Desarrollo de un prototipo funcional de etiquetado de informes médicos con una terminología clínica reconocida mediante técnicas de PLN

Universidad Internacional de Valencia

Marzo 2023

Titulación:

Máster de Big Data y Ciencia de Datos

Curso académico

2022 – 2023

Alumno/a:

Medina Fernández, Fernando

D.N.I:

Director/a de TFM: Igual Pérez, Román

Convocatoria:

Primera

De:
Planeta Formación y Universidades

Índice

Índice de ilustraciones

Índice de tablas

Resumen

En España la mayor parte de los informes radiológicos tienen una estructura definida, pero se confeccionan en texto libre no estructurado. Esto hace que sean imprecisos, dificulta su búsqueda posterior y que la codificación de los juicios clínicos que la legislación española va a exigir en un breve plazo tenga que hacerse manualmente a partir de catálogo.

El procesamiento del lenguaje natural (PLN) ha evolucionado mucho de forma paralela a la IA. Empezó con algoritmos que implementaban reglas lingüísticas, posteriormente se introdujo algoritmos de aprendizaje automático para resolver problemas de clasificación a partir de los datos, más tarde se crearon las técnicas de aprendizaje profundo como las redes neuronales densas prealimentadas recurrentes y, en la actualidad, se utiliza las arquitecturas de *Transformers* como *BERT* y los métodos generativos *GPT*. Las aplicaciones actuales más relevantes donde se utilizan estas técnicas son: la traducción automática de textos entre idiomas, la respuesta ante preguntas, los sistemas de diálogo, el análisis de sentimiento, la clasificación de tópicos y la generación de contenido de texto.

Con el fin de resolver las dificultades mencionadas previamente se propone realizar un prototipo funcional que permita la búsqueda y la clasificación de informes radiológicos a partir de un método de vinculación automática (*semantic annotation* o *named entity linking*), así como la recuperación posterior de los informes médicos a partir de una terminología médica reconocida, utilizando técnicas de PLN. Esto requiere en primer lugar analizar el contexto general del PLN, los informes médicos y las terminologías clínicas para elegir una aproximación, datos disponibles y las técnicas más adecuadas. En segundo lugar, se debe crear un motor de inferencia que permita experimentar y validar un método de vinculación automática y de recuperación posterior de los informes médicos. Finalmente, se presenta un interfaz de usuario del prototipo funcional elaborado por el autor que sirva de entorno demostrativo.

El prototipo funcional desarrollado permite realmente la vinculación de los términos clínicos de un vocabulario definido y la localización posterior de los informes ajustándose a los requerimientos planteados. Se ha planteado el uso de la similitud semántica mediante un heurístico definido para vincular los términos con el texto, que es útil y que presenta resultados válidos. Sin embargo, esto no es suficiente, puesto que también se vinculan términos incorrectos dada la ambigüedad inherente a las palabras y grupos de palabras fuera de su contexto. Además, el heurístico utilizado para procesar los documentos no evita que existan términos que debieran vincularse y no lo hacen. Como trabajo futuro se propone ampliar la técnica utilizada con modelos que tengan en cuenta el contexto del documento utilizando los mecanismos de atención y/o el entrenamiento de modelos de clasificación multietiqueta con los datos disponibles.

Abstract

In Spain, most radiological reports have a defined structure, but they are written in unstructured free text. This makes them imprecise, it makes difficult to search for them later and it means that the coding of clinical judgements, which Spanish regulatory will soon require, will have to be done manually from a catalogue.

Natural language processing (NLP) has evolved significantly in parallel to AI. It started with algorithms implementing linguistic rules, then machine learning algorithms were introduced to solve classification problems from data, followed by the use of deep learning techniques such as feed forward dense neural networks, recurrent and nowadays Transformer architectures such as BERT and generative GPT methods are increasingly being used. The most relevant applications where these methods are used are: machine translation of texts between languages, question answering, dialogue systems, sentiment analysis, topic classification and text content generation.

In order to solve difficulties mentioned above, we are going to develop a NLP software prototype that implements named entity linking over medical records using a proposed medical terminology. This requires firstly analysing the general context of PLN, medical clinical reports and clinical terminologies in order to choose an approach, available data and the most appropriate techniques. Secondly, an inference engine must be created to experiment and validate a method of automatic linking and subsequent retrieval of medical reports. Finally, a user interface of the functional prototype is created as a demostrative environment.

The developed functional prototype actually allows linking of clinical terms from a defined vocabulary and subsequent localisation of the reports according to the requirements. The use of semantic similarity using a defined heuristic to link the terms to the text has been proposed, which is useful and shows valid results. However, this is not enough, as incorrect terms are also linked given the inherent ambiguity of words and word groups out of context. In addition, heuristic used to process the documents does not prevent omitting terms that should be linked. As future work, it is proposed to extend technique used with models that take into account the context of the document using attention mechanisms and/or training multi-label classification models with the available data.

Keywords

- Natural language processing (NLP), named entity linking (NEL), semantic annotation inference engine, SNOMED-CT entity linking of Spanish clinical reports.

1. Introducción

La mayor parte de los informes médicos en el ámbito del radiodiagnóstico en España tienen un grado de madurez bajo y están definidos con un conjunto de apartados en texto libre sin codificar que los hace comprensibles, pero en muchas ocasiones imprecisos. Uno de los inconvenientes de no estar clasificados por códigos es que su búsqueda y localización posterior requiere una búsqueda por comparación de textos muy intensiva en el uso de recursos y donde se pierden resultados deseados. Adicionalmente, los informes médicos suelen estar cifrados en su almacén para impedir accesos no deseados. Dicha búsqueda tiene mayor latencia y complejidad(Martí-Bonmatí et al., 2022).

Otro inconveniente es que la ley española que regula el contenido y cuerpo de los informes médicos está en proceso de cambio y va a requerir que los diagnósticos o juicios clínicos estén codificados con una terminología clínica concreta (SNOMED-CT o CIE) con el objetivo de fomentar la interoperabilidad semántica entre aplicaciones sanitarias. Esto requerirá que los médicos busquen en el catálogo la codificación adecuada del juicio clínico manualmente y sin ayuda, lo que ralentizará la redacción de los informes (*Proyecto de Real Decreto Por El Que Se Modifica El Real Decreto 1093/2010, de 3 de Septiembre, Por El Que Se Aprueba El Conjunto Mínimo de Datos de Los Informes Clínicos En El Sistema Nacional de Salud*, n.d.). En este contexto las técnicas de procesamiento de lenguaje natural (PLN) son clave para clasificar automáticamente informes médicos.

La resolución de este problema en el ámbito médico español ya ha sido tratado en el pasado con diferentes enfoques y modelos como, por ejemplo, en las tareas conjuntas de codificación de casos clínicos CodiEsp (Miranda-Escalada et al., 2022). En el presente trabajo se va a tener en cuenta estos enfoques y otros disponibles para diseñar un heurístico y experimentar con un modelo preentrenado disponible de forma que se satisfaga los requerimientos de partida.

El objetivo del presente trabajo es el desarrollo de un prototipo para la experimentación de un sistema de etiquetado o vinculación automática de términos clínicos a los informes médicos del ámbito del radiodiagnóstico. Para ello, utilizaremos técnicas de PLN. Dicho prototipo debe contener un motor de inferencia que pueda ser utilizado en el futuro desde un Sistema de Información Radiológico (RIS) y/o desde otras aplicaciones clínicas, además del interfaz de usuario necesario para realizar el etiquetado y la búsqueda posterior de los informes médicos etiquetados. La investigación de este trabajo se centrará en las siguientes preguntas: ¿qué es, evolución, y para qué se puede aplicar el PLN? ¿cómo es la estructura de un informe médico y cuál es la terminología clínica más adecuada en España? ¿cuáles son las técnicas de PLN actuales más avanzadas? ¿qué modelos preentrenados y datos disponibles de PLN existen en el ámbito clínico en español? ¿qué heurístico se debe seguir para vincular los términos clínicos al informe? ¿cómo desarrollar el prototipo funcional de vinculación y localización de informes médicos?

Para responder a estas preguntas se trabajará revisando información disponible en línea, como páginas web, imágenes y videos. Inicialmente se consultará recursos más generales profundizando posteriormente, con fuentes de expertos en la materia, fabricantes/proveedores del ámbito, publicaciones en universidades, así como revistas y artículos de investigación específicos. Con la información disponible se probará y validará inicialmente los algoritmos, modelos y heurísticos necesarios para llevar a cabo la vinculación de entidades (NEL) y, posteriormente se construirá el prototipo funcional con el motor de inferencia junto a su interfaz de usuario correspondiente. Además, se validará el uso del prototipo con un conjunto disponible de datos de informes médicos y términos clínicos enlazados.

El autor de este trabajo dirige el proyecto desarrollo de un sistema de información radiológico RIS que contiene informes de pruebas de imagen reales. El RIS es propiedad de la empresa BABEL Sistemas de Información (*Sistema de Información Radiológico EOS Del Servicio de Salud Del Principado de Asturias - BABEL Sistemas de Información*, n.d.) y está actualmente implantado en 13 hospitales españoles del ámbito del Servicio de Salud del Principado de Asturias (SESPA).

El detonante del presente TFM ha sido un conjunto de hechos relevantes: el conocimiento y la experiencia del autor en el ámbito de los informes médicos de radiodiagnóstico, la posibilidad de influir en la hoja de ruta del desarrollo del RIS mencionado, la disponibilidad de un cliente final donde implantar la solución y la detección de una necesidad real a solventar con interés clínico.

Los apartados de la memoria se enumeran a continuación:

- Se establece un apartado de **objetivos generales y específicos** para conseguir el resultado deseado, así como un apartado con el **glosario** de la memoria para facilitar la comprensión de los términos utilizados.
- Se incluye un apartado de **conceptos generales** del ámbito, como son el campo del PLN, los informes médicos de prueba de imagen en España, y las terminologías clínicas más utilizadas.
- Se incluye un apartado específico del **estado del arte y del marco teórico** del PLN, presentando las redes neuronales artificiales, las técnicas de PLN generales y las técnicas, iniciativas y recursos de PLN en español médico.
- Se incluye un apartado de **desarrollo del proyecto y resultados**, donde se presenta la metodología, el planteamiento del problema, el desarrollo del proyecto y los resultados.
- Se incluye un apartado de **conclusión y trabajos futuros**, en el que se exponen las conclusiones sobre la investigación del estado del arte, el desarrollo realizado, las dificultades encontradas y los trabajos futuros propuestos y previstos.
- Finalmente, se incluyen los apartados de **referencias** bibliográficas consultadas para la redacción de la memoria y desarrollo del trabajo, **agradecimientos**, **apéndices** con información sobre los artefactos de software y los resultados de validación generados.

2. Objetivos

El **objetivo general** que se persigue mediante la vinculación automática de términos clínicos a informes es agilizar el proceso de diagnóstico de la imagen médica facilitando la recuperación de los informes a través de dichos términos. Los especialistas que basen el diagnóstico de una imagen radiológica en la revisión comparativa de los casos previos lo harán de forma más ágil. Para ello no tendrán la necesidad de recordar de memoria qué casos son similares, ni de realizar una búsqueda documental en bruto sobre el contenido de los informes previos. Una vez que los informes estén etiquetados se podrán localizar a partir de dichas etiquetas.

Adicionalmente, en el futuro y cuando lo exija la legislación, esta vinculación automática permitirá reducir el tiempo de redacción del informe médico al vincular los diagnósticos a partir del texto libre introducido, evitando la búsqueda manual de juicios clínicos a partir de catálogos.

Para ello se plantea desarrollar un prototipo funcional con un método experimental para etiquetar los informes médicos con las enfermedades que se mencionan utilizando una terminología médica reconocida mediante técnicas PLN de aprendizaje profundo y además permitir su identificación y búsqueda posterior a partir de dichas etiquetas.

La solución debe etiquetar los informes que se vayan generando progresivamente en base a una terminología médica estándar para que dichas etiquetas puedan ser utilizadas en la selección y búsqueda de casos comparativos previos.

Para la consecución del objetivo principal, son necesarios cumplir los siguientes objetivos específicos:

- Llevar a cabo un análisis de la evolución y del estado del arte del PLN, de las técnicas utilizadas y de sus aplicaciones, para situar el contexto general de realización del prototipo.
- Analizar el formato y el contenido de los informes de pruebas de imagen en España y localizar informes de muestra en español para validar el resultado.
- Analizar las terminologías clínicas estándar más utilizadas para determinar la que se debe utilizar en este caso de uso y localizar el catálogo disponible en español.
- Realizar una revisión del estado del arte del PLN para el ámbito clínico y/o biomédico en español y localizar modelos de aprendizaje automático preentrenados disponibles.
- Llevar a cabo un análisis de las utilidades, librerías y herramientas disponibles de PLN para la implementación del caso de uso.

- Construcción de un motor de inferencia para experimentación con el caso de uso utilizando los modelos de aprendizaje automático, el catálogo y las herramientas seleccionadas.
- Construcción de un interfaz de usuario amigable que resuelva el caso de uso a través del motor de inferencia mencionado.
- Desarrollo de la presente memoria que recoja todo el trabajo realizado.

3. Glosario

Se mencionan los siguientes conceptos (López Rubio, 2019d) (Donnelly et al., 2022):

Biomarcadores de la imagen: Características medibles extraídas de las imágenes médicas que indican un proceso biológico normal, una enfermedad o una respuesta a una intervención terapéutica.

***Byte-pair Encoding* (BPE)**: Algoritmo de segmentación en subpalabras que codifica las palabras en secuencias de unidades específicas (letras o sílabas). El algoritmo combina cada secuencia diferente como un nuevo símbolo, utilizando símbolos más cortos para las secuencias más comunes, con lo que el texto se comprime. Esto permite obtener una representación con mayor significado de las palabras en los procesos de PLN, especialmente para palabras desconocidas, dado que cada nueva palabra se codificará con secuencias conocidas.

Corpus/Corpora: Conjunto de datos de textos de lenguaje natural acompañados con metadatos.

Etiquetado de *tokens* (clase de palabra): Consiste en asignar una etiqueta concreta o clase a cada *token*/palabra, por ejemplo, tipo de palabra, función gramatical o entidad. Esto permite comprobar si una secuencia de *tokens* tiene sentido gramaticalmente en función del orden en el que aparecen. Esta operación se realiza de forma automática, aunque puede haber *tokens* con etiquetado ambiguo.

Instancia: Texto en lenguaje natural junto con sus metadatos asociados que lo identifican.

Interoperabilidad semántica: Capacidad de sistemas de información heterogéneos de compartir información comprensible y con el mismo significado inequívoco.

Lematización (*lemmatization*): Conversión de palabras a formas lingüísticas base, o lemas. Facilita el PLN al disponer de la raíz de la palabra independientemente de la variación gramatical.

Lexicón: Tipo de diccionario que define la categoría y la variabilidad léxica de las palabras. Establece el conjunto de reglas por el que se construyen y modifican las palabras.

Lingüística computacional: La lingüística computacional es un campo interdisciplinario que se ocupa del desarrollo formal del funcionamiento del lenguaje natural, desarrollos que pueden ser transformados en programas ejecutables para un ordenador.

Metatesauro: Base de datos de términos con varios usos multilingüe que contiene información de conceptos, sus nombres y las relaciones entre sí.

N-grama: Secuencia de N *tokens* que aparecen consecutivamente. También denominados: Unigrama, bigrama, trigrama, tetragrama en función del número de *tokens*.

Obtención de palabras troncales/raíz (*stemming*): Simplificación de palabras que provienen de una raíz común con un significado o valor similar a un grupo de palabras (por ejemplo, - "encontrar, hallar, hallado"). Facilita el PLN al trabajar con la raíz independientemente de la variación gramatical.

Palabras léxicas (*content words*): Palabras que tienen contenido semántico, como por ejemplo los nombres, verbos, adjetivos y adverbios.

Palabras gramaticales (*stopwords*): Palabras que no tienen un contenido semántico, como pueden ser los pronombres, artículos, preposiciones y conjunciones, aparecen comúnmente en los textos creando ruido al tratar de obtener el significado.

Partes de la oración (POS): Clases de palabras de una lengua (sustantivo, verbo, adjetivo, adverbio, pronombre, preposición, conjunción, artículo, interjección y signos de puntuación).

Radiómica: La radiómica es una ciencia “ómica” que extrae, por medio de algoritmos computacionales, parámetros cuantitativos en las imágenes médicas para detectar y medir aquellas características inapreciables a la observación directa, llamadas características radiómicas, con el objetivo de asociarlas a estados fisiológicos concretos. Constituye una fuente de información muy relevante de cara a profundizar en el conocimiento de la diversidad biológica y funcional de los tejidos, la heterogeneidad de los fenómenos patológicos y sobre la evolución previsible de las enfermedades.

Sistema de Información Radiológico (*RIS*): Aplicación informática que gestiona el proceso y flujos de trabajo del departamento o servicio de Radiodiagnóstico para gestionar la actividad de realización, informado y difusión de las pruebas de imagen médica y de los informes médicos asociados.

Test de Alan Turing: Test diseñado en el año 1950 por el matemático inglés Alan Turing para determinar si una máquina era inteligente. La prueba consistía en establecer una comunicación escrita entre una persona examinadora por un lado y por el otro una computadora examinada y otra persona. La persona examinadora hacía preguntas que eran respondidas tanto por la computadora como por la otra persona. La computadora pasaba dicho Test si no se podía identificar si la respuesta dada era humana o no.

Texto plano: Conjunto de caracteres que componen un texto, normalmente en formato ASCII o UNICODE, aunque existen otros.

Tokens: Agrupación de caracteres alfanuméricos, dependientes del idioma. Es español o inglés son palabras o secuencias numéricas separados por espacios en blanco o por signos de puntuación.

Tokenización: Proceso de descomposición de textos en unidades discretas o *tokens*, para facilitar su análisis matemático posterior.

Tipos: Tokens únicos de un corpus específico.

Vectorización de palabras/textos: Técnica consistente en convertir las palabras/textos en vectores numéricos con números interpretables por los algoritmos de PLN que representan el significado de un *token* o unidad de texto.

Vocabulario o léxico: Conjunto de todos los tipos de un lenguaje.

4. Conceptos Generales

En los siguientes apartados se recoge un resumen de la información más relevante para contextualizar el presente trabajo con respecto a identificar qué es el PLN, cómo ha evolucionado a lo largo de la historia, qué aplicaciones más relevantes tiene, cómo son los informes médicos de pruebas de imagen y qué terminologías clínicas se utilizan habitualmente.

4.1. Introducción al Procesamiento del Lenguaje Natural (PLN)

El lenguaje natural es uno de los aspectos fundamentales del ser humano y sirve como vehículo de comunicación de forma cotidiana para trasladar información perdurable entre personas. Es una cuestión de interés entre los lingüistas, filósofos y también entre los ingenieros (Palomar, 2006).

El lenguaje natural escrito expresa un significado de forma discreta y simbólica, significado que se corresponde con la señal hablada o con la comunicación mediante gestos y en última instancia con el procesamiento de la comunicación continua en el cerebro humano.

El PLN es una de las técnicas analíticas que se enfocan en la interpretación y el tratamiento del lenguaje humano. Es una rama de la inteligencia artificial que utiliza los computadores y las técnicas computacionales para reconocer, comprender y generar las lenguas humanas. En el PLN se integran la lingüística computacional, las ciencias de la computación y las ciencias cognitivas.

El PLN intenta modelar matemáticamente los mecanismos cognitivos de comprensión y producción de las lenguas humanas para la interacción entre ordenadores y seres humanos. A un nivel muy cercano a la máquina, el PLN consiste en la manipulación del texto libre para convertirlo en datos numéricos estructurados que permitan crear información útil. Gracias al auge y la evolución de la tecnología el PLN está avanzando con gran celeridad (Palomar, 2006).

Las aplicaciones principales del PLN son: el reconocimiento del habla, la comprensión del lenguaje, la construcción de sistemas de diálogo, el análisis léxico y sintáctico del texto, la traducción automática, la gestión estructurada del conocimiento, la recuperación de información, la obtención de respuestas a preguntas, el análisis de sentimientos y la generación de lenguaje natural (López Rubio, 2019).

Existen dos categorías principales en lo que respecta a la aproximación para extraer información mediante PLN: la primera que se utilizó históricamente es la simbólica o basada en reglas que imitan mediante algoritmos las reglas diseñadas por el ser humano, y la segunda, con más proyección futura y uso actual es la que hace uso de técnicas estadísticas o de aprendizaje automático que se basan en la evaluación de un

conjunto de datos de entrenamiento evaluados y anotados por expertos previamente (Donnelly et al., 2022)

Parte del problema y de la complejidad del PLN tanto en la comprensión como en la generación del lenguaje es que la comunicación escrita es discreta mientras que la comunicación en el cerebro humano es continua, lo que lleva a la ambigüedad y a la necesidad de un contexto del lenguaje para la correcta interpretación del mensaje.

