Sistema informático para el reconocimiento y normalización de entidades biomédicas basado en reglas heurísticas y búsquedas semánticas en lenguaje español

Llano Chinchero, Christopher Fabricio y Rugel Tiaguaro, Cristopher Alexis

Departamento de Ciencias de la Computación

Carrera de Ingeniería en Software

Trabajo de titulación, previo a la obtención del título de Ingeniero en Software

Uyaguari Uyaguari, Alvaro Danilo Msc.

22 de agosto del 2023

Latacunga

Copyleaks

Plagiarism and AI Content Detection Report

Sistema informático para el reconoci...

Scan details

Scan time:	Total Pages:	Total Words:
August 21th, 2023 at 17:52 UTC	30	7312

Plagiarism Detection

7.9%

Types of plagiarism		Words
Identical	0.7%	49
Minor Changes	0.1%	5
Paraphrased	5.8%	422
Omitted Words	17.9%	1308

AI Content Detection

Text coverage		Words
AI text	33%	16446
Human text	67%	24297

Learn more

Plagiarism Results: (19)

download 2.6%

https://repository.ucc.edu.co/server/api/core/bitstreams/7b0...

daniel vergel SB

1 FORMULACION DE UNA GUIA DE LA METODOLOGÍA ÁGIL SCRUM PARA LAS EMPRESAS MEDIANAS Y PEQUEÑAS DE COLOMBIA Daniel Enrique Vergel...

SCRUMstudy-SBOK-Guide-3rd-edition-Spanish.pdf 2.4%

https://primeconsultores.com.pe/wp-content/uploads/2021/...

Gaurav Garg

Una guía para el CUERPO DE CONOCIMIENTO DE SCRUM (Guía SBOKTM) 3ra Edición Una guía integral para la entrega de proyectos utilizando Sc...

Scrum Roles | Ole Business Agility incorporated 2.3%

https://obainc.es/scrum-roles/

Inicio OBAINC Certificaciones Servicios Blog Contac...

Certified by Copyleaks

About this report
help.copyleaks.com

copyleaks.com

Firma:

Uyaguari Uyaguari, Alvaro Danilo Msc.

C. C: 0103411112

Departamento de Ciencias de la Computación

Carrera de Ingeniería en Software

Certificación

Certifico que el trabajo de titulación: "**Sistema informático para el reconocimiento y normalización de entidades biomédicas basado en reglas heurísticas y búsquedas semánticas en lenguaje español**" fue realizado por los señores **Rugel Tiaguaro, Cristopher Alexis y Llano Chinchero, Christopher Fabricio;** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Latacunga, 22 de agosto del 2023

Firma:

Uyaguari Uyaguari, Alvaro Danilo Msc.

C. C: 0103411112

Departamento de Ciencias de la Computación

Carrera de Ingeniería en Software

Responsabilidad de Autoría

Nosotros, **Rugel Tiaguaro, Cristopher Alexis y Llano Chinchero, Christopher Fabricio**, con cédulas de ciudadanía n° 1725278053 y n° 0503424046, declaramos que el contenido, ideas y criterios del trabajo de titulación**: "Sistema informático para el reconocimiento y normalización de entidades biomédicas basado en reglas heurísticas y búsquedas semánticas en lenguaje español."** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Latacunga, 22 de agosto del 2023

Firma

.....

Rugel Tiaguaro, Cristopher Alexis

C.C.: 1725278053

Firma

.....

Llano Chinchero, Christopher Fabricio

C.C.: 0503424046

Departamento de Ciencias de la Computación

Carrera de Ingeniería en Software

Autorización de Publicación

Nosotros **Rugel Tiaguaro, Cristopher Alexis y Llano Chinchero, Christopher Fabricio**, con cédulas de ciudadanía n° 1725278053 y n° 0503424046, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Sistema informático para el reconocimiento y normalización de entidades biomédicas basado en reglas heurísticas y búsquedas semánticas en lenguaje español."** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi/nuestra responsabilidad.

Latacunga, 22 de agosto del 2023

Firma

.....

Rugel Tiaguaro, Cristopher Alexis

C.C.: 1725278053

Firma

.....

Llano Chinchero, Christopher Fabricio

C.C.: 0503424046

Dedicatoria

Me dedico esta tesis como testimonio de mi incansable dedicación, perseverancia y pasión por alcanzar las metas que me propuse. Estoy agradecido por la determinación que me impulsa a nunca rendirme y siempre esforzarme por ser la mejor versión de mí mismo.

A mi madre Monica, mi heroína, quien con su fortaleza, sabiduría y amor incondicional me ha guiado en cada paso de mi vida. Por festejar conmigo las medallas y trofeos que conseguía en algún deporte desde niño, quiero que seas tu nuevamente la que este a mi lado cuando reciba este nuevo logro, mis triunfos siempre será de los dos. Gracias por ser mi luz en momentos oscuros y por creer en mí siempre. Eres mi roca y mi motivación para seguir. Estoy orgulloso de llamarte madre.

A mi padre Arturo por darme la fuerza y los ánimos para seguir adelante. Gracias por haberme enseñado a ser la persona que soy hoy, mis principios, mis valores, mi perseverancia y mi empeño todo te lo debo a ti.

A mis hermanos Leonel y Patricia, por los momentos buenos y malos que estuvieron a mi lado. Por ser mi constante inspiración para convertirme en un mejor hermano. Espero este logro les sirva como ejemplo de que en esta vida todo se puede alcanzar.

A mis primos Alexander, Javier, Cristian, Alex y Dennis, por su tiempo, apoyo y orientación fueron fundamentales. Su ayuda me proporciono la confianza necesaria para superar los desafíos y crecer académicamente.

A mi abuelita Maria por tu constante apoyo y palabras alentadoras. Gracias por siempre creer en mí y decirme que “serás la primera en estar presente el día de mi graduación”.

Finalmente, a mi familia, gracias por estar siempre a mi lado, por apoyarme en todo momento. A pesar de no poder mencionar a cada uno individualmente, cada miembro ha dejado una huella en mi corazón y ha contribuido a mi éxito de manera única. A todos ustedes, les agradezco de todo corazón por haber formado parte de este camino que hoy culmina con la

presentación de mi tesina. Sin su amor, apoyo y motivación, este logro no hubiera sido posible. Les agradezco por ser mi familia y amigos, los amo profundamente.

Llano Chinchero, Christopher Fabricio

Dedicatoria

Primeramente, a Dios que gracias a su sabiduría y su amor incondicional he podido llegar a esta importante etapa en vida que es la presentación de mi tesina.

A mis padres Luis y Paola que han estado conmigo desde el principio de todo mi periodo académico, por su amor incondicional, sus consejos y ha como salir adelante. Gracias por enseñarme que soy capaz de hacer cualquier cosa, sin duda son un ejemplo para mí.

A mi tío Kevin que desde tengo uso de razón, me ha cuidado y ha estado en todos los momentos más importantes en mi vida. A mis tías gracias por sus consejos y amor que me brindan. A mis abuelitos José y María y también a mi abuelita Pilar, que con sus consejos y sabiduría me han hecho ser una mejor persona. A mis primos gracias por sacarme una sonrisa y estar para mi siempre. A mi chiquilín por darme el amor y el cariño que una mascota puede dar.

Finalmente, a mis amigos Dayana, Darwin, William y Bryan por apoyarme en todo momento y escucharme en los peores momentos.

A todos ustedes, les agradezco de lo más profundo de mi corazón, por ser parte de este momento que hoy finaliza con mi presentación de tesina, sin su amor y apoyo que me brindaron no hubiera sido esto posible.

Les agradezco familia y amigos infinitamente, que Dios les cuide siempre.

Rugel Tiaguaro Cristopher Alexis

Agradecimiento

Deseo expresar mi más sincero agradecimiento a mis compañeros de universidad, por su amistad y compañía durante el transcurso de mi vida académica. Gracias por hacer que mi experiencia universitaria fuera más enriquecedora y placentera. Su contribución a un ambiente de crecimiento constante y competencia sana ha sido invaluable. Cada uno de ustedes ha sido un motor para superarme y aprender de manera continua. ¡Gracias por convertir mi camino académico en algo memorable y gratificante!

Asimismo, agradezco a mis docentes, cuya guía y enseñanzas fueron fundamentales durante mi trayectoria académica. Su dedicación en proporcionarme los conocimientos y las herramientas necesarias ha sido un pilar esencial en este camino.

