DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE INGENIERÍA DE SOFTWARE

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SOFTWARE

TEMA:

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

AUTORES:

LLANO CHINCHERO, CHRISTOPHER FABRICIO Y
RUGEL TIAGUARO, CRISTOPHER ALEXIS

DIRECTOR:
ING. UYAGUARI UYAGUARI, ALVARO DANILO

LATACUNGA AGOSTO, 2023

Orden del día

- Introducción
- Marco
- Implementación del Sistema
- Validación
- Conclusiones

Orden del día

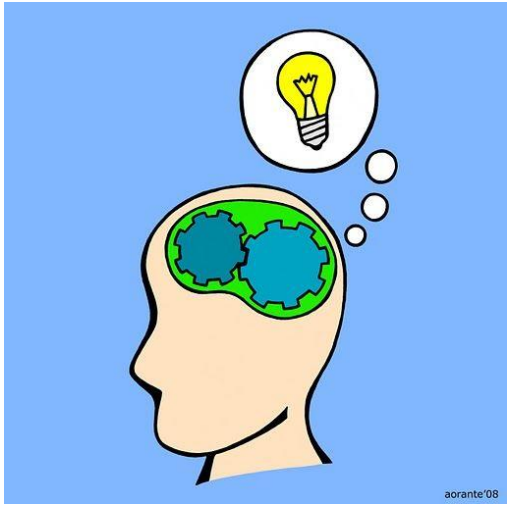
- Introducción
- Marco
- Implementación del Sistema
- Validación
- Conclusiones

Problema

- 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).

Planteamiento de la solución

- 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).
- Se decidió llevar a cabo una investigación para crear un sistema basado en la web para la búsqueda semántica e identificar entidades biomédicas.



Objetivo General

Desarrollar un sistema informático para el reconocimiento y normalización de entidades biomédicas basado en reglas heurísticas y búsquedas semánticas en lenguaje 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.

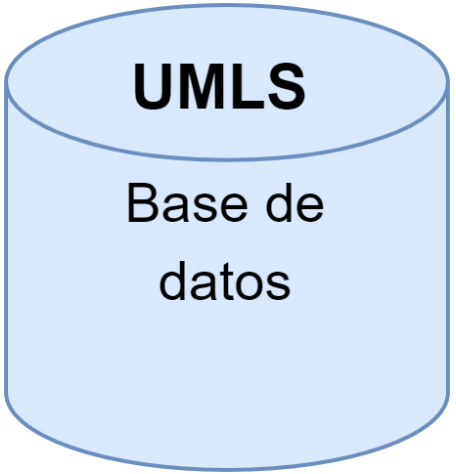
Adaptar prácticas eficientes para implementar sistemas con arquitecturas actuales.

Orden del día

- Introducción
- Marco
- Implementación del Sistema
- Validación
- Conclusiones

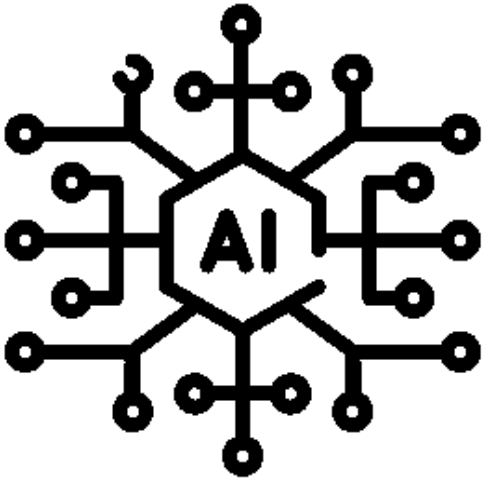
UMLS

- 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).



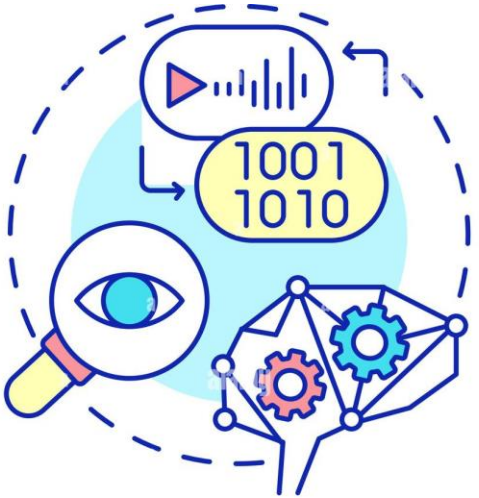
Reconocimiento de entidades biomédicas con modelos NER

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).



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).



Herramientas para la realización de una búsqueda semántica con embeddings

- ***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).

