[30] B. Manaris, "Natural Language Processing: A Human-Computer Interaction Perspective," 1998, pp. 1–66. doi: 10.1016/S0065-2458(08)60665-8.

[31] E. Ovchinnikova, *Integration of World Knowledge for Natural Language Understanding*, vol. 3. Paris: Atlantis Press, 2012. doi: 10.2991/978-94-91216-53-4.

[32] M. E. Vicente, C. Barros, F. Agulló, F. S. Peregrino, and E. Lloret, "La generacion de lenguaje natural: análisis del estado actual," *Comput. y Sist.*, vol. 19, no. 4, Dec. 2015, doi: 10.13053/cys-19-4-2196.