Existen en la actualidad diversas aproximaciones tecnológicas al PLN, avances recientes en estas técnicas y muchas aplicaciones. Los procesos de aprendizaje automático del PLN han evolucionado desde el proceso de vectorización dispersa *Bag-of-Words*, el uso de modelos de vectorización densa *Word2Vec*, los modelos de redes neuronales recurrentes *LSTM,* el marco de trabajo de transformadores como BERT o los modelos generativos como *GPT* en la actualidad (López Rubio, 2019b).

4.2. Historia y evolución del Procesamiento del Lenguaje Natural (PLN)

El PLN surgió de forma paralela a la Inteligencia Artificial (IA) en la década de los años cincuenta para simular y reproducir la interacción entre las máquinas y los humanos. En esa época se diseñó una prueba para determinar si una máquina era inteligente (Test de Alan Turing) (González, 2007).

Se pueden distinguir tres etapas en el ámbito del PLN (López Rubio, 2019c):

Etapa racionalista 1950-1989

En esa etapa los racionalistas como Noam Chomsky (*Biografia de Noam Chomsky*, n.d.), suponían que el lenguaje estaba prefijado por la herencia genética de los seres humanos.

Siguiendo este razonamiento los modelos de interpretación del lenguaje se basaban en la estructura sintáctica y semántica de este. Los sistemas de PLN integraban el conocimiento y razonamiento de los seres humanos mediante reglas sintácticas y lógicas diseñadas manualmente.

Esta época coincidió con el desarrollo de sistemas expertos en IA que utilizaban reglas simbólicas lógicas fáciles de interpretar. Las construcciones gramaticales y lógicas de reglas eran sencillas de depurar pero también muy regulares y estrictas. En la práctica estos sistemas no eran útiles porque el lenguaje natural humano es irregular y flexible.

Como ejemplos de aquella época en el año 1954 podemos nombrar el experimento Georgetown-IBM (*IBM Archives: 701 Translator*, 1954) que consistía en un sistema de traducción automática de ruso a inglés. Sin embargo, según el informe publicado en el año 1966 para evaluar el estado del arte del PLN denominado *Automatic Language*

Processing Advisory Committee Report ALPAC 166 (Joseph, 2013), la gran cantidad de esfuerzos que se invirtieron en estas actividades no dio el resultado ni los avances significativos esperados y las expectativas eran poco realistas.

En esta época también se desarrolló en el MIT el programa de PLN denominado ELIZA (1964-1966) (*Así Era ELIZA, El Primer Bot Conversacional de La Historia*, n.d.), un sistema de diálogo interactivo que simulaba la respuesta de un psiquiatra en una consulta. La realidad es que no había entendimiento real del lenguaje porque sus respuestas se basaban en frases preconstruidas en base al análisis de patrones de palabras clave identificadas de las propias preguntas.

Etapa empirista 1990-2008

En esta etapa los empiristas como Zellig Harris (*Biografia de Zellig Harris*, n.d.), partían de la suposición de que, aunque la mente humana dispone de habilidades genéricas para reconocer patrones y generalizarlos el conocimiento no es innato y para el aprendizaje del lenguaje se necesita mucho volumen de información de entrada.

Siguiendo este razonamiento se diseñaron modelos probabilísticos y estadísticos basados en asociaciones, reconocimiento de patrones y empezaba a utilizase el aprendizaje automático (*Machine Learning* o ML) haciendo énfasis en los datos en lugar de en los algoritmos.

En esta etapa estaba en auge en IA la estrategia de énfasis en los datos mencionada, como el reconocimiento de patrones y la visión por computador. Con este nuevo enfoque se parte de habilidades importantes como el aprendizaje y la percepción de patrones lo que hace que los sistemas sean flexibles, generalicen mejor y den mejores resultados.

En el ML no se diseñan reglas, sino que se utilizan modelos estadísticos o redes neuronales artificiales sencillas para aprender los parámetros del sistema y detectar patrones a partir de muchos datos de entrenamiento. Esto facilita la generalización entre situaciones y dominios diferentes al gestionar la incertidumbre y admitir flexibilidad en los errores.

Como ejemplo, el ML se puede aplicar en la traducción automática entre dos idiomas. Se puede entrenar un modelo de aprendizaje automático a partir de documentos bilingües, en los que cada documento se encuentra dividido en frases en ambos idiomas. Al no utilizar reglas gramaticales específicas y estar basada en los datos, esta técnica descarta la estructura gramatical de los idiomas y proporciona mejores resultados que otros enfoques previos.

En el año 2005 comenzó el desarrollo el programa Watson de IBM (*IBM Watson*, n.d.) que debutó en el año 2011. Watson era capaz de contestar a preguntas específicas, para ello trabajaba con un volumen masivo de datos de enciclopedias, diccionarios, tesauros y obras literarias, y mediante el uso de cientos de algoritmos de análisis

lingüístico y de aprendizaje automático encontraba correlaciones en las preguntas y predecía las respuestas.

Etapa del aprendizaje profundo 2009-actualidad

Hasta este momento los modelos de aprendizaje automático no podían procesar los datos masivos dado que no había sistemas informáticos con la capacidad suficiente para ello.

En el año 2009 surgió en IA el aprendizaje profundo basado en redes neuronales densas (*Deep learning*). El aprendizaje tradicional exigía experiencia humana y no tenía suficiente capacidad para detectar relaciones complejas. En el aprendizaje profundo sin embargo existen varias capas densas de procesamiento sucesivas que se entrenan mediante algoritmos específicos hasta el punto de inflexión del rendimiento.

El aprendizaje profundo permite extraer automáticamente rasgos característicos de los datos sin la necesidad de un humano experto. Se delega a las primeras capas detectar rasgos de bajo nivel (mayor concreción) y a las últimas capas los rasgos de mayor nivel (mayor abstracción). Con este enfoque las redes neuronales aprenden una jerarquía de conceptos, de menor a mayor nivel de abstracción.

En el año 2010 se utilizó el PLN de forma industrial en el reconocimiento del habla con redes neuronales de aprendizaje profundo o densas que eran capaces de aislar el contenido del habla del contexto vocal. Algunos ejemplos son, los asistentes virtuales como Apple Siri en 2011 (*10 Years of Siri: The History of Apple's Voice Assistant | TechRadar*, n.d.), Amazon Alexa en 2013 (*Amazon Gets Into Voice Recognition, Buys Ivona Software To Compete Against Apple's Siri | TechCrunch*, n.d.) y Microsoft Cortana en 2014 (*What Is Microsoft Cortana? Everything You Need to Know*, n.d.).

4.3. Aplicaciones de Procesamiento de Lenguaje Natural (PLN).

A continuación, se recogen a grandes rasgos las aplicaciones del PLN más relevantes y las técnicas utilizadas para resolver cada una de las tareas mencionadas.

Traducción automática

La aplicación original del PLN ha sido la traducción automática de un texto de una lengua a otra. Para llevar a cabo esta tarea se puede plantear la traducción de dos formas, o bien basada en componentes (por palabras o grupo de palabras) o bien mediante la traducción de frases/sentencias completas del texto.

La traducción se realiza con algoritmos de aprendizaje automático supervisados que aprenden a realizar las traducciones a partir de los textos en los idiomas a tratar junto

con su traducción correspondiente (etiquetas). Se puede llevar a cabo con los algoritmos de clasificación clásicos (*SVM*, *kNN*, etc.) o como en la actualidad mediante técnicas de aprendizaje profundo más avanzadas (transformadores basados en *BERT*).

Contestación a preguntas

En esta aplicación de PLN el sistema es capaz de comprender el significado de una pregunta, extraer de una base de datos de conocimiento los hechos relevantes relacionados con la pregunta y finalmente elegir la respuesta apropiada.

Para dar respuesta a esta aplicación es necesario disponer de una base de datos de conocimiento preconstruida que contenga los diferentes conceptos y su relación. La obtención de la respuesta se consigue obteniendo el significado de las preguntas, buscando los conceptos de significado similar en la base de datos para posteriormente acceder a las posibles respuestas relacionadas. Todo esto se consigue mediante técnicas de aprendizaje automático.

Sistemas de diálogo

Los sistemas de diálogo denominados asistentes virtuales (*chatbots*) son sistemas interactivos que mantienen una conversación con los usuarios con el fin de asistir en el desarrollo de una tarea.

Los *chatbots* reconocen patrones del texto de varias maneras, se pueden reconocer mediante reglas lingüísticas o mediante sistemas de obtención del significado entrenados previamente con aprendizaje automático o profundo. La generación del texto se puede hacer mediante reglas lógicas en la conversación, mediante sistemas de búsqueda entrenados previamente o con métodos generativos, que son capaces de mantener una conversación realista.

Búsqueda semántica de palabras y clasificación de documentos

Otra de las aplicaciones del PLN es la construcción de sistemas de búsqueda semántica de palabras y la búsqueda/clasificación de documentos según su contenido semántico.

Para la búsqueda semántica de palabras se puede utilizar aprendizaje profundo con redes neuronales que se entrenan para representar las palabras mediante una codificación específica que representa el significado.

Para la clasificación de documentos se obtiene inicialmente el significado de las palabras buscadas y se compara con el significado del resto de palabras de los documentos de búsqueda. Esto nos permite buscar textos similares en base a su similitud o diferencia semántica de las palabras que los componen.

Este concepto de codificar las palabras se puede ampliar a la representación de oraciones, párrafos o documentos completos para extraer el significado semántico de estos elementos, lo que permite la búsqueda y clasificación a niveles más amplios.

Clasificación de tópicos, análisis de sentimiento y generación textual

Otra aplicación es la clasificación y generación de documentos mediante aprendizaje profundo.

La clasificación de tópicos y análisis de sentimiento se puede hacer a partir de documentos clasificados o etiquetados previamente. El enfoque es entrenar una red neuronal para que clasifique documentos similares con el fin de detectar estilos o sentimientos expresados en los textos.

La generación de documentos se puede llevar a cabo aplicando modelos generativos que permiten generar un documento de estilo similar a los del conjunto de entrenamiento. Para ello se entrenan modelos de aprendizaje automático inicialmente para reconocer patrones y estructuras en el texto que posteriormente se ajustan para predecir palabras partiendo de un texto inicial dado, en base a la distribución de probabilidad de dichas palabras en el corpus.

Otra aplicación posible es el entrenamiento de modelos para generar una respuesta automática a un correo electrónico, para hacer una traducción a otro idioma, o para generar un resumen a partir de un documento.

4.4. Informes Médicos de Pruebas de Imagen

El informe médico de prueba de imagen es un documento final de texto que contiene la información necesaria y relevante relacionada con la interpretación del radiólogo de todos los hallazgos de una prueba de imagen médica. Es el mecanismo primario y oficial para la comunicación entre radiólogos y los médicos responsables del caso de cada paciente (Donnelly et al., 2022).

Los informes de prueba de imagen han evolucionado desde el texto libre hasta incluir información estructurada y codificada de protocolos, observaciones, datos y guías/ayudas. Esto minimiza la incertidumbre, el uso de un lenguaje evasivo y permite su explotación posterior mediante ciencia de datos e inteligencia de negocio. El objetivo de este informe es proporcionar un diagnóstico diferencial a partir de la interpretación de los hallazgos de la imagen que debe ser comunicado de forma precisa y efectiva (Martí-Bonmatí et al., 2022).

Los informes han ido evolucionando a lo largo del tiempo en su redacción en cuatro fases o grados de madurez.

- **Informe organizado**: informe que requiere los apartados de juicio clínico, técnica empleada, hallazgos y conclusiones. Estos informes mejoran la comprensibilidad del contenido con respecto a los informes sin estructura, pero al ser redactados en texto libre pueden contener imprecisiones, errores o confusiones.
- **Informe predefinido**: informe basado en plantillas de sociedades profesionales orientados a una técnica concreta en general (ecografía, resonancia magnética…). Estos informes mejoran la completitud y la consistencia, pero aún contienen texto libre no estructurado y por lo tanto imprecisiones.
- **Informe estructurado**: informe basado en preguntas/respuestas pactadas y/o en plantillas específicas en lo referente a una patología concreta que contiene información gráfica. Estos informes pueden ser explotables para cuadros de mando o de forma científica al contener información estructurada.
- **Informe estructurado cuantitativo**: informe estructurado en el que se incluye información cuantitativa sobre el resultado radiómico o de biomarcadores de la imagen. Estos informes mejoran la información objetiva sobre el grado de avance, de gravedad de las patologías o condiciones adversas (grado de expresión biológica de las enfermedades).

El informe estructurado en el ámbito de la imagen responde a la necesidad de un lenguaje uniforme, precisión en el diagnóstico, estandarización de términos, mejor comprensión e incorporación de datos clave y parámetros clínicos. El objetivo es poder realizar posteriormente minería de datos e inteligencia de negocio. Desafortunadamente no se ha avanzado al mismo nivel en todos los tipos de informe ni en todas las especialidades. La gran mayoría de los informes están actualmente en el nivel de informe organizado y es lo que plantea la necesidad del uso del PLN para su explotación posterior (Martí-Bonmatí et al., 2022).

En España, en la actualidad, los campos mínimos recomendables que debe contener un informe de pruebas de imagen en formato texto libre son (*BOE.Es - BOE-A-2010-14199 Real Decreto 1093/2010, de 3 de Septiembre, Por El Que Se Aprueba El Conjunto Mínimo de Datos de Los Informes Clínicos En El Sistema Nacional de Salud.*, n.d.):

- **Información Clínica** (datos clínicos que justifican la prueba y establecen la sospecha diagnóstica)
- **Descripción de la exploración** (prioridad, medios utilizados, reacciones, incidentes, limitaciones y otras exploraciones comparadas)
- **Hallazgos** (descripción detallada de qué se observa en la imagen)
- **Diagnóstico** (identificación de enfermedad, lesión o afección)
- **Recomendaciones** (cuidados, tratamientos y otras pruebas posteriores)

Actualmente esta estructura está en proceso de cambio para dar respuesta a la necesidad de interoperabilidad semántica, de normalización de la representación de los datos y de su vinculación terminológica con un nuevo proyecto de Real Decreto. En

esta propuesta de cambio ya se plantea la utilización de diagnósticos codificados de forma estructurada (SNOMED-CT/CIE). (*Proyecto de Real Decreto Por El Que Se Modifica El Real Decreto 1093/2010, de 3 de Septiembre, Por El Que Se Aprueba El Conjunto Mínimo de Datos de Los Informes Clínicos En El Sistema Nacional de Salud*, n.d.).

4.5. Terminologías Clínicas

Citando la documentación disponible en el Ministerio de Sanidad de España al respecto, "una terminología clínica es un conjunto de términos estructurados y normalizados que busca servir de instrumento para el registro de datos clínicos, como base para posibles investigaciones o como medio de intercambio de información clínica entre profesionales para la atención de la salud de los pacientes".

Una parte clave de la necesidad del uso de las terminologías clínicas es el objetivo de conseguir la interoperabilidad semántica para permitir que diferentes sistemas informáticos puedan intercambiarse automáticamente información comprensible manteniendo el mismo significado. Para ello, es necesario que los datos y su contexto, sean normalizados antes de ser intercambiados (*Ministerio de Sanidad - Profesionales - Preguntas Frecuentes Sobre SNOMED CT*, n.d.).

A continuación, se incluyen algunas de las terminologías clínicas analizadas y más reconocidas para el ámbito médico y especialmente para el ámbito del Radiodiagnóstico.

RadLex©

El conjunto de términos RadLex© ha sido desarrollado en Estados Unidos por la Sociedad Radiológica de Norte América (*RSNA*) (*RadLex Term Browser*, n.d.). Es un conjunto completo de términos radiológicos para su uso en informes de radiología, soporte a la toma de decisiones, minería y registro de datos, educación e investigación. Se proporcionan las bases de los datos utilizados, elementos comunes, plantillas de informes, y un sistema para nombrar los procedimientos en radiología.

Está disponible en inglés y en alemán actualmente.

Current Procedural Terminology (CPT)

La terminología CPT es una terminología propiedad de la Asociación Médica Americana (*AMA*) que contiene de forma codificada una lista de códigos y descripciones de términos utilizados por los profesionales sanitarios para la facturación de los servicios y procedimientos médicos en los programas de salud públicos y privados (*CPT - CPT Codes - Current Procedural Terminology - AAPC*, n.d.).

Está disponible en inglés y en español.

Descriptores en Ciencias de la Salud (DeCS)

DeCS es un vocabulario estructurado multilingüe desarrollado por el Centro Latinoamericano y Caribeño en Información de Ciencias de la Salud (*BIREME*) a partir de otra codificación *Medical Subject Headings (MeSH)* norteamericana. DeCS se ha creado como un lenguaje único en la indexación de artículos para la búsqueda de temas de literatura científica disponibles en la Biblioteca Virtual en Salud (BVS).

Los conceptos de DeCS tienen una estructura jerárquica de conceptos más amplios a más específicos, y está en idioma español (*Descriptores En Ciencias de La Salud - Wikipedia, La Enciclopedia Libre*, n.d.).

Clasificación internacional de enfermedades (ICD/CIE-10)

La clasificación CIE-10 es la 10ª versión de la clasificación *ICD*, en español. *ICD* determina la clasificación y codificación de enfermedades y sus manifestaciones, síntomas, hallazgos anormales, reclamaciones, circunstancias sociales y origen de daños y patologías (*International Classification of Diseases (ICD)*, n.d.). CIE fue publicada por la Organización Mundial de la Salud (OMS) para trabajar con fines estadísticos relacionados con la morbilidad y mortalidad, con sistemas de control de los reingresos de pacientes y para el soporte a la toma de decisiones clínicas.

CIE-10 contiene actualmente la traducción de la clasificación de diagnósticos y de procedimientos en español y dispone de la posibilidad de mapear sus términos a otras codificaciones (*ECIE-Maps - CIE-10-ES Diagnósticos*, n.d.).

ACR Common

La terminología estándar informática ACR Común, es una colección de términos radiológicos y estructuras semánticas comunes desarrollados por el Colegio Americano de Radiología (ACR) para facilitar la interacción con los productos y servicios ACR. Incluye ontologías y esquemas de codificación existentes como RadLex©, SNOMED, CPT e ICD y se organiza en ejes fundamentales y derivados como los escenarios, procedimientos y hallazgos clínicos (*Informatics Standard Terminology* | *American College of Radiology* | *American College of Radiology*, n.d.).

Ontología de fenotipos humanos (HPO)

El proyecto HPO proporciona una ontología de fenotipos relevantes médicos, anotaciones de fenotipos de enfermedades y algoritmos para operar con ellos. Cuenta con 13.000 términos organizados en un grafo acíclico dirigido que conecta dichos términos como subclases unos de otros de forma general a particular (Köhler et al., 2021).

Esta ontología está siendo desarrollada por un consorcio internacional apoyado por el *NIH* (Institutos Nacionales de Salud de EEUU), dedicado a la integración semántica de datos y de los modelos biomédicos para mejorar la investigación biomédica.

La ontología está en inglés, aunque existen traducciones finalizadas o en progreso para varios idiomas (excluido el español).

Nomenclatura de medicina sistematizada – Términos clínicos (SNOMED CT)

SNOMED CT (*Systematized Nomenclature of Medicine – Clinical Terms*) es una terminología clínica codificada completa desarrollada en varios idiomas con gran difusión e importancia que puede ser utilizada para la codificación y análisis de datos clínicos. Es un estándar internacional distribuido por la *International Health Terminology Standards Development Organisation* (*IHTSDO*).

SNOMED CT está constituida por conceptos (organizados en jerarquías), descripciones y relaciones entre los conceptos para representar información y conocimiento clínico para la asistencia sanitaria. Permite la representación e interpretación automática de los documentos clínicos dado que la información se introduce de forma estandarizada asociada a códigos.

El Ministerio de Sanidad de España proporciona una serie de recursos semánticos en el ámbito sanitario, entre ellos esta terminología. Esta terminología es la que se ha seleccionado de referencia para la Historia Clínica Digital del Sistema Nacional de Salud (compendio de información clínica relevante para la atención sanitaria de los pacientes en España y que está disponible en formato electrónico) (*Ministerio de Sanidad - Profesionales - SNOMED CT*, n.d.).

Sistema de lenguaje médico unificado UMLS

El sistema de lenguaje médico unificado UMLS ha sido diseñado y es mantenido por la Biblioteca Nacional de EEUU de Medicina. UMLS es un conjunto de ficheros y de software que contiene vocabularios y estándares de salud y biomedicina para permitir la interoperabilidad semántica (*Unified Medical Language System (UMLS)*, n.d.).

UMLS contiene un metatesauro con términos codificados de distintos vocabularios, jerarquías, definiciones, atributos y relaciones. Los vocabularios pueden ser: CIE-10-CM, LOINC, MeSH, RxNorm y SNOMED-CT. UMLS contiene una red semántica, con categorías generales y sus relaciones entre sí.

Además, UMLS proporciona herramientas léxicas para normalizar cadenas de texto, generar variantes léxicas, crear índices, y contiene un lexicón especializado biomédico.

5. Estado del Arte y Marco teórico

En los siguientes apartados se recoge un resumen de la información del estado del arte más relevante para la preparación del prototipo funcional. Se hace referencia a los tipos y características de las arquitecturas basadas en redes neuronales más utilizadas en PLN, las técnicas de PLN generales y su utilización y las técnicas y recursos de PLN disponibles en el dominio biomédico en español: los modelos pre entrenados, las utilidades de software, los datos y las iniciativas de resolución de tareas.

5.1. Redes Neuronales Artificiales

Se exponen a continuación varias arquitecturas diferentes basadas en redes neuronales dado que estas constituyen la base de las técnicas más utilizadas y avanzadas del estado del arte para el uso en PLN.

Redes neuronales prealimentadas (*feed-forward neural network* o FFNN)

Las redes neuronales prealimentadas son redes neuronales artificiales que procesan la información en una única dirección y sin ciclos de realimentación (sin memoria interna) (López Rubio, 2019f).

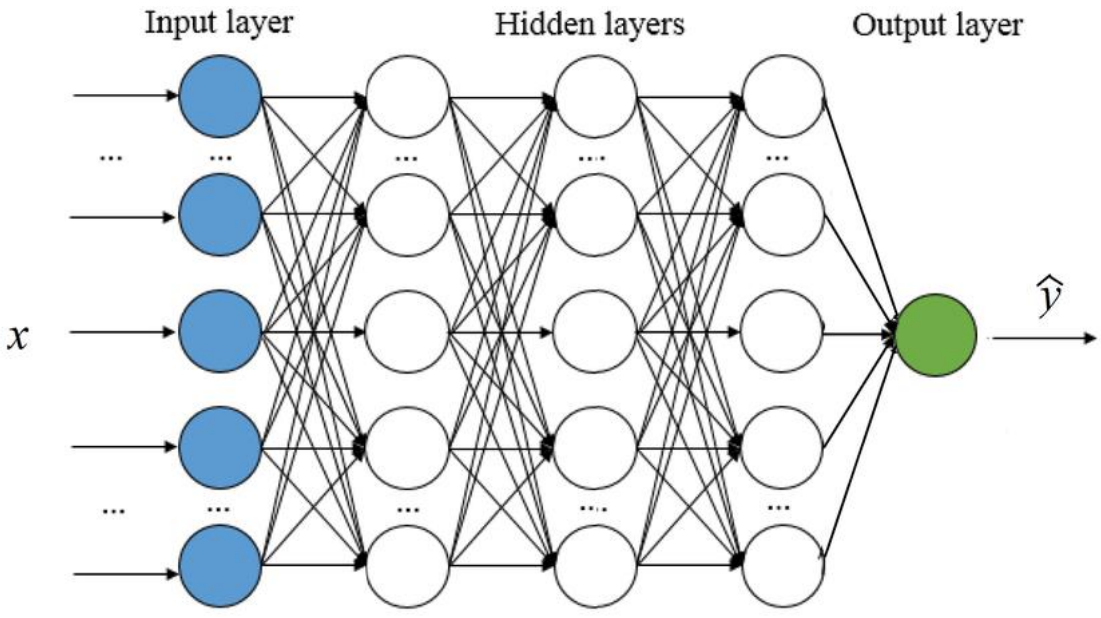


Ilustración 1. Red neuronal prealimentada profunda. Fuente: https://www.researchgate.net

Esta arquitectura de red se utiliza en PLN para detectar en cada capa rasgos característicos del texto desde los más concretos a los más abstractos, por ejemplo, para resolver problemas de clasificación en PLN como la clasificación de frases para el análisis de sentimiento o la clasificación de tópicos.

Para poder utilizar estas redes los textos se deben dividir primero en *tokens* y convertir cada *token* en un valor real único que lo identifique y distinga dentro del vocabulario.

La capa de entrada de esta red contiene el mismo número de neuronas que valores posibles del vocabulario.

Los operadores utilizados habitualmente son:

- **Operador de convolución (red neuronal convolucional prealimentada):** En esta técnica primero se establece un vector núcleo con un grupo de tokens consecutivos a identificar en un texto. Posteriormente se recorre secuencialmente el texto de entrada en grupos de tokens del mismo tamaño que el núcleo, comprobando la similitud de cada grupo de tokens con dicho vector núcleo. El resultado de operador es un vector de salida con los valores de similitud obtenidos entre el texto original y el vector núcleo.
  Tanto el texto de entrada como el texto del vector núcleo deben haber sido tokenizados y vectorizados previamente.
- **Operador de resumen (*pooling*)**. El funcionamiento del operador de resumen consiste en tomar un vector de entrada y calcular un resumen de él para reducir la cantidad de información disponible. Normalmente se coge el máximo de las componentes del vector (que será un valor alto cuando se detecte el rasgo característico, es decir la similitud con el vector núcleo, y cero, cuando no).

Una capa que utiliza el operador de convolución suele ir seguida de otra capa que utiliza el operador de resumen para mantener el tamaño de los datos ajustado.

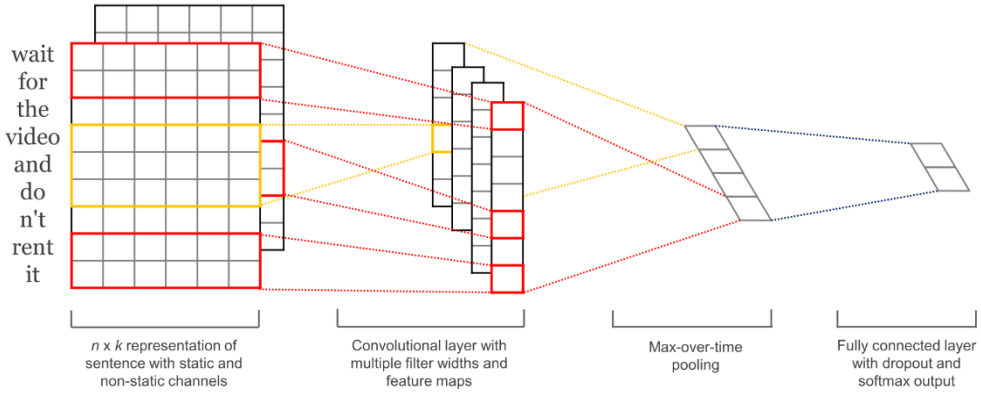


Ilustración 2. Red neuronal prealimentada multicanal con operador de convolución y de resumen para clasificación de frases. Fuente: https://dennybritz.com

Normalmente se trabaja con **redes neuronales prealimentadas multicanal** cuyo funcionamiento consiste en aplicar el operador de convolución a una misma entrada de texto con distintos núcleos, especializando cada núcleo en la detección de un rasgo concreto y asociando el resultado de cada comparación a un canal concreto.

Es destacable elegir un tamaño de núcleo (*kernel size*) adecuado en función de la longitud del patrón a comparar (bigramas, trigramas…) porque los núcleos con longitudes cortas pueden identificar patrones que se reconozcan frecuentemente mientras que los núcleos con longitudes largas identificarán patrones menos frecuentes. En función de la tarea a realizar puede ser más significativo identificar unos patrones u otros.

El tamaño de paso (*stride*) indica cuántos *tokens* se deben obviar del texto estableciendo saltos en la comparación para reducir la cantidad de datos obtenidos resumiendo así la información obtenida.

Como en todos los algoritmos de aprendizaje automático supervisado, se deben separar los datos en un grupo de entrenamiento y otro grupo de validación, con el fin de evaluar el rendimiento posterior del modelo. Se utiliza como función de coste el error cuadrático medio (*MSE*) o la Entropía cruzada (*CE*). Durante el entrenamiento, el algoritmo va modificando los pesos sinápticos de las neuronas artificiales para que la función de coste vaya disminuyendo. Para ello se sigue el mecanismo de descenso de gradiente.

Es importante en las redes neuronales evitar el sobreajuste (*overfitting*) deteniendo el entrenamiento cuando el ajuste deja de mejorar (*early stopping*) o, bien, modificando un conjunto distinto de neuronas en cada paso del algoritmo (*dropout*).

Redes neuronales realimentadas o recurrentes (*recurrent neural networks*, RNN y *long short-term memory*, LSTM)

Las redes neuronales prealimentadas con operado de convolución (CNN) se desarrollaron a partir de los primeros modelos de redes neuronales multicapa para su uso de datos espaciales y para imágenes. Las redes neuronales recurrentes fueron una evolución de las redes neuronales convolucionales para su uso con datos secuenciales como el texto o lenguaje natural.

Una de las limitaciones de las redes neuronales prealimentadas es que no tienen memoria interna y eso les impide conocer las relaciones que hay entre palabras no consecutivas en el texto (por ejemplo, conocer la relación entre el sujeto y el verbo en oraciones largas).

Para paliar esta limitación se utilizan un tipo de redes neuronales derivadas de las RNN que tienen una larga memoria a corto plazo (*long short-term memory*, LSTM). Estas redes LSTM guardan en dicha memoria información semántica de los *tokens* que aparecieron previamente en el texto con el propósito de descubrir dichas relaciones entre palabras no consecutivas.

La arquitectura está basada en unidades/celdas de memoria conectadas entre sí por diferentes puertas por donde pasa la información del texto (*tokens*) y se preserva la información a largo plazo. En este tipo de redes la capa de entrada contiene el vector de cada *token* del texto, la capa oculta combina la información de la capa de entrada con la de la memoria interna y la capa de salida transforma el resultado de la capa oculta para obtener la salida final (López Rubio, 2019a).

Esta arquitectura es más compleja porque debe decidir qué información debe olvidar de la memoria, qué información debe recordarse de cada entrada que se procesa y cuál es la salida que se debe producir. Para ello cada red neuronal se entrena para aprender por separado cómo tomar estas decisiones (que se denominan redes de compuertas: *forget gate, candidate gate y output gate*).

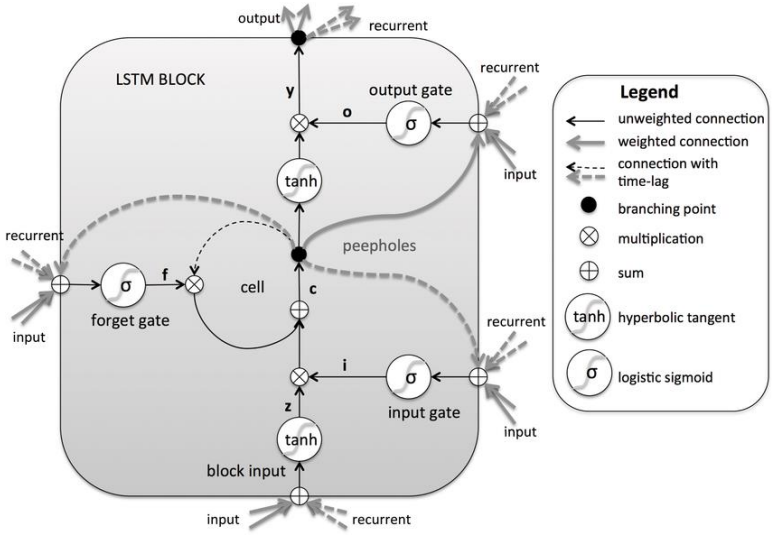


Ilustración 3. Red neuronal LSTM. Fuente: https://www.researchgate.net

En la ilustración se observa un bloque de memoria de la red LSTM en la que los datos de entrada se procesan con los datos previos de la memoria para pasar por las diferentes puertas que determinarán la salida y los nuevos datos de la memoria.

Este tipo de redes se pueden utilizar en PLN por ejemplo para las tareas de clasificación como el análisis de sentimiento o la clasificación de tópicos. La ventaja de capturar las relaciones entre palabras distantes es que mejoran su rendimiento.

El texto debe tokenizarse y vectorizarse para que pueda ser utilizado como entrada para cada celda LSTM. El entrenamiento se hace con descenso de gradiente mediante aprendizaje supervisado como se ha indicado anteriormente. Este entrenamiento es lento en este caso por la complejidad de la arquitectura por lo que se han propuesto versiones simplificadas como la Unidad Recurrente de Compuertas (*Gated Recurrent Unit, GRU*).

Transformers

Uno de los problemas de estas redes recurrentes es que el peso en la salida de las palabras más lejanas es menor que el de las más cercanas porque la red va olvidando las primeras palabras de los textos. Esto puede ser un problema cuando interesa predecir la relación entre palabras alejadas de un texto.

Para solucionar el problema indicado, una estrategia utilizada consiste en entrenar redes neuronales que aprendan las relaciones que existen entre las palabras con los **mecanismos de atención** (Vaswani et al., 2017)**.**

Estos mecanismos consisten en procesar los textos a partir de frases completas y entrenar varias redes neuronales para obtener la relación entre las palabras de cada frase.

Para aprender las relaciones entre las palabras es necesario entrenar dos redes neuronales. Se entrena una red neuronal para codificar una palabra un vector de búsqueda (*query*) y otra red neuronal para generar un vector de identificación (*key*) que codifican cada palabra con dos vectores complementarios entre sí, de manera que la operación sobre ambos vectores identifica cómo de relacionadas están entre sí las palabras (dando lugar a vectores o matriz de vectores de atención).

Para contextualizar cada palabra de una frase con respecto al resto de las palabras y calcular la atención es necesario entrenar una tercera red neuronal para generar un vector de valor (*value*) para cada palabra y emparejarla con la matriz de atención del resto de palabras.

Con este mecanismo no es necesaria una memoria a largo plazo dado que con este heurístico se reporta mayor importancia a aquellas palabras que tienen mayor atención.

En el año 2017 surgieron los ***Transformers*** para la realización de traducción automática de texto entre varios idiomas. Esta arquitectura se basa en mecanismos de atención que tienen mayor calidad y permiten mayor capacidad de paralelización del procesamiento, por lo que se minimiza el tiempo de entrenamiento necesario.

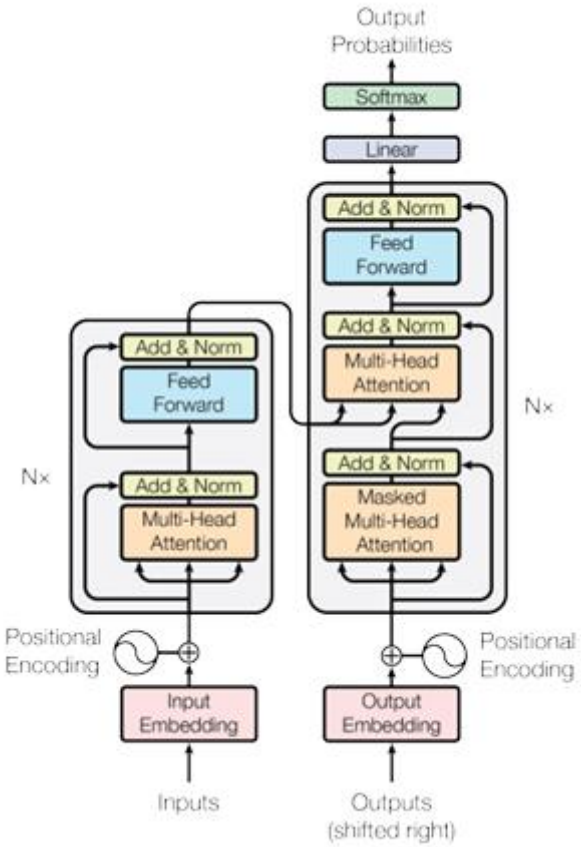


Ilustración 4. Arquitectura de Transformers. Fuente: https://arxiv.org

A continuación, se indica el flujo de información que se observa en la figura anterior:

La **entrada** se obtiene a partir de una red neuronal que genera una representación densa de cada token de un documento (tokenización y vectorización). El transformador es capaz de procesar todos los tokens en paralelo añadiendo información posicional de cada uno de ellos (la posición/orden que ocupa en la frase).

La **codificación** se compone de un bloque atencional y una FFNN que se repite durante seis iteraciones. Los vectores densos pasan por un bloque atencional a varios niveles donde se localizan las relaciones entre las palabras obteniendo las palabras más importantes. Posteriormente una FFNN procesa todos los vectores de la secuencia y los consolida. En ambos casos hay un bloque residual que suma y normaliza los datos.

En el **bloque atencional** se parte de los vectores de cada palabra y como se ha mencionado previamente se les hace pasar por tres redes neuronales (*Queries, Keys y Values*) que obtienen representaciones alternativas de los *tokens* para obtener la matriz de atención. Esto permite identificar qué tokens tienen la información de contexto más relevante con respecto a cada uno de los tokens de la secuencia procesada. Esto se hace utilizando varios bloques atencionales para encontrar las relaciones entre palabras y grupos de palabras a diferentes niveles.

Los pasos del **decodificador** son muy similares al codificador previamente indicado con algunas salvedades. En este caso se dispone de un bloque atencional con enmascaramiento de datos para que no se tengan en cuenta las palabras posteriores en la secuencia (la decodificación del texto es secuencial). Este bloque atencional se retroalimenta de la salida de la decodificación (la salida decodificada influye en las siguientes decodificaciones).

Al final de los decodificadores hay una capa lineal que consiste en una red neuronal con tantas neuronas como palabras del vocabulario y una capa *softmax* de **salida** (convierte cada elemento en probabilidad entre 0 e 1) que recoge a partir del vector decodificado la palabra resultado con mayor probabilidad.

Los *transformers* se pueden utilizar en las diferentes tareas del ámbito del PLN al igual que las arquitecturas anteriormente mencionadas, aunque el procesamiento en paralelo de todas las palabras de una frase y el uso de mecanismos de atención los hace más eficientes en las tareas que las arquitecturas previas.

Los *transformers* también se han utilizado para la generación de texto en iniciativas como GPT-3 *Generative Pretrained Transformers-3* de la empresa OpenAI en el año 2020, o el modelo *Megatron-Nuring* (530B) de Microsoft y NVIDIA en el año 2021 (Smith et al., 2022).

A continuación, se exponen algunos de los modelos relacionados con la arquitectura de *Transformers*.

BERT

En el año 2018 Google entrenó BERT (*Bidirectional Encoding Representation from Transformers*). BERT consiste en un modelo base que interpreta el lenguaje general al que se le pueden añadir capas adicionales para particularizarlo a un problema específico (en inglés). Para su particularización se reentrena posteriormente con conjuntos de datos más reducidos aplicando el aprendizaje por transferencia (Devlin et al., 2018).

BERT nace de los *transformers* y utiliza los codificadores para obtener la información más relevante del texto mediante una representación numérica. Existe una versión de BERT Básico con 12 codificadores, y otra de BERT Extenso con 24 codificadores.

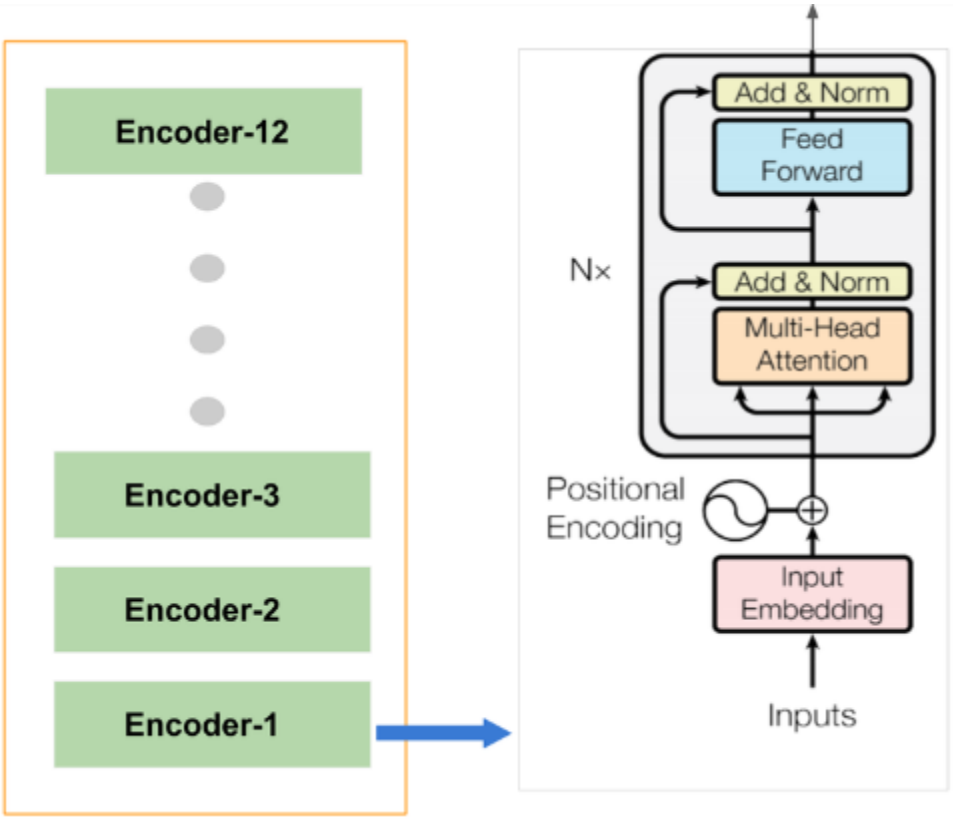


Ilustración 5. Arquitectura de BERT. Fuente: https://pysnacks.com

El texto de entrada se compone de frases que deben ser separadas entre sí mediante un token específico que marca el inicio de cada frase. A cada token de la frase se le asocia una representación (*embedding*), una posición dentro de cada frase (orden), y un número que representa al número de frase al que pertenece (para separar varias frases de entrada).

El entrenamiento se realizó utilizando los recursos de *Wikipedia* y de *Google Books* codificando cada palabra junto con todo su contexto. Este entrenamiento se realizó en dos fases, la primera a nivel de palabras entrenando el modelo para completar palabras enmascaradas de forma bidireccional (completando palabras anteriores y posteriores), y la segunda a nivel de frases entrenando el modelo para que prediga qué frase viene después de cada una de las frases de entrada. Este entrenamiento hace que el modelo comprenda las palabras del lenguaje y sus relaciones junto con el contexto donde se ubican.

Como se ha indicado previamente BERT es un modelo pre entrenado que se debe afinar posteriormente para cada tarea aplicando aprendizaje por transferencia (análisis sentimientos, pregunta/respuesta, …) manteniendo las primeras capas de la red neuronal, una red *softmax* utilizada habitualmente en los clasificadores de redes neuronales para convertir valores en probabilidades y añadiendo finalmente las capas específicas necesarias para la tarea.

BETO

En el año 2020 en la Universidad de Chile se pre entrenó un modelo basado en BERT específicamente para español para el uso de público abierto (Cã et al., n.d.).

El entrenamiento partió de datos de *Wikipedia* y de otras fuentes en español, el tokenizado del texto se llevó a cabo a nivel de subpalabras mediante *Byte-pair encoding BPE* y el entrenamiento para la predicción de palabras enmascaradas del texto se realizó mediante la técnica de enmascaramiento dinámico (técnica que utiliza diferentes máscaras simultáneamente para la misma frase).

MarIA

MarIA es una familia de modelos del lenguaje en español puestos a disposición de la industria y la comunidad en el año 2020 dentro del ámbito del Plan de Tecnologías del Lenguaje de España (*Plan de Tecnologías Del Lenguaje - Página Principal Del Plan de Impulso de Las Tecnologías Del Lenguaje*, n.d.) para tareas de entendimiento del lenguaje natural (*MarIA: Spanish Language Models | Gutiérrez-Fandiño | Procesamiento Del Lenguaje Natural*, n.d.).

Actualmente se incluyen modelos de lenguaje extensos como *RoBERTa-base/large* y *GPT2-base/large* (modelos generativos). Estos modelos han sido entrenados con corpus masivos de textos de un archivo web en español obtenidos por la Biblioteca Nacional Española entre 2009-2019, los textos han sido tokenizados mediante *BPE* (*[1508.07909v5] Neural Machine Translation of Rare Words with Subword Units*, n.d.).

Estos textos se evaluaron mediante el desarrollo de las tareas de, clasificación del texto en categorías, reconocimiento y clasificación de entidades (NERC), identificación de paráfrasis (expresiones lingüísticas diferentes con el mismo significado), *POS tagging*, similitud textual semántica, vinculación textual (relación entre fragmentos del texto) y generación de respuestas a preguntas (*QA*).

5.2. Técnicas de PLN

Se exponen a continuación varias técnicas de PLN basadas en redes neuronales al considerar analizando la literatura que estas constituyen las técnicas más utilizadas y/o avanzadas del estado del arte.

Los documentos y textos a tratar mediante PLN se procesan habitualmente con un enfoque híbrido, aplicando métodos lingüísticos (reglas) y técnicas de aprendizaje automático/profundo de forma conjunta.

Tokenización y representación de documentos (*embedding*)

Como se ha mencionado previamente los textos se deben dividir en *tokens* que pueden englobar palabras, subpalabras como sílabas o caracteres, o incluso, frases completas en función del caso de uso. Dichos tokens deben ser codificados con valores numéricos para que sean tratados por los algoritmos utilizados de aprendizaje automático o profundo.

Una primera aproximación podría ser codificar cada *token* diferente con un número natural. Esto no es lo más adecuado porque el algoritmo de aprendizaje puede tratar de establecer una relación por diferencia numérica entre los *tokens* codificados y además confundir los textos que contienen números con su codificación numérica.

La aproximación más adecuada es la vectorización, técnica que consiste en asignar a cada token una secuencia o vector de números (*embedding*).

Una primera aproximación es utilizar la técnica de **representación dispersa de documentos** que permite representar los textos convirtiendo los *tokens* incluidos (palabras típicamente) en una lista de números. La representación numérica consiste en un vector de números enteros de tamaño del vocabulario de tokens posibles. Todos los números se inicializan con “0” y solo se rellenan con valores positivos los *tokens*, que sí se encuentran en el documento. Esto tiene la peculiaridad de que los documentos que contienen los mismos *tokens* se representan mediante vectores similares. Las representaciones se denominan dispersas porque contienen muchos ceros en su representación (López Rubio, 2019e).

Representación densa de documentos (*dense embedding*) y semántica distribucional

La vectorización de textos mediante las representaciones dispersas mencionadas anteriormente tiene el inconveniente de que las dimensiones de los vectores son excesivas conforme el vocabulario es mayor y los vectores ocupan espacio innecesario con información redundante (López Rubio E, 2019).

Para resolver este problema, se utilizan representaciones densas que presentan menor dimensionalidad y redundancia al ser más compactos. Para que estos vectores sean óptimos para la tarea de representación de *tokens*, estas representaciones

densas se diseñan mediante redes neuronales de aprendizaje profundo a partir de los datos de los documentos.

El entrenamiento de estas redes neuronales se realiza de forma no supervisada mediante arquitecturas FFNN o *Transformers* utilizando como etiquetas de salida los *tokens* del texto inicial.

Este entrenamiento se lleva a cabo resolviendo tareas auxiliares como la predicción de *tokens* en el texto. Algunas estrategias pueden ser: predecir el anterior o siguiente *token* de una secuencia de *tokens* de un texto, predecir el *token* intercalado en una secuencia de tokens de una frase o predecir *tokens* dentro de una ventana específica de *tokens* de un texto independientemente de la posición.

Este aprendizaje fuerza a la red a mantener una representación compacta de los *tokens* y una vez entrenada se puede utilizar para codificar cada palabra del vocabulario en un vector de dimensionalidad reducida. Las dimensiones más habituales de los vectores son entre 50 y 300 pero hay autores que indican que este valor se debería elegir en base a datos estadísticos del corpus (Patel & Bhattacharyya, 2017).

Con este enfoque la arquitectura de la red neuronal **aprende el significado, la idea o la semántica de los *tokens*** que se encuentran en el texto y los codifican con vectores densos que los representan (Ferrando Javier, n.d.).

La base de esto es la **semántica o hipótesis distribucional** que se resume en que los elementos lingüísticos con distribuciones similares tienen significados similares. Esto implica que: “las palabras que ocurren en contextos similares tienden a tener significados similares” (Harris, 2015), “el significado de una palabra no se debe obtener de forma aislada sino en el contexto de una frase” (Firth J.R., 1957) y “una palabra se define por sus palabras acompañantes” (*Frege, Contextuality and Compositionality on JSTOR*, n.d.).

Teniendo en cuenta lo anterior la red neuronal aprende a agrupar en cercanía en el espacio de n-dimensiones los vectores de *tokens* con semántica similar. Dichos *tokens* se disponen entre sí de manera que se establece una relación semántica extrapolable a otros *tokens* del mismo grupo. Dichas agrupaciones se disponen con respecto al resto de agrupaciones de manera que se puedan establecer relaciones entre dichos grupos. Esto permite que se pueden aplicar operaciones aritméticas con sentido semántico entre *tokens* (Ferrando Javier, n.d.).

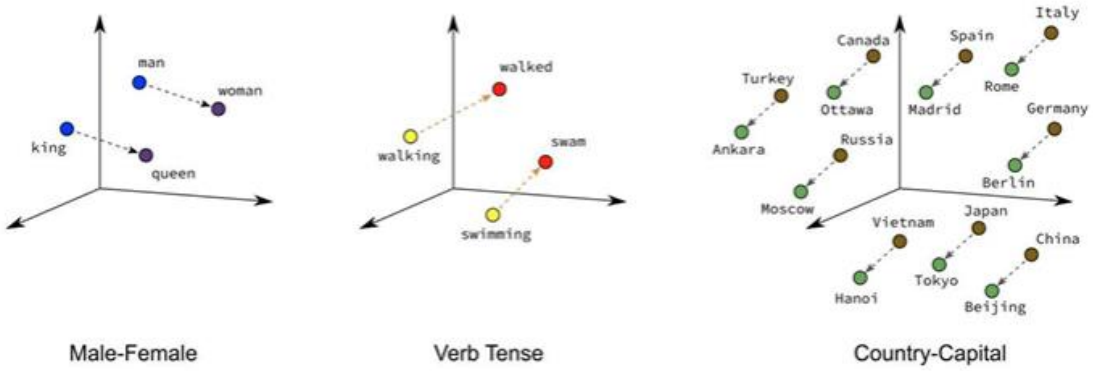


Ilustración 6. Representación visual 2D/3D de la relación semántica entre vectores densos. Fuente: https://www.cs.upc.edu

En estos casos, es necesario aplicar un algoritmo de reducción de dimensionalidad como PCA para la visualización gráfica en 2D/3D de la ubicación de las palabras y, por tanto, de las relaciones semánticas. Una vez hecho esto, es posible representar gráficamente los tokens y la proximidad entre ellos.

Como se observa en la ilustración de forma a visual se puede llegar de la palabra "king" a "queen", de "walking" a "walked" o de "Russia" a "Moscow" de similar manera en distancia/ángulo y dirección al resto de palabras del mismo ámbito semántico.

Por ejemplo, si a la representación densa de la palabra "King" se le resta la representación densa de "man" y se le suma la representación densa de "woman" el resultado es la representación densa de la palabra "Queen".

Actualmente existen algoritmos de representaciones de vectores densos ya pre entrenados como por ejemplo "*Word2vec*" que se pueden aplicar a tokens o, incluso, a documentos completos. Estos modelos pueden servir de partida para generar nuevos modelos mediante aprendizaje por transferencia para documentos o textos más específicos evitando tener que partir de nuevo de los datos de partida que son más extensos y generales.

A continuación, se indican varios métodos ampliamente utilizados para la representación densa de *tokens*.

- **Word2Vec**

Word2Vec es un método publicado en el año 2013 por Google para generar vectores densos de palabras. Estos vectores obtienen características relevantes de las palabras en relación con el texto completo (relaciones semánticas, definiciones, contexto…) para representar las palabras junto con su similitud entre ellas (Vatsal P, 2021).

A partir de un conjunto suficientemente extenso de palabras *Word2Vec* puede estimar el significado de cada palabra y asociarlas entre sí basándose en su ocurrencia en el texto.

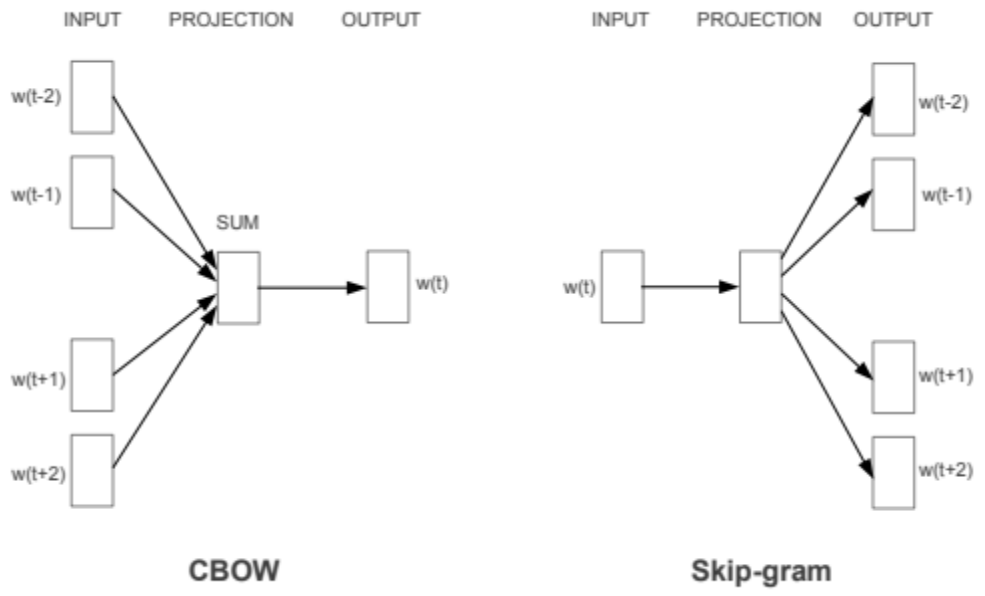


Ilustración 7. Arquitecturas CBOW y Skip-gram para predecir la palabra actual o las palabras del contexto respectivamente. Fuente: https://arxiv.org

Word2Vec trabaja con dos enfoques y arquitecturas diferentes: *CBOW* y *skip-gram*. La arquitectura del enfoque *CBOW* es una red neuronal prealimentada que funciona tratando de predecir una palabra enmascarada intercalada en una lista de palabras del contexto (frase). La arquitectura del enfoque *skip-gram* es una red neuronal simple con una capa oculta entrenada para predecir palabras enmascaradas anteriores y posteriores del contexto (frase) a una palabra dada.

Las representaciones densas de *Word2vec* están dispuestas en el espacio n-dimensional de manera que palabras con similares características están en proximidad unas de otras (Mikolov et al., 2013).

- **GloVe**

GloVe es una librería publicada en el año 2014 por la Universidad de Standford para obtener la representación densa de palabras manteniendo la medida de similitud semántica entre ellas (Pennington et al., 2014).

GloVe utiliza la ocurrencia estadística de las palabras de un corpus, en concreto teniendo en cuenta el ratio de coocurrencia de las palabras (en pares de palabras) lo que permite identificar las palabras que se relacionan más frecuentemente en el corpus.

- **FastText**

FastText es una librería publicada en el año 2015 por *Facebook* para el aprendizaje de representación mediante vectores densos de palabras para la clasificación del texto (Bojanowski et al., 2016).

FastText trabaja explotando la información a nivel de subpalabras para construir los vectores densos de las palabras (Bojanowski et al., 2016). La representación densa se aprende para cada uno de los n-gramas de caracteres, que posteriormente se suman entre sí para representar cada palabra. Una vez que se dispone del vector denso de las palabras completas se entrena el modelo en una segunda fase con el enfoque *skip-gram*.

Este enfoque de subpalabras: permite un entrenamiento más rápido, hace que el modelo trabaje bien con prefijos y sufijos de palabras y que mejore la predicción de palabras nuevas que no estaban en los datos de entrenamiento inicial.

A continuación, se indica como es la secuencia de pasos para desarrollar un modelo de clasificación de aprendizaje automático de extracción de términos como base para el cumplimiento del objetivo principal perseguido.

Pipeline de procesado de los modelos de clasificación

Los pasos propuestos típicamente para conseguir realizar la tarea objetivo para la que se realiza el PLN son los siguientes:

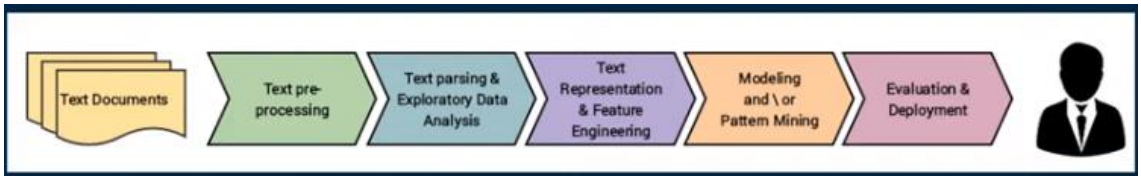


Ilustración 8. Pipeline de procesado del modelo. Fuente: https://suneelpatel18.medium.com

1) Preprocesado del texto para su homogeneización y para eliminar símbolos o caracteres irrelevantes.
2) Análisis exploratorio de los datos del texto para encontrar inconsistencias y solventar problemas de calidad de los datos.
3) Representación del texto mediante representación dispersa o densa para identificar los *tokens* dentro del vocabulario y las características semánticas o significado de los mismos.
4) Preparación y entrenamiento del modelo de clasificación mediante aprendizaje automático supervisado.
5) Evaluación y despliegue del modelo.

- **Preprocesado del texto previo**

El preprocesado del texto previo común a la aplicación de las técnicas de aprendizaje automático consiste en: conversión del texto en minúsculas, eliminación de las palabras irrelevantes (*stopwords*), lematización de las palabras derivadas, clasificación de las palabras en función de su tipo para filtrar las palabras que deben ser analizadas del resto (*POS tagging*), la reducción de caracteres especiales y el tokenizado del texto (Allender, 2021).

Para realizar estas acciones se requiere aplicar un enfoque híbrido, por un lado, reglas lingüísticas conocidas del lenguaje concreto (idioma y ámbito), y por otro, modelos de aprendizaje profundo basados en redes neuronales.

- **Representación de las palabras**

La representación de las palabras se suele realizar con los métodos indicados previamente como *BOW, TF-IDF, Word embedding (ELMo, GloVe, Word2Vec, Fast Text)* o *Transformers (BERT)* (Allender, 2021).

Para implementar estos métodos de forma efectiva se debe conocer el idioma, el vocabulario con el que se va a trabajar, el corpus utilizado, y, por tanto, el ámbito del texto.

- **Modelos de clasificación automática**

Es posible utilizar modelos supervisados de aprendizaje automático para clasificar los textos (*SVM, kNN,* …), aunque actualmente los modelos de aprendizaje profundo, son los que mejores resultados están dando (*FFNN, CNN, RNN/LSTM y Transformers*).

Extracción de términos en los textos

A continuación, se identifica someramente un enfoque planteado para desarrollar la vinculación de entidades.

Las fases necesarias para poder extraer términos de un texto y mapearlos en entidades de un vocabulario conocidas son las siguientes (Krauthammer & Nenadic, 2004):

1. **Reconocimiento o extracción de entidades (*NER*),** en esta fase se deben reconocer aquellas palabras o grupos de palabras del texto que son términos potenciales. Para ello se clasifica las palabras en categorías (por ejemplo, personas, expresiones de tiempo, cantidades, términos, etc.) y se filtra aquellas categorías más relevantes.
2. **Clasificación de términos**, en esta fase se clasifica las entidades reconocidas (términos) de un dominio específico filtrando el resto. Para la clasificación, se analiza cómo de relevantes son dichos términos en el dominio específico en contraste con su utilización en un dominio general en base a su ocurrencia estadística en ambos corpus del dominio y general. Si el término candidato es muy utilizado en el corpus del dominio en contraste con un uso reducido en un corpus general, se clasifica en dicho dominio para ser vinculado.

3. **Emparejamiento o vinculación del término (*NEL*)**, en esta fase se vinculan los términos clasificados en el dominio deseado con conceptos definidos en vocabularios o bases de datos de términos conocidos.

De lo indicado anteriormente, es de interés para el autor la idea planteada de que para poder vincular los términos, es necesario previamente extraer las entidades reconocidas del texto y clasificarlas en base a su grado de pertenencia al corpus de los términos a vincular.

5.3. Técnicas y recursos de PLN en el dominio biomédico en idioma español

Algunas tareas que el autor ha encontrado en la literatura de este tipo en el dominio biomédico en idioma español son: la clasificación automática de notas clínicas en base a una codificación médica, la detección del diagnóstico de cáncer en los informes médicos, la detección automática de términos médicos en un texto y la detección de entidades reconocidas del ámbito clínico.

Se indican a continuación algunas de las aproximaciones analizadas para resolver la vinculación de entidades (NEL) utilizando terminologías clínicas estándar para conocer diferentes enfoques para resolver esta tarea.

Tarea conjunta de codificación de casos clínicos CodiEsp

Esta tarea fue promovida por el Plan TL nacional de España en el año 2020 (Plan TL España, 2020) consistente en la asignación de códigos CIE-10 de diagnósticos, procedimientos y la exploración de cómo de explicable es el resultado, a documentos de casos clínicos en español (1000 casos clínicos anotados) (Miranda-Escalada et al., 2022).

Se extractan algunas de las iniciativas de esta tarea conjunta cuya aproximación es de utilidad para la resolución del problema planteado.

- **FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding**

  De forma resumida, la aproximación en este caso para resolver las tareas consistió en (García-Santa & Cetina, 2020):

1. Utilizar una base de datos de conocimiento basada en grafos para almacenar los términos clínicos de la codificación y sus relaciones.
2. Entrenar un modelo NER basado en el modelo pre entrenado BERT *Multilingual* para reconocer las entidades o términos clínicos anotados en el conjunto de entrenamiento.

3. Enlazar los términos reconocidos con los términos de la base de datos de conocimiento mediante el análisis de la distancia semántica con los mismos.
4. Procesar la salida posteriormente eliminando las entidades negadas en el texto con una herramienta de negación existente.

- **LSI UNED team on ICD-10 coding based on semantic distance**

La aproximación en este caso se basaba en un método no supervisado basado en la similitud conceptual teniendo como partida una terminología basada en SNOMED-CT, combinado con métodos de *Gradient Boosting* (Almagro, 2020).

1. Entrenamiento de un modelo con la tarea *NER* para asignar el texto a los conceptos de SNOMED-CT mediante comparación léxica y descripciones CIE, computando la similitud entre los conceptos en función de su afinidad.
2. En paralelo, se utilizaba el algoritmo de multi etiquetado *Gradient Boosting* basado en clasificadores binarios para predecir los códigos CIE en el texto.
3. Fusión de ambos resultados para obtener la salida definitiva.

- **IXA-AAA Multi-label Classifiers Similarity Match Coders**

En el caso planteado, la estrategia seguida fue la utilización de clasificadores multi etiqueta basados en árboles (*XGBoost*) junto con los sistemas de similitud de cadenas de texto. Los fragmentos de texto y la definición de los códigos CIE se comparaba mediante varias métricas: la distancia (*Levenshtein, Jaro Winkler y Auto*) y la distancia coseno entre vectores densos obtenidos con *BERT* (Blanco et al., 2020).

Un aspecto de interés para el autor de la presente memoria en los trabajos previos es el uso de la distancia semántica para localizar grupos de palabras con el mismo significado en un texto, y la utilización de la distancia coseno calcular esta similitud. Esto es, enlazar términos comparando las palabras de una terminología clínica con las palabras del informe.

Reconocimiento y vinculación de entidades (*NER* y *NEL*) utilizando BERT y *embeddings* en el ámbito del español médico

Este artículo plantea un enfoque para el reconocimiento de enfermedades en textos clínicos y su relación con términos de SNOMED-CT (patologías) (Reyes-Aguillón et al., 2022).

Está basado en el uso de un modelo BERT entrenado con una red neuronal de clasificación utilizando para la vinculación de las patologías un modelo de representación densa de palabras (*embedding*) generado a partir de documentos científicos de medicina en español.

1. En la primera etapa de reconocimiento de entidades (*NER*) se entrena un modelo BERT para la clasificación a partir de entidades anotadas del ámbito biomédico y se aplica para la predicción sobre los textos a reconocer.
2. En la etapa de vinculación de entidades (*NEL*), se realiza en primer lugar el preprocesado tanto de las entidades reconocidas como de la base de datos de conocimiento (SNOMED-CT). El preprocesado consiste en la conversión a minúsculas, *tokenización*, eliminación de palabras irrelevantes y signos de puntuación.
   En segundo lugar, se obtiene la representación densa del texto de entrada y del vocabulario y se compara mediante similitud coseno. Las entidades con más de una palabra obtienen un único vector denso, promediando los vectores de cada una de las palabras.
3. Finalmente se normalizan los resultados.

Un aspecto de interés para el autor de la presente memoria en este trabajo previo es el heurístico utilizado para el procesamiento previo del texto, además del uso de nuevo de la distancia semántica para localizar grupos de palabras con el mismo significado en un texto, el promediado de los vectores de las palabras y la utilización de la distancia coseno calcular esta similitud.

A continuación, se añaden los recursos disponibles en base a modelos pre entrenados y artefactos software analizados para resolver la vinculación de entidades (NEL) utilizando terminologías clínicas estándar para disponer de diferentes herramientas para resolver esta tarea a la hora de desarrollar el prototipo.

Plan de Impulso de las Tecnologías del Lenguaje (Plan TL)

El Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) es una iniciativa del Ministerio de Asuntos Económicos y Transformación Digital de España cuyo objetivo es fomentar el desarrollo del procesamiento del lenguaje natural en lengua española y lenguas cooficiales (*Plan de Tecnologías Del Lenguaje - Página Principal Del Plan de Impulso de Las Tecnologías Del Lenguaje*, n.d.).

El Plan TL establece medidas para incrementar las infraestructuras lingüísticas en español, impulsar la industria el lenguaje y el sector de forma coordinada y evitando duplicidad de esfuerzos, fomentando la internacionalización y participación en proyectos de I+D+i, especialmente con Latinoamérica.

Los aspectos relevantes del Plan que son de interés del autor de esta memoria, son las infraestructuras lingüísticas desarrolladas con herramientas comunes y los recursos de acceso público en el ámbito clínico o biomédico que se mencionan a continuación (*Plan de Tecnologías Del Lenguaje - Gobierno de España*, n.d.):

- **Modelos entrenados para la generación de vectores densos generados a partir de corpora biomédico/clínico en español (Biomedical-Word-Embedding-for-Spanish).**
  Se proporciona modelos listos para utilizar, basados en *Fasttext Word embedding* y *BPE subword embedding* entrenados a partir de documentos biomédicos y clínicos con diferentes enfoques (*CBOW* y *Skip-gram*) y dimensiones (50, 100 y 300).

Modelos de codificación clínica automática en español

En el estudio se analizaron modelos basados en *Transformers* para la codificación clínica automática en español (CIE-10) utilizando el enfoque basado en aprendizaje por transferencia con una estrategia de clasificación de frases multi etiqueta (Lopez-Garcia et al., 2021).

Se partió de los modelos *multilingual BERT, BETO y XML-RoBERTa* pre entrenados en un corpus de casos clínicos de oncología para adaptarlos a las particularidades de los textos médicos, con las tareas de predicción de siguiente frase (*NSP*) y modelo de lenguaje enmascarado (*MLM*), a los que se les hizo un ajuste fino para la tarea de codificación clínica con informes etiquetados con CIE-10 de la iniciativa CodiEsp.

Se generaron y se proporcionaron como disponibles para ser utilizados los modelos preentrenados: *mBERT-Galen*, *BETO-Galen* y *XML-R-Galen* (Lopez-Garcia et al., 2021).

El interés del autor por este estudio es la disponibilidad de un modelo basado en *Transformers* orientado al ámbito clínico en español. Con este modelo se pueden obtener los vectores densos de los grupos de palabras y términos clínicos con esta arquitectura avanzada.

6. Desarrollo del proyecto y resultados

6.1. Metodología

Para el desarrollo de los objetivos mencionados en el apartado 2 de la memoria, se ha propuesto seguir una metodología de desarrollo en cascada.

Para ello se ha descompuesto el problema en tareas y actividades más manejables y se ha planificado su desarrollo en base al tiempo disponible hasta la entrega del TFM ajustando la profundidad de cada tarea en base a ese parámetro.

Se ha confeccionado una estructura de descomposición del trabajo (EDT) o *Work Breakdown Structure* (*WBS*), descomponiendo el objetivo en grupos de trabajo siguiendo el enfoque en cascada mencionado.

Ilustración 9. Cronograma de actividades y tareas del TFM desde noviembre 2022 hasta marzo 2023, desarrollado con la herramienta xmind. Elaboración propia.

La dedicación inicial prevista era de entre 10 horas semanales, lo que implicaba 150 horas totales de dedicación.

6.2. Planteamiento del problema

Los informes médicos de pruebas de imagen tienen una estructura definida, pero están desarrollados con texto libre. Por tanto, los informes no están clasificados de ninguna forma codificada que permita su búsqueda y localización posterior. Los

especialistas que redactan los informes médicos necesitan localizar casos previos a partir de palabras clave. Actualmente eso requiere una búsqueda a partir de texto extensiva y compleja en recursos y en tiempo. En el futuro se requerirá que los profesionales codifiquen los diagnósticos en los informes. Esto obligará a los médicos a codificar esos datos manualmente a partir de catálogos, lo que llevará mucho tiempo.

Para resolver el problema mencionado se deben resolver las siguientes cuestiones: identificar las técnicas y modelos de PLN adecuados, seleccionar una terminología clínica reconocida, buscar un heurístico apropiado para vincular la terminología clínica del texto, desarrollar un motor de inferencia que permita etiquetar los informes y recuperarlos a partir de dichas etiquetas y construir un interfaz de usuario para validar dicho motor de inferencia.

El campo del PLN ha avanzado considerablemente con el aprendizaje profundo, las redes neuronales densas y otras estructuras como los *Transformers*, para resolver los problemas relacionados con la interpretación, obtención de significado y generación de texto. En el ámbito clínico en idioma español se ha avanzado fruto de algunas iniciativas nacionales de países como España u otros de habla hispana, y se dispone de modelos preentrenados disponibles para ser utilizados en tareas concretas como el reconocimiento de entidades, la representación densa de palabras o en otras tareas mediante al aprendizaje por transferencia.

Después de analizar el estado del arte y las diferentes alternativas para resolver el problema el autor opta por una propuesta de prototipo que contemple los siguientes requisitos:

- No debe necesitar demasiados recursos de computación dado que el entorno de producción no tiene grandes recursos (no hay *GPUs*).
- Debe dar una respuesta interactiva a los usuarios ágil (menor a un minuto).
- Debe ser capaz de trabajar con datos no vistos durante el entrenamiento dado que no se tiene acceso a datos de informes reales el entorno de producción (restricciones legales).
- Debe ser desarrollable en 4 o 5 meses dado que existen restricciones temporales para la entrega y defensa de la presente memoria.

Por todo lo expuesto para construir una posible solución al problema planteado se llevará acabo lo siguiente:

1. Se experimentará con las técnicas, algoritmos y modelos seleccionados desarrollando cuadernos *Jupyter* mediante Google *Colab* (*Te Damos La Bienvenida a Colaboratory - Colaboratory*, n.d.), con el objetivo de seleccionar un método de establecer la anotación semántica de los informes y validarlo con respecto a los requisitos de partida.
2. Se desarrollará el motor de inferencia como un *API REST* en *Python* siendo este uno de los lenguajes funcionales preferidos para el desarrollo de

soluciones de aprendizaje automático, se dispone de amplias librerías, permite desarrollos rápidos.

El motor de inferencia estará alojado en un servidor accesible a través de un API para su acceso, mediante el *framework microweb Flask* , para dar servicio al prototipo. Las pruebas del motor se llevarán a cabo inicialmente con la aplicación *Postman* (*Postman API Platform | Sign Up for Free*, n.d.) para simular las llamadas al API y comprobar los resultados obtenidos.

3. Para el desarrollo y la interacción con los usuarios se desarrollará el interfaz de usuario como una aplicación web *HTML/CSS/Javascript*, por su simplicidad en el desarrollo, madurez y por estar muy extendido en la actualidad.

   El prototipo se propone realizarlo para su edición en un entornos de desarrollo basados en el navegador *Stackblitz* (https://stackblitz.com/) y en escritorio como *Visual Studio Code* (https://code.visualstudio.com/), que permiten el desarrollo de aplicaciones web con *HTML/CSS/JS*. En ejecución, no obstante, se utilizará un servidor web como librería Python *http.server* (*Http.Server — HTTP Servers — Python 3.11.2 Documentation*, n.d.), será accesible a través del navegador web, y utilizará el motor de inferencia mediante su *API REST* expuesto.
4. Tanto el motor de inferencia, como el prototipo, utilizarán una base de datos relacional *mysql* (*MySQL*, n.d.), para gestionar la información necesaria (informes y catálogo de términos).
5. Se utilizará la librería *Spacy* (*SpaCy · Industrial-Strength Natural Language Processing in Python*, n.d.), uno de los principales *frameworks* PLN, por su potencia, velocidad, soporte del castellano e incorporación de modelos integrados pre entrenados *(parsing, tagging, NER, lemmatizer, tok2vec*…), para el procesamiento del contenido del informe y la división del texto en frases.
6. Se utilizará la librería *Gensim* (*Gensim: Topic Modelling for Humans*, n.d.), con el fin de utilizar los modelos entrenados de generación de *embeddings* y determinar la similitud entre tokens/palabras. *Gensim* es la librería más rápida para entrenar los vectores densos (*embeddings*), para trabajar con corpus amplios. Esta librería es muy utilizada por la comunidad de ingenieros.

   Para que la respuesta a la predicción sea ágil e interactiva, y dada las restricciones temporales del TFM, el motor de inferencia se planteará buscando la similitud semántica entre los informes médicos y los términos del vocabulario. Para ello se utilizarán modelos pre entrenados ya existentes como el desarrollado dentro de la iniciativa del Plan de Tecnologías del Lenguaje, de tipo *FastText*, *Biomedical Word Embeddings for Spanish* (*PlanTL-GOB-ES/Biomedical-Word-Embeddings-for-Spanish: Biomedical Word Embeddings Generated from Spanish Biomedical Corpora.*, n.d.). Adicionalmente se hará alguna prueba con modelos de basados en transformadores adaptados a la codificación clínica en Español *BETO-Galen* (*Guilopgar/ClinicalCodingTransformerES: Clinical Coding in Spanish Using Transformers*, n.d.).

7. Los conjuntos de datos a utilizar para probar el correcto funcionamiento del motor son los registros del corpus DisTEMIST (*DisTEMIST Corpus: Detection and Normalization of Disease Mentions in Spanish Clinical Cases | Zenodo*, n.d.). DisTEMIST es una colección de 1000 casos clínicos con anotaciones de enfermedades con conceptos SNOMED-CT.
8. Con respecto a las terminologías médicas a utilizar se plantea el uso de una terminología en español que se pueda utilizar en el ámbito de las pruebas de imagen, como es la terminología clínica SNOMED-CT (*Navegador SNOMED CT SNS*, n.d.). Existen muchas del ámbito clínico como se ha mencionado en el apartado 4.5 de la memoria, pero SNOMED CT (*Systematized Nomenclature of Medicine – Clinical Terms*) es la terminología clínica integral, multilingüe y codificada de mayor amplitud, precisión e importancia desarrollada en el mundo. SNOMED-CT es también un producto terminológico que puede usarse para codificar, recuperar, comunicar y analizar datos clínicos para representar la información de forma adecuada, precisa e inequívoca. Está basada en conceptos, descripciones y relaciones para representar con precisión información y el conocimiento clínico en el ámbito de la asistencia sanitaria.
9. Como repositorio de todo el código fuente generado y los datos utilizados se utilizará Github (*GitHub*, n.d.) .

Todo el código generado se ubica en el proyecto público con la siguiente ruta: https://github.com/fmedinafernandez/14MBID-TFM-NEL .

6.3. Desarrollo del proyecto

A continuación, explicamos el desarrollo de las diferentes fases del proyecto mostradas en el cronograma de la Ilustración 14.

Recopilación de conceptos generales

En esta primera fase de recopilación de conceptos generales el objetivo perseguido ha sido obtener y afianzar los conocimientos generales sobre en qué consiste, evolución y aplicaciones del PLN, cómo se confeccionan y qué estructura tienen los informes de pruebas de imagen y cuáles son las terminologías clínicas reconocidas.

Se ha trabajado revisando información disponible en línea, como páginas web, imágenes y videos, especialmente con recursos más generales, de expertos en la materia, fabricantes/proveedores en el ámbito, publicaciones en universidades, y también revistas y artículos de investigación específicos (www.sciencedirect.com, http://pubs.rsna.org , www.ncbi.nlm.nih.gov, http://arxiv.org , entre otras).

La metodología seguida ha sido la de buscar, analizar y extraer información de la documentación partiendo de ideas generales y profundizando en los aspectos claves o

más relevantes. Es decir, de lo general a lo particular pero siempre dejando de particularizar cuando la información era suficiente (uno de los riesgos detectados es que el particularizado detallado de la información puede ser demasiado extenso en tiempo lo que pone en riesgo el cumplimiento de los objetivos).

Esta primera fase se ha planificado para ejecutarse en dos semanas y se ha cumplido en plazo. Se debe indicar, no obstante, que el autor había empezado a analizar estos conceptos generales sobre PLN en octubre de 2022 por lo que, en la práctica, se han empleado dos semanas adicionales en la ejecución de esta tarea.

Se ha avanzado relativamente rápido al ser conceptos más generales tratando de empezar en un enfoque más general e ir profundizando ligeramente, aunque también se ha descartado mucha información redundante en diferentes sitios, ha sido necesario contrastar información para determinar el rigor de la información mostrada al proceder de medios más generales.

Revisión del estado del arte

En esta segunda fase, el objetivo perseguido ha sido concentrar el conocimiento específico sobre el PLN en el ámbito médico, los modelos pre entrenados disponibles en español clínico, las terminologías clínicas disponibles en español, los datos de informes de pruebas de imagen y las librerías y utilidades disponibles para este ámbito y caso de uso.

En relación a los grupos de actividades de revisión del estado del arte, se ha enfocado la búsqueda al ámbito de las publicaciones en universidades, revistas y artículos de investigación específicos de forma preferente. Una vez sentadas las bases ha sido necesario enfocarse con mayor rigor en los aspectos más científicos. También se ha trabajado revisando información disponible en línea, como páginas web, imágenes y videos, con recursos más generales y/o técnicos.

La metodología seguida ha sido la de buscar, analizar y extraer información de la documentación, partiendo de ideas generales, y profundizando en los aspectos claves o más relevantes. En este caso, también se ha partido de aspectos generales del PLN, hasta llevar a aspectos concretos relevantes (p.e. PLN para reconocer entidades en informes médicos).

Esta segunda fase ha sido planificada en tres semanas y se ha cumplido en plazo. Se debe indicar, no obstante, que el autor había empezado a analizar estos conceptos generales sobre PLN en octubre de 2022 y que en la redacción de la memoria ha sido necesario retomar o profundizar en algunas fuentes por lo que en la práctica se han empleado tres semanas adicionales en la ejecución de esta tarea de revisión del estado del arte.

Se ha avanzado de forma más lenta al ser conceptos más específicos y concretos, tratando de empezar en un enfoque más general e ir profundizando. También ha sido necesario limitar el profundizado en los temas para no exceder el tiempo planificado. Afortunadamente, se ha encontrado gran parte de la información buscada (modelos pre entrenados disponibles, métodos y técnicas, datos disponibles, herramientas y terminologías clínicas para desarrollar el caso de uso).

En relación a la parte práctica de revisión del estado del arte, se ha llevado a cabo el desarrollo y pruebas de los diferentes algoritmos y estrategias encontrados para llevar a cabo la vinculación de los elementos de la terminología a los informes médicos, utilizando *Google Colab* y los cuadernos *Jupyter*.

Desarrollo del motor de inferencia

En relación a los grupos de actividades de desarrollo del motor de inferencia, se ha seguido una metodología de desarrollo en cascada, seleccionando las herramientas más adecuadas, analizando los requisitos y funcionalidades a proporcionar por el motor, diseñando posteriormente la arquitectura del software correspondiente para proceder a la construcción del motor y finalmente las pruebas técnicas de funcionamiento correcto y validación de los resultados.

Esta tercera fase ha sido planificada en cuatro semanas por ser la tarea que requiere mayor complejidad técnica. Se ha cumplido en el plazo programado.

El análisis y diseño se ha realizado conforme a la experiencia personal del autor en el desarrollo de software.

Se ha avanzado adecuadamente en la construcción a partir de la información de las pruebas y experimentos relacionados con la revisión del estado del arte. Adicionalmente, el autor tiene experiencia previa en el desarrollo de motores con *API REST* en *Python* para otros casos de uso.

Las dificultades encontradas principalmente han sido relativas al uso del lenguaje *Python* (sintaxis) y a la arquitectura elegida de exposición de un *API REST*, y de división de la lógica de forma modular (interfaz/*API*, gestores de negocio, capa de acceso a datos y utilidades).

Con respecto a las pruebas y validación, la depuración de un programa desarrollado en *Python* como lenguaje funcional requiere tiempo y un cuidado importante a la hora de desarrollar el código, porque los errores no se detectan hasta la ejecución del programa. Por ejemplo, una mala indentación (o uso del sangrado) en el código en Python, hace que un módulo no pueda ser interpretado. Se ha seguido la experiencia en la construcción de este tipo de artefactos por parte del autor.

Se han realizado inicialmente pruebas unitarias de cada uno de los módulos y de sus funciones para asegurar el funcionamiento correcto del código desarrollado.

Posteriormente se han llevado a cabo pruebas de integración probando los flujos de la información completos, desde la entrada/solicitud de datos hasta la salida de los mismos, en todas las funcionalidades expuestas.

La validación de los resultados se ha llevado a cabo de forma automática con utilidades de carga, a partir de extractos de informes médicos completos junto con las terminologías clínicas a ser reconocidas. La validación del algoritmo experimental se ha hecho también de forma automática con utilidades de comprobación y ha consistido en revisar los términos identificados en base al texto de origen, y comprobar cuántos términos han sido correcta o erróneamente vinculados, y cuántos términos vinculables se han omitido.

Construcción del prototipo

En relación a los grupos de actividades de la construcción del prototipo, se ha seguido una metodología de desarrollo en cascada, seleccionando las herramientas más adecuadas, analizando los requisitos y funcionalidades a proporcionar por el prototipo, diseñando posteriormente la arquitectura del software correspondiente, para proceder a la construcción del prototipo y, finalmente, a las pruebas técnicas y de usuario de funcionamiento correcto.

Esta cuarta fase ha sido planificada en tres semanas, y se ha cumplido en el plazo programado.

El análisis y diseño se ha realizado conforme a la experiencia personal del autor en el desarrollo de software.

Se ha avanzado adecuadamente en la construcción, a partir del análisis y diseño previo y partiendo del motor de inferencia ya desarrollado. Adicionalmente, el autor tiene experiencia previa en el desarrollo de interfaces Web en *HTML/JS/CSS* para otros casos de uso.

Las dificultades encontradas principalmente han sido relativas al uso del lenguaje *HTML* y *Javascript*, así como a la interpretación del documento por parte de los navegadores web, específicamente las de errores de sintaxis, de código inerte (que no realiza la función esperada) y las de errores de colocación de los elementos en pantalla.

Con respecto a las pruebas y validación, la depuración de un programa desarrollado en *HTML/JS/CSS* como lenguaje funcional requiere tiempo, y un cuidado importante a la hora de desarrollar, ya que los errores no se detectan hasta la interpretación del documento y la ejecución del código *Javascript*.

Se han realizado inicialmente pruebas unitarias de cada una de las funciones para asegurar el funcionamiento correcto del código desarrollado. Posteriormente se han llevado a cabo pruebas de integración probando los flujos de la información completos,

desde la entrada/solicitud de datos hasta la salida de los mismos, en todas las funcionalidades expuestas.

Confección de la memoria del TFM

En relación a los grupos de actividades de la construcción de la memoria del TFM, se ha seguido una metodología de confección iterativa y progresiva.

Los diferentes apartados de la memoria se han ido redactando conforme se finalizaban las diferentes fases de desarrollo, de forma progresiva, pero refinando el contenido en varias iteraciones. Las primeras iteraciones más burdas únicamente servían para añadir el contenido a desarrollar a grandes rasgos, posteriormente se completaba la información, y, en otras iteraciones se añadían las referencias y se revisaba. En muchos casos ha sido necesario redactar de nuevo, corregir información incorrecta, y/o reestructurar apartados.

Esta quinta fase ha sido planificada en tres semanas, aunque en lugar de dejarla para el final, como se ha mencionado, se ha realizado en cada una de las fases previas, aprovechando algunos adelantos de tiempo. Esto es, el autor dedicó cuatro semanas previas a adelantar el trabajo de las dos primeras fases (conceptos generales y revisión del estado del arte). Algunas partes del desarrollo del motor partían de experiencias previas del autor por lo que tardaron menos tiempo en realizarse, así como algún enfoque con mayor complejidad fue descartado del motor por no dar resultados aceptables (uso de *Transformers*) lo que ahorró tiempo en la tercera fase de desarrollo del motor de inferencia.

La redacción de la introducción y objetivos, los conceptos generales, estado del arte, desarrollo/resultados del proyecto se realizó de forma progresiva como se ha mencionado.

La redacción de la conclusión, trabajos futuros y del resumen, además de las referencias, se realizó posteriormente al resto de apartados.

Las dificultades encontradas han sido principalmente las inherentes a las tareas de sintetizar, explicar de forma coherente y rigurosa la información y los resultados obtenidos, sin ambigüedades ni incoherencias con la realidad.

Duración del trabajo

La duración del trabajo total, que estaba prevista en 150 horas, ha requerido la dedicación hasta la redacción de la presente oferta de **210 horas** aproximadamente. La distribución de la dedicación por cada fase o grupo de tareas ha sido la siguiente.

Fase	*Duración Prevista (h)*	*Duración Real (h)*
Información sobre el TFM y planteamiento de idea original	*-*	*5*
Recopilación de conceptos generales	*20*	*30*
Revisión del estado del arte	*30*	*45*
Desarrollo del motor de inferencia	*40*	*60*
Construcción del prototipo	*30*	*30*
Redacción de la memoria	*30*	*40*
TOTAL	*150*	*210*

Tabla 1 Tabla de fases, duración prevista y real de la realización del TFM. Elaboración propia.

6.4. Resultados

Cuadernos de Google Colab

Se ha llevado a cabo el desarrollo y pruebas de los diferentes algoritmos y estrategias encontrados en la revisión del estado del arte para llevar a cabo la vinculación de los elementos de la terminología a los informes médicos, utilizando Google Colab y los cuadernos Jupyter.

A continuación, se indican los resultados obtenidos de las pruebas realizadas, aunque el nombre de los cuadernos utilizados junto con información más detallada sobre las actividades realizadas se encuentra en los Apéndices del actual documento.

Todo el código generado se ubica en el proyecto público con la siguiente ruta: https://github.com/fmedinafernandez/14MBID-TFM-NEL .

Después de analizar los resultados obtenidos en la ejecución de los cuadernos, se determina que la obtención de vectores densos mediante los modelos pre entrenados del PlanTL (*PlanTL-GOB-ES/Biomedical-Word-Embeddings-for-Spanish: Biomedical Word Embeddings Generated from Spanish Biomedical Corpora.*, n.d.), y el cálculo de similitud entre los vectores densos de grupos de palabras con significado parecido o igual (en el ámbito clínico en español), puede servir con el propósito de la experimentación para encontrar grupos de palabras similares entre sí.

Se constata que el cálculo de similitud entre los grupos de palabras se puede hacer mediante la medición del ángulo o mediante la medición de distancia euclidiana entre los vectores densos, aunque es necesario ajustar adecuadamente el umbral de valor de la similitud o distancia que se tiene en cuenta para que la vinculación sea adecuada.

El preprocesado del texto previo a la obtención de la representación densa debe consistir principalmente en la conversión a minúsculas, la eliminación de palabras irrelevantes, lematización de palabras, así como eliminación de tildes y de signos de puntuación.

El algoritmo con el que se experimenta y que puede dar los resultados deseados a la hora de vincular los términos clínicos a los informes consiste en lo siguiente: se itera secuencialmente sobre las palabras del texto obteniendo el vector denso de diferentes n-gramas y comparándolos con los vectores densos de los términos del catálogo para calcular la similitud. De esta manera se obtienen los términos clínicos candidatos a ser vinculados, que son más similares a los diferentes n-gramas del texto.

En este punto se empiezan a detectar limitaciones en la aproximación. La vinculación funciona bien para encontrar palabras similares en el texto y en el vocabulario para proceder al etiquetado, pero al ejecutar el algoritmo de vinculación se vinculan al texto etiquetas similares en contenido desde el punto de vista morfológico pero que no se corresponden realmente, por ejemplo, por ser de partes anatómicas diferentes del cuerpo. Por ejemplo, no es lo mismo un “carcinoma en el hígado” que un “carcinoma en la mama”, sin embargo, ante la palabra “carcinoma” en el texto se devuelven ambas etiquetas erróneamente. Esto se acentúa cuando se compara la similitud de palabras atómicas con grupos de palabras del vocabulario.

A título ilustrativo, mostramos a continuación el informe tipo utilizado, y los resultados de términos vinculados obtenidos.

“*Hay evidencia de hipoatenuación hepática difusa compatible con* ***infiltración grasa*** *.* *No hay dilatación de los conductos biliares intra o extrahepáticos.* *El paciente se encuentra en estado post* ***colecistectomía****.* *El bazo es normal.*

El páncreas es de contorno y características de atenuación normales.

*No hay evidencia de **masa suprarrenal**.*

*Hay una **hernia supraumbilical** de tamaño moderado que contiene grasa.*

Los riñones son normales en tamaño, forma y configuración.

*No se identifican **cálculos renales** ni **ureterales**.*

*No hay **hidrouréter** ni **hidronefrosis**.*

*No hay evidencia de **apendicitis**.*

*Hay varias asas de intestino delgado llenas de líquido, compatibles con una **enteritis** leve.*

*No hay **engrosamiento de la pared intestinal**.*

*No hay evidencia de **obstrucción del intestino delgado o grueso**.*

*No hay evidencia de **ascitis** abdominal o **linfadenopatía**.*

*No hay evidencia de **masa vesical** intrínseca o extrínseca .*

*No hay **ascitis** pélvica ni **linfadenopatía** .*

El útero y los ovarios no presentan ninguna anomalía.

*Las imágenes de las bases pulmonares no muestran evidencia de **masa pleural** o **parenquimatosa** .*

*No hay **derrames pleurales**.*

*Hay **cicatrices** en el lóbulo medio derecho y en la língula, así como en ambas bases pulmonares.*

Las estructuras óseas están libres de lesiones líticas o blásticas.

Se observan cambios degenerativos multinivel en la columna toracolumbar.

*Se observan **calcificaciones** dispersas en la aorta y en sus principales ramas, compatibles con la **aterosclerosis**."*

Fuente preprocesada	Término	Distancia/similitud
apendicitis	apendicitis	100%
ascitis	ascitis	100%
lingular base	bronquiectasias en ambas bases	92.4848%
cambio degenerativo multinivel	cambios degenerativos en columna	90.245600000000001%
cicatriz	cicatriz	100%
calculo renal	cálculo renal	100%
derrame pleural	derrame pleural	100%
dilatacion conducto biliar	dilatación del conducto de Wirsung	90.6955%
dilatacion conducto biliar intro	dilatación, incluso quística, de los ductos biliares	90.9952%
asa intestino delgado lleno	edemas en asas de intestino delgado	90.2751%
enteritis leve	escleritis leve	93.3063%
degenerativo multinivel	espondilolistesis degenerativa	90.984%
enteritis	gastroenteritis	90.5493%
hernia supraumbilical	hernia paraumbilical	98.5847%
hidronefrosis	hidronefrosis	100%
obstruccion intestino delgado	invaginación de intestino delgado	95.757%
lesion	lesiones	100%
lesion litico	lesiones líticas	100%
linfadenopatia	linfadenopatía	100%
masa vesical	masa prostática	91.0735%
masa suprarrenal	masa suprarrenal	100%
evidenciar masa	masa testicular	93.2371%
vesical intrinseco	mullerianosis vesical	90.2918%
evidenciar obstruccion intestino delgado	obstrucción del intestino delgado	96.3749%
obstruccion intestino	obstrucción intestinal	95.4534%

Tabla 2. Lista de términos clínicos vinculados al informe tipo proporcionado. Elaboración propia.

En la tabla se aprecia cómo el modelo no se comporta correctamente en algunas de las etiquetas vinculadas a algunas partes del texto.

Por otro lado, a la hora de obtener la representación densa de los grupos de palabras (n-gramas), se debe obtener un vector que representa la media de los vectores densos de las palabras aisladas (por ejemplo, para “abdomen distendido” se debe obtener la media entre el vector de “abdomen” y el vector de “distendido”).

Con respecto al modelo probado basado en *Transformers* para obtener la representación densa de las palabras (*BETO-Galen*) utilizando el heurístico propuesto los resultados obtenidos no han sido satisfactorios. Las pruebas realizadas con el informe tipo determinaron que se vinculaban erróneamente gran cantidad de términos, no encontrando ningún umbral adecuado de similitud coseno para minimizar este problema sin perder términos relevantes.

Se ha experimentado con varios modelos de la librería *Spacy* para las operaciones de procesado del texto y de separación del texto en sentencias concluyendo que el modelo más extenso o *large* denominado “es_core_news_lg” proporciona un mejor pre procesado del texto y el tiempo de carga es similar que los modelos más reducidos.

Se han obtenido diversas mediciones de tiempos, de umbrales de distancia/similitud y de proporción de vinculación correcta de términos con los 12 modelos disponibles dentro de los modelos basados en *FastText* mencionados previamente. Los parámetros que diferencian unos modelos de otros son el corpus con los que han sido entrenados (clínico o biomédico), la estrategia de *embedding* (a nivel de palabras o subpalabras), la tarea de aprendizaje (CBOW o Skip-gram) y el número de dimensiones de los vectores (50, 100 o 300). Por otro lado, la similitud coseno y la distancia euclidiana han sido utilizadas para obtener los resultados. Las conclusiones obtenidas son que los modelos con vectores con más dimensiones utilizando como medición la similitud coseno, son más efectivos al identificar los términos unívocamente, aunque no se ha apreciado gran diferencia en los tiempos en los diferentes modelos y enfoques, más allá de que los modelos más grandes tardan más tiempo en cargar por primera vez y que la medición de la distancia euclidiana es más rápida de calcular que la medición de la similitud coseno.

La similitud coseno se basa en medir la dirección del vector calculando el coseno del ángulo de dos vectores, y esto es muy útil cuando se trabaja con vectores de alta dimensionalidad porque no depende de la magnitud sino en la dirección de estos (*9 Medidas de Distancia En Ciencia de Datos*, n.d.). Por otro lado, la distancia euclidiana mide la distancia entre dos puntos, pero no es adecuada en entornos de alta dimensionalidad porque puede dar resultados erróneos (Aggarwal et al., 2001).

Desarrollo del motor de inferencia

La infraestructura tecnológica del motor, es un entorno Python 3.8, con múltiples librerías para la funcionalidad perseguida, con base de datos *MySQL Community Server 8.0* para la persistencia de datos.

Los detalles concretos se incluyen en los Apéndices.

- **Vocabulario o terminología clínica disponible**

El vocabulario que se ha cargado para probar el correcto funcionamiento del motor ha sido el conjunto de referencias en español de "Problemas de salud en atención hospitalaria" (https://snomedsns.es), de la extensión para España del SNS del 01/12/2022, que contiene 3.678 términos clínicos de SNOMED-CT, agrupados para un propósito concreto.

El conjunto de referencias mencionado de SNOMED-CT es un catálogo apropiado para codificar los diagnósticos de los informes médicos de pruebas de imagen del ámbito hospitalario, dado que contienen habitualmente referencias a problemas de salud.

Adicionalmente para probar el correcto funcionamiento del motor y sobre todo, para poder validar los resultados se han cargado los términos clínicos etiquetados del corpus DisTEMIST (*DisTEMIST Corpus: Detection and Normalization of Disease Mentions in Spanish Clinical Cases* | *Zenodo*, n.d.).

- **Lógica de vinculación de entidades**

El motor contiene la lógica clave para conseguir uno de los objetivos del TFM: experimentar con un modelo para etiquetar los informes con términos relacionados en base a una terminología clínica. Para ello a continuación se indican de forma esquemática las operaciones realizadas para conseguirlo. Posteriormente, en otros módulos se detallará algunas de las operaciones mencionadas de forma concreta.

Los pasos que se llevan a cabo para procesar un informe son los siguientes:

1. Se cargan en memoria todos los términos clínicos del catálogo, junto con el *embedding* de cada término (transformando los *embeddings* de formato texto *json* a lista de números decimales para poder ser procesados).
2. Se divide el texto en frases o sentencias para su proceso independiente.
   Por cada frase o sentencia:
   2.1. Se pre procesa el texto de la frase.
   2.2. Se establecen diferentes agrupaciones de la frase en n-gramas, empezando por grupos de 6 palabras (posteriormente 5, 4… hasta una única palabra).
   Por cada grupo de palabras:

2.2.1. Se obtiene el *embedding* correspondiente a dicho grupo de palabras utilizando el modelo disponible.
2.2.2. Se itera por todos los términos clínicos disponibles.
Por cada término clínico del mismo tamaño que el grupo de palabras:
2.2.2.1. Se compara la similitud del *embedding* del término clínico con respecto al *embedding* del grupo de palabras correspondiente.
2.2.2.2. Si la similitud es mayor que el umbral determinado, se guarda el término clínico, su similitud y el modelo utilizado como término candidato para ser vinculado.
2.2.3. De todos los términos candidatos obtenidos en la iteración, en la búsqueda del catálogo, se descartan aquellos menos similares y se guarda el de mayor similitud únicamente, como término a vincular.

3. Del resultado de todos los términos a vincular obtenidos, se eliminan todos los duplicados con menor similitud.

El resultado que se devuelve es el de términos a vincular obtenidos, que incluyen la porción del texto origen, el identificador del término, la descripción del término, la similitud obtenida y el identificador del modelo utilizado.

- **Validación de los resultados**

Para validar los resultados del motor de inferencia se han llevado a cabo las siguientes acciones.

En primer lugar, se han cargado y tenido en cuenta para la validación los términos clínicos de SNOMED-CT vinculados en DisTEMIST. Para la carga se ha desarrollado una herramienta a medida para cargar los datos en lotes utilizando el motor de inferencia.

En segundo lugar, se han cargado los informes del corpus DisTEMIST. DisTEMIST es una colección de 1000 casos clínicos con anotaciones de enfermedades con conceptos SNOMED-CT. Para la carga se ha desarrollado una herramienta a medida para cargar los datos en lotes utilizando el motor de inferencia.

En tercer lugar, se han descargado los datos de los informes vinculados en la base de datos del prototipo, y mediante una herramienta hecha a medida se han validado los datos.

La validación ha consistido en comparar los términos clínicos vinculados mediante el prototipo con respecto a los términos clínicos vinculados en DisTEMIST. Se han identificado cuántos se han etiquetado correctamente, cuántos de forma incorrecta y cuántos términos no han sido etiquetados.

A continuación, se muestran los resultados de forma agregada. En los Apéndices se dispone de los datos concretos para cada informe.

Informes	Correctos	Incorrectos	Faltantes	Términos etiquetados	Términos DISTEMIST
580	5,54	8,05	3,27	13,59	8,81

Tabla 3. Resultados agregados en números absolutos de la validación de los datos. Elaboración propia.

Informes: Esta columna nuestra el número de informes utilizados para validar los resultados.

Correctos: Promedio de etiquetas correctamente vinculadas en relación con los datos de DISTEMIST.

Incorrectos: Promedio de etiquetas incorrectamente vinculadas en relación con los datos de DISTEMIST, es decir, que son etiquetas que no pertenecen al informe.

Faltantes: Promedio de etiquetas no vinculadas en relación con los datos de DISTEMIST, es decir, que son etiquetas que no se han detectado en el informe.

Informes	%Correctos sobre DISTEMIST	%Faltantes sobre DISTEMIST	%Incorrectos sobre etiquetados
580	64%	36%	58%

Tabla 4. Resultados agregados porcentuales de la validación de los datos. Elaboración propia.

Informes: Esta columna nuestra el número de informes utilizados para validar los resultados.

%Correctos sobre Distemist: Porcentaje promedio de etiquetas correctamente vinculadas con respecto al total de etiquetas de DISTEMIST.

%Faltantes sobre Distemist: Porcentaje promedio de etiquetas que no han sido vinculadas con respecto al total de etiquetas de DISTEMIST.

%Incorrectos sobre etiquetados: Porcentaje promedio de etiquetas vinculadas incorrectamente con respecto al total de etiquetas vinculadas por el prototipo.

De los datos anteriores se observa como el prototipo con el algoritmo indicado vincula correctamente el 64% de las etiquetas esperadas y se vinculan erróneamente casi seis de cada diez etiquetas propuestas. Podemos concluir como punto positivo que el prototipo vincula un porcentaje alto de las etiquetas esperadas. Como punto negativo, el prototipo vincula también un porcentaje alto de etiquetas que no tienen relación con el informe.

Esto es debido a la ambigüedad de las palabras, al uso de un modelo que no tiene en cuenta el contexto que rodea a dichas palabras y al ajuste del heurístico propuesto.

Este heurístico debe balancear un grado de vinculación adecuado: si este es muy alto existe el riesgo de vincular más términos incorrectos y si este es muy bajo existe el riesgo de vincular menos términos correctos.

Debido a lo que se ha mencionado previamente el resultado es el esperado de forma realista, especialmente una vez el autor ha ido profundizando en el conocimiento de PLN para resolver este problema.

Desarrollo del interfaz de usuario prototipo de vinculación de informes médicos

El interfaz de usuario se ha desarrollado como una aplicación web alojada en un servidor de aplicaciones.

El lenguaje de maquetado es *HTML.* Se utilizan hojas de estilo *CSS* para gestionar la apariencia y código *Javascript* para que el interfaz sea interactivo y se integre con el motor de integración. El servidor de aplicaciones se ha implementado utilizando una librería desarrollada en *Python*.

Las funcionalidades principales proporcionadas por el prototipo son la vinculación de términos clínicos a los informes médicos y la localización posterior de dichos informes a través de dichas etiquetas que permiten cumplir el objetivo principal del presente TFM. No obstante, de forma ilustrativa se han incluido algunas otras funcionalidades adicionales como la creación, consulta y filtrado de términos clínicos.

Ilustración 10. Interfaz de usuario del prototipo. Elaboración propia.

Todos los detalles de este y de los anteriores módulos están incluidos en los Apéndices de la presente memoria:

- Apéndice I. Cuadernos de *Google Colab*.
- Apéndice II. Desarrollo del motor de inferencia.
- Apéndice III. Base de Datos.
- Apéndice IV. Desarrollo del interfaz de usuario prototipo de vinculación de informes médicos.
- Apéndice V. Validación de resultados DISTEMIST.

7. Conclusión y trabajos futuros

El campo del PLN ha sufrido muchos avances recientes, con la exploración y uso de las técnicas avanzadas de aprendizaje profundo.

El uso del PLN en el ámbito médico se justifica por sí mismo. El objetivo final es asistir a los profesionales del ámbito clínico en la toma de mejores decisiones. Estas decisiones redundarán en la calidad de la asistencia y eficacia de las acciones, para mejorar la salud de las personas usuarias.

Existe mucha literatura, artículos e iniciativas realizadas al respecto, además de específicamente en el ámbito clínico en idioma español.

Queda demostrado que es posible utilizar técnicas de PLN y modelos preentrenados para el ámbito clínico en español. Con ello es posible crear una herramienta y experimentar con el etiquetado de los informes médicos en texto libre con diferentes terminologías clínicas.

En este caso se ha experimentado con modelos de lenguaje natural en español para preprocesar los textos y generar vectores densos. Con los vectores se ha calculado la similitud semántica entre las palabras de los textos y las terminologías proporcionadas, midiendo la distancia euclidiana o similitud coseno entre ellos.

Este algoritmo ha dejado patente que la utilización de *embeddings* es una técnica necesaria para vincular términos de los informes. Sin embargo, esta técnica no es suficiente por sí sola. Se han encontrado muchos términos erróneamente enlazados, especialmente al gestionar monogramas (la misma patología puede darse en diferentes partes anatómicas: “carcinoma de mama” vs “carcinoma de vulva”). Es necesario un postprocesado posterior de los términos a vincular y una mejora del algoritmo para evitar este problema.

También se ha probado cómo es posible aplicar dichos modelos encapsulándolos en un motor de inferencia accesible mediante un *API REST* de manera que pueda ser utilizada desde cualquier aplicación clínica que necesite aplicar este etiquetado.

Finalmente se ha desarrollado un pequeño prototipo de aplicación médica para poder realizar dicho etiquetado a partir de un interfaz de usuario sencillo demostrador.

Las dificultades encontradas en el desarrollo de la investigación han sido:

1. La identificación y validación de técnicas y modelo existentes ya entrenados en el ámbito médico en español, con el requisito de la búsqueda de un planteamiento ágil que permita una respuesta interactiva a los usuarios sin disponer de acceso a los datos reales y que sea realizable en un marco temporal muy acotado.

2. La validación de los resultados obtenidos sin la participación de personal facultativo para determinar la idoneidad de los resultados en el uso cotidiano.

Como trabajo futuro previsto el autor tiene como objetivo implementar el motor de inferencia en un entorno producto real e integrarlo junto con un sistema de información radiológico RIS para el etiquetado y localización posterior de informes médicos. El RIS contiene informes de pruebas de imagen reales. El producto se denomina EOS (https://www.babelgroup.com), es propiedad de la empresa BABEL Sistemas de Información SL y está actualmente implantado en 13 hospitales del ámbito del Servicio de Salud del Principado de Asturias (SESPA) en España.

Algunos otros trabajos futuros propuestos relacionados con este tema son los siguientes:

- Validar con usuarios facultativos reales el algoritmo completo de procesamiento del texto médico, como, por ejemplo, la necesidad real de normalizar los datos filtrando términos clínicos similares entre sí o la necesidad de trabajar a nivel de grupos de palabras o de frases completas.
- Validar con usuarios facultativos reales los datos de parametrización del algoritmo como la distancia de similitud apropiada y el número máximo de palabras conjuntas a procesar.
- Validar el motor de inferencia y sus resultados con los informes médicos reales del SESPA.
- Ampliar el algoritmo actual añadiendo la gestión de abreviaturas, los términos negados o las expresiones de incertidumbre en el texto.
- Ampliar el catálogo actual utilizado al requerido por los usuarios finales, por ejemplo, añadir otros datos además de las patologías hospitalarias como alergias, antecedentes del paciente, o procedimientos realizados.
- Entrenar y/o utilizar un modelo basado en *Transformers*, para realizar la tarea de etiquetado (clasificación multi etiqueta), con datos reales de informe médicos reales del SESPA convenientemente etiquetados comprobar los resultados obtenidos con respecto al planteamiento actual.

8. Agradecimientos

Agradecimiento especial a mi familia y amigos por todo el tiempo que les habría podido dedicar y que tendré que compensar cuando finalice el presente TFM.

De forma especial a mi tutor Román Igual Pérez, en primer lugar, por aceptar este reto, y después por su ayuda, su seguimiento constante, interés y buenos consejos que me ha brindado.

Al director y a todos los profesores del Máster por los conocimientos, experiencia, dedicación e interés que han compartido con nosotros en este camino.

Al resto de compañeros que han estado ahí para compartir los esfuerzos y las dificultades.

Finalmente, a mis compañeros de trabajo del proyecto de Soporte de Imagen Digital de BABEL y a las personas de la Subdirección de Infraestructuras y Servicios Técnicos del SESPA que conocen este trabajo, creen en que tiene sentido dedicar este esfuerzo y han hecho aportaciones al mismo.

9. Referencias

9 medidas de distancia en ciencia de datos. (n.d.). Retrieved March 13, 2023, from https://ichi.pro/es/9-medidas-de-distancia-en-ciencia-de-datos-159983401462266

10 years of Siri: the history of Apple's voice assistant | TechRadar. (n.d.). Retrieved February 26, 2023, from https://www.techradar.com/news/siri-10-year-anniversary

[1508.07909v5] Neural Machine Translation of Rare Words with Subword Units. (n.d.). Retrieved February 26, 2023, from https://arxiv.org/abs/1508.07909v5

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1973*, 420–434. https://doi.org/10.1007/3-540-44503-X_27/COVER

Allender, Dr. H. (2021). *Procesamiento del lenguaje natural en medicina: clasificación automática de notas clínicas - YouTube*. [Vídeo] YouTube. https://www.youtube.com/watch?app=desktop&v=czqe8tP70pc

Almagro, M. (2020). ICD-10 coding based on semantic distance: LSI UNED at CLEF eHealth 2020 Task 1. *CLEF EHealth 2020*.

Amazon Gets Into Voice Recognition, Buys Ivona Software To Compete Against Apple's Siri | TechCrunch. (n.d.). Retrieved February 26, 2023, from https://techcrunch.com/2013/01/24/amazon-gets-into-voice-recognition-buys-ivona-software-to-compete-against-apples-siri/?guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAEC_kf8XB65f32bCK-KmPM10NFbP9fJcuuP6qsaI1TkahkQDvXwbHzv_AVPwt4oKx-qJfR4imRt9HK00RcXDQo2GxTJBHtd0v0QqxQPn7W_YvIKdIc38BWjyTIL-zEq1LzdPfq5hsL-HIzDAKdmcwETGDSihwssE8Tw_6XKZSeiA&guccounter=2

Así era ELIZA, el primer bot conversacional de la historia. (n.d.). Retrieved February 26, 2023, from https://www.xataka.com/historia-tecnologica/asi-era-eliza-el-primer-bot-conversacional-de-la-historia

Biografia de Noam Chomsky. (n.d.). Retrieved February 26, 2023, from https://www.biografiasyvidas.com/biografia/c/chomsky.htm

Biografia de Zellig Harris. (n.d.). Retrieved February 26, 2023, from https://www.biografiasyvidas.com/biografia/h/harris.htm

Blanco, A., Pérez, A., & Casillas, A. (2020). IXA-AAA at CLEF eHealth 2020 CodiEsp Automatic classification of medical records with Multi-label Classifiers and Similarity Match Coders. *CLEF EHealth 2020*.

BOE.es - BOE-A-2010-14199 Real Decreto 1093/2010, de 3 de septiembre, por el que se aprueba el conjunto mínimo de datos de los informes clínicos en el Sistema Nacional de Salud. (n.d.). Retrieved December 6, 2022, from https://www.boe.es/buscar/doc.php?id=BOE-A-2010-14199

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.48550/arxiv.1607.04606

Cã, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (n.d.). *SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA*. Retrieved February 26, 2023, from https://github.com/josecannete/spanish-corpora

CPT - CPT Codes - Current Procedural Terminology - AAPC. (n.d.). Retrieved February 26, 2023, from https://www.aapc.com/resources/medical-coding/cpt.aspx

Descriptores en Ciencias de la Salud - Wikipedia, la enciclopedia libre. (n.d.). Retrieved February 26, 2023, from https://es.wikipedia.org/wiki/Descriptores_en_Ciencias_de_la_Salud

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://doi.org/10.48550/arxiv.1810.04805

DisTEMIST corpus: detection and normalization of disease mentions in spanish clinical cases | Zenodo. (n.d.). Retrieved March 14, 2023, from https://zenodo.org/record/7614764#.ZBBGcnbMIuV

Donnelly, L. F., Grzeszczuk, R., & Guimaraes, C. v. (2022). Use of Natural Language Processing (NLP) in Evaluation of Radiology Reports: An Update on Applications and Technology Advances. *Seminars in Ultrasound, CT and MRI*, *43*(2), 176–181. https://doi.org/10.1053/J.SULT.2022.02.007

eCIE-Maps - CIE-10-ES Diagnósticos. (n.d.). Retrieved February 26, 2023, from https://eciemaps.mscbs.gob.es/ecieMaps/browser/metabuscador.html

Ferrando Javier. (n.d.). *Mining Unstructured Data - 8. Word Embeddings*. Universidad Politécnica de Cataluña BarcelonaTech - Máster de Data Science. Retrieved February 26, 2023, from https://www.cs.upc.edu/~turmo/mud/lectures/08-Word-embeddings.pdf

Firth J.R. (1957). *A Synopsis of Linguistic Theory*. Studies in Linguistic Analysis (Pp. 1-31). Special Volume of the Philological Society. Oxford Blackwell. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1846447

Frege, Contextuality and Compositionality on JSTOR. (n.d.). Retrieved February 26, 2023, from https://www.jstor.org/stable/40180264

García-Santa, N., & Cetina, K. (2020). FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding. *CLEF EHealth 2020*. http://www.fujitsu.com/emea/about/fle/

Gensim: Topic modelling for humans. (n.d.). Retrieved March 14, 2023, from https://radimrehurek.com/gensim/

GitHub. (n.d.). Retrieved March 14, 2023, from https://github.com/

González, R. (2007). EL TEST DE TURING: DOS MITOS, UN DOGMA. *Revista de Filosofía*, *63*, 37–53. https://doi.org/10.4067/S0718-43602007000100003

guilopgar/ClinicalCodingTransformerES: Clinical Coding in Spanish using Transformers. (n.d.). Retrieved March 14, 2023, from https://github.com/guilopgar/ClinicalCodingTransformerES

Harris, Z. S. (2015). Distributional Structure. *Http://Dx.Doi.Org/10.1080/00437956.1954.11659520*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

http.server — HTTP servers — Python 3.11.2 documentation. (n.d.). Retrieved March 14, 2023, from https://docs.python.org/3/library/http.server.html

IBM Archives: 701 Translator. (1954). IBM Archives. https://www.ibm.com/ibm/history/exhibits/701/701_translator.html

IBM Watson. (n.d.). Retrieved February 26, 2023, from https://www.ibm.com/es-es/watson

Informatics Standard Terminology | American College of Radiology | American College of Radiology. (n.d.). Retrieved February 26, 2023, from https://www.acr.org/Practice-Management-Quality-Informatics/Informatics/Terminology

International Classification of Diseases (ICD). (n.d.). Retrieved February 26, 2023, from https://www.who.int/classifications/classification-of-diseases

Joseph. (2013). *The Lingua File from TLF Translation: The ALPAC Report: The Failings of Machine Translation*. TheLinguaFile. http://www.thelinguafile.com/2013/11/the-alpac-report-failings-of-machine.html#.Y_sQsHbMJPZ

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., … Robinson, P. N. (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, *49*(D1), D1207–D1217. https://doi.org/10.1093/NAR/GKAA1043

Krauthammer, M., & Nenadic, G. (2004). *Term identification in the biomedical literature*. https://doi.org/10.1016/j.jbi.2004.08.004

Lopez-Garcia, G., Jerez, J. M., Ribelles, N., Alba, E., & Veredas, F. J. (2021). Transformers for Clinical Coding in Spanish. *IEEE Access*, *9*, 72387–72397. https://doi.org/10.1109/ACCESS.2021.3080085

López Rubio E. (2019). *Curso de Inteligencia Artificial Aplicada. Procesamiento del lenguaje natural. 07 Represe densa docs - YouTube*. [Vídeo] YouTube. https://www.youtube.com/watch?app=desktop&v=QJ7HpatxXgI&list=PLxdfokX9ltve7G09uBtl3wP_Mc4lMx67W&index=19

López Rubio, E. (2019a). *Curso de Inteligencia Artificial Aplicada. Procesamiento del lenguaje natural. 08 Redes Realimentada - YouTube*. [Vídeo] YouTube. https://www.youtube.com/watch?v=eJV79JnUMS4&list=PLxdfokX9ltve7G09uBtl3wP_Mc4lMx67W&index=21

López Rubio, E. (2019b, November 27). *Curso de Inteligencia Artificial Aplicada. Procesamiento del lenguaje natural. 01 Presentación - YouTube*. [Vídeo] YouTube. https://www.youtube.com/watch?v=GqUoIIchthE&list=PLxdfokX9ltve7G09uBtl3wP_Mc4lMx67W&index=14

López Rubio, E. (2019c, November 28). *Curso de Inteligencia Artificial Aplicada. Procesamiento del lenguaje natural. 03 Historia - YouTube*. [Vídeo] YouTube. https://www.youtube.com/watch?v=X6qIbRsnKUs

López Rubio, E. (2019d, November 28). *Curso de Inteligencia Artificial Aplicada. Procesamiento del lenguaje natural. 04 Conceptos básicos - YouTube*. [Vídeo] YouTube. https://www.youtube.com/watch?v=IEKTHyXn9rc&list=PLxdfokX9ltve7G09uBtl3wP_Mc4lMx67W&index=16

López Rubio, E. (2019e, November 28). *Curso de Inteligencia Artificial Aplicada. Procesamiento del lenguaje natural. 05 Rep dispersa docs - YouTube*. [Vídeo] YouTube. https://www.youtube.com/watch?v=Bhm0ampwsaE&list=PLxdfokX9ltve7G09uBtl3wP_Mc4lMx67W&index=18

López Rubio, E. (2019f, November 28). *Curso de Inteligencia Artificial Aplicada. Procesamiento del lenguaje natural. 06 Redes prealimentad - YouTube*. [Vídeo] YouTube. https://www.youtube.com/watch?v=mN1ML33ERrA&list=PLxdfokX9ltve7G09uBtl3wP_Mc4lMx67W&index=20

MarIA: Spanish Language Models | Gutiérrez-Fandiño | Procesamiento del Lenguaje Natural. (n.d.). Retrieved February 26, 2023, from http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405

Martí-Bonmatí, L., Alberich-Bayarri, & Torregrosa, A. (2022). El informe radiológico. Estructura, estilo y contenido. *Radiología*, *64*, 186–193. https://doi.org/10.1016/J.RX.2022.01.013

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://doi.org/10.48550/arxiv.1301.3781

Ministerio de Sanidad - Profesionales - Preguntas frecuentes sobre SNOMED CT. (n.d.). Retrieved February 26, 2023, from https://www.sanidad.gob.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/preguntas.htm

Ministerio de Sanidad - Profesionales - SNOMED CT. (n.d.). Retrieved February 26, 2023, from https://www.sanidad.gob.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/home.htm

Miranda-Escalada, A., Farré, E., Gasco, L., Lima, S., & Krallinger, M. (2022). *DisTEMIST corpus: detection and normalization of disease mentions in spanish clinical cases*. https://doi.org/10.5281/ZENODO.7614764

MySQL. (n.d.). Retrieved March 14, 2023, from https://www.mysql.com/

Navegador SNOMED CT SNS. (n.d.). Retrieved March 14, 2023, from https://snomedsns.es/

Palomar, M. (2006). *Tecnologías del Lenguaje aplicadas al aprendizaje de segundas lenguas*. Universidad de Alicante. http://www.artic.ua.es/sites/u38/sitio171/PresentacionEducacion.pdf

Patel, K., & Bhattacharyya, P. (2017). Towards Lower Bounds on Number of Dimensions for Word Embeddings. *Proceedings of the The 8th International Joint Conference on Natural Language Processing, Taipei, Taiwan AFNLP*, 31–36. https://aclanthology.org/I17-2006.pdf

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543. https://doi.org/10.3115/V1/D14-1162

Plan de Tecnologías del Lenguaje - Gobierno de España. (n.d.). Plan TL. Retrieved February 28, 2023, from https://github.com/PlanTL-GOB-ES

Plan de Tecnologías del Lenguaje - Página principal del Plan de Impulso de las tecnologías del Lenguaje. (n.d.). Retrieved March 13, 2023, from https://plantl.mineco.gob.es/Paginas/index.aspx

Plan TL España. (2020). *CodiEsp: Clinical Case Coding in Spanish Shared Task*. EHealth CLEF 2020. https://temu.bsc.es/codiesp/

PlanTL-GOB-ES/Biomedical-Word-Embeddings-for-Spanish: Biomedical Word embeddings generated from Spanish Biomedical corpora. (n.d.). Retrieved March 14, 2023, from https://github.com/PlanTL-GOB-ES/Biomedical-Word-Embeddings-for-Spanish

Postman API Platform | Sign Up for Free. (n.d.). Retrieved March 14, 2023, from https://www.postman.com/

Proyecto de Real Decreto por el que se modifica el Real Decreto 1093/2010, de 3 de septiembre, por el que se aprueba el conjunto mínimo de datos de los informes clínicos en el Sistema Nacional de Salud. (n.d.). Retrieved February 26, 2023, from https://www.sanidad.gob.es/normativa/audiencia/docs/DG_57_22_PRD_modificacion_RD_1093_2010.pdf

RadLex Term Browser. (n.d.). Retrieved February 26, 2023, from https://radlex.org/

Reyes-Aguillón, J., del Moral, R., Ramos-Flores, O., Gómez-Adorno, H., & Bel-Enguix, G. (2022). *Clinical Named Entity Recognition and Linking using BERT in Combination with Spanish Medical Embeddings*. http://ceur-ws.org

Sistema de información radiológico EOS del Servicio de Salud del Principado de Asturias - BABEL Sistemas de Información. (n.d.). Retrieved February 26, 2023, from https://www.babelgroup.com/es/soluciones/EOS

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R. Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., & Catanzaro, B. (2022). *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model*. https://doi.org/10.48550/arxiv.2201.11990

spaCy · Industrial-strength Natural Language Processing in Python. (n.d.). Retrieved March 14, 2023, from https://spacy.io/

Te damos la bienvenida a Colaboratory - Colaboratory. (n.d.). Retrieved March 14, 2023, from https://colab.research.google.com/

Unified Medical Language System (UMLS). (n.d.). Retrieved February 26, 2023, from https://www.nlm.nih.gov/research/umls/index.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999–6009. https://doi.org/10.48550/arxiv.1706.03762

Vatsal P. (2021, July 21). *Word2Vec Explained. Explaining the Intuition of Word2Vec &… | by Vatsal | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/word2vec-explained-49c52b4ccb71

What is Microsoft Cortana? Everything You Need to Know. (n.d.). Retrieved February 26, 2023, from https://www.techtarget.com/searchenterprisedesktop/definition/Cortana

Apéndice I

Cuadernos de Google Colab

Se ha llevado a cabo el desarrollo y pruebas de los diferentes algoritmos y estrategias encontrados en la revisión del estado del arte para llevar a cabo la vinculación de los elementos de la terminología a los informes médicos, utilizando Google Colab (https://colab.research.google.com/) y los cuadernos Jupyter (https://jupyter.org/).

Se han desarrollado, entre otros, varios cuadernos específicos para comprobar los resultados y los tiempos de los diferentes modelos disponibles de obtención de vectores densos, en el ámbito clínico/biomédico en español. En unos casos, se han utilizado los modelos que trabajan a nivel de palabras completas (*Word embedding*), y en otros casos, de los modelos elegidos se han tomado los modelos basados en cálculo de vectores densos a partir de subpalabras con codificación *Byte-Pair Encoding* (*BPE subwords*).

Se han cargado seis modelos con cada tipo (a nivel de palabras o de subpalabras) y datos de entrenamiento (textos biomédicos y textos clínicos) a partir de los modelos que se han entrenado mediante textos clínicos, en sus variantes de entrenamiento *CBOW* y *Skip-gram*, con generación de vectores densos de 50, 100 y 300 dimensiones.

Después se ha ejecutado con cada modelo el algoritmo de obtención de términos clínicos relacionados para un informe radiológico tipo con el catálogo precargado previamente. De estas operaciones se han tomado tiempos y se han comparado los resultados con los esperados para dicho informe médico.

Se destaca que para la similitud entre palabras o grupo de palabras se han utilizado dos tipos de medidas, la distancia euclidiana y la similitud coseno.

Las frases y etiquetas del informe tipo utilizado son las siguientes:

"*Hay evidencia de hipoatenuación hepática difusa compatible con* ***infiltración grasa*** .

No hay dilatación de los conductos biliares intra o extrahepáticos.

El paciente se encuentra en estado post ***colecistectomía****.*

El bazo es normal.

El páncreas es de contorno y características de atenuación normales.

No hay evidencia de ***masa suprarrenal****.*

Hay una ***hernia supraumbilical*** *de tamaño moderado que contiene grasa.*

Los riñones son normales en tamaño, forma y configuración.

*No se identifican **cálculos renales** ni **ureterales**.*

*No hay **hidrouréter** ni **hidronefrosis**.*

*No hay evidencia de **apendicitis**.*

*Hay varias asas de intestino delgado llenas de líquido, compatibles con una **enteritis** leve.*

*No hay **engrosamiento de la pared intestinal**.*

*No hay evidencia de **obstrucción del intestino delgado o grueso**.*

*No hay evidencia de **ascitis** abdominal o **linfadenopatía**.*

*No hay evidencia de **masa vesical** intrínseca o extrínseca .*

*No hay **ascitis** pélvica ni **linfadenopatía** .*

El útero y los ovarios no presentan ninguna anomalía.

*Las imágenes de las bases pulmonares no muestran evidencia de **masa pleural** o **parenquimatosa** .*

*No hay **derrames pleurales**.*

*Hay **cicatrices** en el lóbulo medio derecho y en la língula, así como en ambas bases pulmonares.*

Las estructuras óseas están libres de lesiones líticas o blásticas.

Se observan cambios degenerativos multinivel en la columna toracolumbar.

*Se observan **calcificaciones** dispersas en la aorta y en sus principales ramas, compatibles con la **aterosclerosis**."*

Los resultados obtenidos de la ejecución del prototipo con este informe tipo son las siguientes etiquetas:

Fuente preprocesada	Término	Distancia/similitud
apendicitis	apendicitis	100%
ascitis	ascitis	100%
lingular base	bronquiectasias en ambas bases	92.4848%
cambio degenerativo multinivel	cambios degenerativos en columna	90.24560000000001%
cicatriz	cicatriz	100%
calculo renal	cálculo renal	100%
derrame pleural	derrame pleural	100%
dilatacion conducto biliar	dilatación del conducto de Wirsung	90.6955%
dilatacion conducto biliar intro	dilatación, incluso quística, de los ductos biliares	90.9952%
asa intestino delgado lleno	edemas en asas de intestino delgado	90.2751%
enteritis leve	escleritis leve	93.3063%
degenerativo multinivel	espondilolistesis degenerativa	90.984%
enteritis	gastroenteritis	90.5493%
hernia supraumbilical	hernia paraumbilical	98.5847%
hidronefrosis	hidronefrosis	100%
obstruccion intestino delgado	invaginación de intestino delgado	95.757%
lesion	lesiones	100%
lesion litico	lesiones líticas	100%
linfadenopatia	linfadenopatía	100%
masa vesical	masa prostática	91.0735%
masa suprarrenal	masa suprarrenal	100%
evidenciar masa	masa testicular	93.2371%
vesical intrinseco	mullerianosis vesical	90.2918%
evidenciar obstruccion intestino delgado	obstrucción del intestino delgado	96.3749%
obstruccion intestino	obstrucción intestinal	95.4534%

Tabla 5. Resultados de etiquetas vinculadas para el informe tipo. Elaboración propia mediante el interfaz de usuario del prototipo.

La ruta de acceso a los apéndices detallados es: https://github.com/fmedinafernandez/14MBID-TFM-NEL/blob/main/14MBID_TFM_Medina_Fernandez_Fernando_Ap%C3%A9ndices.pdf

Los cuadernos de Google Colab están incluidos en el repositorio de *github* correspondiente: https://github.com/fmedinafernandez/14MBID-TFM-NEL

Apéndice II

Desarrollo del motor de inferencia

Como se ha mencionado previamente la infraestructura tecnológica del motor, es un entorno Python 3.8, con múltiples librerías para la funcionalidad perseguida, con base de datos *MySQL Community Server 8.0* para la persistencia de datos.

La arquitectura del software está basada en capas (multicapa), separando y agrupando las funciones en cada una de ellas:

1) Capa de entrada (API REST).
2) Capa de gestores (*managers*).
3) Capa de acceso a datos.
4) Capa general de utilidades.

A continuación, se extracta algunas de las partes más relevantes del motor de inferencia, como son el interfaz o capa de entrada (API REST), y el módulo con las funciones de NLP utilizadas.

- **Capa de entrada (API REST)**

La capa de entrada contiene un módulo con funciones que gestionan las operaciones de solicitud y entrega de información a través de un *API REST*, encargándose únicamente de la validación de los datos de entrada, y de la sesión de los usuarios. Esta capa utiliza la siguiente capa, de gestores, para ejecutar acciones u obtener la información solicitada.

El módulo se denomina *ClinicalNELEngineREST.py*, y las operaciones implementadas actualmente se indican a continuación.

operación	URL	parámetros	acción	retorno
GET	/application	user	Obtener la versión del motor.	Versión del motor.
GET	/users	user, password	Validar al usuario (contra directorio activo).	Token de acceso.
PUT	/terminosclinicos	user, token, idtermino, termino	Alta de nuevo término clínico.	Término clínico creado.
POST	/terminosclinicos	user, token	Inicialización y carga del catálogo de términos clínicos.	-
GET	/terminosclinicos	user, token, id	Consultar los térmi-	Términos

		(opcional), termino (opcional)	nos clínicos disponibles en el motor.	clínicos.
PUT	/reports	user, token, reportid, informe (*body*)	Creación de nuevo informe clínico.	Términos clínicos relacionados.
GET	/reports	user, token, idtermino (opcional), report_id (opcional), owner (opcionesl)	Consulta de los informes disponibles o filtrados por un término clínico o un identificador.	Informes clínicos.

Tabla 6. Operaciones implementadas en el motor de inferencia desarrollado. Elaboración propia.

- **Módulo de PLN**

Algunas de las funciones implementadas en este módulo son las siguientes:

loadmodel_NLP

Carga del modelo utilizado por la librería *spacy*, en este caso el modelo "*es_core_news_lg*" (https://spacy.io/models/es).

loadmodel_FastText

Carga del modelo, en este caso el modelo "*Biomedical-Word-Embedding-for-Spanish*" (https://github.com/PlanTL-GOB-ES/Biomedical-Word-Embeddings-for-Spanish), en este caso un modelo que genera *FastText embeddings* de 300 dimensiones, entrenado mediante *CBOW*, en base a casos clínicos.

preprocesaDoc

Función que transforma el texto utilizando para ello la librería *Spacy*, el modelo cargado previamente de esta librería, con el fin de realizar un proceso que lo prepare para las operaciones posteriores.

Las operaciones de preprocesado realizadas son las siguientes:

- Conversión del texto a minúsculas (independiente del caso).
- Tokenizado del texto en palabras.
- Eliminación de signos de puntuación del texto.
- Eliminación de palabras irrelevantes (*stopwords*) del texto.
- Lematización de palabras.
- Eliminación de tilde o diéresis (sustituyendo por caracteres sin estos signos).
- Eliminación de espacios previos/posteriores.

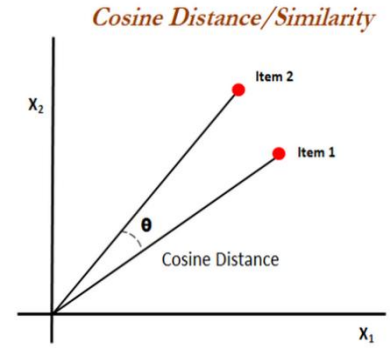
getEmbedding

Función que obtiene el vector denso (*embedding*) de un palabra o grupo de palabras utilizando el modelo cargado de obtención de *embeddings* mencionado.

Para ello, se obtiene el vector denso de cada una de las palabras por separado y después los vectores se promedian para obtener un único vector denso que represente el grupo de palabras.

calculateCosineSimilarity

Función que calcula la distancia o similitud coseno entre dos vectores densos, mediante la siguiente operación matemática específica.



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Ilustración 11. Representación gráfica y fórmula matemática para obtener la distancia o similitud coseno. Fuente: https://www.tyrrell4innovation.ca

calculateEuclideanDistance

Función que calcula la distancia euclídea entre dos vectores densos, mediante la siguiente operación matemática específica.

$$\|\overrightarrow{AB}\| = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + \ldots + (b_n - a_n)^2} \text{ siendo } \overrightarrow{OA} = (a_1, a_2, \ldots, a_n) \text{ y } \overrightarrow{OB} = (b_1, b_2, \ldots, b_n).$$

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} = \sqrt{\sum_{i=1}^{n} v_i^2}$$

Ilustración 12. Fórmula matemática para obtener la distancia euclidiana o norma euclídea. Fuente: https://es.wikipedia.org

La ruta de acceso a los apéndices detallados es: https://github.com/fmedinafernandez/14MBID-TFM-NEL/blob/main/14MBID_TFM_Medina_Fernandez_Fernando_Ap%C3%A9ndices.pdf

El código fuente está incluido en el repositorio de *github* correspondiente: https://github.com/fmedinafernandez/14MBID-TFM-NEL

Apéndice III

Base de Datos

La base de datos del prototipo almacena la siguiente información:

- **Términos clínicos**. Catálogo de términos clínicos que se utilizan para la vinculación con los informes médicos, junto con los vectores densos obtenidos en base al modelo utilizado.
- **Informes**. Repositorio de informes médicos que se han procesado mediante el prototipo, incluyendo el cuerpo del informe como información principal.
- **Términos clínicos asociados/enlazados**. Repositorio de términos clínicos asociados a los informes médicos procesados, incluyendo la similitud que se ha determinado en base al modelo y el algoritmo propuesto.

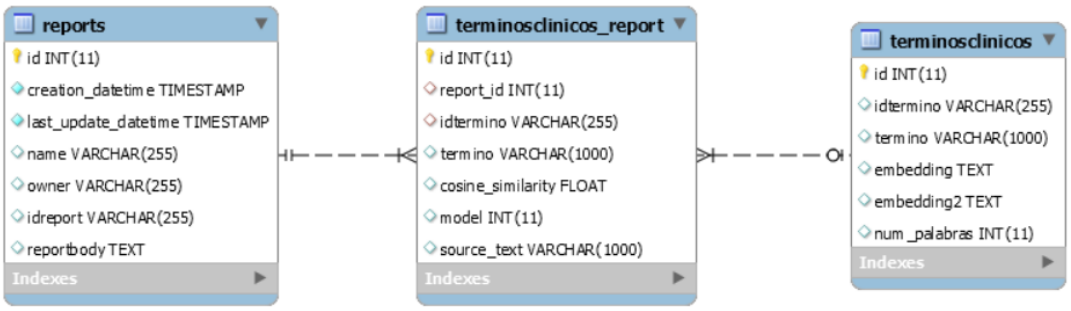


Ilustración 13. Diagrama físico de la base de datos. Elaboración propia mediante MySQL Workbench.

La ruta de acceso a los apéndices detallados es: https://github.com/fmedinafernandez/14MBID-TFM-NEL/blob/main/14MBID_TFM_Medina_Fernandez_Fernando_Ap%C3%A9ndices.pdf

El "script" de creación de la base de datos está incluido en el repositorio de *github* correspondiente: https://github.com/fmedinafernandez/14MBID-TFM-NEL

Apéndice IV

Desarrollo del interfaz de usuario prototipo de vinculación de informes médicos

Como se ha mencionado previamente el lenguaje de maquetado es HTML, se utilizan hojas de estilo *CSS* para gestionar la apariencia y código *Javascript* para que el interfaz sea interactivo y se integre con el motor de integración. El servidor de aplicaciones se ha implementado utilizando una librería desarrollada en *Python*.

A continuación, se muestra una pantalla de la aplicación donde se dispone de un menú en la parte superior con acceso a las funcionalidades de la misma.

Ilustración 14. Menú de "acerca de" del prototipo. Elaboración propia.

También se muestra un informe tipo junto con la información obtenida:

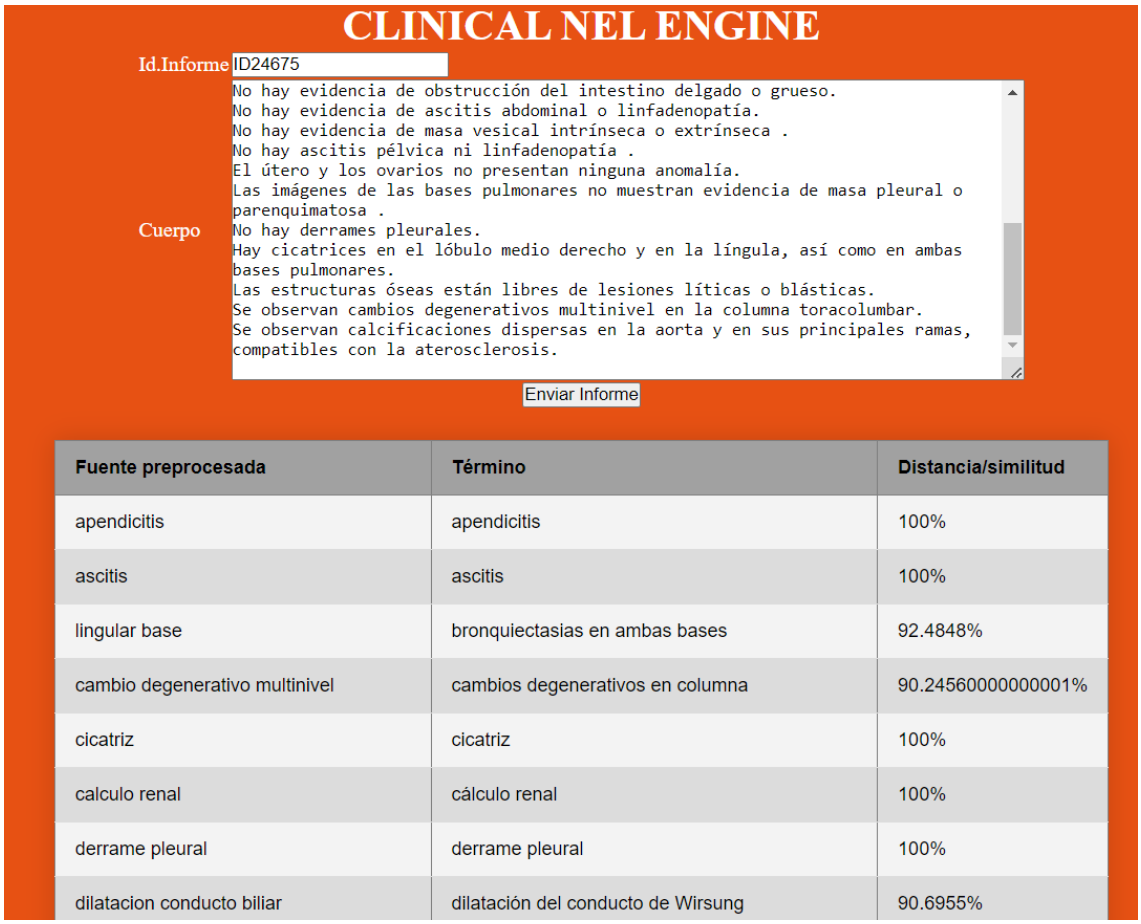


Ilustración 15. Resultado del etiquetado de un nuevo informe. Elaboración propia.

La ruta de acceso a los apéndices detallados es: https://github.com/fmedinafernandez/14MBID-TFM-NEL/blob/main/14MBID_TFM_Medina_Fernandez_Fernando_Ap%C3%A9ndices.pdf

El código fuente está incluido en el repositorio de *github* correspondiente: https://github.com/fmedinafernandez/14MBID-TFM-NEL

Apéndice V

Validación de resultados DISTEMIST

A continuación, se muestran los resultados de forma agregada. En los Apéndices se dispone de los datos concretos para cada informe.

Informes	Correctos	Incorrectos	Faltantes	Términos etiqueta-dos	Términos DISTE-MIST
580	5,54	8,05	3,27	13,59	8,81

Tabla 7. Resultados agregados en números absolutos de la validación de los datos. Elaboración propia.

Informes: Esta columna nuestra el número de informes utilizados para validar los resultados.

Correctos: Promedio de etiquetas correctamente vinculadas en relación con los datos de DISTEMIST.

Incorrectos: Promedio de etiquetas incorrectamente vinculadas en relación con los datos de DISTEMIST, es decir, que son etiquetas que no pertenecen al informe.

Faltantes: Promedio de etiquetas no vinculadas en relación con los datos de DISTEMIST, es decir, que son etiquetas que no se han detectado en el informe.

Informes	%Correctos sobre DISTE-MIST	%Faltantes sobre DISTE-MIST	%Incorrectos sobre etiquetados
580	64%	36%	58%

Tabla 8. Resultados agregados porcentuales de la validación de los datos. Elaboración propia.

Informes: Esta columna nuestra el número de informes utilizados para validar los resultados.

%Correctos sobre Distemist: Porcentaje promedio de etiquetas correctamente vinculadas con respecto al total de etiquetas de DISTEMIST.

%Faltantes sobre Distemist: Porcentaje promedio de etiquetas que no han sido vinculadas con respecto al total de etiquetas de DISTEMIST.

%Incorrectos sobre etiquetados: Porcentaje promedio de etiquetas vinculadas incorrectamente con respecto al total de etiquetas vinculadas por el prototipo.

La ruta de acceso a los apéndices detallados es: https://github.com/fmedinafernandez/14MBID-TFM-NEL/blob/main/14MBID_TFM_Medina_Fernandez_Fernando_Ap%C3%A9ndices.pdf