De manera especial, deseo expresar mi reconocimiento a mi tutor, Msc Alvaro Uyaguari, por su constante orientación y valiosos consejos a lo largo de este proceso. Le agradezco sinceramente por dedicar su tiempo, paciencia y esfuerzo en contribuir a alcanzar el éxito de este trabajo.

Gracias a todos por ser parte de mi vida y hacer posible este logro.

Llano Chinchero, Christopher Fabricio

Agradecimiento

Quiero agradecer a mis compañeros de universidad por su apoyo y comentarios durante el proceso de desarrollo de esta tesina. Sus contribuciones han sido invaluables para el éxito de este trabajo.

También quiero agradecer a mis docentes por brindarme los conocimientos y herramientas necesarios para llevar a cabo este proyecto. Su dedicación y compromiso con la enseñanza son dignos de admirar y agradezco por su tiempo y esfuerzo.

Quiero agradecer especialmente a mi tutor, Msc Álvaro Uyaguari, por su orientación y consejos a lo largo de todo el proceso. Su tiempo, paciencia y compromiso en ayudarme a alcanzar mis objetivos han sido invaluables. Sus conocimientos y experiencia han sido una gran influencia en mi formación académica y personal.

Estoy muy agradecido por la oportunidad de haber trabajado con él y espero continuar aprendiendo de él en el futuro.

Rugel Tiaguaro Cristopher Alexis

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Resumen

Actualmente se dispone de pocos algoritmos, especialmente en el ámbito de la medicina, para el reconocimiento y normalización de entidades biológicas basados en criterios heurísticos y búsquedas semánticas. Por ello, se ha creado un sistema que utiliza búsquedas semánticas en lengua española para identificar y normalizar entidades biomédicas. Se utiliza métodos de Procesamiento del Lenguaje Natural (PLN). El sistema informático desarrollado utiliza criterios específicos para reconocer y categorizar entidades biomédicas de importancia, tales como proteínas, químicos y sustancias farmacológicas. Además, incorpora métodos de búsqueda semántica con el objetivo de mejorar la exactitud de los resultados obtenidos. Esta solución computacional brinda una herramienta altamente eficaz para el análisis de textos biomédicos en español, simplificando de esta manera la extracción y organización de información en el ámbito médico. En conclusión, el sistema desarrollado en esta tesina es una herramienta valiosa para la investigación y la práctica médica. El sistema utiliza un enfoque basado en el procesamiento del lenguaje natural (PLN) para identificar entidades biomédicas en texto, y luego utiliza búsquedas semánticas para ampliar su alcance y precisión.

Palabras clave: Sistema Informático, Procesamiento del lenguaje natural (PLN), Búsqueda Semántica.

Abstract

Few algorithms are currently available, especially in the field of medicine, for the recognition and standardization of biological entities based on heuristic criteria and semantic searches. Therefore, a system has been created that uses semantic searches in Spanish language to identify and normalize biomedical entities. Natural Language Processing (NLP) methods are used. The developed computer system uses specific criteria to recognize and categorize biomedical entities of importance, such as proteins, chemicals and pharmacological substances. In addition, it incorporates semantic search methods to improve the accuracy of the results obtained. This computational solution provides a highly efficient tool for the analysis of biomedical texts in Spanish, thus simplifying the extraction and organization of information in the medical field. In conclusion, the system developed in this dissertation is a valuable tool for medical research and practice. The system uses a natural language processing (NLP) approach to identify biomedical entities in text, and then uses semantic searches to extend its scope and accuracy.

Key words: Computer system, Natural Language Processing (NLP), Semantic Search.

Capítulo I

Introducción

La extracción y organización de información contenida en textos biomédicos en español es un desafío debido a la complejidad del lenguaje técnico y la diversidad de términos utilizados. Actualmente, existen grandes cantidades de información no estructurada en forma de artículos científicos, informes clínicos y bases de datos médicos en español, los cuales contienen valiosos conocimientos y datos relevantes para la investigación y la práctica médica.

Sin embargo, el reconocimiento y la normalización de entidades biomédicas, como enfermedades, tratamientos, genes y proteínas, en estos textos presentan dificultades. Esto se debe a la falta de un sistema informático especializado que utilice reglas heurísticas y búsquedas semánticas en lenguaje español. La falta de este sistema limita la capacidad de los profesionales de la salud y los investigadores para acceder y utilizar de manera eficiente la información biomédica disponible en su idioma nativo.

Por lo tanto, surge la necesidad de desarrollar un sistema informático que pueda reconocer y normalizar de manera precisa las entidades biomédicas en textos en español, aprovechando reglas heurísticas y técnicas de búsqueda semántica. Este sistema permitiría una extracción y organización automatizada de información, facilitando la investigación médica, la toma de decisiones clínicas y la generación de conocimiento en el ámbito biomédico.

En resumen, la extracción y organización de información contenida en textos biomédicos en español es un desafío importante que debe ser abordado. El desarrollo de este sistema sería una contribución valiosa a la comunidad científica y médica, este sistema podría ayudar a mejorar la calidad de la investigación biomédica y la atención médica.

Esto se debe a que los textos biomédicos suelen ser complejos y contienen una gran cantidad de información, lo que puede dificultar la identificación de las entidades relevantes. Un sistema informático especializado podría ayudar a identificar y normalizar las entidades biomédicas, lo que facilitaría la búsqueda y el análisis de información biomédica.

Basándonos en la problemática se formula la siguiente pregunta:

¿Cuáles son las ventajas de utilizar reglas heurísticas y búsquedas semánticas para el reconocimiento y normalización de entidades biomédicas?

Antecedentes

Con el uso de la inteligencia artificial (IA), la búsqueda semántica supone un gran avance en la búsqueda de nueva generación al ofrecer a los usuarios resultados más precisos y relevantes (Moreno, 2013).

Una entidad con nombre es una palabra o frase que distingue un elemento de un grupo de componentes relacionados proporcionando una indicación clara de su identidad (Sharnagat, 2014).

En el contexto de la medicina, las entidades pueden incluir nombres de genes, proteínas, medicamentos y trastornos. El proceso de reconocimiento de entidades nombradas (NER) consiste en encontrar y clasificar entidades con nombre en el texto en categorías predeterminadas (Aguilera Murrell, 2022).

El estudio de la investigación sobre búsqueda semántica revela ciertas técnicas similares. Otras están más estrechamente relacionadas con el ámbito de la búsqueda, mientras que algunas son fundamentales para el formalismo RDF (Resource Description Framework) que se utilizar para representar una amplia gama de información, incluyendo información sobre personas, lugares, eventos, productos y servicios (Mäkelä, 2005).

La planificación de futuras aproximaciones a los problemas de la búsqueda semántica se verá favorecida por el conocimiento y la comprensión de estas metodologías comunes, así como por la forma en que se aplican en los numerosos enfoques reales (Mäkelä, 2005).

Justificación e importancia

Los modelos de cálculo de palabras sirven de base a los métodos de búsqueda tradicionales, que el análisis de enlaces se encarga de mejorar. La búsqueda semántica, por su

parte, amplía el uso de los paradigmas convencionales de recuperación de información (Wei et al., 2008).

Para abordar mejor estas características del lenguaje biomédico en español es imprescindible crear reglas heurísticas hechas expresamente para este lenguaje. Para mejorar la comprensión y el reconocimiento de la información incluida en los textos médicos (Aguilera Murrell, 2022).

Para abordar estos requisitos, se decidió llevar a cabo una investigación para crear un sistema basado en la web para la búsqueda semántica e identificar entidades biomédicas.

Objetivos

Objetivos General

Desarrollar una interfaz web que permita a los usuarios buscar entidades biomédicas en español.

Objetivos Específicos

- Explorar nuevos enfoques para identificar y estandarizar conceptos biomédicos.
- Validar el sistema mediante la aplicación en casos de uso reales en el ámbito biomédico. Esto implicaría probar el sistema en situaciones prácticas, como la extracción de información de artículos científicos, registros médicos electrónicos u otras fuentes de datos biomédicos en español.
- Adaptar prácticas eficientes para implementar sistemas con arquitecturas actuales.

Variables de Investigación

Variable Independiente

- Algoritmo de búsqueda semántica.
- Herramientas o recursos biomédicos.
- Parámetros de configuración.
- Precisión el umbral de similitud semántica

Variable Dependiente

- Eficiencia de la búsqueda semántica.

Hipótesis

La combinación de técnicas de procesamiento del lenguaje natural (NLP) y aprendizaje automático (machine learning) con ontologías y bases de conocimiento biomédicas puede mejorar significativamente la eficiencia y precisión de las búsquedas semánticas de entidades biomédicas.

Capítulo II

Marco Teórico

UMLS

UMLS es una estructura de metáfora conceptual en la que términos de diferentes vocabularios se agrupan en conceptos unificados y se basan en relaciones semánticas entre términos. Su objetivo es integrar y armonizar una amplia gama de términos médicos y terminología utilizada en diversos campos de la medicina (Bodenreider, 2004).

Esto permite que las aplicaciones y los sistemas médicos accedan a la información y realicen búsquedas más precisas, así como mapear y traducir varios términos médicos (Aronson, 2001).

Átomos

Los átomos en UMLS son unidades básicas que representan términos médicos específicos y sus relaciones semánticas, y juegan un papel fundamental en la estructura y función de los sistemas que facilitan el acceso y la comprensión de la información médica (Pires & Ruiz, 2010).

Los átomos en UMLS se pueden dividir en dos categorías principales:

Átomos de término: estos átomos representan términos específicos y sus sinónimos que se encuentran en el léxico médico integrador de UMLS (Pires & Ruiz, 2010).

Átomos de relación: estos átomos describen relaciones semánticas entre diferentes términos médicos (Pires & Ruiz, 2010).

Relaciones entre conceptos

Las relaciones conceptuales se construyen sobre las relaciones semánticas entre términos médicos que describen sus interacciones y conexiones en medicina y salud. La interoperabilidad, la precisión y la comprensión de la información médica dependen de estas relaciones para proporcionar una navegación semántica más rica y precisa (Ortega et al., 2008).

Hay muchas otras relaciones semánticas que brindan información sobre la información médica y hacen que el sistema sea más fácil de usar para mejorar la atención al paciente, la investigación biomédica y otras aplicaciones médicas (López-Úbeda et al., 2018).

Reconocimiento de entidades biomédicas con modelos NER

Los modelos para el Reconocimiento de entidades nominales (NER) se refiere al uso de técnicas de procesamiento de lenguaje natural (PNL) y aprendizaje automático para identificar y extraer entidades biomédicas específicas de textos médicos, como nombres de medicamentos, productos farmacéuticos, productos farmacéuticos, ingredientes activos y otras entidades farmacológicamente relacionadas (Buttigieg et al., 2013).

PharmacoNER.

PharmacoNER es un corpus de texto etiquetado que contiene unidades de medicamentos de una variedad de fuentes, incluidas publicaciones científicas, artículos de noticias y sitios web de medicamentos. El corpus se utiliza para entrenar modelos de aprendizaje automático que pueden identificar unidades de medicamentos en el texto. Estos modelos se pueden utilizar para mejorar la precisión de la búsqueda de medicamentos(Gonzalez-Agirre et al., 2019).

Normalización de entidades biomédicas

La normalización de entidades biomédicas (BME) es el proceso de convertir entidades biomédicas, como enfermedades, medicamentos y genes, en una representación única y consistente en un conjunto de datos. Esto es importante para la investigación biomédica y la

atención médica, ya que permite a los investigadores y los proveedores de atención médica comparar y analizar datos de diferentes fuentes (Aguilera Murrell, 2022).

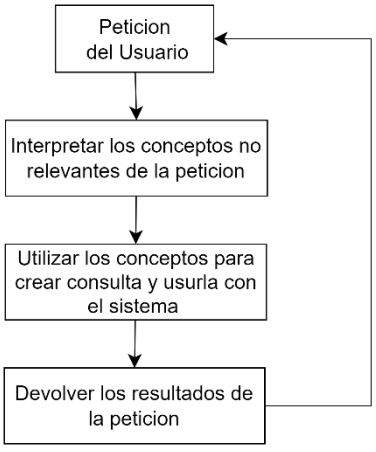
La estandarización de las unidades biomédicas es necesaria para diversas aplicaciones de la informática biomédica, como el análisis de datos clínicos, la gestión de bases de datos biomédicas, la interoperabilidad entre los sistemas sanitarios, la farmacovigilancia y la investigación biomédica. Estandarizar la representación de entidades mejora la calidad y consistencia del uso de la información biomédica, facilitando su análisis tanto en aplicaciones clínicas como de investigación (Aguilera Murrell, 2022).

Búsqueda semántica

La búsqueda semántica es una estrategia avanzada que utiliza significados y relaciones semánticas entre conceptos para mejorar la recuperación de información y brindar a los usuarios resultados más relevantes y útiles, la búsqueda semántica puede proporcionar resultados más completos y precisos (Hidalgo Delgado & Rodríguez Puente, 2013).

Figura 1

Arquitectura de la búsqueda semántica genérica



Nota. En la figura 1 se presenta la estructura general de una búsqueda semántica en la que mediante una petición del cliente se obtienen los conceptos más relevantes de la petición para

crear la consulta y usarla con la ontología del sistema para finalmente devolver al usuario los resultados obtenidos.

Búsqueda semántica para normalizar entidades biomédicas

En el campo de la informática biomédica, la búsqueda semántica para estandarizar entidades biomédicas utiliza significados semánticos y relaciones entre conceptos biomédicos para proporcionar una representación uniforme y estandarizada de entidades (López-Úbeda et al., 2018).

Búsqueda semántica con embeddings

Las embeddings son representaciones numéricas densas que convierten palabras o conceptos en vectores de números reales en un espacio multidimensional. Estos vectores están diseñados para que palabras o conceptos similares se representen juntos en el espacio vectorial, lo que ayuda a medir la similitud semántica entre ellos (Delgado, 2022).

Mediante la representación semántica, los embeddings captan el significado de una palabra o concepto basándose en su uso y contexto en el lenguaje natural. Los términos o ideas similares se representarán en el espacio mediante un vector (Delgado, 2022).

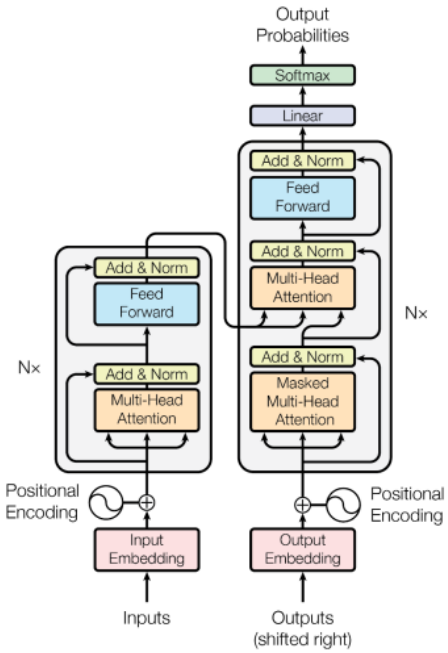
Para tareas como encontrar información en grandes corpus, extraer documentos científicos o biomédicos, hacer recomendaciones de contenido y otros usos que requieren una comprensión más profunda del significado del lenguaje natural, la búsqueda semántica integrada es particularmente útil para mejorar la precisión y relevancia de los resultados de búsqueda (Jolodkow & Romani, 2021).

Generación de embeddings con modelos transformers.

La técnica de entrenamiento de una red transformers y un modelo en un conjunto de datos es un enfoque muy sofisticado para generar incrustaciones en el ámbito del procesamiento del lenguaje natural. Los modelos transformadores, como BERT (Bidirectional Encoder Representations from Transformers), han revolucionado la representación de palabras y frases en el aprendizaje automático basado en texto (Pérez Josende, 2022).

Figura 2

Diagrama de Arquitectura Transformers



Nota. Es importante señalar que la Figura 2 ilustra una red neuronal de tipo transformador (GPT y BERT) y demuestra cómo utilizarla para construir modelos únicos para el análisis (Pérez Josende, 2022).

Herramientas para la realización de una búsqueda semántica con embeddings

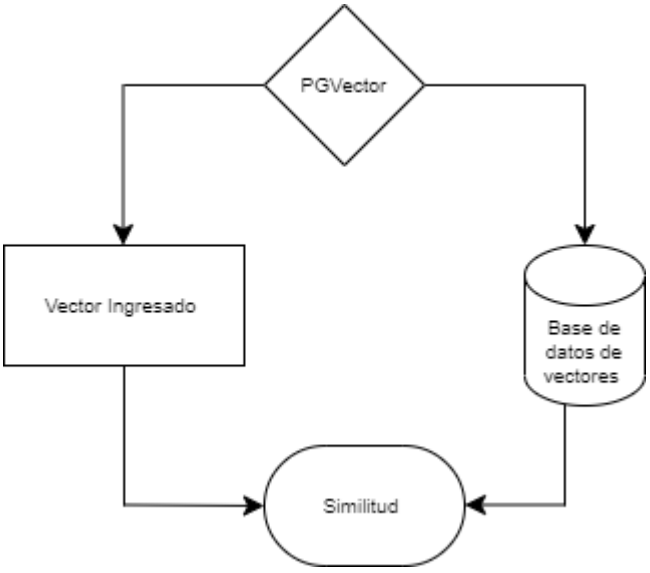
El uso de herramientas y bibliotecas específicas que permiten el cálculo y la búsqueda de similitud semántica entre vectores de embeddings es necesario para la búsqueda semántica que utiliza embeddings. Para identificar entidades y conceptos relacionados en el contexto de la búsqueda semántica, estas herramientas son cruciales para comparar y relacionar el significado de las palabras y frases representadas por los vectores (Smilkov et al., 2016).