- ***Sentence transformers***

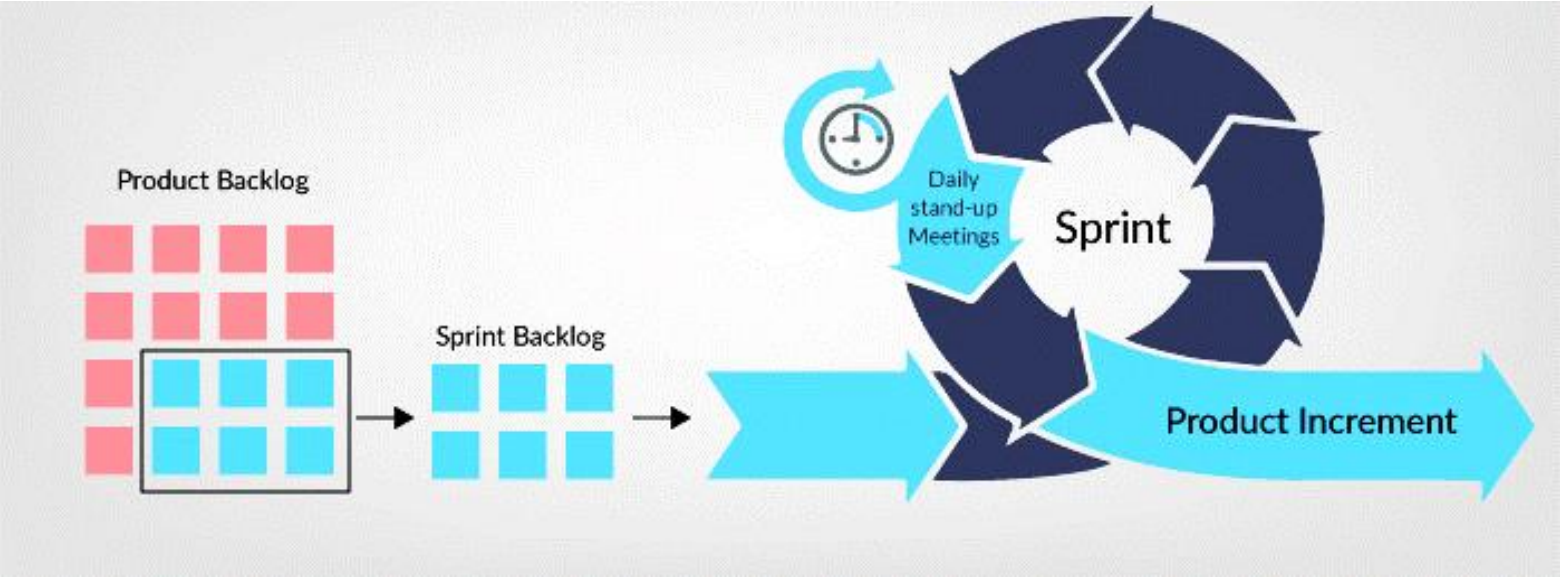
Determina la similitud semántica de las oraciones del texto. El modelado de pares de oraciones, la similitud del texto y el modelado del lenguaje son algunas tareas importantes en la NLP (Mayil & Jeyalakshmi, 2023).

Orden del día

- Introducción
- Marco
- Implementación del Sistema
- Validación
- Conclusiones

Metodología de desarrollo

- **Esquema de la metodología Scrum**



Recuperado de (Abdrakhmanov, 2018)

Análisis del sistema

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

Historia de usuario 1
Como programador
Quiero desarrollar un algoritmo que pueda identificar y extraer entidades biomédicas de un texto.
Para el etiquetado automático de entidades biomédicas.

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

Historia de usuario 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 extraer el concepto al que pertenece la entidad.