PgVector

PgVector es una extensión que proporciona búsqueda de similitud de vectores y almacenamiento integrado para PostgreSQL. Esto es particularmente útil para aplicaciones relacionadas con el procesamiento del lenguaje natural (*pgvector*, 2021/2023).

Figura 3

Diagrama PgVector



Nota. Es una extensión que ofrece almacenamiento integrado PostgreSQL y búsqueda de similitud vectorial. (*pgvector*, 2021/2023).

Sentence transformers

El procesamiento del lenguaje natural (NLP) es determinar la similitud semántica de las oraciones del texto. El modelado de pares de oraciones, la similitud del texto y el modelado del lenguaje son algunas tareas importantes en la NLP. Los algoritmos de aprendizaje automático tradicionales requieren una gran cantidad de datos de entrenamiento, pero este es un proceso que requiere mucho tiempo (Mayil & Jeyalakshmi, 2023).

Sentence Transformers es una biblioteca de Python, utiliza modelos de transformadores pre entrenados para generar representaciones semánticas densas de oraciones o párrafos completos.

Capítulo III

Desarrollo del sistema

Este capítulo describe el desarrollo de un sistema web para identificar y normalizar entidades biomédicas utilizando reglas heurísticas y búsqueda semántica. El sistema recibe como entrada texto relevante para el contexto médico y genera como salida un extracto de datos, como proteínas, químicos y sustancias farmacológicas. El sistema funciona de la siguiente manera: i) El sistema utiliza un proceso de reconocimiento de entidades nombrado (NER) para identificar las entidades biomédicas en el texto. ii) Luego, el sistema extrae las entidades biomédicas identificadas. iii) A continuación, el sistema transforma las entidades biomédicas a vectores utilizando Sentence Transformers. iv) Finalmente, el sistema busca la similitud de los vectores en una base de datos para obtener el código biomédico de la entidad biomédica.

El sistema ha sido evaluado en un conjunto de datos de texto biomédico y se ha encontrado que es capaz de identificar y normalizar entidades biomédicas con un alto grado de precisión.

El sistema se está desarrollando utilizando el método ágil de desarrollo de software, que es un enfoque de desarrollo de software que se centra en el trabajo en equipo y la colaboración. El equipo de desarrollo trabaja en iteraciones cortas, llamadas sprints, y entrega un producto funcional al final de cada sprint. Esto permite al equipo obtener comentarios de los usuarios temprano y a menudo, y realizar cambios en el sistema según sea necesario (Trigás Gallego, 2012).

Scrum es una metodología ágil, lo que significa que se enfoca en el trabajo en equipo, la colaboración y la entrega rápida de productos. Scrum se divide en iteraciones cortas, llamadas sprints, y al final de cada sprint, el equipo de desarrollo entrega un producto funcional. Esto permite al equipo obtener comentarios de los usuarios temprano y a menudo, y realizar cambios en el sistema según sea necesario (Deemer et al., 2009).

El flujo de trabajo Scrum involucra al Scrum Master, el Product Owner y el equipo Scrum. El Scrum Master tiene un rol clave responsable de administrar el proceso y eliminar los obstáculos que pueden afectar la entrega del producto. El Scrum Master es responsable de garantizar que el equipo Scrum siga los principios y prácticas de Scrum. El Scrum Master también es responsable de ayudar al equipo Scrum a resolver problemas y superar obstáculos (Trigás Gallego, 2012).

El Product Owner es responsable de asegurar que el producto satisfaga las necesidades de los usuarios. El Product Owner trabaja con el equipo Scrum para definir las características del producto y priorizar el trabajo. El equipo Scrum es responsable de desarrollar el producto. El equipo Scrum trabaja en iteraciones cortas, llamadas sprints, y al final de cada sprint, el equipo entrega un producto funcional (Trigás Gallego, 2012).

Scrum es un marco de trabajo ágil que puede adaptarse a diferentes proyectos con diferentes requisitos. Scrum proporciona flexibilidad al equipo de desarrollo para elegir los requisitos que se deben completar en cada sprint. Esto permite al equipo de desarrollo adaptarse a los cambios en los requisitos del proyecto y entregar el producto lo más rápido posible (Trigás Gallego, 2012).

Sprint: Cada sprint de Scrum dura entre una y tres semanas. Durante este tiempo, el equipo de desarrollo completa una tarea específica. El objetivo de cada sprint es entregar un producto potencialmente entregable (Trigás Gallego, 2012).

Sprint backlog: Es una herramienta importante para el equipo de desarrollo. Ayuda al equipo a mantenerse enfocado en los objetivos del sprint y a garantizar que el trabajo se complete a tiempo y dentro del presupuesto (Trigás Gallego, 2012).

Product backlog: es una lista de requisitos para el producto. El backlog del producto se crea y se mantiene por el propietario del producto, con la ayuda del equipo de desarrollo. El

backlog del producto se divide en historias de usuario, que son historias breves y fáciles de entender sobre lo que el usuario quiere que haga el producto (Trigás Gallego, 2012).

Análisis y diseño del sistema

Las historias de usuarios se utilizarán para delinear los deberes de cada miembro del equipo involucrado en el desarrollo del proyecto. Esto se llevará a cabo tras un examen exhaustivo de los componentes de la metodología Scrum para la selección de requisitos (Izaurralde, 2013).

La definición de los roles de cada uno de los partícipes en el proyecto se visualiza en la Tabla 1. El Scrum Master es responsable de asignar roles a los miembros del equipo. El Scrum Master también es responsable de explicar el rol de cada miembro del equipo y la función que deben realizar durante el proceso del desarrollo del proyecto.

Tabla 1

Descripción del Scrum Team

N°	**Rol**	**Participante**	**Responsabilidades**
1	Product Owner	Msc. Álvaro Uyaguari	Encargado de establecer los requisitos del sistema, administrar el proceso de desarrollo del sistema y asignar listas de tareas.
2	Scrum Master	Christopher llano	Líder del equipo que va a dirigir y guiar el desarrollo del sistema.
3	Developers	Cristopher Rugel y Christopher Llano	Desarrolladores cuya misión es realizar las tareas para llevar a cabo el sistema de reconocimiento y normalización de entidades biomédicas

Historias de Usuarios

Las metodologías ágiles aprovechan las historias de usuario para acortar los documentos formales, agilizar la gestión del tiempo y facilitar la especificación de las necesidades. Además, permiten dar respuestas rápidas a medida que cambian las necesidades (Villamizar Suaza et al., 2015).

La Tabla 2 se utiliza para todo el proceso de desarrollo del proyecto. Pueden utilizarse para definir los requisitos del producto, para supervisar el proceso de desarrollo del producto y para asignar las tareas al equipo de desarrollo.

Tabla 2

Descripción de las Historias de Usuario empleadas

N° de historia de usuario	**Nombre**	**Rol**	**Característica**	**Razón**
1	S.R.N.E.B.1	Como programador	Quiero desarrollar un sistema que pueda identificar y extraer entidades biomédicas de textos.	Para el reconocimiento automático de entidades biomédicas.
2	S.R.N.E.B.2	Como programador	Quiero normalizar textos a través de un algoritmo para la obtención de conceptos biomédicos.	Para identificar entidades biomédicas a partir del texto de entrada y extraiga el concepto al que pertenece la entidad.

N° de historia de usuario	Nombre	Rol	Característica	Razón
3	S.R.N.E.B.3	Como programador	Implementar el algoritmo desarrollado dentro de un sistema web.	Para la representación visual del algoritmo dentro de una interfaz intuitiva y agradable para el usuario.

Product Backlog del proyecto

La ejecución del backlog del producto, que consiste en una lista de requisitos que el propietario del producto quiere que se cumplan en el producto. La ejecución del backlog del producto es el proceso de convertir estos requisitos en realidad (Izaurralde, 2013).

En la Tabla 3 se muestra el Product Backlog, el cual contiene una lista de todas las historias de usuario que se desarrollarán durante el proyecto. Cada historia de usuario tiene una estimación de tiempo en días, una fecha de inicio, una fecha final y el número de sprint al que pertenece.

Tabla 3

Descripción del Product Backlog empleado

N° de historia de usuario	Nombre	Estimación (días)	fecha inicio	Fecha final	N° de Sprint
1	S.R.N.E.B.1	21	03/05/2023	31/05/2023	1

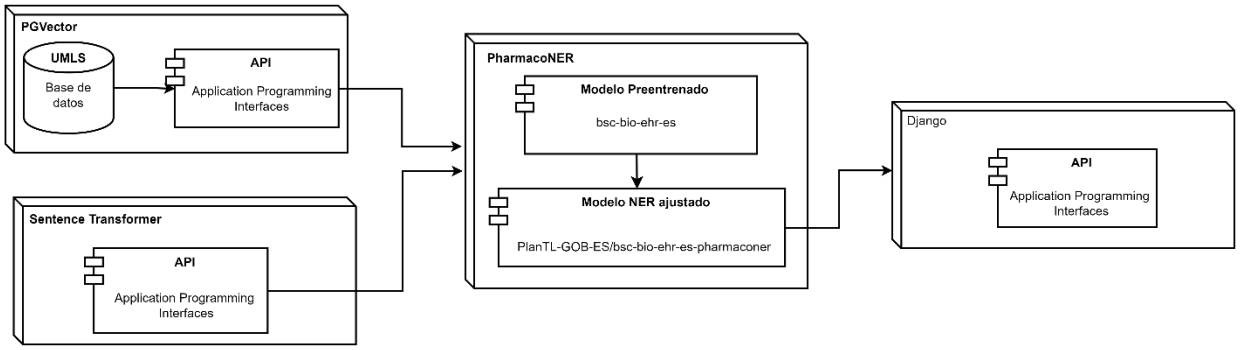
2	S.R.N.E.B.2	25	01/06/2023	05/07/2023	2
3	S.R.N.E.B.3	20	06/07/2023	02/08/2023	3

Arquitectura

La arquitectura que se aplicó en el sistema de búsquedas semánticas en lenguaje español es la arquitectura de componentes. Esta arquitectura se enfoca en la descripción del diseño en componentes funcionales que tengan interfaces bien definidas (Garlan, 2008). El objetivo principal de esta arquitectura es descomponer el diseño en componentes lógicos o funcionales que expongan interfaces de comunicación bien definidas.

Figura 4

Diagrama de los componentes del sistema

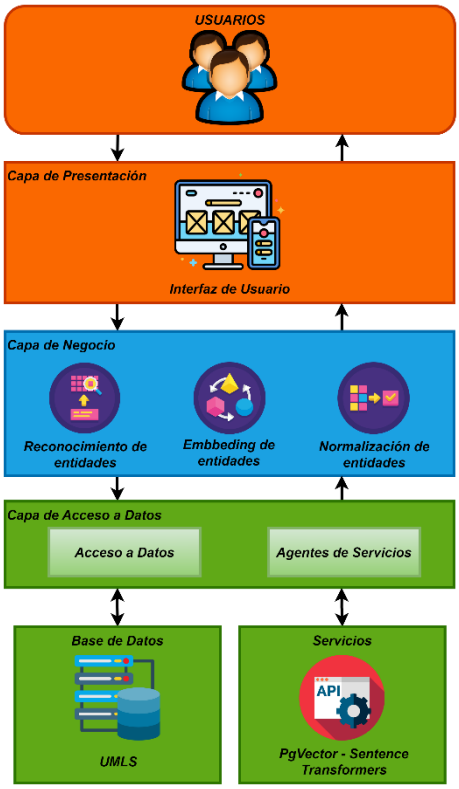


Nota. Diseño del sistema web para reconocimiento de entidades biomédicas basado en reglas heurísticas y búsquedas semánticas en lenguaje español. Utiliza el diagrama mostrado en la figura 4, en el cual se detalla la arquitectura utilizada, las cuales contienen las herramientas del PgVector, Sentence Transformers y el modelo PharmaCoNER.

Arquitectura del sistema web basada en capas.

Figura 5

Diagrama de la arquitectura basada en capas.



Nota. La figura 5 nos muestra el diagrama de la arquitectura basada en capas que tendrá el sistema. Donde estará compuesta por una capa de presentación que contiene la funcionalidad relacionada con la interfaz de usuario, una capa de negocios que presenta las funciones que realiza el sistema como normalizar, embeddings y normalización de entidades y por último cuenta con una capa de acceso a datos donde se la realiza las consultas a la base de datos UMLS y también tiene acceso a los servicios como el PgVector y el Sentence Transformers.

Herramientas empleadas para el desarrollo del sistema

Para desarrollar el sistema, se seleccionaron diferentes herramientas que nos ayudarán a lograr nuestro objetivo. Estas herramientas se explicarán más adelante, incluyendo sus características, funcionamiento y otros componentes necesarios para su correcto funcionamiento.

Tabla 4

Descripción de las herramientas empleadas

Nombre	Descripción
Python	Lenguaje de programación utilizado para el desarrollo del proyecto (Versión 3.11.3).
Visual Studio Code	Editor de código utilizado.
Docker	Software utilizado para crear la imagen de la base de datos UMLS.
PgAdmin	Herramienta de administración de base de datos utilizada para gestionar y trabajar con la base de datos UMLS.

Nota. En la tabla 4 se indican las herramientas que fueron utilizadas para el desarrollo del sistema.

Definición e implementación de modelos.

De acuerdo con la metodología utilizada en el diseño del sistema, la siguiente fase del sistema implicará la planificación de cada sprint y la clasificación de las tareas más cruciales en orden de importancia utilizando el Sprint Backlog. Se realizaron reuniones presenciales y videoconferencias utilizando la plataforma Google Meet para cumplir con los objetivos antes mencionados (Izaurralde, 2013).

Sprint 1: Desarrollar un algoritmo para el etiquetado automático de entidades médicas

Para el desarrollo del Sprint 01, se utilizó la Historia de Usuario (S.R.N.E.B.1), ubicada en la Tabla 5, en donde se detalla que se va a desarrollar un algoritmo para el etiquetado automático de entidades médicas.

Descripción de la Historia de Usuario 1

Tabla 5

Historia de Usuario para desarrollar un algoritmo para el etiquetado automático de entidades médicas.

Historia de Usuario

Número: S.R.N.E.B.1	**Usuario:** Administrador
Característica: investigar e implementar el reconocimiento de entidades biomédicas.	**Número de Sprint:** 1
Prioridad: Alta	**Riesgo de desarrollo:** Bajo
Duración: 21 días	**Interacción asignada:** 1

Responsables en el desarrollo: Cristopher Rugel, Christopher Llano

Descripción: Como programador quiero investigar e implementar el reconocimiento de entidades biomédicas.

Validación:

- Se elaboró un algoritmo que permite el reconocimiento automático de entidades biomédicas.
- Se realizaron pruebas del funcionamiento del algoritmo ingresando textos y validando las entidades a partir del corpus PharmaCoNER.

Sprint Backlog 1

En la tabla 6 se muestra el sprint Backlog 1, donde se especifican las tareas realizadas en el sprint 1, horas empleadas, fechas de inicio y fin, responsable y el estado.

Tabla 6

Sprint Backlog 1

NOMBRE	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
S.R.N.E.B.1	Selección, configuración y testeo de las herramientas	15	03/05/ 2023	09/05/ 2023	Cristopher Rugel, Christopher Llano	Completado
S.R.N.E.B.1	Instalar la imagen de UMLS en Docker	18	10/05/ 2023	17/05/ 2023	Cristopher Rugel, Christopher Llano	Completado
S.R.N.E.B.1	Análisis e implementación de la clasificación de tokens utilizando el corpus de PharmaCoNER	30	18/05/ 2023	31/05/ 2023	Cristopher Rugel, Christopher Llano	Completado

Sprint 2: Normalizar textos a través de un algoritmo para la obtención de conceptos biomédicos.

Para el desarrollo del Sprint 2, se utilizó la Historia de Usuario (S.R.N.E.B.2), ubicada en la Tabla 7, en donde se detalla que se va a desarrollar un algoritmo para la obtención de conceptos biomédicos.

Descripción de la Historia de Usuario 2

Tabla 7

Historia de Usuario para normalizar textos a través de un algoritmo para la obtención de conceptos biomédicos.

Historia de Usuario	
Número: S.R.N.E.B.2	**Usuario:** Administrador
Característica: *normalizar textos a través de un algoritmo para la obtención de conceptos biomédicos*	**Número de Sprint:** 1
Prioridad: Alta	**Riesgo de desarrollo:** Alto
Duración: 25 días	**Interacción asignada:** 5

Responsables en el desarrollo: Cristopher Rugel, Christopher Llano

Descripción: Como programador quiero normalizar textos a través de un algoritmo para la obtención de conceptos biomédicos.

Validación:

- Se realizará la conversión de la base de datos UMLS a vectores
- Se realizará el desarrollo del algoritmo.
- Se realizará la conexión a la base de datos para buscar la similitud entre el texto ingresado y la base de datos UMLS.
- Se realizaron pruebas del algoritmo.

Sprint Backlog 2

En la tabla 8 se muestra el sprint Backlog 2, donde se especifican las tareas realizadas en el sprint 2, horas empleadas, fechas de inicio y fin, responsable y el estado.

Tabla 8

Sprint Backlog 2

NOMBRE	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
S.R.N.E.B.2	Utilización y conversión de la base de datos UMLS a vectores con la herramienta Sentence Transformers.	36	01/06/ 2023	16/06/ 2023	Cristopher Rugel, Christopher Llano	Completado
S.R.N.E.B.2	Desarrollo del algoritmo	30	19/06/ 2023	30/06/ 2023	Cristopher Rugel, Christopher Llano	Completado
S.R.N.E.B.2	Aplicación y verificación de la funcionalidad del algoritmo	6	03/07/ 2023	05/07/ 2023	Cristopher Rugel, Christopher Llano	Completado

Sprint 3: Implementar el algoritmo desarrollado dentro de un sistema web.

Para el desarrollo del Sprint 3, se utilizó la Historia de Usuario (S.R.N.E.B.3), ubicada en la Tabla 8, en donde se detalla que se va a implementar el algoritmo desarrollado dentro de un sistema web.

Descripción de la Historia de Usuario 3

Tabla 9

Historia de Usuario para implementar el algoritmo desarrollado dentro de un sistema web.

Historia de Usuario	
Número: S.R.N.E.B.3	**Usuario:** Administrador
Característica: *implementar el algoritmo desarrollado dentro de un sistema web.*	**Número de Sprint:** 1
Prioridad: Alta	**Riesgo de desarrollo:** bajo
Duración: 20 días	**Interacción asignada:** 3
Responsables en el desarrollo: Cristopher Rugel, Christopher Llano	
Descripción: Como programador quiero implementar el algoritmo desarrollado dentro de un sistema web.	
Validación: - Se desarrollará una interfaz adecuada para el uso del algoritmo. - Se realizará la optimización del código para mejorar la satisfacción del usuario. - Se realizaron las pruebas del sistema.	

Sprint Backlog 3

En la tabla 10 se muestra el sprint Backlog 3, donde se especifican las tareas realizadas en el sprint 3, horas empleadas, fechas de inicio y fin, responsable y el estado.

Tabla 10

Sprint Backlog 3

NOMBRE	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
S.R.N.E.B.3	Desarrollo de una interfaz utilizando el framework de Django.	21	06/07/ 2023	14/07/ 2023	Cristopher Rugel, Christopher Llano	Completado
S.R.N.E.B.3	Integración del algoritmo a la interfaz web.	15	17/07/ 2023	21/07/ 2023	Cristopher Rugel, Christopher Llano	Completado
S.R.N.E.B.3	Optimización del algoritmo y pruebas del sistema.	24	24/07/ 2023	02/08/ 2023	Cristopher Rugel, Christopher Llano	Completado

Programación en Python: Algoritmo e Interfaz web

Algoritmo para la normalización de entidades biomédicas

Para el desarrollo de nuestro proyecto, utilizamos un algoritmo hecho a medida, mediante el uso del lenguaje de programación Python. Como primer paso consiste en reconocer una entidad biomédica de un texto anteriormente ingresado, esto mediante un

algoritmo NER. La entidad puede ser una palabra o un grupo de palabras que corresponden a la misma categoría.

El segundo paso consiste en extraer la entidad detectada y transformarla a vector mediante el uso Sentence Transformers. Una vez convertidas en vectores todas las entidades detectadas, posteriormente se busca la similitud de estos vectores en una base de datos para determinar el código biomédico al que pertenece utilizando PgVector.

Figura 6

Algoritmo para la lectura de texto y reconocimiento de entidades biomédicas

```
        texto_ingresado = request.POST.get('texto')

        # Cargar el modelo y el tokenizador
        tokenizer = AutoTokenizer.from_pretrained("PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer")
        model = AutoModelForTokenClassification.from_pretrained("PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer")
        pipe = pipeline(task='token-classification', model=model, tokenizer=tokenizer, aggregation_strategy='max')

        # Realizar el tratamiento del texto utilizando el modelo
        results = pipe(texto_ingresado)

        # Obtener la forma original de las palabras de las entidades reconocidas
        combined_words = []
        current_word = ""
        last_end = -1
        for res in results:
            start_idx, end_idx = res['start'], res['end']
            entity_text = texto_ingresado[start_idx:end_idx]

            if start_idx == last_end:
                current_word += entity_text
            else:
                if current_word:
                    # Agregamos la palabra al array solo si no está duplicada
                    if current_word not in combined_words:
                        combined_words.append(current_word)
                current_word = entity_text

            last_end = end_idx

        if current_word:
            # Agregamos la última palabra al array solo si no está duplicada
            if current_word not in combined_words:
                combined_words.append(current_word)
```

Figura 7

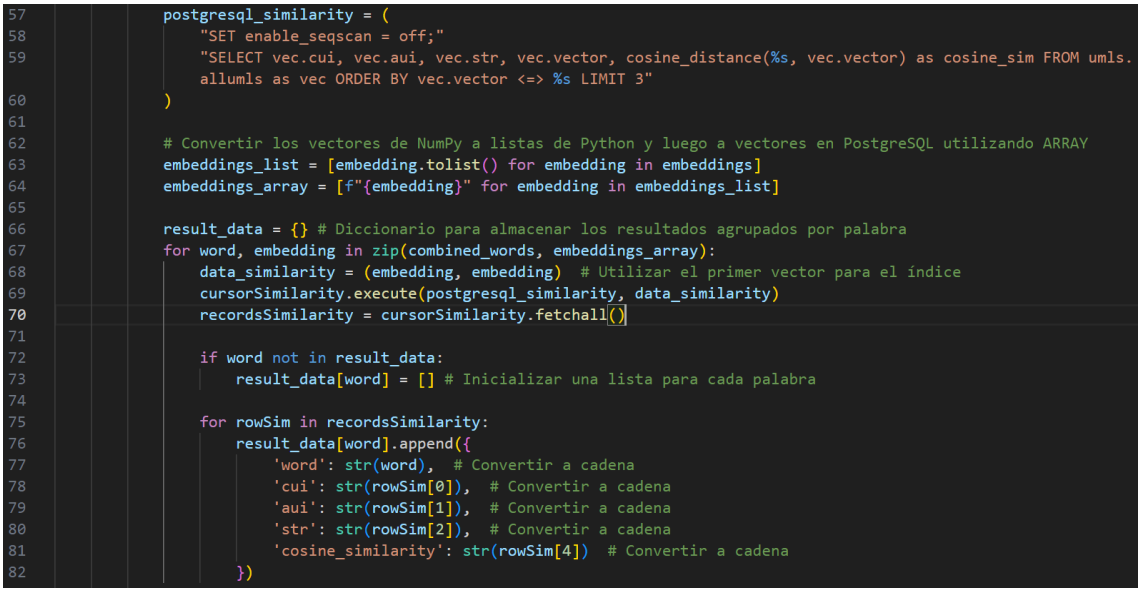
Algoritmo para convertir a vector las entidades detectadas.

```
        # Creamos el modelo
        Sentence_Transformer_model = SentenceTransformer('paraphrase-multilingual-mpnet-base-v2')

        # Ahora vamos a convertir cada palabra en combined_words en un vector utilizando el modelo
        embeddings = Sentence_Transformer_model.encode(combined_words)
```

Figura 8

Algoritmo para la normalización de las entidades detectadas.



```
            postgresql_similarity = (
                "SET enable_seqscan = off;"
                "SELECT vec.cui, vec.aui, vec.str, vec.vector, cosine_distance(%s, vec.vector) as cosine_sim FROM umls.
                allumls as vec ORDER BY vec.vector <=> %s LIMIT 3"
            )

            # Convertir los vectores de NumPy a listas de Python y luego a vectores en PostgreSQL utilizando ARRAY
            embeddings_list = [embedding.tolist() for embedding in embeddings]
            embeddings_array = [f"{embedding}" for embedding in embeddings_list]

            result_data = {} # Diccionario para almacenar los resultados agrupados por palabra
            for word, embedding in zip(combined_words, embeddings_array):
                data_similarity = (embedding, embedding)  # Utilizar el primer vector para el índice
                cursorSimilarity.execute(postgresql_similarity, data_similarity)
                recordsSimilarity = cursorSimilarity.fetchall()

                if word not in result_data:
                    result_data[word] = [] # Inicializar una lista para cada palabra

                for rowSim in recordsSimilarity:
                    result_data[word].append({
                        'word': str(word),  # Convertir a cadena
                        'cui': str(rowSim[0]),  # Convertir a cadena
                        'aui': str(rowSim[1]),  # Convertir a cadena
                        'str': str(rowSim[2]),  # Convertir a cadena
                        'cosine_similarity': str(rowSim[4])  # Convertir a cadena
                    })
```

Algoritmo para la interfaz web

Para la composición de la interfaz web, combinamos el marco de desarrollo de Django con una capa frontal basada en HTML, CSS y JavaScript para crear la interfaz web. En cuanto a la estructura interna, se utilizará la programación Python para implementar los elementos de soporte. Para la visualización de los resultados de desarrollo un apartado ‘main’ para mostrar tanto el texto ingresado, como las entidades biomédicas detectadas por el sistema como muestra la Figura 9. Además, para mostrar la normalización de las entidades biomédicas detectadas se desarrolló un apartado ‘content’ en el que se detalla mediante una tabla los resultados de la normalización por cada entidad biomédica encontrada como muestra la Figura 10.

Figura 9

Algoritmo para mostrar las entidades biomédicas detectadas en una interfaz web.

```
<div class="main">
    <div class="cuadro">
        <h1>Resultado del análisis</h1>
        <hr/>
        <div class="analisis">{{ texto_ingresado|safe }}</div>
    </div>
    <div class="cuadro">
        <h1>Busquedas semánticas</h1>
        <hr/>
        <div class="analisis-2">
            <ul>
                {% for enlace_html in enlaces_html %}
                    {{ enlace_html|safe }}
                {% endfor %}
            </ul>
        </div>
    </div>
</div>
```

Figura 10

Algoritmo para mostrar la normalización de las entidades biomédicas detectadas en una interfaz web.

```
<div class="content">
    {% for word, results in result_data.items %}
        <div class="table">
            <h1>{{ word }}</h1>
            {% if results %}
                <table class="content-table">
                    <thead>
                        <tr>
                            <th>Palabra</th>
                            <th>CUI</th>
                            <th>AUI</th>
                            <th>STR</th>
                            <th>Similitud Coseno</th>
                        </tr>
                    </thead>
                    <tbody>
                        {% for result_row in results %}
                            <tr>
                                <td>{{ result_row.word }}</td>
                                <td>{{ result_row.cui }}</td>
                                <td>{{ result_row.aui }}</td>
                                <td>{{ result_row.str }}</td>
                                <td>{{ result_row.cosine_similarity }}</td>
                            </tr>
                        {% endfor %}
                    </tbody>
                </table>
        </div>
    {% endfor %}
</div>
```

Capítulo IV

Resultados del Sistema

Las búsquedas semánticas son una parte muy importante en la búsqueda de entidades biomédicas, por lo que es importante contar con una búsqueda casi perfecta y automatizada de estas entidades médicas. En este capítulo hablaremos de las actividades realizadas durante todo el periodo de este proyecto para la obtención de resultados más precisos en la búsqueda semántica.

Para el funcionamiento del sistema se toma como entrada un lenguaje pertinente al contexto médico que como resultado nos mostrará los datos extraídos, que en este caso pueden estar categorizados como proteínas, químicos y sustancias farmacéuticas. Enumeramos brevemente los procedimientos que utiliza el sistema para su función: i) El primer paso consiste en reconocer una entidad mediante un algoritmo NER; la entidad puede ser una palabra o un grupo de palabras que corresponden a la misma categoría. ii) Después de extraer la entidad, se procederá a normalizar la entidad detectada. Las categorías de una entidad pueden ser de cualquier asunto, en este caso estarán comprendidas en entidades de tipo (proteínas, químicos y sustancias farmacéuticas).

Selección de las herramientas para la validación del sistema

Para la validación del sistema, se seleccionaron varias herramientas que nos ayudarán a alcanzar nuestro objetivo. Estas herramientas se explicarán en detalle más adelante, incluidas sus características, funcionamiento y otros componentes necesarios para su correcto funcionamiento.

EMEA (Del Clef Gold Corpus)

El corpus EMEA (Del Clef Gold) es un conjunto de datos de etiquetas de medicamentos de la Agencia Europea de Medicamentos (EMA). Fue creado como parte de la campaña de evaluación CLEF (Conferencia y Laboratorios del Foro de Evaluación) en 2013. El corpus

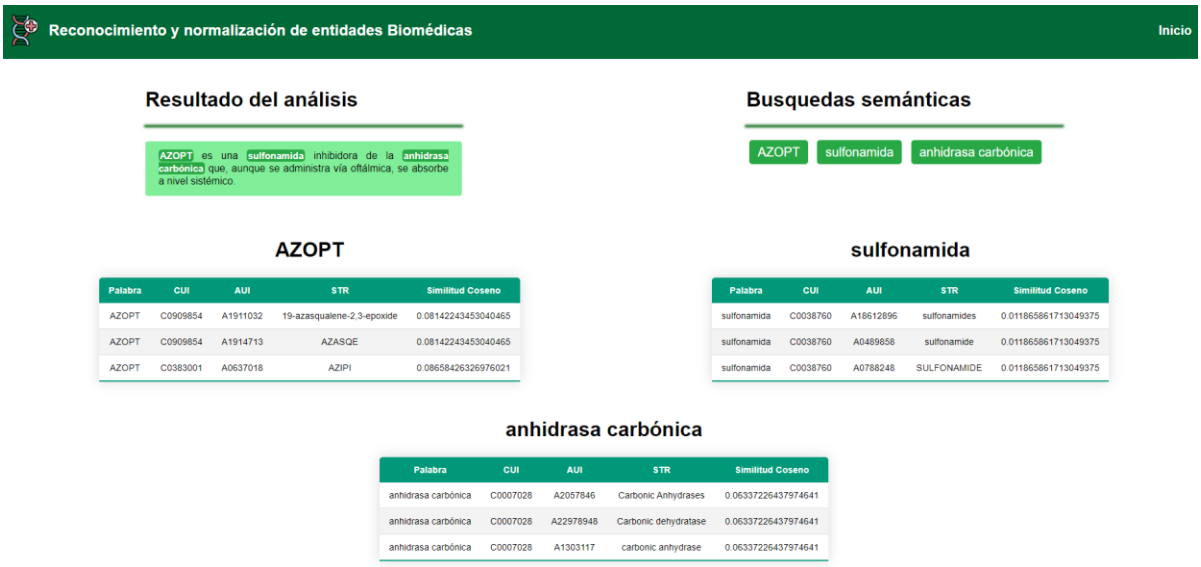
consta de 1.000 etiquetas de medicamentos en inglés, francés, alemán, italiano y español (Liu et al., 2014).

Resultados Alcanzados

La interfaz web para el reconocimiento y normalización de entidades biomédicas basado en reglas heurísticas y búsquedas semánticas en lenguaje español ha sido puesta a prueba y se presenta a continuación los resultados alcanzados.

Figura 11

Interfaz de usuario del Sistema de reconocimiento y normalización de entidades biomédicas



Nota. En la figura 11 se observa el sistema de reconocimiento y normalización de entidades biomédicas, donde se ve integradas las herramientas de Sentence Transformers, PgVector y por último el Corpus de etiquetado de PharmaCoNER, además de poder evidenciarse que gracias al uso de los mismos se puede obtener los resultados de la normalización de las entidades reconocidas.

Validación del PharmaCoNER

Tabla 11

Resultados del reconocimiento de entidades biomédicas con PharmaCoNER

Corpus	*Reconocimiento de entidades biomédicas con el corpus PharmaCoNER*	*% de efectividad*
EMEA	44/53	83.02%

Nota. En la tabla 11 se presenta el resultado del reconocimiento de entidades biomédicas usando el corpus de PharmaCoNER, el mismo que muestra que fue capaz de reconocer 44 entidades de las 53 que contiene el corpus EMEA.

Validación de la normalización

Tabla 12

Resultados de la efectividad de la normalización de entidades biomédicas.

Corpus	*Normalización de entidades biomédicas*	*% de efectividad*
EMEA	23/53	43.39%

Nota. En la tabla 12 se presenta el resultado de la normalización de entidades biomédicas, el mismo que muestra que fue capaz de reconocer 23 entidades de las 53 que contiene el corpus EMEA.

Tabla 13

Ejemplo 1 del resultado utilizando el sistema en el corpus EMEA

Texto ingresado	*Entidad biomédica*	*CUI*	**Reconocimiento de entidad biomédica**	***Normalización de entidad biomédica***
AZOPT es una sulfonamida	anhidrasa carbónica	C0007028	1	1

Texto ingresado	*Entidad biomédica*	*CUI*	Reconocimiento de entidad biomédica	*Normalización de entidad biomédica*
inhibidora de la anhidrasa carbónica que, aunque se administra vía oftálmica, se absorbe a nivel sistémico.	AZOPT	C0673967	1	0
	sulfonamida	C0038760	1	1

Nota. En la tabla 13 se presenta un ejemplo de cómo actúa el algoritmo para el reconocimiento y normalización de entidades biomédicas donde el “1” representa que es acertada el reconocimiento o normalización y el “0” indica que no fue acertada el reconocimiento o normalización.

Tabla 14

Ejemplo 2 del resultado utilizando el sistema en el corpus EMEA

Texto ingresado	***Entidad biomédica***	***CUI***	**Reconocimiento de entidad biomédica**	***Normalización de entidad biomédica***
Neupro 4 mg/ 24 h parche transdérmico	Neupro	C1949346	1	1
Un parche libera 4 mg de rotigotina cada 24 horas.	rotigotina	C1700683	1	1

Nota. En la tabla 14 se muestra un ejemplo de cómo actúa el algoritmo para el reconocimiento y normalización de entidades biomédicas donde el “1” representa que es acertada el reconocimiento o normalización y el “0” indica que no fue acertada el reconocimiento o normalización.

Capítulo V

Conclusiones y Recomendaciones

Conclusiones

- Se cumplió con el objetivo de desarrollar un Sistema informático para el reconocimiento y normalización de entidades biomédicas basado en reglas heurísticas y búsquedas semánticas en lenguaje español, logrando una precisión del 83.02% en el NER y 43.39% en la normalización de la búsqueda semántica de entidades biomédicas sea más efectiva.
- El uso de las herramientas Sentence Transformers, PgVector y el Corpus de etiquetado de PharmaCoNER permite obtener resultados de normalización de las entidades reconocidas, lo que conduce a resultados más efectivos.
- La herramienta PgVector ayudo al momento de la ejecución del algoritmo implementado, ya que utilizando los índices que tiene PgVector se optimiza de una mejor manera, y así la búsqueda sea más rápida.
- El modelo implementado ayudó a la búsqueda semánticas de entidades biomédicas en su mayoría como proteínas, químicos y sustancias farmacológicas
- El uso del NER permitió generar un porcentaje estable de reconocimiento de entidades correctas, lo cual ayudó en el proceso de la normalización
- El algoritmo implementado para el reconocimiento y normalización de entidades biomédicas, obtuvo un resultado aceptable, durante la evaluación del Corpus EMEA.

Recomendaciones

- Es importante saber de los temas a tratar, tener conocimiento de las búsquedas semánticas de entidades biomédicas, esto ayudará a que se tenga un conocimiento previo.
- Para que el sistema tenga una mejor oportunidad de realizar una búsqueda semántica de entidades biomédicas en este caso, es crucial formular una buena pregunta que desee abordar.
- Para determinar si el sistema tiene éxito en el logro de sus objetivos, es fundamental evaluar su desempeño. Identificar las deficiencias del sistema y las posibles áreas de mejora es crucial si el sistema es ineficaz. Esto se puede lograr examinando los resultados del sistema y hablando con el personal apropiado acerca de las limitaciones del sistema.
- Se debe investigar sobre temas parecidos o sistemas para que así sea se tenga una noción del aplicativo que se realizará.
- El sistema debe ser fácil de usar, lo que significa que debe ser fácil de usar y comprender. El sistema debe proporcionar una interfaz clara y concisa, y debe utilizar una terminología simple y fácil de entender.

Bibliografías

Aguilera Murrell, K. (2022). *Extracción de entidades nombradas en artículos de prensa en español* [B.S. thesis]. Universidad de las Ciencias Informáticas. Facultad de Ciencias y Tecnologías ….

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proceedings of the AMIA Symposium*, 17.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic acids research*, *32*(suppl_1), Article suppl_1.

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., & the ENVO Consortium. (2013). The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, *4*(1), Article 1. https://doi.org/10.1186/2041-1480-4-43

Deemer, P., Benefield, G., Larman, C., & Vodde, B. (2009). Información básica de SCRUM. *California: Scrum Training Institute*.

Delgado, B. H. (2022). *Desarrollo de una solución de Procesamiento de Lenguaje Natural en Inteligencia Artificial para la búsqueda semántica en repositorios académicos* [PhD Thesis]. Universidad Central de Venezuela.

Garlan, D. (2008). *Software architecture*.

Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., & Krallinger, M. (2019). Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 1-10.

Hidalgo Delgado, Y., & Rodríguez Puente, R. (2013). La web semántica: Una breve revisión. *Revista Cubana de Ciencias Informáticas*, *7*(1), Article 1.

Izaurralde, M. P. (2013). Caracterización de Especificación de Requerimientos en entornos Ágiles: Historias de Usuario. *Trabajo de especialidad, Febrero*.

Jolodkow, N., & Romani, T. (2021). *Procesamiento de lenguaje aplicado a herramientas de*

búsqueda de información.

Liu, H.-T., Ning, C.-G., Huang, D.-L., & Wang, L.-S. (2014). Vibrational Spectroscopy of the Dehydrogenated Uracil Radical by Autodetachment of Dipole-Bound Excited States of Cold Anions. *Angewandte Chemie International Edition*, *53*(9), 2464-2468.

López-Úbeda, P., Díaz Galiano, M. C., Montejo Ráez, A., Martínez Santiago, F., Andreu-Marín, A., Martín Valdivia, M. T., & Ureña López, L. A. (2018). *Buscador Semántico Biomédico*. https://doi.org/10.26342/2018-61-29

Mäkelä, E. (2005). Survey of semantic search research. *Proceedings of the seminar on knowledge management on the semantic web*.

Mayil, V. V., & Jeyalakshmi, T. R. (2023). Pretrained Sentence Embedding and Semantic Sentence Similarity Language Model for Text Classification in NLP. *2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP)*, 1-5.

Moreno, M. C. (2013). *Dificultades lingüísticas del español para los estudiantes sinohablantes y búsqueda de soluciones motivadoras*.

Ortega, J. M. P., Valdivia, M. T. M., Ráez, A. M., & Galiano, M. C. D. (2008). Categorización de textos biomédicos usando UMLS. *Procesamiento del lenguaje Natural*, *40*, Article 40.

Pérez Josende, A. (2022). *Machine learning basado en modelos transformer aplicado a la traducción automática*. Universitat Politècnica de Catalunya.

Pgvector. (2023). [C]. pgvector. https://github.com/pgvector/pgvector (Obra original publicada en 2021)

Pires, D. F., & Ruiz, E. E. S. (2010). Interoperabilidade terminológica em sistemas de informação em saúde: Problemas e soluções com a UMLS. *Journal of Health Informatics*, *2*(2), Article 2.

Sharnagat, R. (2014). Named entity recognition: A literature survey. *Center For Indian Language Technology*, 1-27.

Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F. B., & Wattenberg, M. (2016).

Embedding Projector: Interactive Visualization and Interpretation of Embeddings (arXiv:1611.05469; Número arXiv:1611.05469). arXiv. http://arxiv.org/abs/1611.05469

Trigás Gallego, M. (2012). *Metodologia scrum*.

Villamizar Suaza, K., Tabares García, J. J., & Zapata Jaramillo, C. M. (2015). Mejora de historias de usuario y casos de prueba de metodologías ágiles con base en TDD. *Cuaderno activa*, *7*, Article 7.

Wei, W., Barnaghi, P. M., & Bargiela, A. (2008). Search with meanings: An overview of semantic search systems. *International journal of Communications of SIWN*, *3*, 76-82.

Anexos