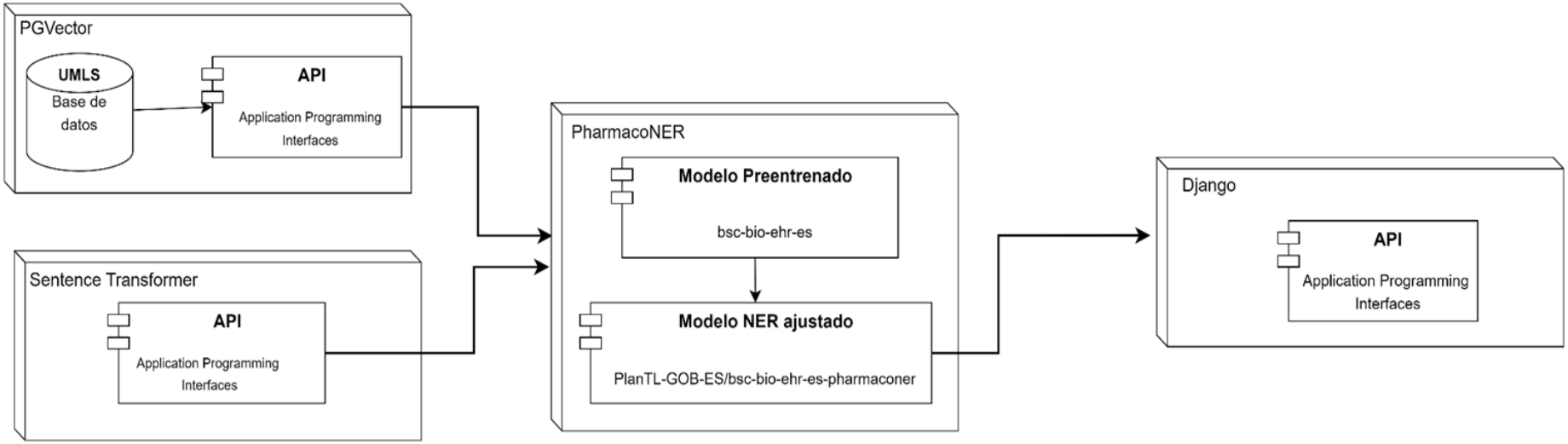
Análisis del sistema

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

Historia de usuario 3
Como programador
Quiero 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.

Diseño del sistema

Arquitectura

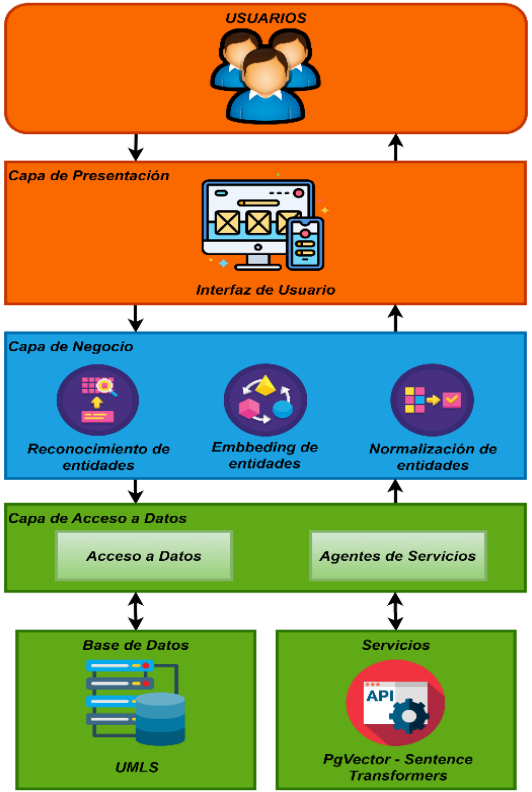


Elaboración propia

Diseño del sistema

Arquitectura funcional del sistema

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.



Elaboración propia

Diseño del sistema

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

Elaboración propia

Orden del día

- Introducción
- Marco
- Implementación del Sistema
- Validación
- Conclusiones

Validación del Sistema

Ejemplo 1 del resultado del uso del sistema en el corpus EMEA

Texto ingresado	Entidad biomédica	CUI	Reconocimiento de entidad biomédica	Normalización de entidad biomédica
AZOPT es una sulfonamida inhibidora de la anhidrasa carbónica que, aunque se administra vía oftálmica, se absorbe a nivel sistémico.	anhidrasa carbónica	C0007028	1	1
	AZOPT	C0673967	1	0
	sulfonamida	C0038760	1	1

Validación del Sistema

Ejemplo 2 del resultado del uso del 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 Un parche libera 4 mg de rotigotina cada 24 horas.	Neupro	C1949346	1	1
	rotigotina	C1700683	1	1

Validación del Sistema

Validación del PharmaCoNER

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

Validación del Sistema

Validación de la normalización

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

Análisis de resultados

- Resultados obtenidos 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.

- Resultados obtenidos 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.

Orden del día

- Introducción
- Marco
- Implementación del Sistema
- Validación
- Conclusiones

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 algoritmo implementado para el reconocimiento y normalización de entidades biomédicas, obtuvo un resultado aceptable, durante la evaluación del Corpus EMEA.

Conclusiones

- El uso del NER permitió generar un porcentaje estable de reconocimiento de entidades correctas, lo cual ayudó en el proceso de la normalización.

- 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.

Bibliografía

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

- 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 ….

Bibliografía

- 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

Bibliografía

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

- 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.

Bibliografía

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

- 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.

Software
DOMINA EL MUNDO

Gracias por su atención

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA