uc3m | Universidad Carlos III de Madrid

Máster Universitario en Ciencia y Tecnología Informática
2019-2020

Trabajo Fin de Máster

Codificación automática de historias de casos clínicos con códigos de diagnóstico ICD-10

Paula Queipo Álvarez

Tutores
Paloma Martínez Fernández
Israel González Carrasco
Leganés, Octubre 2020

[Incluir en el caso del interés en su publicación en el archivo abierto]
Esta obra se encuentra sujeta a la licencia Creative Commons **Reconocimiento - No Comercial - Sin Obra Derivada**

ABSTRACT

In Healthcare domain, clinical case studies are provided in a narrative way with an unstructured format. They describe the patient's conditions with natural language, making the automated processing of such texts hard and challenging. In order to analyze and transform medical narratives into a structured or coded format, clinical coding is required.

Clinical coding is a crucial task for standardizing medical texts, monitor health trends and medical reimbursement. It is very critical for hospitals, insurance companies and governments.

This work addresses the task of automatically assigning codes from the International Classification of Diseases, 10th version (ICD-10) for Diagnostics to unstructured Spanish clinical case studies that were also translated into Spanish and evaluating the results.

This document presents an approach based on Named Entity Recognition (NER) to detect diagnoses and semantic linking relying on a terminological resource to extract the labels. Each label is an ICD-10 code.

Precision results have been the best results obtained in the CodiEsp 2020 competition with a result of 86.6 %. The results of recall and Mean of Average Precision were 0.066 and 0.115 respectively.

The results are promising, especially precision and the evaluation regardless of the codes' sub-category.

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (DeepEMR project TIN2017- 87548-C2-1-R).

Keywords: ICD-10-CM * Clinical case studies * Multilabel classification * Named-Entity Recognition * Dictionary based approach * Fuzzy matching

RESUMEN

En el ámbito de la asistencia sanitaria, los casos clínicos se proporcionan de forma narrativa con un formato no estructurado. Describen las condiciones del paciente con lenguaje natural, lo que hace que el procesamiento automatizado de dichos textos sea difícil y desafiante. Para analizar y transformar narrativas médicas en un formato estructurado o codificado, se requiere codificación clínica.

La codificación clínica es una tarea crucial para estandarizar textos médicos, monitorizar las tendencias de salud y reembolsos médicos. Es muy importante para hospitales, compañías de seguros y gobiernos.

Este trabajo aborda la tarea de asignar automáticamente códigos de la Clasificación Internacional de Enfermedades, décima versión (CIE-10), para el diagnóstico de estudios de casos clínicos españoles no estructurados que también fueron traducidos al español y evaluar los resultados.

En el presente documento se presenta un enfoque basado en el Reconocimiento de Entidades Nombradas (NER) para detectar diagnósticos y vinculación semántica apoyándose en un recurso terminológico para extraer las etiquetas. Cada etiqueta es un código CIE-10.

Se han obtenido los mejores resultados de precisión en la competición CodiEsp 2020 con un resultado del 86.6 %. Los resultado de exhaustividad y Media de Precisión Media han sido 0.066 y 0.115 respectivamente.

Los resultados son prometedores, especialmente la precisión y la evaluación de los códigos sin tener en cuenta la subcategoría.

Este Trabajo Fin de Máster ha sido financiado por el Programa de Investigación del Ministerio de Economía y Competitividad del Gobierno de España. (Proyecto DeepEMR TIN2017-87548-C2-1-R).

Palabras clave: CIE-10-Diagnósticos * Estudios de caso clínico * Clasificación multi-etiqueta * Reconocimiento de Entidade Nombradas * Enfoque basado en diccionario * Coincidencia difusa

DEDICATORIA

Quiero agradecer a todos los miembros de HULAT la oportunidad de formación que me han otorgado. En primer lugar a Paloma Martínez Fernández e Israel González Carrasco, por ser mis tutores y compañeros de proyecto, tan atentos y críticos, generosos con su tiempo y su conocimiento. Y también a Jose Luis López Cuadrado y Maria Luisa Arjonilla López por su paciencia y tenacidad.

A Belén Ruiz Mezcua y a mis compañeros del CESyA, por incluirme en su clima laboral lleno de humanidad y cariño.

También quiero mencionar a mis compañeros Alejandro, Miguel, Marcos, Javier y Lucía y a mis profesoras Maribel Sánchez e Isabel Segura por apoyarme este curso.

A mi familia, por ayudarme siempre.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Motivación personal del proyecto

La motivación de este TFM surge de una situación que se está dando en gran cantidad de países: la sobrecarga de trabajo en muchas profesiones relacionadas con el campo de la sanidad. Este es un tema de la máxima actualidad debido a que este problema de sobrecarga se ha visto acentuado por la pandemia que ha surgido a finales del 2019.

Debido a esta crisis sanitaria, los hospitales se han visto saturados y los especialistas están sometidos a fuertes cargas de trabajo. Muchas de las tareas administrativas, los análisis epidemiológicos, la atención primaria y muchas otras funciones clínicas requieren del análisis de textos clínicos que faciliten la toma de decisiones. Las necesidades de datos nunca han sido mayores en la Industria de la Salud. En un entorno sometido a mucha presión y en el que el contacto personal se ha visto limitado, la informática clínica se posiciona como un importante campo de innovación que facilita enormemente la gestión intra e interhospitalaria.

La mayoría de la información de los textos clínicos se encuentra desestructurada. Se necesita estructurar para extraer conocimiento de los datos. Por ejemplo, para ayudar en la toma de decisiones clínicas, en la estratificación en las cohortes de pacientes, la supervisión de eventos médicos, o el manejo de la salud de una población. Por lo tanto, es muy importante estructurar textos clínicos para ofrecer mayor información a los profesionales del sector de la salud.

Entre los beneficios económicos de utilizar codificación con textos médicos, se encuentran la reducción de la carga de trabajo de los expertos, del tiempo de estructuración de la información, un mejor ajuste de las tasas pago por servicio hospitalario y las ventajas asociadas a una toma de decisiones adecuada sobre los recursos disponibles.

Hay diversas formas de estructurar la información. Una de las más importantes es la codificación clínica. La codificación y los informes correctos de los diagnósticos y servicios sanitarios se han vuelto cada vez más críticos a medida que han evolucionado las necesidades de datos sanitarios[1].

Las aplicaciones de los códigos clínicos son variadas[2]. Por ejemplo, reflejar la forma de proceder de los clínicos, ofrecer procesos que faciliten la toma de decisiones, analizar la efectividad del tratamiento de los servicios clínicos, facilitar los estudios epidemiológicos y la investigación en servicios médicos, la selección de los pacientes y los indicadores clínicos. También favorecen la estandarización, los estudios estadísticos y la creación de sistemas bioinformáticos. En la actualidad, se emplean para fijar la tasa de los pagos, los reembolsos de la asistencia hospitalaria y la monitorización de los recursos y de las horas de trabajo para asegurar la calidad de los centros médicos. La toma de decisiones en

base a códigos facilita la asignación de recursos disponibles. También se utilizan códigos para realizar estudios sobre el acceso, la calidad, los costos y la efectividad de la atención médica, con variaciones temporales, transversales y ajustadas al riesgo.

Un texto clínico que posee de especial relevancia son las historias clínicas. Éstas son analizadas cuidadosamente por los expertos, que interpretan los resultados y realizan los diagnósticos clínicos que correspondan.

El diagnóstico de enfermedades es una tarea complicada, que requiere de unos niveles de concentración elevados y una cantidad significativa de tiempo de la cual es mejor no prescindir. Además, en entornos de urgencia, el manejo de información sensible exige un control de calidad que evite los fallos. La fatiga de los profesionales sanitarios, las largas jornadas de trabajo, el aumento de pacientes y otros problemas comunes pueden perjudicar el rendimiento de su trabajo.

Por tanto, la carga de trabajo en los centros médicos puede reducirse gracias a sistemas informáticos que realicen tareas de Procesamiento del lenguaje natural (NLP) en el dominio clínico.

Para analizar textos clínicos se necesita conocimiento del dominio médico, una elevada cualificación y entender la lengua en la que están escritos los textos a analizar. Esto supone un problema, puesto que no se puede extrapolar fácilmente la información obtenida por sistemas construidos en otra lengua. Actualmente se están logrando muchos avances en la estructuración de la narrativa clínica en inglés. Además, existe una mayor cantidad de recursos para tareas de extracción de información clínica en la lengua inglesa. Sin embargo, otras lenguas deben de ser tomadas en cuenta a la hora de procesar información clínica y es necesario promover los recursos en el resto de lenguajes.

Este problema es importante, ya que es necesario desarrollar codificadores de enfermedades que puedan trabajar en diferentes lenguas para reducir la carga de trabajo en los centros médicos, y permitir que diferentes países puedan extraer información de textos médicos en lenguas no inglesas.

1.2. Objetivo del estudio

El objetivo final que persigue este trabajo es el siguiente: «Desarrollar una herramienta de apoyo al diagnóstico médico en historias de caso clínico en español que codifique enfermedades con la terminología ICD10-CM».

Para ello, se deben de realizar diferentes tareas:

- Normalizar las historias de caso clínico. El dominio clínico tiene muchas variantes léxicas a la hora de referirse a una enfermedad.
- Reconocer entidades nombradas (NER) en el dominio clínico en español.

- Codificar los diagnósticos con la terminología estándard ICD-10-CM en español, utilizando recursos en esa lengua.
- Obtener los mismos resultados que los codificadores nativos de lengua española.

El presente trabajo se centra en la lengua española, además de la lengua inglesa. Esto es relevante, ya que no hay tantos avances en español como en inglés.

Para ello se ha participado en una tarea de clasificación denominada CodiEsp 2020, que busca mejorar la asignación automática de códigos ICD10 para clasificar diagnósticos en historias clínicas. Esta competición es organizada por la Conferencia CLEF 2020 y eHealth, cuya finalidad es analizar textos médicos en varios idiomas y promover la investigación en Extracción de la Información (IE), Recuperación de la Información (IR), y Manejo de la Información (IM).

Este sistema tiene las siguientes ventajas que permiten definir su alcance:

- Automatizar la codificación. Al introducir una historia de caso clínico, se obtiene los códigos de diagnósticos presentes en la historia.
- Independiente del lenguaje. No se necesita leer el contenido de la historia clínica para codificar, por lo que no se requiere entender la lengua nativa de las historias clínicas.
- Utiliza recursos en español. Esto permite que más de 20 países de habla hispana puedan extraer información de historias clínicas.
- Disminuye carga de trabajo en los centros médicos. Reduce las horas de etiquetado de las notas clínicas.
- Etiquetado multi-etiqueta. Es capaz de clasificar una historia clínica con más de un código.
- Mejora la calidad del etiquetado. El sistema reduce el factor de error humano y mejora la precisión en la codificación de enfermedades.

Por lo tanto, el alcance del presente Trabajo de Fin de Máster es proponer un sistema que codifique enfermedades en historias clínicas en español, automatizando la tarea de los codificadores, utilizando recursos en español, clasificando de forma multi-etiqueta y mejorando la precisión del etiquetado respecto a otros sistemas similares. Esto se ha podido lograr durante la evaluación del sistema.

Método de investigación y Enfoque

Se ha optado por un trabajo científico aplicado con un objetivo exploratorio o piloto. Debido a la necesidad de validar la hipótesis, se ha elegido realizar una investigación

experimental con una metodología empírico-analítica, controlada, rigurosa, válida, verificable, empírica, crítica y sistemática [3].

Se ha comenzado identificando el tema de trabajo, el enfoque elegido, la obtención y estudio de los recursos necesarios para resolver la hipótesis de investigación. Finalmente se evalúan los resultados y se exponen las conclusiones principales y mejoras del trabajo. Por ello se ha estudiado la bibliografía y se han documentado las fuentes externas necesarias siguiendo el índice del contenido de la memoria. A continuación, se ha decidido participar en una competición que ofrece un corpus adecuado para la hipótesis propuesta. Finalmente se ha descrito la tarea, su resolución, su evaluación y las conclusiones.

Para abordar el objetivo del presente trabajo, se enuncian diferentes propósitos, siendo éstos:

- Estudio teórico: enfocar el objetivo, definir la hipótesis, plantear el problema, trabajos relacionados y definir el enfoque de la solución.
- Desarrollo del sistema.
- Evaluación y análisis de los resultados con diferentes métricas
- Conclusiones, mejoras y alternativas del trabajo.

Resultados de investigación y conclusiones principales

Con la investigación, se ha probado a automatizar la codificación de enfermedades con la terminología Internacional ICD10-CM. Se ha realizado un programa que obtiene una gran precisión (0.866), pero menores valores de recuperación (0.066). Sin embargo, es un avance en los sistemas de codificación de enfermedades en lengua española, que ofrecen ventajas prometedores en su implementación en hospitales.

1.3. Organización de la memoria

En el capítulo 2 se ofrece un marco de estudio tanto en el ámbito clínico como en las herramientas técnicas. Se incluye la información sobre la codificación con la terminología ICD-10-CM.

En el capítulo 3 se detalla el Estado de la Arte en la codificación de texto clínico. Además, se explica la participación en la competición CodiEsp y los trabajos relacionados.

Los capítulos 4 y 5 se centran en los pasos metodológicos, la justificación del enfoque principal y secundarios y la explicación del Sistema desarrollado.

En el capítulo 6 se aporta información sobre la Planificación del Proyecto, incluyendo la planificación temporal, el presupuesto estimado y el marco legal.

En el capítulo 7 se hace referencia a la Experimentación, las métricas de evaluación, los resultados obtenidos y su correspondiente análisis.

Finalmente, el ultimo capítulo incluye una discusión final sobre el trabajo.

Además, se incluye la bibliografía y un anexo con la lista de abreviaturas.

2. MARCO CLÍNICO Y TÉCNICO

En este capítulo, se incluye el Marco clínico, el Marco técnico y el sistema Internacional de Codificación de enfermedades ICD-10-CM.

2.1. Marco clínico

En este trabajo se plantea codificar enfermedades con códigos ICD-10-CM a partir de historias de caso clínico. En la Figura 2.1 se observa el esquema clínico con las cuatro propiedades que se han abordado en el presente trabajo.

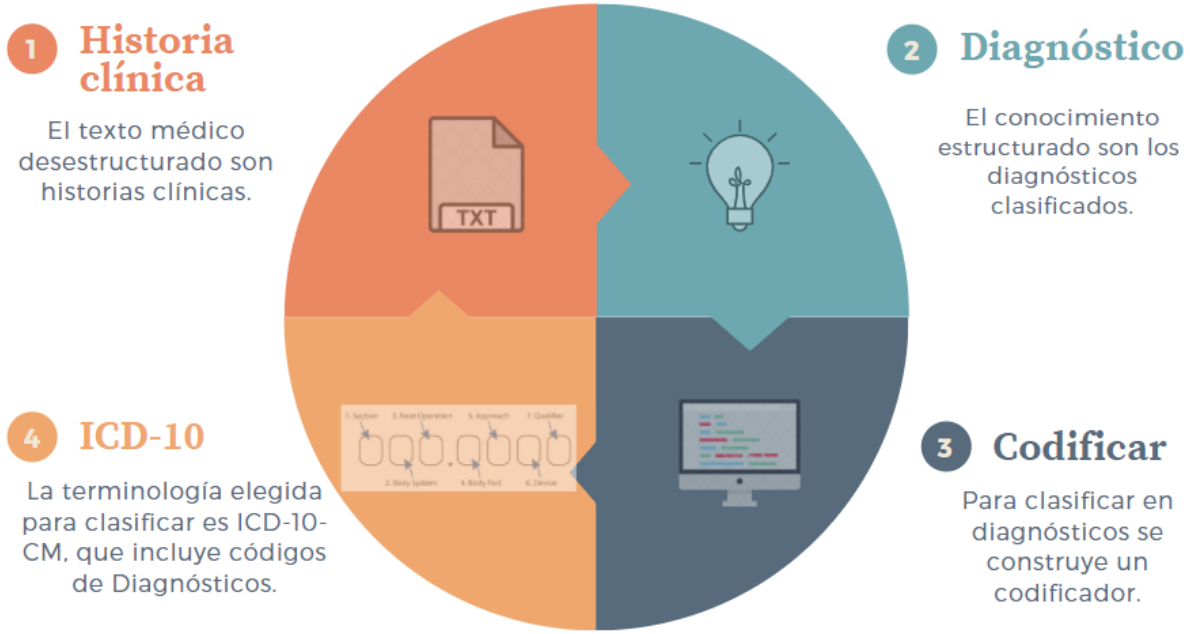

FIG. 2.1. ESQUEMA DEL MARCO CLÍNICO

En primer lugar, se ha elegido un texto clínico desestructurado del que se desea extraer información. Las historias clínicas son documentos adecuados ya que ofrecen datos relevantes de interés médico.

En segundo lugar, se ha propuesto estructurar el conocimiento realizando una clasificación de diagnósticos clínicos presentes en el cuerpo de historias clínicas. El diagnóstico permite valorar el tipo de enfermedad presente en la historia clínica y facilita la recuperación de historias clínicas con diagnósticos similares.

En tercer lugar, se ha decidido clasificar los diagnósticos con códigos estandarizados. La codificación facilita la normalización y la ramificación de enfermedades relacionadas, y permite identificar una enfermedad con un código común independientemente del idioma.

Finalmente, se ha seleccionado la terminología estándar que codifique los diagnósticos en las historias de caso clínico. En este caso, ICD-10-CM permite abordar esta tarea al presentar códigos internacionales para la clasificación de enfermedades.

A continuación se incluye información sobre las historias clínicas, la clasificación de enfermedades, la codificación y la terminología ICD-10-CM.

2.1.1. Historias clínicas

Las historias clínicas son documentos que constituyen un registro de la atención prestada a un paciente clínico. En un sistema de información sanitario es imprescindible a nivel legal, asistencial y administrativa.

Estas historias pueden tener un soporte físico o electrónico. Su información es muy sensible, por lo cual los datos clínicos deben ser anonimizados para garantizar una mayor confidencialidad y seguridad. Además, se deben de almacenar y gestionar mediante diferentes bases de datos.

Las funciones de las historias clínicas son variadas y necesarias para el desarrollo de funciones de los profesionales de la salud. A nivel legal, regulan la relación paciente-médico. A nivel administrativo, permiten la gestión de los asistencia prestada por los organismos sanitarios. También son utilizadas para evaluar el nivel de calidad de los servicios ofrecidos. Además, aportan información que facilita la docencia, la investigación sanitaria y estudios epidemiológicos locales o internacionales. Algunos ejemplos de usos de las notas clínicas son ayudar en la toma de decisiones, estratificar las cohortes de pacientes, supervisar eventos médicos, o gestionar la información sanitaria de una población.

Sin embargo, la mayor parte de la información de las historias clínicas (y otros textos médicos), se encuentra desestructurada. Sin una apropiada estructura no es posible extraer, visualizar o analizar al máximo la información de estos documentos. Uno de los objetivos del presente trabajo consiste en estructurar historias clínicas para facilitar información a los profesionales del sector sanitario.

El dominio clínico tiene muchas variantes léxicas ya que en el lenguaje natural hay muchas formas de referirse a una enfermedad. Se deben de tratar los sinónimos y las abreviaturas, y las conjunciones aditivas.

Las historias clínicas pueden tener un identificador, además del contenido desestructurado sin ningún campo. En la Figura 2.2 se muestra el contenido de una historia clínica procedente de la competición de codificación médica CodiEsp [2]. El id de la historia es «S0004-06142010000100015-1» y el contenido se observa en la Figura 2.2.

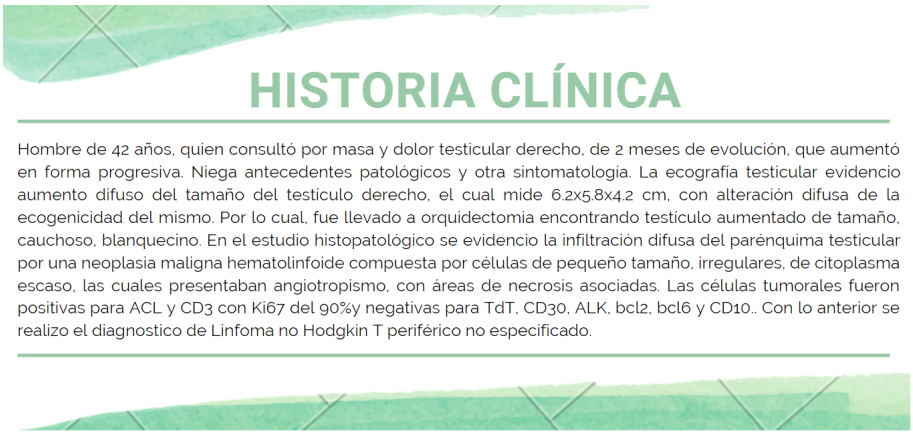

FIG. 2.2. EJEMPLO DE HISTORIA CLÍNICA DE LA COMPETICIÓN CODIESP

2.1.2. Clasificación de enfermedades

La Nosología es el estudio de la clasificación sistemática de enfermedades y ocupa un lugar central en el área de la salud médica.

A medida que la industria sanitaria busca una mejora general de los servicios asistenciales, el uso de datos estructurados continúa creciendo. La indexación y la codificación clínica se ven fomentadas por ello, ya que incrementa el volumen de los datos estructurados.

Puede ser difícil discernir la diferencia entre indexar y aplicar códigos clínicos, ya que pueden estar relacionadas ambas tareas. Al realizar un estudio de indexación se analizan documentos y se puede incluir un código de salida de la clasificación. Lo que no se debe confundir con la codificación de términos clínicos que pueden o no utilizarse con el propósito de indexar o recuperar información.

Existen diferentes clasificaciones, vocabularios, diccionarios y terminologías. Algunos sistemas de clasificación en base a códigos son [4]:

- International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM)
- Current Procedural Terminology (CPT)
- International classification of functioning, disability and health (ICF)
- Unified Medical Language System (UMLS)
- Medical Subject Headings (MeSH)
- MedLEE's controlled vocabulary (MED)
- HICDA (Mayo modification of ICD-8)
- RxNorm (clinical drugs)

- SNOMED (International Health Terminology Standards Development Organisation)
- SNOP (Systematized Nomenclature of Pathology)
- American Medical Association's Current Procedural Terminology, 4th Edition (CPT-4)
- Health Care Common Procedural Coding System (HCPCS)
- American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, 4th Revision (DSM-IV)
- Europe's Classification of Surgical Operations and Procedures, 4th Revision (OPCS-4)
- Agency for Healthcare Research and Quality's Clinical Classification Software (CCS)

2.1.3. Codificación clínica

Al comienzo los códigos servían para clasificar la mortalidad y la morbilidad de los pacientes, pero su uso se ha extendido en una gran variedad de aplicaciones[4]. Al agrupar a los pacientes en función de su diagnóstico, los epidemiólogos pueden estudiar los patrones de los tratamientos y la evolución de la enfermedad. Los códigos ICD sirven como criterios de inclusión y exclusión para definir marcos de muestreo, documentar las comorbilidades de los pacientes e informar sobre la incidencia de complicaciones.

En la actualidad, también se emplean códigos para fijar la tasa administrativa de los pagos, los reembolsos de la asistencia hospitalaria y la monitorización de los recursos y de las horas de trabajo. Esto permite asegurar la calidad administrativa y médica de los centros sanitarios. También se utilizan en el rastreo de las tasas de utilización del acceso, la calidad, los costos y la efectividad de los servicios[4].

Generalmente la codificación se realiza mediante un proceso manual. Ello implica la revisión humana de la documentación clínica para identificar los códigos aplicables. La asignación del código puede ser realizada por médicos, pero a menudo es realizada por otro personal, como los profesionales de la codificación[1].

Un grupo de trabajo de la American Health Information Management Association (AHIMA) informó que este flujo de trabajo de codificación manual es caro e ineficaz. «La industria necesita soluciones automatizadas para permitir que el proceso de codificación se vuelva más productivo, eficiente, preciso y consistente» [5].

Sistemas automatizados de codificación y clasificación clínicos

Los sistemas automatizados de codificación y clasificación clínicos o codificación asistida por ordenador son aplicaciones informáticas diseñadas para generar cualquier

tipo de código clínico o clasificación a partir de documentos clínicos de texto libre. Estos sistemas están siendo construidos y evaluados por los investigadores. Actualmente se encuentran disponibles pero no se usan ampliamente, probablemente porque los sistemas aún están en desarrollo y el rendimiento en producción no está comprobado[1]. Sin embargo, la industria dedicada a recopilar datos de atención médica se puede beneficiar por los sistemas automatizados de codificación y clasificación para reducir el tiempo, los costes y la carga de trabajo asociada al análisis de la información.

2.1.4. Códigos ICD-10

Al aplicar un esquema de codificación complejo, el proceso puede ser asistido por el uso de libros de códigos, la selección de listas abreviadas o el uso de aplicaciones de software que facilitan las búsquedas alfabéticas y proporcionan ediciones y consejos[1].

Es la Clasificación Internacional de Enfermedades, Modificación Clínica (ICD-CM) es uno de los estándares de clasificación de diagnósticos más utilizados[4] para aplicaciones clínicas y de investigación en el dominio médico. Tiene diversas ventajas: (i) está aceptado internacionalmente, (ii) está bien diseñado y (iii) permite una asignación inequívoca de diferentes enfermedades. Además, (iv) están disponibles diversas herramientas online [6]-[8].

La primera edición se remonta a 1893, cuando el médico francés Jacques Bertillon introdujo la Clasificación Bertillon de las causas de muerte. Esta primera edición tuvo 179 códigos. Se recomendó que este sistema de clasificación se revisara cada 10 años. Con cada revisión aumentaba el número de códigos y el atractivo de su uso en diversas aplicaciones[4].

En 1978, la Organización Mundial de la Salud (OMS) publicó la novena revisión. El desarrollo de una décima revisión en 2003, introdujo códigos alfanuméricos y una mayor especificidad que la ICD-9 e incluye más de 21.800 códigos en total.

Para hacer que el ICD sea más útil, el Centro Nacional de Estadísticas de Salud (NCHS) ha desarrollado la Modificación Clínica (ICD-10-CM). En la Modificación Clínica (CM) se pretende que los códigos fueran más precisos [9] y se utiliza para codificar y clasificar diagnósticos y procedimientos de registros de pacientes hospitalizados y ambulatorios.

Jerarquía

Los códigos de ICD-10-CM se parecen a sus antecesores. Se compone de tres a siete caracteres alfanuméricos de longitud. Los primeros tres indican la categoría del diagnóstico. Los últimos son opcionales e indican la subcategoría. El nivel final de subdivisión es el código final válido. En la documentación oficial [10], [11] se describe la estructura en forma de árbol de los códigos en categorías y subcategorías. Se corresponde con:

- Capítulo: de forma general, cada capítulo está identificado por este primer carácter alfabético. Hay excepciones como capítulos que incluyen más de una letra (capítulo 1 incluye las letras A y B) y capítulos que comparte la misma letra (los capítulos 7 y 8 con la letra H). El capítulo no se incluye en el código.
- Categoría específica: constan de los primeros tres caracteres. Son obligatorios. El primero de ellos es alfabético e indica el capítulo. El segundo carácter es numérico. Finalmente, el tercero puede ser numérico o alfabético. Por ejemplo S72. Para referirse a un grupo de categorías relacionadas, se utilizan códigos genéricos que conectan diferentes categorías como en O03-O06.
- Subcategoria: los siguientes cuatro o cinco caracteres alfanuméricos. Por ejemplo 109B.

En la Figura 2.3 su muestra un esquema de un código. La categoría específica se compone de tres caracteres. Tras un punto se encuentra la subcategoría.

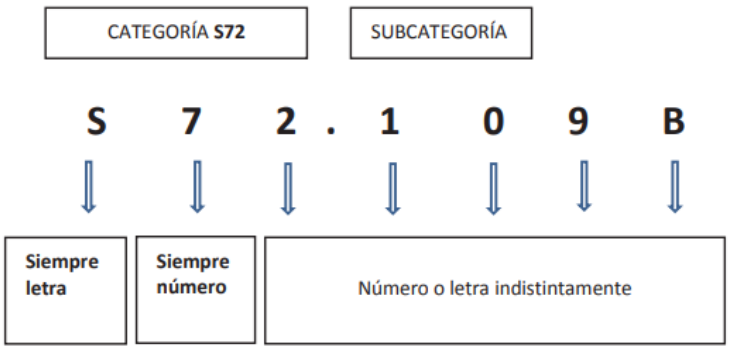

FIG. 2.3. JERARQUÍA DEL CÓDIGO ICD-10-CM

2.2. Marco técnico

Inteligencia Artificial

La Inteligencia Artificial es un campo que incluye conocimientos de otros campos, como la Ciencia informática y la Lingüística. Incluye herramientas que se puede aplicar en muchos entornos como la robótica, la lingüística computacional, los sistemas de apoyo a la decisión o la medicina.

Algunas ramas de la Inteligencia Artificial son:

- Minería de Texto: la minería de texto es el proceso de descubrir y extraer conocimiento de datos no estructurados [12]. Se sirve de diferentes técnicas y herramientas, como el aprendizaje automático, aprendizaje profundo, extracción de la información, etc... que no son excluyentes.
- Aprendizaje Automático: es un conjunto de técnicas estadísticas y modelado matemático que utiliza una variedad de enfoque para aprender de forma automática y mejorar la predicción de un estado sin programarlo de forma explícita. Mientras que en la programación por computadores se utilizan sistemas basados en reglas en las cuales se escribe una función para transformar el fichero de entrada en una salida, el Aprendizaje Automático se compone de dos fases. En el entrenamiento, el ordenador aprende la función que transforma la entrada en la salida a partir de ejemplos. En la fase de inferencia, se utiliza la función o modelo entrenado para predecir el resultado de un conjunto de validación. Algunos modelos clásicos son Bosques aleatorios o Máquinas de Soporte vectorial.
- Aprendizaje Profundo: es una técnica de Aprendizaje automático que utiliza redes neuronas con un número elevado de capas. Un ejemplo de aprendizaje profundo son las redes neuronales convolucionales.

En la Figura 2.4 se muestra un diagrama de Venn con Inteligencia artificial, Aprendizaje Automático y Aprendizaje profundo.

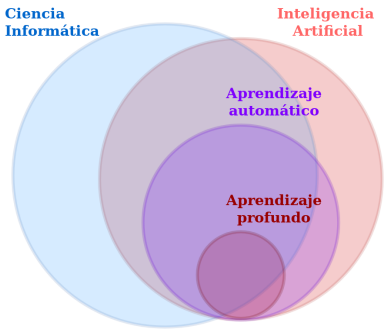

FIG. 2.4. DIAGRAMA DE VENN DE IA, ML Y DL.

Se aprecia que el Aprendizaje Profundo está incluido dentro del Aprendizaje Automático, y ambos dentro de la Inteligencia Artificial. Las aplicaciones de la Inteligencia Artificial en el dominio médico tiene gran variedad de aplicaciones. Por ejemplo, el análisis de datos estructurados como imágenes, señales biofísicas o genes. Otra aplicación es el análisis de datos desestructurados como notas clínicas o historias clínicas[13].

En la siguiente figura 2.5 se aprecia otro diagrama de Venn.

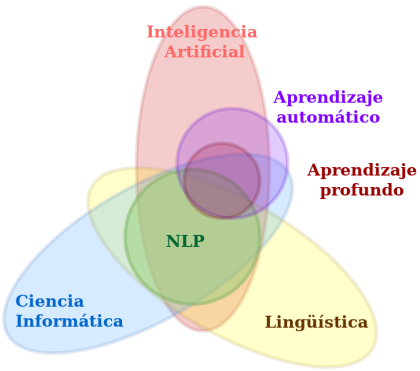

FIG. 2.5. DIAGRAMA DE VENN DE IA, ML, DL Y NLP.

Se muestra que el Aprendizaje Profundo está incluido dentro del Aprendizaje Automático, y ambos dentro de la Inteligencia Artificial. La Inteligencia Artificial es un subcampo de la ciencia informática, pero es transversal. El Procesamiento del Lenguaje Natural es un campo interdisciplinar que combina conocimientos de ciencia informática con conocimientos lingüísticos, y puede utilizar modelos de aprendizaje profundo, aprendizaje automático y/o Inteligencia Artificial.

Minería de texto

La minería de texto es el proceso de descubrir y extraer conocimiento de datos no estructurados [12]. Se sirve de diferentes técnicas y herramientas, como el Aprendizaje Automático, Aprendizaje Profundo, Extracción de la Información, etc... que no son excluyentes. En la Figura 2.6 se observa el conjunto de técnicas de Minería de texto utilizadas.

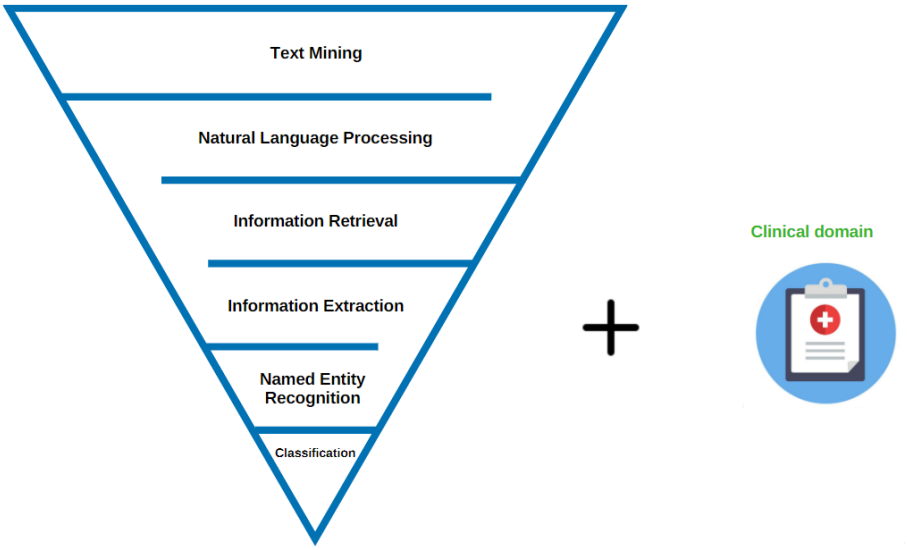

FIG. 2.6. TÉCNICAS DE MINERÍA DE TEXTO

Generalmente, se consigue clasificar utilizando Reconocimiento de Entidades Nombradas, que a su vez es una tarea de Extracción de Información, que es un subconjunto de la Recuperación de Información, que es una tarea importante del Procesamiento de Lenguaje Natural que se enmarca en la minería de texto. Además se aplican estas técnicas al dominio clínico, haciéndolas más específicas y dependientes del dominio.

En la Figura 2.7 se observan una serie de técnicas utilizadas en el dominio clínico.

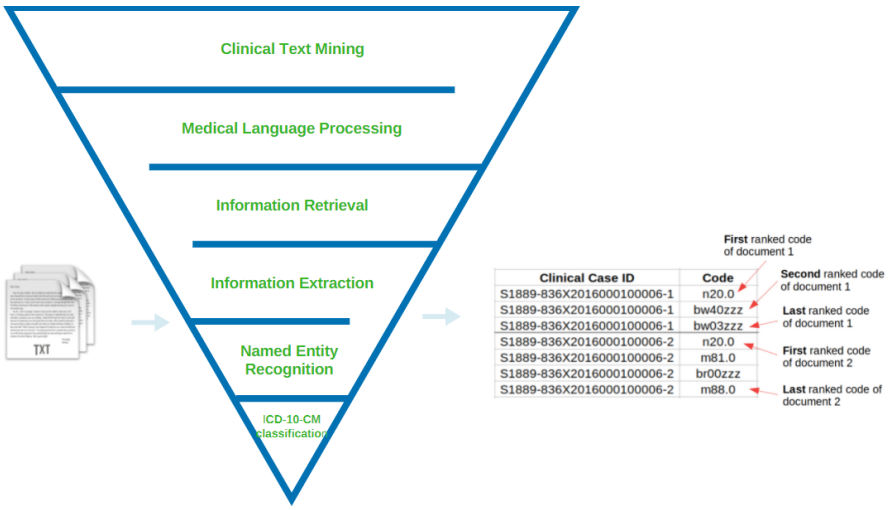

FIG. 2.7. TÉCNICAS DE MINERÍA DE TEXTO DEL DOMINIO CLÍNICO

Generalmente, la clasificación en códigos clínicos requiere de Reconocimiento de entidades biomédicas, que a su vez es una tarea de Extracción de Información, que a su vez es un subconjunto de la Recuperación de Información, que es una tarea importante del Procesamiento de Lenguaje médico que se enmarca en la minería de texto clínico. El presente trabajo parte de historias clínicas, y a través de estas técnicas se va a predecir las clases de códigos a las que pertenecen las historias.

Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. De acuerdo a la Enciclopedia de Inteligencia Artificial[14], «Natural Language Processing (NLP) es un subcampo de la Inteligencia Artificial que se centra en la generación y la comprensión automática de los lenguajes naturales»[12]. La investigación se centra en construir modelos computacionales para entender el lenguaje natural que utilizan los seres humanos. El lenguaje natural se distingue de los lenguajes de programación y representación de datos utilizados por computadoras y descrito como artificial.

El Procesamiento del Lenguaje Médico en el dominio clínico presenta dificultades adicionales pero es muy necesario. Gran parte de los datos clínicos disponibles están en forma narrativa como resultado de la transcripción de dictados, la entrada directa de proveedores o el uso de aplicaciones de reconocimiento de voz. Este formulario de texto libre es difícil de buscar, resumir, respaldar decisiones o realizar análisis estadísticos[12].

Recuperación de la Información

También llamado Information retrieval (IR), es una técnica enfocada en encontrar documentos, no información o hechos dentro del texto. Un ejemplo son los motores de búsqueda como Google[15] o PubMed[16].

Extracción de la Información

En inglés se llama Information Extraction (IE) y es un subdominio del procesamiento del lenguaje natural cuyo objetivo es extraer información de un texto concreto. Por ejemplo, documentos del registro electrónico de salud o las historias clínicas electrónicas. Los sistemas de extracción de la información engloban un concepto más amplio que la codificación y clasificación.

Reconocimiento de Entidades Nombradas

Named Entity Recognition (NER) es un subcampo de Extracción de Información que se enfoca en reconocer expresiones designadas en documentos de texto libre. Estas expresiones que denotan entidades pueden ser de diversos tipos, como enfermedades, tratamientos, agentes causales o medicamentos[12]. Los sistemas NER pueden ser muy efectivos, pero requieren un poco de esfuerzo manual. Los enfoques de aprendizaje automático pueden exitosamente extraer entidades nombradas pero requieren grandes corpus anotados de entrenamiento. Las ventajas de los enfoques de aprendizaje automático son que no

requieren intuición humana y se pueden volver a entrenar sin reprogramación para ningún dominio.

En la siguiente Figura 2.8 [13] se puede observar la definición del problema de clasificar enfermedades en historias clinícas.

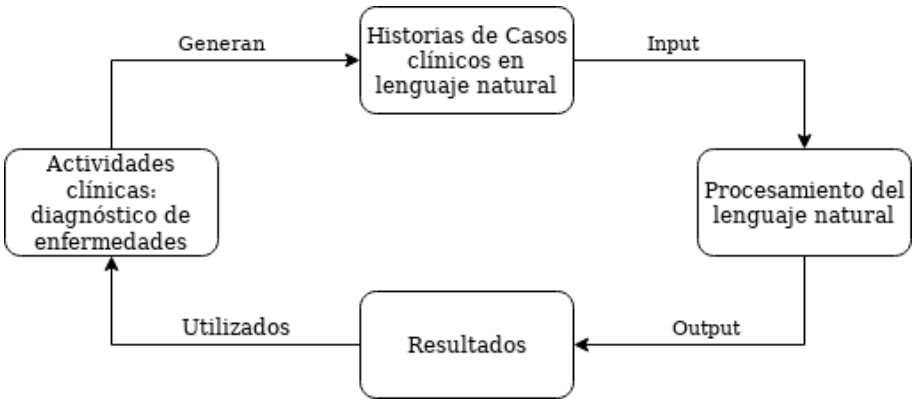

FIG. 2.8. DIAGRAMA DE LA DEFINICIÓN DEL PROBLEMA

Se generan historias clínicas que se estructuran mediante técnicas de Procesamiento de Lenguaje Natural. Luego, el resultado de las técnicas se utilizan para realizar actividades clínicas como el diagnóstico de enfermedades. El producto de estas actividades sirve para enriquecer de nuevo las historias clínicas, etc. Es un proceso iterativo a lo largo del tiempo que busca ofrecer la información más refinada y estructurada posible para facilitar las decisiones y actividades clínicas del personal sanitario.

3. ESTADO DEL ARTE

En primer lugar se describen los antecedentes mediante revisiones y trabajos que incorporan un estado del Arte sobre el entorno tecnológico en el que se lleva a cabo el trabajo. En segundo lugar, se realiza una síntesis de trabajos previos relacionados con la tarea. En tercer lugar, se explica la participación en la competición CodiEsp 2020 y se mencionan los enfoques significativos utilizados tras el desarrollo del sistema de este trabajo.

3.1. Antecedentes

Se han consultado dos revisiones de Extracción de la Información. Una de ellas [17] sobre texto clínico y otra de ellas [12] sobre historias clínicas.

En un Trabajo de Fin de Máster [18] se introduce una arquitectura híbrida de Aprendizaje Automático y Aprendizaje Profundo para el Reconocimiento de Entidades Nombradas en textos biomédicos. El proceso de reconocimiento es interpretado como un proceso de asignación de secuencias de etiquetas en base al contexto de las palabras. La arquitectura consiste de dos redes bidireccionales LSTM (Long Short-Term Memory) y una última red basada en Conditional Random Field (CRF) capaces de obtener información semántica, sintáctica, morfológica, ortográfica y la secuencia en la que éstas ocurren.

En la Figura 3.1 se muestra un diagrama procedente de una revisión [13] sobre modelos de Aprendizaje Automáticos aplicados a la literatura médica.

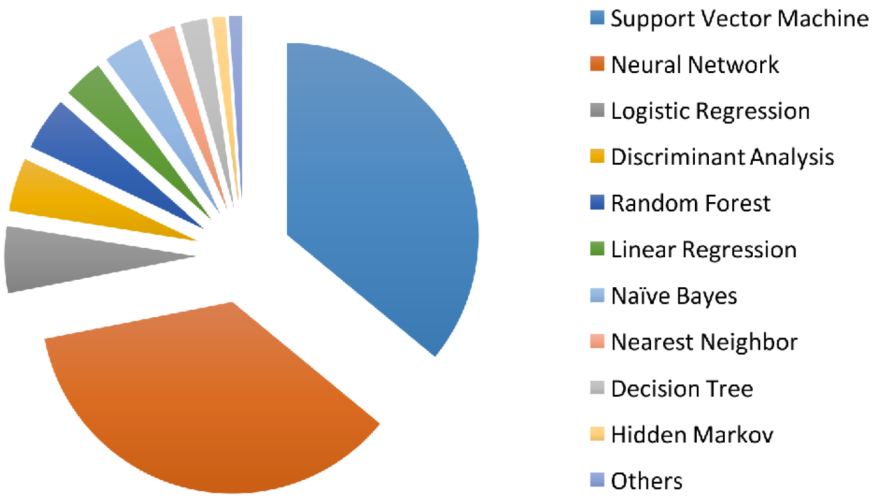

FIG. 3.1. ALGORITMOS DE APRENDIZAJE AUTOMÁTICO MÁS USADOS EN LITERATURA MÉDICA

Modelos tradicionales

En los métodos de aprendizaje automático se realiza ingeniería de características mediante técnicas de aprendizaje automático, como clasificadores por K-vecinos o bayesianos.

En el artículo [19] los autores obtuvieron mejores resultados con un ensemble.

En el artículo [20] se experimentó con técnicas de reconocimiento de patrones, que pueden tomar la forma de expresiones regulares.

En el artículo [21] se utilizan sistemas híbridos que combinan la especificidad de las reglas y la flexibilidad de los métodos de aprendizaje automático y se comparó un aprendizaje basado en árboles de decisión con un algoritmo de regresión logística multinomial.

En el artículo [22] se utilizó una máquina de soporte vectorial y featurización con enegramas.

En el artículo [23] se demostró que el tamaño de los corpus y el número de códigos diferentes afecta a la hora de obtener el mejor mecanismo de entrenamiento. En los corpus de menor tamaño es más importante seleccionar las características relevantes.

En el artículo [24] se ha demostrado que integrando la selección de características en los datos desestructurados y estructurados se mejora el proceso de clasificación.

Modelos de Aprendizaje profundo

En los métodos con aprendizaje profundo no se realiza ingeniería de características debido a que el método aprende las características relevantes desde los datos en crudo.

En la revisión [25] se han explicado las técnicas de Aprendizaje Profundo aplicado a notas clínicas.

Existe un estudio [26] cuyos resultados muestran que la tokenización y la segmentación jerárquica del documento original permiten a una arquitectura con Atención jerárquica GRU (HA-GRU) mejorar otros modelos como SVM, CBOW y redes convolucionales. Además, mantiene una transparencia efectiva al preservar el texto original.

En el trabajo llevado a cabo por Zhang et. al[27] se aborda una clasificación multietiqueta extrema.

Métodos basado en diccionarios

En el artículo [28] se realiza una aproximación a la coincidencia de patrones. La mayor desventaja es la pérdida de generalización que limita la extensión a otros dominios.

En el artículo [29] se desarrolló un enfoque basado en la identificación y extracción de frases que corresponden a patrones Hearst [30]. Se construyó un diccionario basado en

los nombres de medicamentos y sus sinónimos de DrugBank y se combinó con los corpus de texto del medicamento específico.

En el artículo [31] se propuso un sistema combinando información de diferentes diccionario UMLS, MeSH y DrugBank en textos no estructurados. Se procesaron los diccionarios y aplicaron filtrado basado en reglas, la revisión manual de términos frecuentes y la aplicación de reglas de desambiguación. El modelo fue evaluado sobre el un corpus de resúmenes de MEDLINE obteniendo un valor F del 50 %.

En el artículo [32] se utiliza una línea de procesos con preprocesador, etiquetador semántico, analizador de contexto e identificador de términos biomédicos. Obtuvo una medida de 90.6 % sobre textos no estructurados.

En el artículo [33] se extrajo información de resúmenes de MedLine sobre medicamentos y genes utilizando el recurso léxico UMLS Methasaurus. Consta de un etiquetador estocástico que analiza la sintaxis y usa información semántica y pragmática para construir sus afirmaciones.

En el artículo [34] hace uso de RxNorm, un diccionario de referencia cruzada de la nomenclatura clínica de medicamentos y la herramienta Metaphone que es un algoritmo que codifica cadenas de caracteres a una aproximación fonética de su pronunciación en inglés. La idea es aproximar cadenas de caracteres con la misma o similar codificación considerando que una codificación similar da como resultado una pronunciación similar. El sistema obtuvo un 92.2 % de sensibilidad y un 95.7 % de especificidad global.

En el artículo [35] se utiliza un enfoque basado en diccionarios combinando información de diferentes fuentes de datos farmacológicos como DrugBank, MeSH, RxNorm y ATC y un enfoque basado en ontologías utilizando los analizadores Metamap y Mgrep para mapear cada unidad de texto de un texto fuente en uno o más conceptos específicos del dominio. Los resultados de ambos enfoques son combinados y los resultados muestran que la aplicación de los enfoques propuestos individualmente es menor a la combinación de ambos enfoques. Los enfoques basados en diccionarios y en ontologías obtuvieron medidas F1 de 56.8 % y 58.5 % respectivamente mientras que el sistema combinado obtiene una medida F1 de 66.7 % con respecto a la evaluación de coincidencia exacta.

3.2. Tareas anteriores organizadas por eHealth

La importancia de la codificación clínica ha contribuido a la organización de competiciones y tareas en común para promover los sistemas de codificación automatizados que utilizan técnicas de aprendizaje automático, como la serie de laboratorios CLEF eHealth[2].

La Iniciativa CLEF (Conference and Labs of the Evaluation Forum, antes conocido como Cross-Language Evaluation Forum) es un organismo autoorganizado cuya misión principal es promover la investigación, la innovación y el desarrollo de sistemas de acceso a la información. Pone énfasis en la información multilingüe y multimodal con varios niveles de estructura. CLEF se estructura en dos partes principales: los laboratorios de evaluación y las conferencias. Las conferencias revisadas por pares incluyen investigación en metodologías de evaluación, experimentos con datos multilingües y multimodales y desafíos[36]. CLEF promueve la investigación y el desarrollo proporcionando una infraestructura para la:

- Prueba, ajuste y evaluación de sistemas multilingües y multimodales.
- Investigación del uso de datos no estructurados, semiestructurados, altamente estructurados y semánticamente enriquecidos en el acceso a la información.
- Creación de colecciones de pruebas reutilizables para evaluación comparativa.
- Exploración de nuevas metodologías de evaluación y formas innovadoras de utilizar datos experimentales.
- Discusión de resultados, comparación de enfoques, intercambio de ideas y transferencia de conocimientos.

Desde el año 2000, CLEF ha estimulado la investigación en áreas como la Recuperación de la Información (IR) y metodologías de evaluación. A lo largo de los años, se ha construido una comunidad de investigación que abarca diferentes áreas de especialización necesarias para hacer frente a la expansión de las actividades de CLEF.

Desde 2010, CLEF ha tomado la forma de un evento independiente, constituido por una conferencia revisada por pares organizada con un conjunto de laboratorios de evaluación.

CLEF eHealth [37] es un reto de evaluación en el dominio médico y biomédico. El objetivo es ofrecer recursos a los investigadores como bases de datos, marcos de evaluación y eventos. CLEF eHealth fue establecido en 2012 para preparar un Laboratorio de evaluación y desde 2013 ha estado realizando campañas anuales de evaluación en los siguientes dominios:

- Extracción de la Información (IE): clasificación de texto, detección de entidades nombradas, normalización de acrónimos, modelos multilenguaje, etc.

- Manejo de la Información (IM): visualización de los datos eHealth, manejo de textos médicos, etc.
- Recuperación de la Información (IR): basadas en sesiones, de tipo multilenguaje, etc.

La Figura 3.2 [37] representa las tareas organizadas por eHealth desde 2013 hasta 2019:

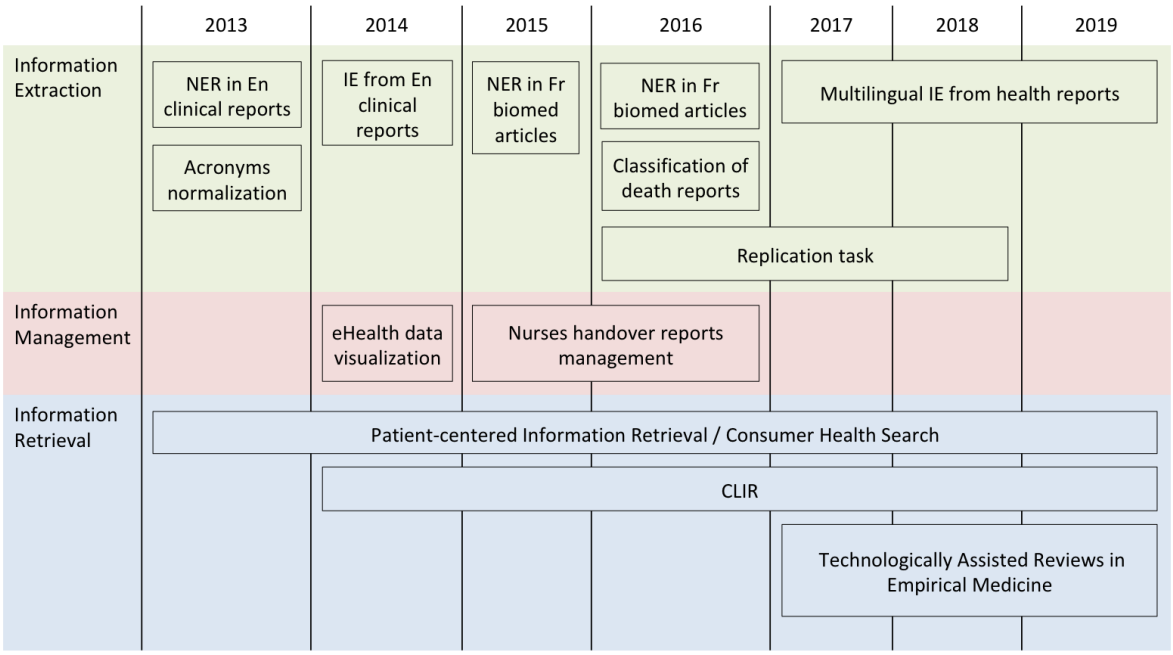

FIG. 3.2. TAREAS ORGANIZADAS POR CLEF

En la tarea eHealth del año 2017 [38] se utilizaron mayoritariamente fuentes léxicas como diccionarios y otras ontologías (UMLS). Se utilizaron concordancia exacta de cadenas entre términos, NER con búsqueda de diccionario basada en reglas y coincidencia aproximada usando Lucene Solr, tecnologías de semántica web, secuencias de modelos de aprendizaje profundo basados en diccionario, Tf-idf para recuperar términos de mayor peso, reglas para traducir acrónimos con un enfoque de ponderación binaria y clasificación con máquinas de soporte vectorial con bolsas de palabras, bigramas, distancia Dirichlet, ontologías Wodnet y UMLS.

En 2018, CLEF eHealth organizó una tarea de indexar certificados de defunción franceses y en 2019 se codificaron resúmenes no técnicos de experimentos de animales en alemán. La tarea del 2019 evolucionaba los enfoques de la tarea del año anterior.

En el artículo de la competición de eHealth 2018 [39], se presentó una tarea de Extracción de la Información y Normalización cuyo objetivo era impulsar los modelos adaptados al lenguaje y los multilenguaje en una tarea que abordaba varios idiomas. Se consiguió un método global que reconocía francés, húngaro e inglés a la vez. Otros métodos utilizados

fueron coincidencia con diccionario con diferentes niveles de variación léxica, técnicas de Recuperación de la Información y clasificación con redes convolucionales. El corpus se constituye de certificados de defunción codificados con ICD-10.

En el artículo [40] se utiliza aprendizaje automático estadístico y algoritmos simbólicos, expresiones matemáticas regulares para mapear datos test y atributos como el género y la edad como característica para entrenar el bosque aleatorio y el modelo Xgboost. Se clasifica en 26 categorías.

En el artículo [41] se transforma el texto en bruto en word-embeddings. Se utilizan redes convolucionales con múltiples filtros, diferentes tamaños de ventana además de los vectores de palabras obtenidos como la primera capa oculta de clasificación. Debido a la representación muy débil de algunos de los códigos ICD, se predice utilizando un clasificador que se basa en el reconocimiento de palabras de una base de conocimientos de un diccionario.

En el artículo [42] se utilizó una aproximación por diccionario, asignando un código por cada línea y detectando errores utilizando la distancia de Levenshtein.

En el artículo [43] se basaron en reglas para trasladar acrónimos, con un enfoque de ponderación binaria para seleccionar los ítem más similares del diccionario a la porción de texto y la clase con mayor ponderación.

En el artículo [44] se ha utilizado un modelo neuronal que intenta mapear los fragmentos de texto de entrada con los códigos ICD10 de salida. Su solución no hace suposiciones sobre el contenido de los datos de entrada y salida, tratándolos mediante un enfoque de aprendizaje automático que asigna un conjunto de etiquetas a cualquier línea de entrada. La solución es independiente del idioma, el tratamiento de un nuevo idioma solo necesita un conjunto de ejemplos, sin utilizar información específica del idioma aparte de los recursos terminológicos como los diccionarios ICD10, cuando estén disponibles.

En el artículo [45] se empleó un enfoque neuronal codificador-decodificador a nivel de documento. El codificador convolucional opera a nivel de carácter y el decodificador es recurrente. El enfoque de coincidencia de cadenas se basa en los diccionarios proporcionados y utiliza una representación de n-gramas de palabras (1-5) para buscar coincidencias.

El articulo [46] se utiliza aprendizaje supervisado usando perceptrón multicapa y estrategia One-vs-Rest. El entrenamiento fue llevado a cabo con los datos de entrenamiento y un diccionario, estimando la frecuencia de los términos ponderados con separación binormal. También ae limita el bias, generando modelos de aprendizaje para los códigos que aparecen más de 100 veces en el conjunto de entrenamiento. Las enfermedades sin clasificar por esos modelos son usadas para construir llamadas y aplicar los motores de búsqueda con las descripciones de los códigos.

En el artículo [47] se aprovechó el gran tamaño y la naturaleza textual de los datos train mediante un enfoque de aprendizaje basado en instancias. Las 360.000 oraciones

anotadas contenidas en los datos de entrenamiento se indexaron con un motor de búsqueda estándar. Luego, los k-Vecinos más cercanos de una oración de entrada fueron explotados para inferir códigos potenciales, gracias a la votación por mayoría. También se utilizó un enfoque basado en diccionario para mapear códigos directamente en oraciones, y ambos enfoques se combinaron linealmente.

En el artículo [48] se utilizaron diferentes algoritmos de aprendizaje automático. Primero, su sistema encontró todos los posibles códigos ICD10 buscando cuántas palabras de cada código existen en el texto. A continuación, se calcularon varias medidas de calidad de estos códigos. Con estas métricas se entrenaron diferentes algoritmos de aprendizaje automático y finalmente se seleccionó el mejor modelo para usar en el sistema. La mayoría de las técnicas utilizadas son independientes del idioma, por lo que el sistema es fácilmente adaptable a otros idiomas.

En el artículo [49] se elige un enfoque de aprendizaje profundo a partir de los datos train proporcionados. Se utilizó OpenNMT-py, un marco de código abierto para traducción automática neuronal implementado en PyTorch.

3.3. Trabajos participantes en CodiEsp

CodiEsp es una competición organizada por CLEF eHealth en el año 2020. Se ha decidido participar en la tarae de Extracción Multilingue de Información aplicada a diagnósticos (Task1). Gracias a ello, se ha utilizado un corpus Gold Standard llamado igualmente CodiEsp anotado manualmente y comparado los resultados con otros métodos. Esta tarea requiere de un conocimiento transversal a diversas áreas:

- Medicina, biomedicina, bioinformática e Informática médica.
- Lingüística, ciencias informáticas y Lingüística computacional.
- Minería de datos, Big Data y Minería de texto.
- Inteligencia Artificial, Aprendizaje automático, Aprendizaje profundo.
- Procesamiento del Lenguaje Natural y Recuperación de la Información.

La Figura 3.3 procede de la web de CodiEsp[2] y muestra la codificación de una historia clínica. En primer lugar, se debe de encontrar los términos clínicos referentes a diagnósticos.

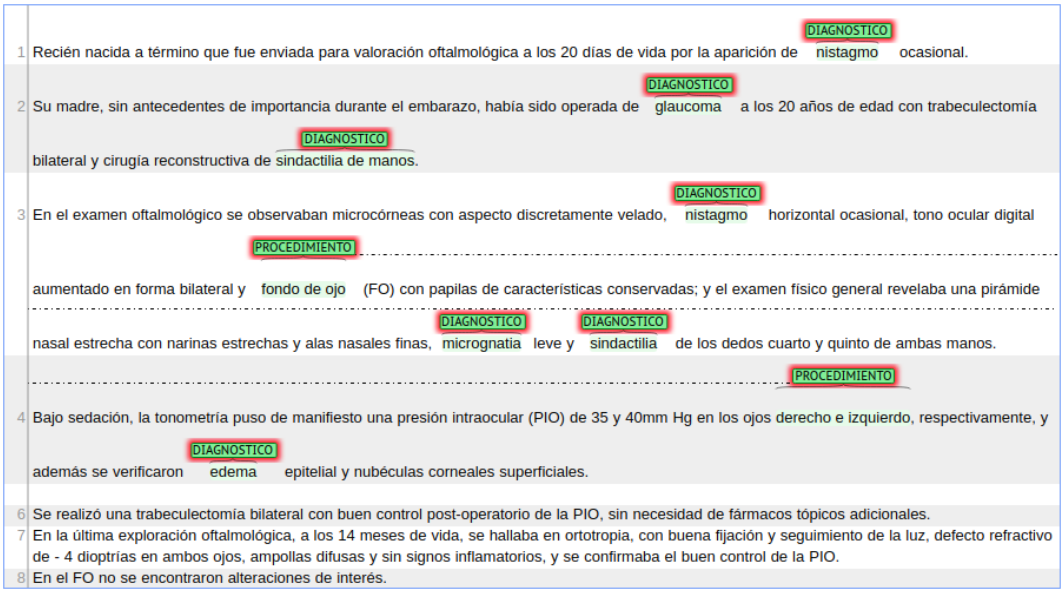

FIG. 3.3. IMAGEN DE ETIQUETADO PROCEDENTE DE CODIESP

En segundo lugar, se debe de predecir los códigos ICD-10-CM correspondientes a los códigos clínicos.

En este apartado se sintetiza los trabajos revisados de la competición cuyos resultados y métodos aparecieron tras desarrollar y presentar el sistema propio en Codiesp.

Según el artículo de visión general de CodiEsp [50] se han detectado tres tipos de enfoques principales:

- Clasificación: se considera que un conjunto de documentos deben de ser categorizados. Estos documentos pueden estar constituidos a nivel de palabra, oración, párrafo y toda la historia clínica. Existen tantas categorías como códigos ICD-10. El artículo [51] utiliza este enfoque.

- Reconocimiento de Entidades Nombradas: se centra en detectar si cada palabra de una historia clínica forma parte de un diagnóstico. Los artículos [52] y [42] lo utilizan.

- Combinación: se combinan los sistemas de Reconocimiento de Entidades Nombradas y clasificación. El artículo [53] es un ejemplo de ello.

Tanto en clasificación como en Reconocimiento de Entidades nombradas se pueden utilizar diferentes tipos de tecnologías. Por ejemplo:

- Sin Aprendizaje Automático: por ejemplo, en [42] se utiliza un enfoque basado en diccionario para realizar Reconocimiento de Entidades Nombradas.

- Con Aprendizaje Automático: por ejemplo, en [53] se utiliza XGBoost para realizar una clasificación de documentos.

- Modelos de Lenguaje: en el artículo [52] se perfeccionó el modelo de lenguaje BERT Multilingual y en [54] se utilizó BETO.

Algunos de los métodos utilizados para la misma tarea[50] son lo que se explican a continuación.

En el artículo [53] se ha combinado un motor de aprendizaje automático con un sistema de similitud de cadenas. Primero, se entrenó un clasificador binario XGBoost para cada etiqueta y se ampliaron los textos concatenando las entidades médicas extraídas de los propios documentos. En segundo lugar, el sistema de similitud de cadenas comparó fragmentos de texto y definiciones de códigos utilizando la distancia de Levenshtein, el algoritmo Jaro Winkler y la similitud de coseno en representaciones BERT multilingües. El mejor sistema obtuvo un MAP igual a 0.593 con una combinación de XGBoost y el algoritmo Jaro Winkler.

En el artículo [42] se explica un sistema basado en un diccionario con una estructura de datos de árbol construida a partir de anotaciones CodiEsp y terminología ICD-10. Las entidades se detectan en los documentos nuevos si coinciden con alguna de las entradas almacenadas del diccionario. La coincidencia se realiza teniendo en cuenta la distancia de Levenshtein y el tratamiento de abreviaturas. Para mejorar aún más la precisión del sistema, eliminaron los términos que conducen a muchos falsos positivos. Su sistema obtuvo la puntuación f1 más alta con 0,687.

En el artículo [54] se entrenó un modelo de lenguaje, se agregó una capa de clasificación lineal y se ajustó el modelo para realizar NER en el corpus. Además, se probó

un sistema basado en un campo aleatorio condicional (CRF) para realizar la misma tarea NER, pero concluyeron que era demasiado conservador en la detección de entidades.

En el artículo [52] se utilizó un enfoque de dos pasos para detectar códigos. En primer lugar, se identificaron las entidades de diagnóstico utilizando un sistema NER basado en un BERT multilingüe previamente entrenado. En segundo lugar, las entidades reconocidas se emparejaron con las definiciones del código ICD-10 utilizando la distancia de Levenshtein. En tercer lugar, se utilizó un algoritmo de aumento del corpus de entrenamiento. El algoritmo de aumento de texto se entrenó con el corpus oficial y se tradujeron al español ejemplos de la base de datos PubMed y MIMIC. En cuarto lugar, se probaron diferentes métodos de pos-procesamiento para detectar y eliminar entidades negadas y superpuestas. Finalmente se clasificaron los códigos según la confianza utilizando la frecuencia y posición de las entidades. Su resultados fueron buenos.

En el artículo [50] se han combinado las predicciones de diferentes enfoques. Se unieron los trabajos [52] y [42] que utilizaron por un lado Bert para NER y distancia Levenshtein, y por otro, búsqueda de diccionario sintonizada. Para ello, se consideraron todas las predicciones de ambos sistemas. Se obtuvo un resultado en predicción superior, más balanceado, con mejor exhaustividad y una precisión un poco más baja que los dos sistemas anteriores.

En el artículo [55] se investiga la combinación de técnicas con mejores resultados. Se introduce un componente jerárquico que explota el conocimiento de la taxonomía ICD.

Propuesta

Los métodos basados en diccionario identifican entidades compuestas por múltiples palabras debido a que este tipo de entidades o términos se encuentran presentes en los diccionarios. Así mismo, se identifica palabras homónimas o palabras con múltiples significados. Sin embargo, existen desventajas. Por ejemplo, sólo se identifican entidades presentes en los diccionarios, los errores ortográficos dificultan la coincidencia, la terminología se encuentra en constante crecimiento y los métodos de coincidencia de cadenas de caracteres no logran buenos rendimientos en presencia de datos con ruido.

Debido a que la coincidencia de cadenas tiene diversos inconvenientes, se ha optado por añadir coincidencia difusa, que aumenta la flexibilidad y mejora los resultados a pesar de errores ortográficos y ruido en los datos.

Además, en la tarea de Codiesp se debe afrontar un problema de clasificación multi-etiqueta extremo. La tarea de clasificación de CodiEsp 2020 es verdaderamente complicada respecto a tareas previas. Existen varios trabajos que abordan cada una de los retos o características críticas de la clasificación multi-etiqueta. Sin embargo, hasta ahora no se habían presentado todas las limitaciones a la vez, particularmente en el dominio clínico en español. Algunas características críticas en la clasificación multi-etiqueta de la tarea son[26]:

I. Configuración multi-etiqueta. Significa que un texto puede tener más de una clase, tantas clases como etiquetas. En este caso, hay bastantes etiquetas por instancia.

   En la Figura 3.4 se muestra un gráfico que relaciona la frecuencia de aparición de una cantidad de clases y el número de clases por historia clínica.

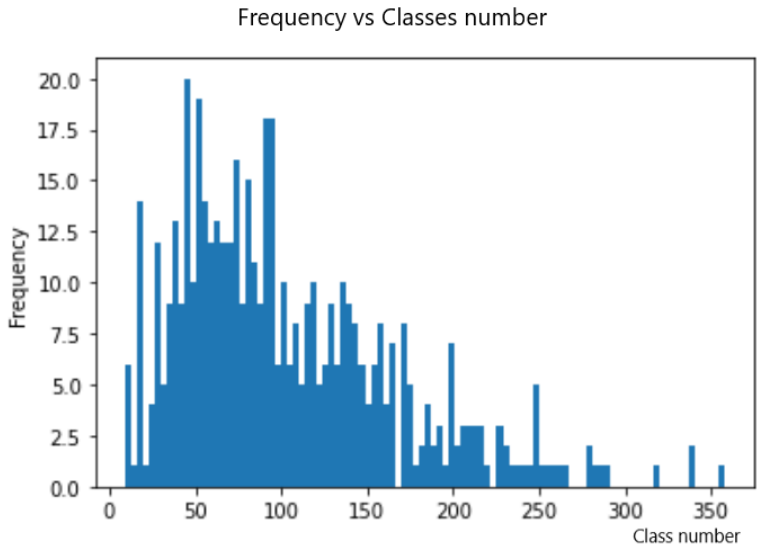

FIG. 3.4. GRÁFICO DE FRECUENCIA VS NÚMERO DE CLASES POR HISTORIA CLÍNICA

II. El número de etiquetas. Cada etiqueta es un código único de ICD-10. En total, el conjunto de etiquetas es muy grande. Sin embargo, en el conjunto train sólo aparecen 1767 códigos únicos y en dev 1158.

III. El número de instancias. En este caso, 500 historias en train y 250 historias en dev.

IV. La longitud de las instancias. Las instancias son las historias de caso clínico. Una característica del domino clínico son documentos largos con un largo número de palabras técnicas y errores de escritura. En este caso son documentos largos, con una longitud media de 1999 palabras.

   En la Figura 3.5 se muestra un gráfico que relaciona la frecuencia de aparición y la longitud de la cadena de una historia clínica.

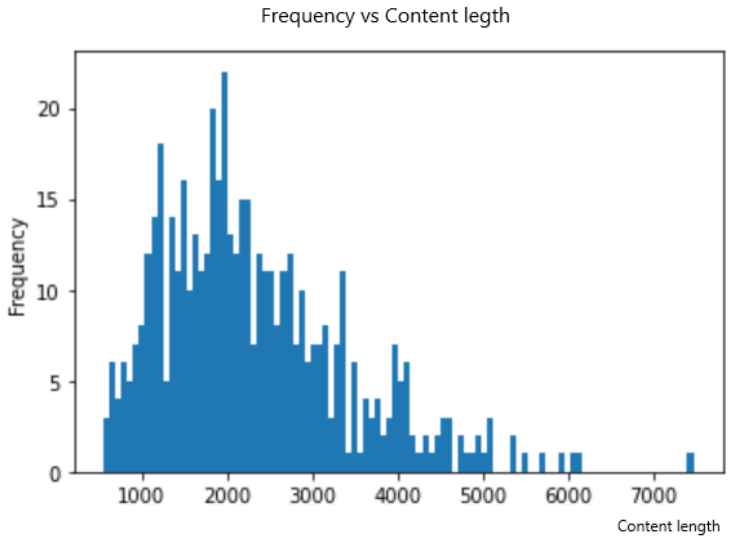

FIG. 3.5. GRÁFICO DE FRECUENCIA VS LONGITUD DE HISTORIAS CLÍNICAS

v. Transparencia. Al utilizar documentos extensos, es útil remarcar los elementos presentes en el documento que apoya y explica la precisión del código o etiqueta. Para ello se puede aportar una referencia del texto relacionada con la etiqueta.

Por tanto, se debe ofrecer una codificación multi-etiqueta que asigne más de un código a una historia clínica. Existe gran variedad de códigos totales y pocos códigos presentes en las historias de entrenamiento. Esto es un reto diferenciador y que se ha resuelto en la tarea de CodiEsp.

Por tanto, la aportación de este trabajo es abarcar historias clínicas en español en un problema de clasificación multi-etiqueta extrema y utilizando coincidencia difusa para codificar.

4. METODOLOGÍA DE INVESTIGACIÓN Y ENFOQUE

El objetivo es desarrollar un clasificador de enfermedades con códigos ICD-10-CM sobre historias de caso clínico en español.

A la hora de describir el trabajo realizado, es importante señalar que se trata de un trabajo aplicado ya que se desarrolla un programa informático. Por ello, la metodología de trabajo consta de varias partes teórico-prácticas.

4.1. Pasos metodológicos

En la Figura 4.1 se muestra un diagrama con la metodología del presente trabajo.

En primer lugar, se ha realizado un estudio teórico del problema. En el primer capítulo se explica la motivación del problema, el síntoma, las causas, el alcance, la importancia y los beneficios esperados tras su resolución. Todo ello permite enfocar el objetivo de trabajo y definir la hipótesis de investigación. En el capítulo segundo se sitúa el trabajo en su área, explicando el marco clínico-técnico necesario para detallar los problemas técnicos que pueden surgir en el trabajo. En el capítulo tercero realizar un estado del arte crítico para analizar las soluciones presentadas en el área de investigación, y más específicamente, en tareas similares de eHealth. También se plantea la participación en la tarea de etiquetado multi-etiqueta CodiEsp 2020 y la propuesta del alcance de este proyecto. En este capítulo, se aborda la metodología de investigación, se describe el corpus CodiEsp y se justifica la elección del enfoque que da lugar a la solución al problema. Se ha probado con 3 aproximaciones diferentes, de las cuales la tercera ha obtenido buenos resultados.

En segundo lugar, se ha procedido a desarrollar el sistema informático que clasifique enfermedades y se ha descrito en el capítulo quinto. Para ello, se ha elegido y estudiado el entorno de desarrollo, se han preparado los datos de entrada y el formato de los datos a generar y se ha desarrollado el clasificador de historias de caso clínico. Además, se ha evaluado el clasificador con diferentes métricas, y finalmente se ha generado una visualización del resultado en una base de datos orientada a grafos.

En tercer lugar, se ha realizado la experimentación, se han evaluado, discutido y comparado los resultados del sistema con otros sistemas similares en la tarea CodiEsp 2020. Esto se describe en el capítulo séptimo.

En cuarto lugar, se han planteado en el capítulo octavo las conclusiones, limitaciones y mejoras del trabajo.

Finalmente, se han documentado todas las fases de la investigación acorde a la metodología presente y se han incluido en el sexto capítulo la planificación temporal por tareas realizadas y el marco legal de los datos clínicos y de las herramientas de trabajo.

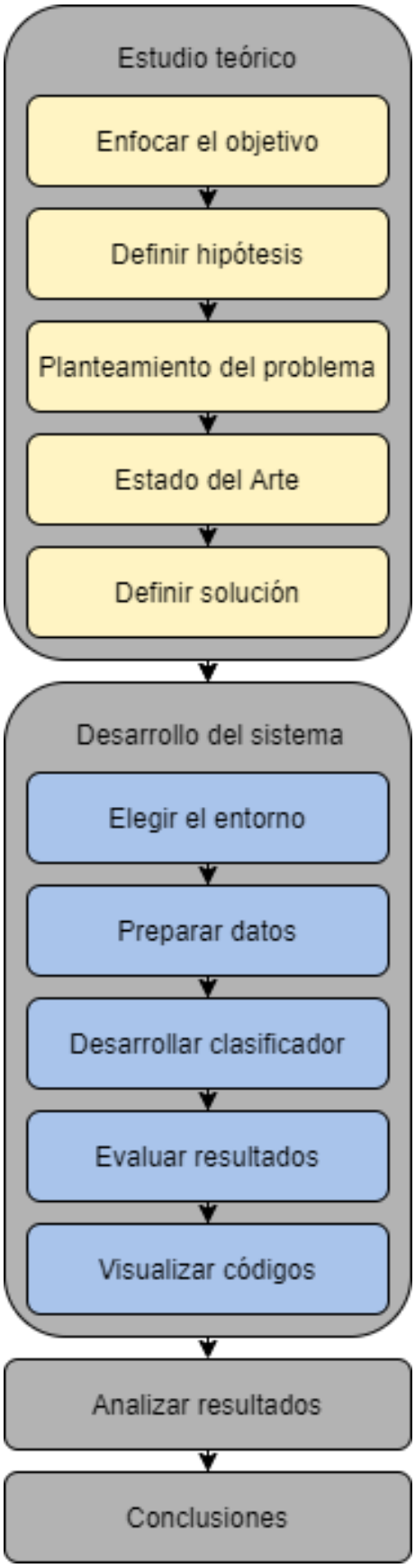

FIG. 4.1. METODOLOGÍA DE TRABAJO

4.2. Descripción del corpus CodiEsp

El corupus CodiEsp está disponible públicamente y se ha utilizado la versión 1.2[56]. Contiene 1000 historias de caso clínico en Español. Su formato es texto libre, están codificado por profesionales de la salud y traducido de forma automática del español al inglés.

Durante el desarrollo de la competición no se encontraban disponibles las anotaciones del conjunto test y sus historias eran las necesarias para la inferencia y predicción de códigos.

Los conjuntos train y dev estuvieron disponibles en todo momento, por lo que el segundo conjunto tiene la función de validar el rendimiento del modelo durante las pruebas de desarrollo. Sin embargo, cuando se quiere realizar una inferencia sobre datos de entrada nuevos, los datos nuevos se consideran el conjunto test. Esto permite estimar el rendimiento del sistema sobre el conjunto de datos sobre el que se desconoce el resultado verdadero. Existen varias versiones de la anotación manual con corpus.

Código de Diagnóstico

Los organizadores ofrecieron un conjunto de códigos válidos ICD-10-CM en español e inglés. El fichero se llama codiesp-D_codes.tsv. Posee 3 columnas y 98288 filas. Esto indica que existen 98288 códigos únicos en la versión modificada de la décima edición de códigos ICD. La primera columna contiene el código ICD-10-CM, con diferente longitud, comprendida en el rango entre 3 y 8 caracteres. La segunda y tercera columna indica la descripción en español e inglés del código y que es crucial para desarrollar el sistema. Existen códigos con una letra desde la A a la Z. A continuación, se encuentran dos dígitos, que indican la categoría específica. Finalmente, se encuentran dígitos tras un punto. En las figuras 4.2 y 4.3 se muestra el comienzo y final del fichero con los códigos.

	Code	Es-description	En-description
0	A00.0	Cólera debido a Vibrio cholerae 01, biotipo cholerae	Cholera due to Vibrio cholerae 01, biovar cholerae
1	A00.1	Cólera debido a Vibrio cholerae 01, biotipo El Tor	Cholera due to Vibrio cholerae 01, biovar eltor
2	A00.9	Cólera, no especificado	Cholera, unspecified
3	A01.00	Fiebre tifoidea, no especificada	Typhoid fever, unspecified
4	A01.01	Meningitis tifoidea	Typhoid meningitis
5	A01.02	Fiebre tifoidea con afectación cardiaca	Typhoid fever with heart involvement
6	A01.03	Neumonía tifoidea	Typhoid pneumonia
7	A01.04	Artritis tifoidea	Typhoid arthritis
8	A01.05	Osteomielitis tifoidea	Typhoid osteomyelitis
9	A01.09	Fiebre tifoidea con otras complicaciones	Typhoid fever with other complications

FIG. 4.2. COMIENZO DE CÓDIGOS DE DIAGNÓSTICOS

	Code	Es-description	En-description
71476	Z98.871	Historia personal de procedimiento intrauterino en etapa fetal	Personal history of in utero procedure while a fetus
71477	Z98.890	Otros estados posprocedimiento especificados	Other specified postprocedural states
71478	Z98.891	Historia de cicatriz uterina por cirugía previa	History of uterine scar from previous surgery
71479	Z99.0	Dependencia de aspirador	Dependence on aspirator
71480	Z99.11	Estado de dependencia de respirador [ventilador]	Dependence on respirator [ventilator] status
71481	Z99.12	Contacto por dependencia de respirador [ventilador] durante falta de electricidad	Encounter for respirator [ventilator] dependence during power failure
71482	Z99.2	Dependencia de diálisis renal	Dependence on renal dialysis
71483	Z99.3	Dependencia a silla de ruedas	Dependence on wheelchair
71484	Z99.81	Dependencia de oxígeno suplementario	Dependence on supplemental oxygen
71485	Z99.89	Dependencia de otras máquinas y dispositivos de apoyo	Dependence on other enabling machines and devices

FIG. 4.3. FIN DE CÓDIGOS DE DIAGNÓSTICOS

Conjunto train (de entrenamiento)

Contiene los ejemplos del modelo utilizados para aprender la función que relaciona los datos de entrada con los de salida. La ruta es train/trainD.tsv. En su interior se encuentra el identificador de las diferentes historias clínicas de train y el código ICD-10 asignado a cada historia clínica. Estos códigos puedes ser generales, como r69, o específicos, como r97.1 o r06.00. El tamaño de trainD es 5639 filas con 2 columnas (el identificador de la nota clínica y los códigos). Se comprueba que no se presentan códigos repetidos en la misma historia clínica. Se encuentran 1767 códigos diferentes en el conjunto de train de las historias clínicas.

Para el entrenamiento, se utilizan:

- Historias train en español e inglés: se componen de 500 archivos de texto en español (e inglés) para entrenar. Cada uno se compone de varias frases. Tiene un identificador que señala el nombre del text file. Este id se repite en las anotaciones. En las anotaciones, las text-references son entidades que hay en cada text-file, y permite identificar qué es una entidad, su código CIE-10 y las etiquetas de diagnóstico.

  En las figuras 4.4 y 4.5 se muestran 500 casos de artículos con su ID correspondiente.

Articles, train, spanish

	article_id	article_content
0	S0004-06142005000700014-1	Describimos el caso de un varón de 37 años con...
1	S0378-48352006000600006-1	Varón de 73 años en su primera visita a nuestr...
2	S0004-06142005000900013-1	Se trata de una mujer de 29 años sometida a un...
3	S0378-48352006000900006-1	Varón de 43 años, fumador activo de unos 15 ci...
4	S0004-06142005000900015-1	Varón de 36 años, sin antecedentes de interés,...
...	...	...
495	S2254-28842012000300010-1	Varón de 72 años en programa de HD desde dicie...
496	S0378-48352006000400005-1	Paciente de 53 años de edad en el momento del ...
497	S2254-28842014000200009-1	Mujer de 33 años, que llega a urgencias con un...
498	S0378-48352006000500005-1	Varón de 63 años sin antecedentes de interés. ...
499	S2340-98942015000100005-1	Presentamos el caso de una paciente de 62 años...

500 rows × 2 columns

FIG. 4.4. HISTORIAS TRAIN EN ESPAÑOL

Articles, train, english

	article_id	article_content
0	S0004-06142005000700014-1	We describe the case of a 37-year-old man with...
1	S0378-48352006000600006-1	A 73-year-old male presented at our department...
2	S0004-06142005000900013-1	We report the case of a 29-year-old woman who ...
3	S0378-48352006000900006-1	A 43-year-old man, an active smoker of about 1...
4	S0004-06142005000900015-1	A 36-year-old male, with no relevant past medi...
...	...	...
495	S2254-28842012000300010-1	A 72-year-old male in HD program from December...
496	S0378-48352006000400005-1	A 53-year-old patient at the time of diagnosis...
497	S2254-28842014000200009-1	A 33-year-old woman presented to the emergency...
498	S0378-48352006000500005-1	A 63-year-old male with no relevant past medic...
499	S2340-98942015000100005-1	We report the case of a 62-year-old patient wh...

500 rows × 2 columns

FIG. 4.5. HISTORIAS TRAIN EN INGLÉS

- Anotaciones de códigos sobre las historias: éstas incluyen las etiquetas para cada Historia. En las figuras 4.6 y 4.7 se muestran las tablas de train en español e inglés.

Joined, train, spanish

	Article_ID	Article_content	ICD10-tags
0	S0004-06142005000700014-1	Describimos el caso de un varón de 37 años con...	['n44.8', 'z20.818', 'r60.9', 'r52', 'a23.9', ...
1	S0378-48352006000600006-1	Varón de 73 años en su primera visita a nuestr...	['r22.2', 'i25.9', 'r69', 'k59.9', 'f17.290', ...
2	S0004-06142005000900013-1	Se trata de una mujer de 29 años sometida a un...	['d30.3', 'r58']
3	S0378-48352006000900006-1	Varón de 43 años, fumador activo de unos 15 ci...	['r11.0', 'r10.13', 'c26.0', 'r91.8', 'c34.9',...
4	S0004-06142005000900015-1	Varón de 36 años, sin antecedentes de interés,...	['r19.00', 'n50.9', 'd49.59', 'r63.4', 'r52', ...
...	...	...	...
495	S2254-28842012000300010-1	Varón de 72 años en programa de HD desde dicie...	['n19', 'k55.9', 'b99.9', 'b96.20', 'k63.5', '...
496	S0378-48352006000400005-1	Paciente de 53 años de edad en el momento del ...	['c77.2', 'c77.9', 'c49.5', 'c78.01', 'c64.9',...
497	S2254-28842014000200009-1	Mujer de 33 años, que llega a urgencias con un...	['d64.9', 'r53.1', 'r63.0', 'r63.4', 'k06.8', ...
498	S0378-48352006000500005-1	Varón de 63 años sin antecedentes de interés. ...	['r16.0', 'r63.4', 'r53.1', 'd49.0', 'c78.7', ...
499	S2340-98942015000100005-1	Presentamos el caso de una paciente de 62 años...	['c78.7', 'r11.10', 'r69', 'r06.00', 'c56.2', ...

500 rows × 3 columns

FIG. 4.6. HISTORIAS TRAIN EN ESPAÑOL CON CÓDIGOS ANOTADOS

Joined, train, english

	Article_ID	Article_content	ICD10-tags
0	S0004-06142005000700014-1	We describe the case of a 37-year-old man with...	['n44.8', 'z20.818', 'r60.9', 'r52', 'a23.9', ...
1	S0378-48352006000600006-1	A 73-year-old male presented at our department...	['r22.2', 'i25.9', 'r69', 'k59.9', 'f17.290', ...
2	S0004-06142005000900013-1	We report the case of a 29-year-old woman who ...	['d30.3', 'r58']
3	S0378-48352006000900006-1	A 43-year-old man, an active smoker of about 1...	['r11.0', 'r10.13', 'c26.0', 'r91.8', 'c34.9',...
4	S0004-06142005000900015-1	A 36-year-old male, with no relevant past medi...	['r19.00', 'n50.9', 'd49.59', 'r63.4', 'r52', ...
...	...	...	...
495	S2254-28842012000300010-1	A 72-year-old male in HD program from December...	['n19', 'k55.9', 'b99.9', 'b96.20', 'k63.5', '...
496	S0378-48352006000400005-1	A 53-year-old patient at the time of diagnosis...	['c77.2', 'c77.9', 'c49.5', 'c78.01', 'c64.9',...
497	S2254-28842014000200009-1	A 33-year-old woman presented to the emergency...	['d64.9', 'r53.1', 'r63.0', 'r63.4', 'k06.8', ...
498	S0378-48352006000500005-1	A 63-year-old male with no relevant past medic...	['r16.0', 'r63.4', 'r53.1', 'd49.0', 'c78.7', ...
499	S2340-98942015000100005-1	We report the case of a 62-year-old patient wh...	['c78.7', 'r11.10', 'r69', 'r06.00', 'c56.2', ...

500 rows × 3 columns

FIG. 4.7. HISTORIAS TRAIN EN INGLÉS CON CÓDIGOS ANOTADOS

Conjunto dev (de desarrollo)

Contiene los ejemplos que no son utilizados para aprender la función. En cambio, se usan para medir el rendimiento generalizado del modelo de forma periódica a medida que se desarrolla el modelo. Este conjunto es opcional, ya que su utilidad es similar al conjunto de test. La ruta es dev/devD.tsv. En su interior se encuentra el identificador de las diferentes historias clínicas del conjunto dev, y el código ICD-10 asignado a cada historia clínica. La estructura es similar al conjunto train. El tamaño de devD es 2676 filas con 2 columnas (el identificador de la nota clínica y los códigos). Se comprueba que no se presentan códigos repetidos en la misma historia clínica. Se encuentran 1158 códigos diferentes en el conjunto dev de las historias clínicas. A los 1158 códigos diferentes del conjunto dev, se le restan los códigos que aparecieron en train, el resultado son 427 códigos en dev que no aparecen en el entrenamiento. Si se le suma a los 1767 códigos diferentes de train los 427 nuevos códigos diferentes del conjunto de test, se contabiliza un total de 2194 códigos ICD-10 presentes en los conjuntos train y dev. Para predecir sobre el conjunto dev, solo se necesitan las historias dev en español o inglés.

Para el desarrollo, se utilizan:

- Historias dev en español e inglés: se componen de 250 archivos de texto en español (e inglés). Cada uno se compone de varias frases. Tiene un identificador que señala el nombre del text file. Este id se repite en las anotaciones. En las anotaciones, las text-references son entidades que hay en cada text-file, y permite identificar que es una entidad, su código CIE-10 y las etiquetas de diagnóstico.

  En las figuras 4.8 y 4.9 se muestran 250 casos destinados al desarrollo.

Articles, dev

	article_id	article_content
0	S0004-06142005000900016-1	Mujer de 29 años con antecedentes de ulcus duo...
1	S0376-78922017000100008-1	Varón de 38 años que trabaja como miembro de l...
2	S0004-06142005001000011-1	Varón de 58 años de edad en el momento del tra...
3	S0378-48352004000200007-1	Presentamos el caso de un varón de 59 años que...
4	S0004-06142006000200011-1	Paciente varón de 22 años de edad, sin anteced...
...	...	...
245	S1889-836X2015000200005-2	Describimos el caso de una paciente de 58 años...
246	S0376-78922015000200010-1	Varón de 3 meses de edad producto de quinta ge...
247	S1889-836X2015000400006-1	Presentamos el caso de un varón de 64 años con...
248	S0376-78922016000200012-1	Varón de 77 años de edad, sin hábitos tóxicos,...
249	S2254-28842013000300009-1	Mujer de 73 años de edad con antecedentes pers...

250 rows × 2 columns

FIG. 4.8. HISTORIAS DEV EN ESPAÑOL

Articles, dev, english

	article_id	article_content
0	S0004-06142005000900016-1	A 29-year-old woman with a history of duodenal...
1	S0376-78922017000100008-1	A 38-year-old man who works as a member of the...
2	S0004-06142005001000011-1	58-year-old male at the time of transplantatio...
3	S0378-48352004000200007-1	We report the case of a 59-year-old man who co...
4	S0004-06142006000200011-1	A 22-year-old male patient, with no history of...
...	...	...
245	S1889-836X2015000200005-2	We describe the case of a 58-year-old woman.\n...
246	S0376-78922015000200010-1	A 3-month-old male, product of fifth pregnancy...
247	S1889-836X2015000400006-1	We report the case of a 64-year-old male with ...
248	S0376-78922016000200012-1	A 77-year-old male, with no toxic habits, with...
249	S2254-28842013000300009-1	A 73-year-old woman with a history of hyperten...

250 rows × 2 columns

FIG. 4.9. HISTORIAS DEV EN INGLÉS

- Anotaciones de códigos sobre las historias: éstas incluyen las etiquetas para cada Historia.

  En las figuras 4.10 y 4.11 se muestra para development.

Joined, dev, spanish

	Article_ID	Article_content	ICD10-tags
0	S0004-06142005000900016-1	Mujer de 29 años con antecedentes de ulcus duo...	['q62.11', 'n28.89', 'n39.0', 'r31.9', 'n23', ...
1	S0376-78922017000100008-1	Varón de 38 años que trabaja como miembro de l...	['i96', 't14.90', 'v29.9xx', 's69.92x', 's62.0...
2	S0004-06142005001000011-1	Varón de 58 años de edad en el momento del tra...	['t86.11', 'i25.10', 'n02.8', 'r25.1', 'n05.1'...
3	S0378-48352004000200007-1	Presentamos el caso de un varón de 59 años que...	['c70.0', 'r22.0', 'c79.51', 'r58', 'd49.7', '...
4	S0004-06142006000200011-1	Paciente varón de 22 años de edad, sin anteced...	['r52', 'f17.200', 'r50.9', 'f12.10', 'd49.59'...
...	...	...	...
245	S1889-836X2015000200005-2	Describimos el caso de una paciente de 58 años...	['r52', 'r60.9', 'e66.9', 'z90.710', 's92.322'...
246	S0376-78922015000200010-1	Varón de 3 meses de edad producto de quinta ge...	['j33.9', 'r56.9', 'q05.9', 'q04.0', 'h04.201'...
247	S1889-836X2015000400006-1	Presentamos el caso de un varón de 64 años con...	['e72.09', 'e83.39', 'm62.81', 'r52', 'r80.9',...
248	S0376-78922016000200012-1	Varón de 77 años de edad, sin hábitos tóxicos,...	['i10', 'e78.00', 'c79.51', 'i25.9', 'z87.442'...
249	S2254-28842013000300009-1	Mujer de 73 años de edad con antecedentes pers...	['f32.9', 'd13.5', 'e27.40', 'z99.2', 'j45.909...

250 rows × 3 columns

FIG. 4.10. HISTORIAS DEV EN ESPAÑOL CON CÓDIGOS ANOTADOS

Joined, dev, english

	Article_ID	Article_content	ICD10-tags
0	S0004-06142005000900016-1	A 29-year-old woman with a history of duodenal...	['q62.11', 'n28.89', 'n39.0', 'r31.9', 'n23', ...
1	S0376-78922017000100008-1	A 38-year-old man who works as a member of the...	['i96', 't14.90', 'v29.9xx', 's69.92x', 's62.0...
2	S0004-06142005001000011-1	58-year-old male at the time of transplantatio...	['t86.11', 'i25.10', 'n02.8', 'r25.1', 'n05.1'...
3	S0378-48352004000200007-1	We report the case of a 59-year-old man who co...	['c70.0', 'r22.0', 'c79.51', 'r58', 'd49.7', '...
4	S0004-06142006000200011-1	A 22-year-old male patient, with no history of...	['r52', 'f17.200', 'r50.9', 'f12.10', 'd49.59'...
...	...	...	...
245	S1889-836X2015000200005-2	We describe the case of a 58-year-old woman.\n...	['r52', 'r60.9', 'e66.9', 'z90.710', 's92.322'...
246	S0376-78922015000200010-1	A 3-month-old male, product of fifth pregnancy...	['j33.9', 'r56.9', 'q05.9', 'q04.0', 'h04.201'...
247	S1889-836X2015000400006-1	We report the case of a 64-year-old male with ...	['e72.09', 'e83.39', 'm62.81', 'r52', 'r80.9',...
248	S0376-78922016000200012-1	A 77-year-old male, with no toxic habits, with...	['i10', 'e78.00', 'c79.51', 'i25.9', 'z87.442'...
249	S2254-28842013000300009-1	A 73-year-old woman with a history of hyperten...	['f32.9', 'd13.5', 'e27.40', 'z99.2', 'j45.909...

250 rows × 3 columns

FIG. 4.11. HISTORIAS DEV EN INGLÉS CON CÓDIGOS ANOTADOS

Conjunto test (de validación)

Contiene los ejemplos que no son incorporados al entrenamiento del modelo. Son necesarios para comprobar el rendimiento de la generalización del modelo al final del desarrollo. Esto conjunto contiene 250 historias anotadas (goldstandard set) y 2751 historias sin anotar (background set). Las historias del conjunto de fondo evita que se conozcan cuáles de las notas del conjunto de test son las historias anotadas que se van a evaluar. Además de las historias se ofrecen las anotaciones del conjunto goldstandar. Las anotaciones tienen los siguientes campos: articleID, con el identificador de las historias, code, con los códigos encontrados en cada historia. Se utilizan para dos tareas:

- Predecir: con las historias test(background+goldstandard) en español e inglés.
- Validación: con las historias test(background+goldstandard) en español e inglés y las anotaciones de códigos manuales goldstandard.

4.3. Análisis estadístico descriptivo del Corpus

La Tabla 4.1 presenta algunos datos estadísticos sobre el corpus[50].

Datos	Total	Train	Dev	Goldstandard
Historias	1000	500	250	250
Anotaciones	14305	7209	3431	3665
Códigos únicos	2557	1767	1158	1143
Frases	16684	8105	4381	4198
Tokens	411067	204815	102719	103533

TABLA 4.1. ANÁLISIS ESTADÍSTICO DE CODIESP

Existen 98288 códigos ICD-10-CM disponibles. En el corpus, la proporción de códigos está poco balanceada. Al sumar todas las etiquetas (aunque repetidas), en las 500 notas de train aparecen 5639 etiquetas. De ellas, sólo aparecen 1767 códigos en el conjunto de train. Por otro lado, en todo el conjunto dev aparecen 2676 (algunos repetidos), de los cuáles 1158 son únicos. Además puede haber códigos repetidos en el conjunto de train (1767) y el conjunto de dev (2676), por lo que en total sólo aparecen 2194 códigos ICD-10 únicos en el conjunto train y dev. Esto contrasta con los 98288 códigos que pueden aparecer en el conjunto test y que hay que clasificar utilizando un corpus train/dev en los que no han aparecido ni una sola vez. En la Figura 4.12 se muestra el esquema con los códigos diferentes ICD-10-CM en los conjuntos train y dev.

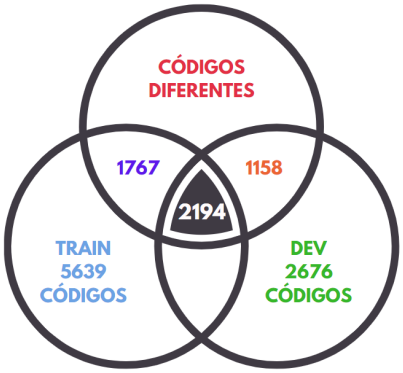

FIG. 4.12. NÚMERO DE CÓDIGOS ICD-10-CM DIFERENTES EN LOS CONJUNTOS TRAIN Y DEV DE CODIESP.

Como consecuencia de ello, el número de población de aprendizajes es muy limitado y el conjunto de aprendizaje es muy reducido, por lo que los modelos de aprendizaje del conjunto train no se ajustan bien. Esto es porque las anotaciones manuales de los códigos no son suficientes para generar buenos modelos que clasifican códigos que han aparecido sólo una vez.

4.4. Definir y justificar el enfoque de la solución

La solución se puede definir de dos maneras diferentes: mediante una clasificación de textos multi-etiqueta o una detección de entidades nombradas (NER), tareas enmarcadas dentro del área de Extracción de Información. También es posible combinar ambas formas para combinar sus ventajas.

Solución como clasificación multi-etiqueta

En primer lugar se consideró el enfoque de clasificación de texto.

La clasificación automática con códigos ICD-10 es un problema multi-clase y multi-etiqueta. Multi-clase porque cada historia de caso clínico se puede clasificar en más de dos tipos de códigos y multi-etiqueta porque porque una historia puede clasificarse con más de un código de diferentes clases.

En la Figura 4.13[57] se muestran diferentes estructuras de clasificadores multi-etiqueta. Todos ellos son posibles en esta tarea.

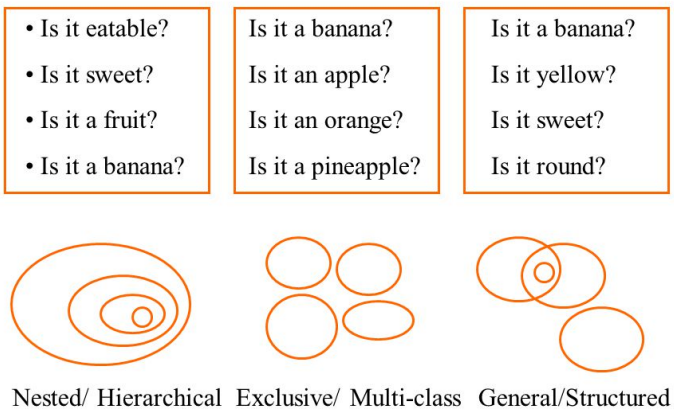

FIG. 4.13. ESTRUCTURAS DE CLASIFICACIÓN MULTI-ETIQUETA

Se obtendría una estructura anidada si dentro de una misma historia se obtuviese un código que pertenece a la subcategoría de otro código con la misma categoría específica. Un ejemplo de ello son los códigos S72.109B y S72. Se obtendría una estructura exclusiva si una historia clínica solo pudiese pertenecer a un código clínico, lo cual no se cumple en todos los casos. Finalmente, se obtiene una estructura general multi-etiqueta si una historia se clasifica por diferentes códigos relacionados o no. Se puede realizar la clasificación a nivel de historia de caso clínico, a nivel de párrafo, a nivel de oración o a nivel de token. Para ello, se requiere tokenizar en oraciones o tokens (palabras y signos de puntuación).

En el artículo[27] se demuestra que el rendimiento de los clasificadores depende de varios factores, como el número de muestras (500 en train), de etiquetas, de clases de etiqueta única (71476 en total, 1767 en train, 1158 en val) o de características. Por tanto, a menor número de muestras y mayor número de etiquetas, peor es el rendimiento de un clasificador. Existen diferentes enfoques de clasificación en Procesamiento de Lenguaje Natural. Además, existen implementaciones preexistentes de diferentes métodos en librerías Python sicikit learn.

Algunas técnicas para clasificación posibles son:

- Reglas semánticas
- Transferencia de conocimiento
- SVM con One-vs-all
- Bolsa de palabras (BOW)
- Bolsa de palabras continua (CBOW)
- Redes neuronales convolucionales (CNN)
- Redes neuronales recurrentes (RNN)
- Basado en modelos BERT

- Técnicas de recuperación de la información multilingüe (MLIR), que recupera documentos en diversos lenguajes con una misma llamada.

4.4.1. Aproximación 1: Enfoque SVM con One-vs-all

En primer lugar se procesaron el contenido de las historias y las etiquetas train. En segundo lugar, se filtraron las etiquetas mas frecuentes para reducir el número de clases. Para realizar una clasificación multi-etiqueta, es necesario transformar los vectores de características en valores numéricos y aplicar los módulos «MultiLabelBinarizer», «Multi-label categorization» y «Text Classification». En tercer lugar, gracias a la librería sklearn.multiclass es posible utilizar el módulo «OnevsRestClassifier». Esta estrategia divide un problema de clasificación multi-clase en tantos problemas de clasificación binaria como clases haya. En cuarto lugar, se tuvo que importar un modelo de clasificador binario implementado en sklearn. Se probó con todos ellos sin éxito:

- SVM (o SVC): con regresión logística como ratio de aprendizaje y regularizador newton-cg
- GaussianMB
- NMF
- MLPCLassifier
- KNeighborsClassifier
- AdaBoostClassifier

En quinto lugar, se utilizaron dos tipos de vectorización, una con Bolsa de palabras (BOW) y otra con Tf-idf («TfidfVectorizer»). En último lugar, se modificó el umbral de probabilidad para que predijesen códigos con menor probabilidad, pero cada historia tenía muchos códigos asociados. Por lo que este ajuste no dio buenos resultados.

En este enfoque se tuvieron varios problemas. La poca población de ejemplos en train, la gran cantidad de códigos no presentes en el conjunto train y el elevando número de códigos totales son grandes impedimentos en este enfoque. Para facilitar el entrenamiento, se podría reducir el número de códigos a clasificar. Esto se podría hacer eliminando los menos frecuentes o reduciendo la granularidad de los códigos eliminando las subcategorías. También se podría aumentar el corpus con otros recursos lingüísticos como historias de caso clínico codificadas o word embeddings preentrenados en el dominio clínico.

Solución como NER

En una de las versiones preliminares del corpus, se ofrecieron referencias de codificación para aumentar la transparencia. Esto quiere decir que cada código etiquetado

aparece asociado a una frase soporte de la codificación. Esto sugiere que la tarea se puede considerar como un problema de Reconocimiento de 71486 Entidades Nombradas.

4.4.2. Aproximación 2: Enfoque de Campos Aleatorios Condicionales

Conditional Random Fields (CRF) o campo aleatorio de Markov (MRF) es un modelo estocástico utilizado habitualmente para etiquetar secuencias de datos. Son modelos de grafos no dirigidos utilizados para calcular la probabilidad condicionada de valores de nodos designados de salida, dados los valores asignados a los nodos designados de entrada[58]. Son adecuados para el análisis de secuencias y se han mostrado útiles en el reconocimiento de entidades[59].

En una prueba de desarrollo, se construyó un Reconocedor de Entidades Nombradas que etiqueta si un término es un diagnóstico o no.

Para ello se realizó un etiquetado en formato IOB (que señala si una palabra del conjunto de entrenamiento pertenece a una entidad o no). El etiquetado en formato IOB es importante pero no necesario al realizar NER, aunque facilita el aprendizaje.

A continuación se implementó un modelo CRF que utilizaba el etiquetado IOB, y categorías morfológicas y sintácticas como características del modelo CRF. Además, los modelos CRF no aprenden solo de las etiquetas, sino también de la probabilidad de las transiciones.

Sin embargo, aunque ese sistema era capaz de detectar entidades con la etiqueta de diagnóstico, no fue capaz de codificar clínicamente considerando cada uno de los códigos ICD-10-CM como etiqueta. Es muy complicado al tratarse de 71486 códigos. Por lo tanto, es necesario incluir un sistema de coincidencia de cadenas basado en diccionario para codificar las entidades detectadas por el sistema NER. También, se puede considerar las entidades nombradas como diagnóstico como características para entrenar un clasificador multi-etiqueta con códigos ICD-10-CM, siendo un enfoque híbrido entre NER y clasificación.

4.4.3. Aproximación 3: Enfoque basado en diccionario

El diccionario de ICD-10-CM contiene 98288 códigos de diagnóstico válidos. Contiene los siguientes campos:

- code: códigos
- es-description: descripción en español
- en-description: descripción en inglés

Se puede reconocer entidades en español e inglés utilizando la relación entre el código, la descripción en español o inglés y el contenido de las historias clínicas.

Se produce un emparejamiento de sintagmas nominales y términos equivalentes del diccionario ICD-10. Estos recursos se utilizan en gran cantidad de trabajos para tareas de reconocimiento de entidades biomédicas. Los enfoques de correspondencia exacta alcanzan alta precisión generalmente, aunque sufren de menores recuperaciones debido a variaciones léxicas, abreviaturas u errores ortográficos.

Para aumentar la recuperación, los enfoques de coincidencia por aproximación utilizan similitud léxica y coincidencia de cadenas de caracteres por aproximación.

Se explica el desarrollo en el siguiente capítulo.

5. DESARROLLO DEL SISTEMA

Tras documentar el estudio teórico en los capítulo anteriores, se procede a detallar el desarrollo del sistema informático que codifique enfermedades en historias clínicas.

Dado el resultado insuficiente de la aproximación 1 «SVM one-vs-all» y aproximación 2 «Conditional Random Forest», se ha elegido utilizar para la competición la aproximación 3 «basado en diccionario» con Reconocimiento de Entidades Nombradas y coincidencia difusa.

En la Figura 5.1 se muestra un diagrama del sistema y sus herramientas. Este sistema requiere cumplir los requerimientos del entorno, importar los ficheros con las historias clínicas y el diccionario con los códigos, preprocesar, utilizar la librería SpaCy para extraer entidades, filtrar los mejores resultados con la distancia de Levenshtein y la librería Fuzzy wuzzy, predecir los códigos a evaluar, calcular las métricas como Mean Average Precision y visualizar los códigos predichos junto a las historias de caso clínico en una base de datos orientada a grafos.

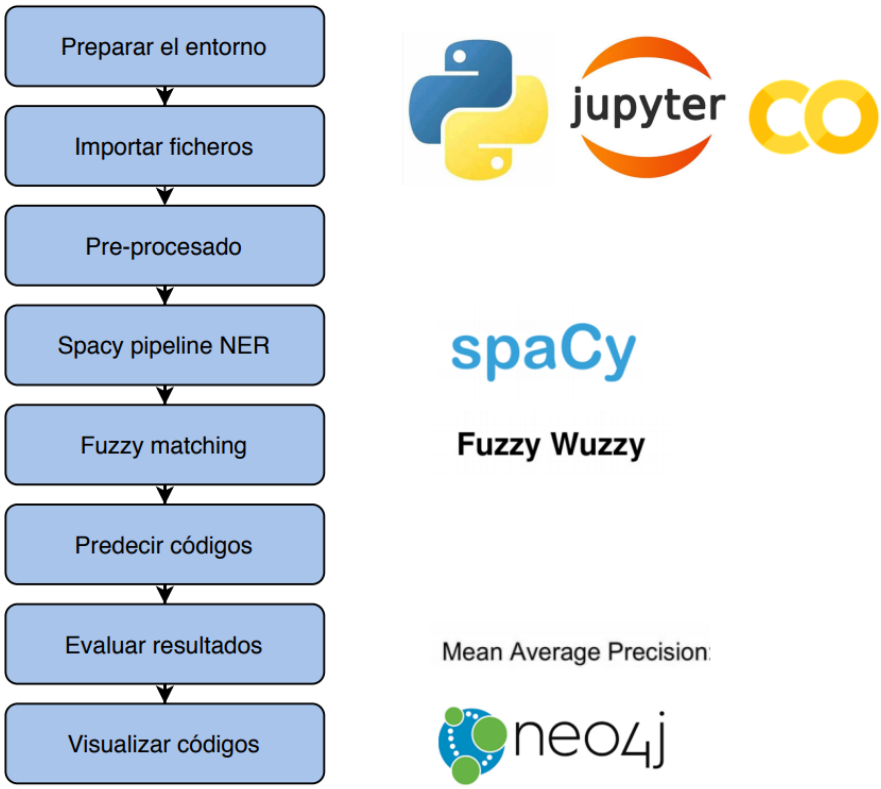

FIG. 5.1. DIAGRAMA DEL SISTEMA

5.1. Preparar el entorno

En primer lugar, se comentan las herramientas técnicas y recursos clínicos utilizados.

Herramientas técnicas utilizadas

- Python 3: El lenguaje de programación utilizado es Python. También se prodría haber utilizado Java o C. Las librerías principales han sido Pandas, NumPy, SciPy, scikit learn, SpaCy (en-core-web-sm 2.2.5). Otras librerías similares son NLTK, NeuroNER, Hugging Face y Tree tagger.
- Cuadernos de Jupyter: es un proyecto de código abierto que se puede acceder utilizado Google Colaborate o de forma local.
- Colaboratory: Google Colaborate [60] permite escribir y ejecutar código de Python en una máquina virtual exclusiva de una cuenta Google. Colab ofrece entornos de procesamiento acelerado opcionales, entre los que se incluyen TPU y GPU (K80, T4, P4 y P100 de NVIDIA). No requiere descarga, instalación ni configuración para usar los cuadernos de Jupyter y ofrece acceso gratuito a recursos computaciones. Entre sus desventajas, se encuentran los límites de uso generales de los recursos, la velocidad de GPU, el tiempo de ejecución de los cuadernos, el tiempo de espera por inactividad, la vida útil de las Máquinas Virtuales y la memoria.
- SpaCy Pipeline de la librería SpaCy.
- Librería Fuzzy Wuzzy.
- Base de datos orientada a grafos Neo4j.

Recursos clínicos utilizados

- Corpus CodiEsp.
- Diccionario ICD-10-CM.

5.2. Importar ficheros

Se han descargado y preparado los directorios del corpus. Se han combinado todos los ficheros de texto presentes en una carpeta como un fichero único en el script «Crear tsv a partir de text files».

En la Figura 5.2 se muestra los ficheros destinados al entrenamiento.

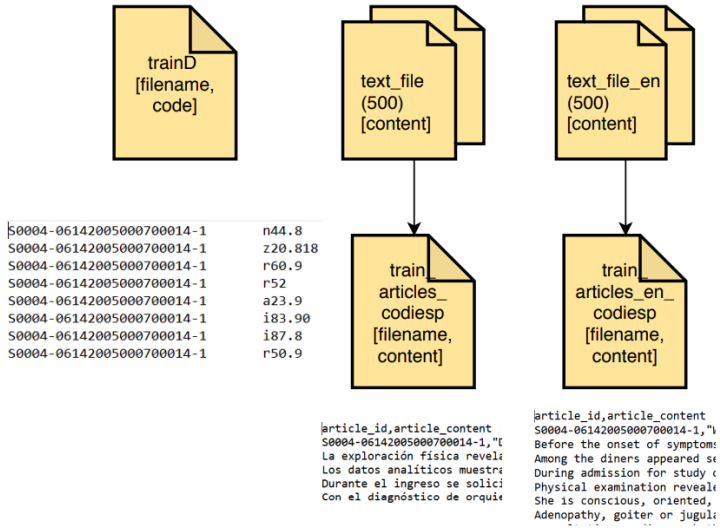

FIG. 5.2. FICHEROS DEL CONJUNTO DE ENTRENAMIENTO

Por un lado, se dispone de dos carpetas con 500 ficheros de texto. Una para las historias originales en español, y otra para las historias traducidas al inglés. Sin embargo, los códigos ICD-10 se mantienen iguales para ambos idiomas. Dentro de las carpetas aparecen como nombre de fichero el identificador de historia clínica y el interior de fichero el contenido de la nota. Se han combinado en otros dos ficheros (uno por idioma) llamados artículos de entrenamiento que incluyen el identificador y el contenido de las 500 historias clínicas. Por otro lado, se dispone de un fichero llamado «anotaciones trainD» que contiene los identificadores de historia clínica y los códigos asignados dentro de esa historia por codificación manual, incluyendo un código por fila.

En la Figura 5.3 se muestra los ficheros destinados a la validación en desarrollo. Sigue una lógica similar al conjunto anterior, pero contiene 250 historias de caso clínico.

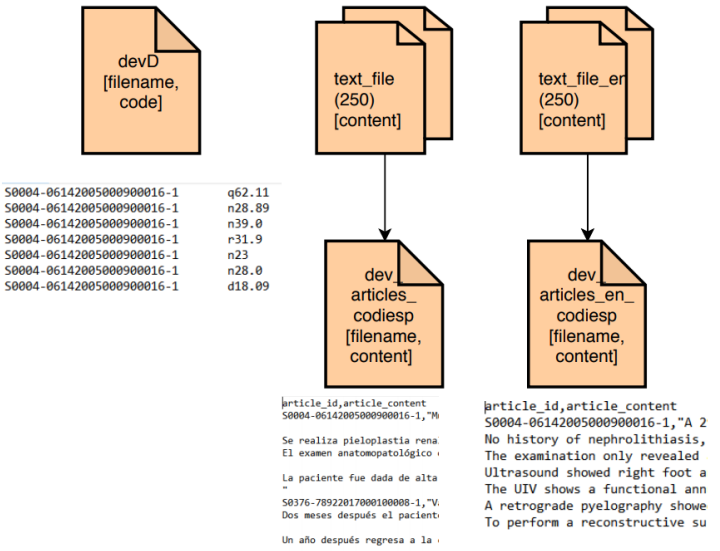

FIG. 5.3. FICHEROS DEL CONJUNTO DEV

En la Figura 5.4 se muestra los ficheros destinados a la validación tras la evaluación de la competición.

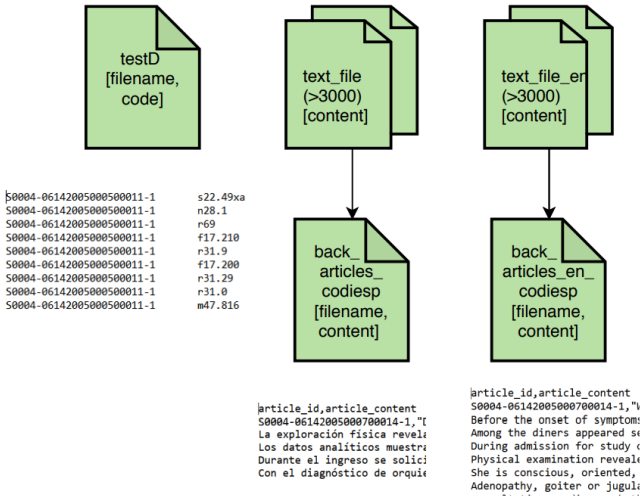

FIG. 5.4. FICHEROS DEL CONJUNTO TEST

En el momento de participar en la tarea sólo se dispone de dos carpetas con 3001 ficheros de texto. Una carpeta para las historias clínicas del conjunto test (que es desconocido) y originales en español, y otra para las historias traducidas al inglés. Sin embargo, los códigos ICD-10 se mantienen iguales para ambos idiomas. Dentro de las carpetas aparecen como nombre de fichero el identificador de historia clínica y el interior de fichero el contenido de la nota. Se han combinado en otros dos ficheros (uno por idioma) llamados artículos de entrenamiento que incluyen el identificador y el contenido de las 3001 historias clínicas.

Existe un fichero extra llamado «testD», sólo disponible en el momento de la validación. Forma el goldstandard, ya que contiene la codificación manual de 250 historias del conjunto de test + background. Sus campos son los identificadores de historia clínica y los códigos asignados dentro de esa historia por codificación manual, incluyendo un código por fila.

5.3. Preprocesado de datos

Es importante preparar los datos de entrada antes y después de aplicar procesos en las historias de caso clínico para obtener el output esperado. Algunas tareas de preprocesamiento son la limpieza de duplicados, normalización, combinación de ficheros y tokenización en palabras y signos de puntuación.

Tokenizar oraciones consiste en dividir un texto en oraciones, mientras que tokenizar palabras consiste dividirlo en palabras. Esto no es trivial, ya que la presencia de abreviaturas y diversos signos de puntuación pueden complicar la división.

En la Figura 5.5[61] se muestra el funcionamiento de un tokenizador que divide en tokens (palabras o signos de puntuación)[61].

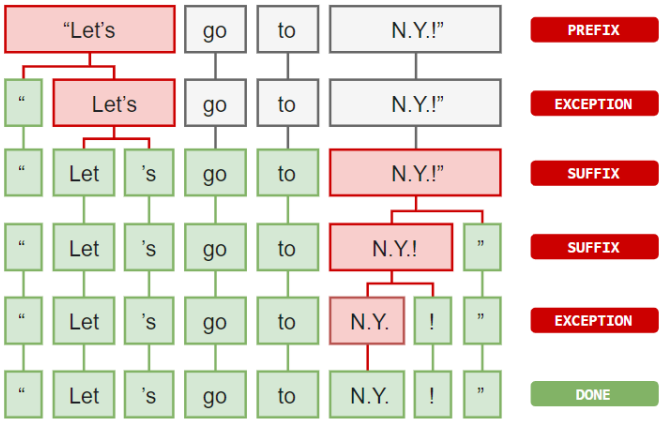

FIG. 5.5. TOKENIZADOR DE SPACY

5.4. Pipeline de procesamiento de lenguaje de SpaCy

Se ha utilizado la librería Spacy para realizar tareas de procesamiento de la nota y reconocimiento de entidades de diagnóstico clínico.

Modelos

Para ello, se ha cargado dos modelos, uno en español y otro en inglés. Se han cargado los modelos, los detalles de los modelos, el identificador de «Language class» y una lista de los componentes del pipeline. «Language class» contiene el vocabulario, reglas de tokenización y esquema de anotación propio del lenguaje.

Tanto el modelo en español es-core-news-sm (2.2.5) y el modelo en inglés encore-websm (2.2.5) incluyen capas convolucionales, conexiones residuales, normalización de capas y maxout non-linearity[62].

Un modelo de SpaCy incluye los pesos, los datos binarios y los nombres de los elementos que componen un pipeline.

Pipeline

Se ha creado una línea de procesos reutilizando elementos de SpaCy que incluye tagger, parser y entity recognizer (reconocedor de entidades nombradas). Se ha iterado sobre las fases del pipeline. Se crean componentes con «create_pipe» y se añade un componente a la línea de procesos con «add_pipe». También se ha creado un función personalizada.

En la Figura 5.6[62] se muestra los componentes de la línea de procesos del reconocedor de entidades de la librería SpaCy.

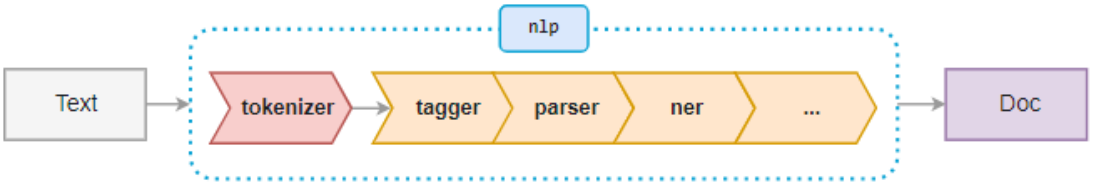

FIG. 5.6. SPACY PIPELINE

- Tokenizer: introduce y segmenta la historia clínica en tokens. Se realizan otras tareas de normalización como eliminar tildes, signos de puntuación, mayúsculas y stopwords (palabras muy frecuentes que no aportan información). Primero se se segmenta el texto en tokens y se producen objetos de tipo documento.
- Tagger: asigna etiquetas morfológicas llamadas part-of-speech-tagging.
- Parser: asigna etiquetas de dependencia sintáctica. Algunas categorías son root, acl, advcl, advmod, amod, appos, aux, case, cc, ccomp, compound, conj, cop, csubj, dep, det, expl:pass, fixed, flat, iobj, mark, nmod, nsubj, nummod, obj, obl, parataxis, punct y xcomp.
- NER: asigna entidades nombradas. Para eliminar solapamientos se eliminan las etiquetas por defecto LOC, MISC, ORG y PER y se añaden las descripciones de los códigos ICD-10-CM como entidades a reconocer.

A continuación, se ha detectado entidades nombradas sobre el conjunto test(background + goldstandard). Sin embargo, se utiliza una comprobación en la siguiente fase para filtrar las entidades más relevantes.

5.5. Comparación difusa de cadenas

Los métodos basados en diccionario son populares a la hora de abordar letras ambiguas. Los métodos de búsqueda en diccionario suelen ser robustos. Sin embargo, pueden incluir cálculos complicados apriori ya que un tamaño grande de diccionario aumenta el coste de búsqueda. [63]. Por ello, la distancia Levenshtein es una métrica sencilla que puede ser una herramienta de aproximación de cadenas efectiva.

Esta distancia es una medida de similaridad entre dos cadenas, la cadena origen (s) y la cadena objetivo (t). La distancia es el número de eliminaciones, inserciones o sustituciones necesarias para transformar s en t. Cuanto mayor sea la distancia, más diferentes son las cadenas. En este caso, la cadena origen (s) es el input de la historia de caso clínico, y la cadena objetivo (t) es una de las cadenas del diccionario ICD-10-CM.

En la Figura 5.7 [64] se muestra un ejemplo de cálculo de la distancia Levenshtein para las palabras Manhattan y su homónima (con errores) Manahaton.

		M	a	n	a	h	a	t	o	n
	0	1	2	3	4	5	6	7	8	9
M	1	0	1	2	3	4	5	6	7	8
a	2	1	0	1	2	3	4	5	6	7
n	3	2	1	0	1	2	3	4	5	6
h	4	3	2	1	1	1	2	3	4	5
a	5	4	3	2	1	2	1	2	3	4
t	6	5	4	3	2	2	2	1	2	3
t	7	6	5	4	3	3	3	2	2	3
a	8	7	6	5	4	4	3	3	3	3
n	9	8	7	6	5	5	4	4	4	3

FIG. 5.7. EJEMPLO DE DISTANCIA DE LEVENSHTEIN PARA MANAHATON Y MANHATTAN**PYTHONDISTANCE**

La librería Fuzzy implementa algoritmos fonéticos en Python. Más específicamente, la librería Fuzzy Wuzzy realiza un mapeo de cadenas difuso. Para ello, utiliza la distancia de Levenshtein para calcular las diferencias entre secuencias[65].

Gracias a esta librería, se pueden comparar las entidades detectadas en la historia clínica y buscar los códigos cuya descripción se aproxima más fielmente al contenido de la nota. La función de fuzzy wuzzy asigna una puntuación de similitud entre cadenas, y es posible configurar el umbral de permisividad, con una puntuación máxima y mínima. En este caso se ha elegido un máximo de -1 y un mínimo de 50.

5.6. Predicción del conjunto test

Gracias a los pasos anteriores es posible realizar la predicción en español e inglés sobre el conjunto test (background + goldstandard) con 3001 historias de caso clínicas. Sin embargo, solo se consideran las predicciones del conjunto goldstandard con 250 historias anotadas para la evaluación.

El formato del fichero de salida del codificador es el sugerido por CodiEsp[66]. En la Figura 5.8 se muestra un ejemplo del formato necesario para evaluar la predicción con las anotaciones manuales con el script desarrollado por CodiEsp [67].

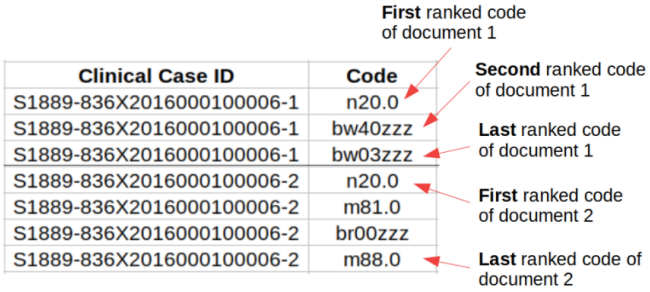

FIG. 5.8. EJEMPLO DEL FORMATO DEL FICHERO A EVALUAR

Las predicciones no tienen encabezados (en la imagen se añade para mayor interpretabilidad) y cada fila del fichero tsv (separado por tabulador) tiene una predicción con dos columnas, el identificador de la historia de caso clínica y el código. Por ejemplo: S1889-836X2016000100006-1 (identificador) y n20.0 (código).

Los códigos asignados a cada documento deben estar ordenados en un ranking por orden de confianza. Por lo que se ofrece mayor relevancia a las predicciones para las cuáles el sistema tiene mayor confianza[68].

5.7. Evaluación

Tras filtrar y obtener la predicción de los códigos ICD-10 de los 250 documentos contenidos en el conjunto test, es necesario comparar las predicciones con las anotaciones manuales asignadas por especialistas codificadores.

La organización ha publicado en Github el script oficial de evaluación bajo el nombre de CodiEsp-Evaluation-Script [67]. Se calcula la media de las precisiones ponderadas (Mean Average Precision o MAP) utilizando una implementación de la herramienta de evaluación TREC llamada trectools [69].

A la hora de ejecutar codiespD_P_evaluate.py, se requiere añadir los siguientes parámetros:

- –gs_path (-g) : gold/gold_standard.tsv. Es la ruta del fichero gold_standard.tsv distribuido por la organización. Gold Standard son 250 historias etiquetadas del conjunto test, e incluye los códigos verdaderos. Dentro del directorio gold se incluyen las historias etiquetadas de test (o train o val).
- –pred_path (-p) : system/prediction_D_official.tsv. Es la ruta del fichero predictions.tsv con las predicciones de los códigos del conjunto de test.
- –valid_codes_path (-c) : codiesp_codes/codiesp-D_codes.tsv. Es la ruta con los códigos válidos, para filtrar los códigos que no formen parte de ICD-10. Dentro del

directorio codiesp_codes se incluye el fichero tsv con los códigos válidos en español e inglés.

Por lo que para evaluar el rendimiento del sistema, se debe comparar el fichero de salida del sistema (predictions_D.tsv) respecto a los códigos verdaderos del conjunto Gold Standard. Se utilizó la siguiente sentencia en línea de comandos para calcular las métricas:

```
$> python3 codiespD_P_evaluation.py -g gold/gold_standard.tsv -p
    system/predictions_D.tsv -c codiesp_codes/codiesp-D_codes.tsv
```

CÓDIGO 5.1. EJECUCIÓN DEL EVALUADOR EN TEST

5.8. Visualización

Es interesante mostrar con un ejemplo la utilidad de la codificación clínica de enfermedades y el interés del alcance del presente TFM. Por ello, se ha realizado una visualización del identificador de las historias clínicas y sus códigos asignados. Se ha elegido utilizar Neo4j, una base de datos orientada a grafos que permite vislumbrar información de manera intuitiva y flexible mediante nodos y relaciones.

En primer lugar, se ha preparado un fichero con los identificadores de las historias clínicas en una columna y los códigos detectados en otra columna. A continuación, se ha transformado el contenido del fichero en lenguaje Cypher para generar las búsquedas y crear los nodos de historia clínica, nodos de códigos ICD-10 y las relaciones entre ambos. También se puede incluir toda la jerarquía de los códigos clínicos y relacionar códigos similares. Tras la creación de los nodos y las relaciones, se pueden crear búsquedas de gran utilidad para los especialistas en el ámbito sanitario.

Por ejemplo, en la Figura 5.9 se muestra todas las relaciones entre el nodo agrupador «Test_set» y los identificadores de historia de caso clínico. Se pueden observar 250 nodos azules, uno por cada historia de caso clínico del conjunto de test.

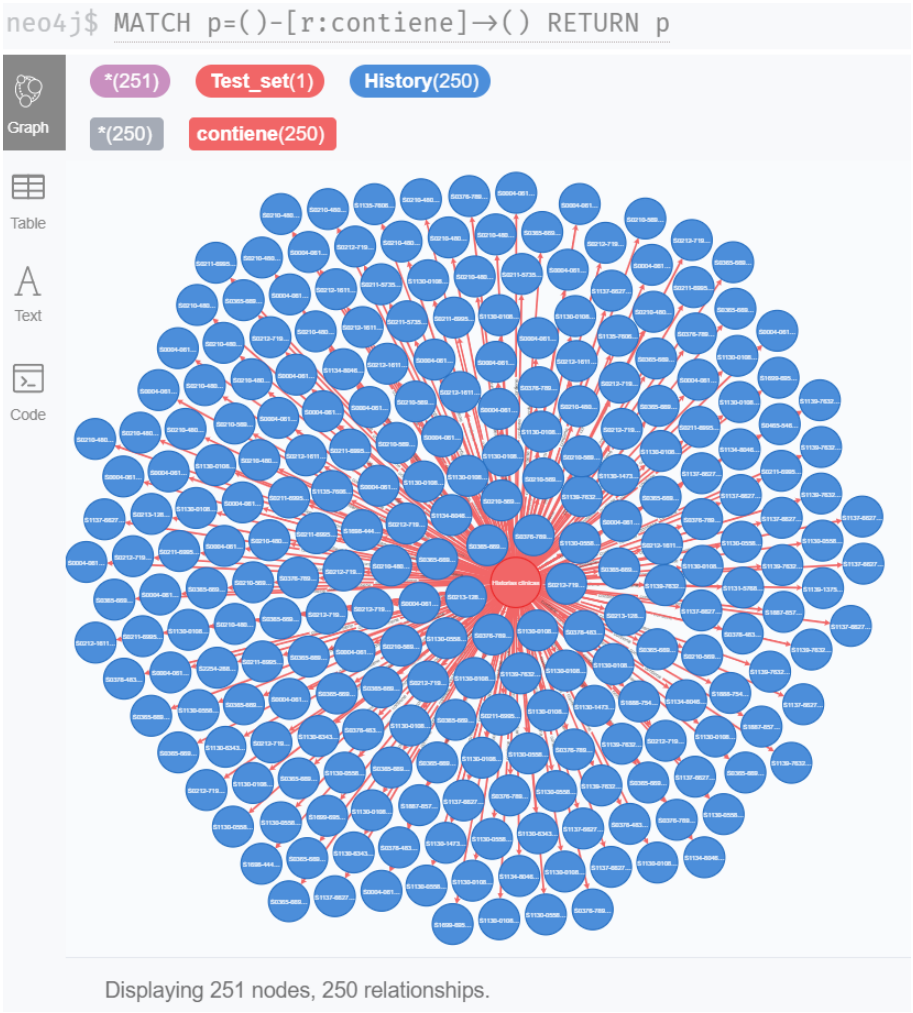

FIG. 5.9. BÚSQUEDA 1. HISTORIAS DEL CONJUNTO DE TEST

Se pueden realizar consultas de interés, como en qué historias de caso clínico se menciona una enfermedad concreta. Esto facilita la consulta de perfiles similares y los estudios epidemiológicos. En la Figura 5.10 se muestra el resultado de una búsqueda de las historias de caso clínico clasificadas con el código n58.

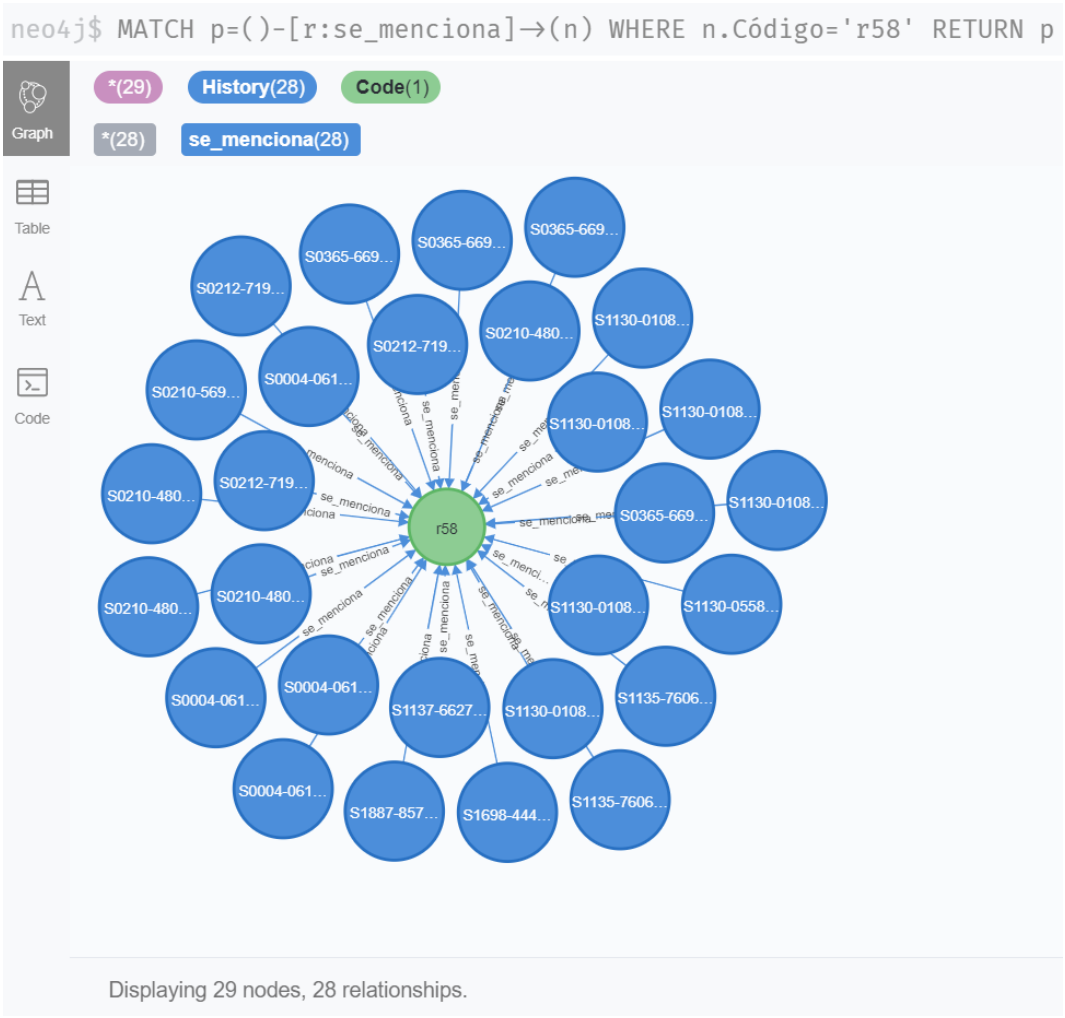

FIG. 5.10. BÚSQUEDA 2. CÓDIGO R58 Y SUS HISTORIAS CLÍNICAS

En la Figura 5.10 se muestra el resultado de una búsqueda de las historias de caso clínico clasificadas con el código n58 y el código r31.9.

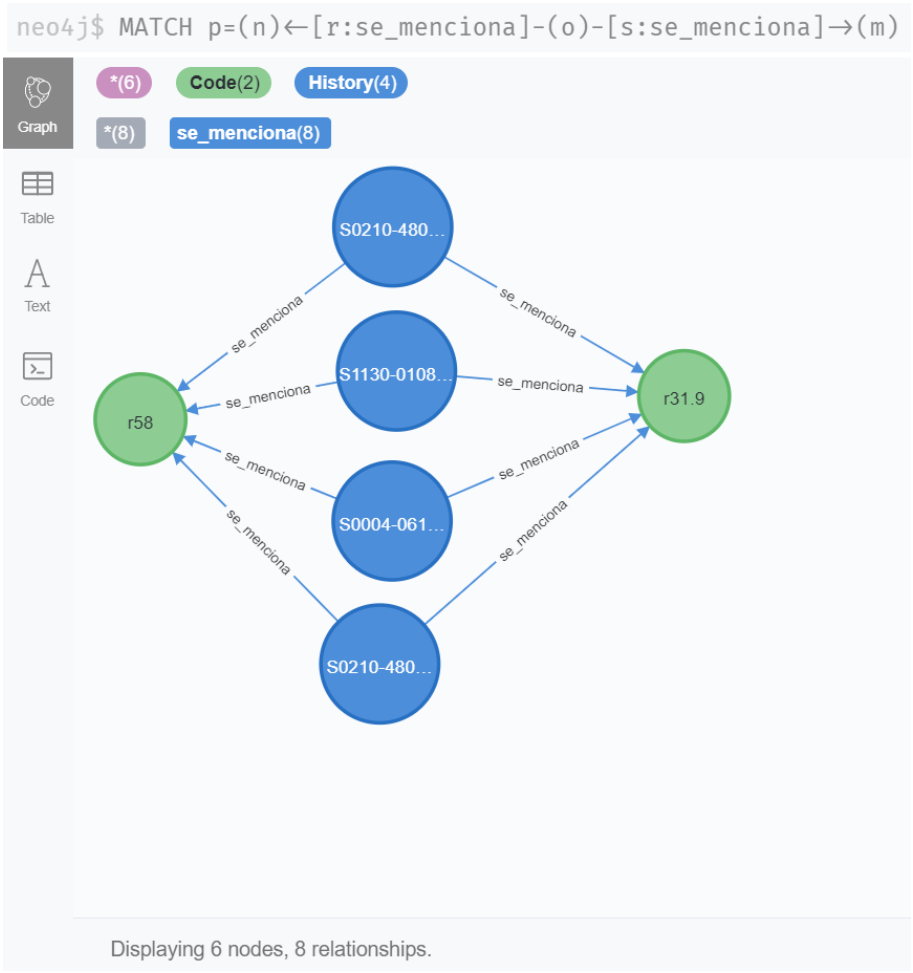

FIG. 5.11. BÚSQUEDA 3. CÓDIGOS R58 Y R31.9 Y SUS HISTORIAS CLÍNICAS

Estas visualizaciones permiten estructurar la información, ahorrar mucho tiempo a los profesionales y disminuir la carga laboral a la que están sometidos.

6. PLANIFICACIÓN DEL PROYECTO

En este capítulo se explica la planificación temporal, el presupuesto y el marco legal del proyecto.

6.1. Planificación temporal

La fecha de comienzo del TFM ha sido el 22 de Enero 2020 y la fecha de fin el 31 de Octubre del 2020. Se ha trabajado aproximadamente 3 horas al día. Sin embargo, hay días en los que se ha trabajado más horas, fines de semana incluidos, y días en los que se ha dedicado menos. Se considera que un mes tiene 22 días laborables.

La Tabla 6.1 muestra un listado de las tareas realizadas y su duración en horas, días de 3 h y meses de 22 días.

Tarea	Horas	Días (3 h)	Meses (22 días)
Planteamiento del problema	30	10	0.45
Estudio del Estado de la cuestión	30	10	0.45
Enfoque de la solución	30	10	0.45
Desarrollo del sistema (útil)	60	20	0.91
Evaluación	30	10	0.45
Discusión de los resultados	24	8	0.36
Documentación	291	97	4.41
Total del proyecto	495	165	7,5
Desarrollo del sistema (no útil)	28	84	1.27

TABLA 6.1. RESUMEN DE TAREAS Y DURACIÓN

Por tanto, se obtiene una duración de 7 meses y medio de trabajo, 165 días de 3 horas que ascienden a 495 horas en total. Además, se ha contabilizado el tiempo dedicado a otros enfoques no exitosos.

En la aproximación 1 «SVM One-vs-all» no se predijeron códigos con suficiente probabilidad.

La aproximación 2 «Conditional Random Forest» el etiqyedado IOB y el modelo de Bosques condicionales aleatorios sólo clasificaba entidades en diagnóstico o no diagnóstico, lo cual se hizo con el modelo de SpaCy también.

Todo ese tiempo no ha sido útil para el desarrollo del sistema y asciende a 28 horas de trabajo.

En la Figura 6.1 se muestra un diagrama Gantt en semanas con las diferentes tareas y subtareas. Se ha realizado con un software específico llamado Gantt Project[70].

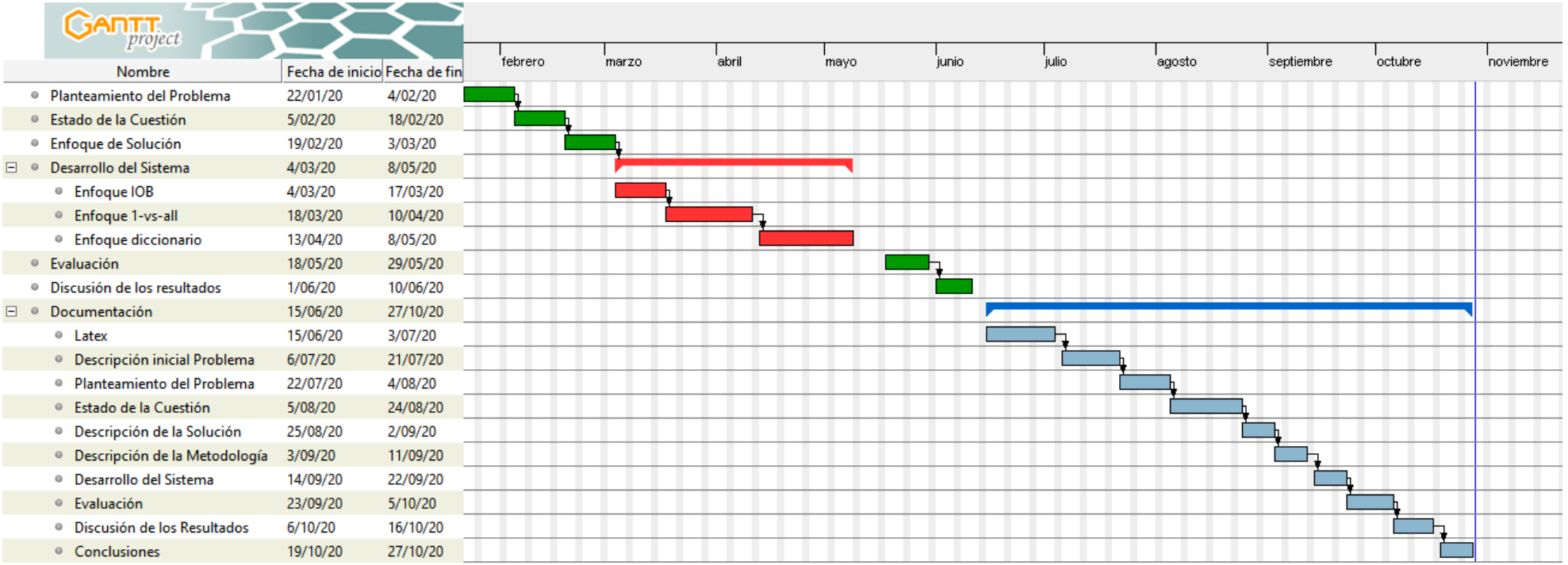

FIG. 6.1. DIAGRAMA GANTT DEL PROYECTO

6.2. Presupuesto estimado

A continuación, se detallan los costes asociados a la realización del proyecto en la Tabla 6.2. Se incluye el coste del personal, la amortización de los equipos y los costes indirectos.

Perfil	Dedicación (mes)	Coste mensual (€)	Coste (€)
Estudio teórico	2.16	775	1674
Desarrollo	0.91	1300	1183
Documentación	4.41	900	3969
Coste de personal total	7.5		6826
Equipos	**Coste (€)**	**Depreciación (meses)**	**Coste imputable (€)**
Portátil Inspiron 157000	1.029	10/60	171.50
HAVIT Ratón Inalámbrico	14	10/60	2.30
Microsoft Office 2016	59	10/60	9.80
Amortización de equipos	1102		183.70
Descripción costes			**Coste (€)**
Coste de personal total			6826
Amortización de equipos			183.70
Costes indirectos			1.040
Coste Total			8049.70

TABLA 6.2. COSTES

El presupuesto total de este proyecto asciende a la cantidad de 8049.70 €.

6.3. Marco legal

En primer lugar, cabe destacar que se ha utilizado software gratuito.

En la Figura 6.2 [71] se muestran los permisos y prohibiciones de la licencia de uso de Google Colaboratory. «BSD 3-Clause (New or Revised) License». Es una licencia permisiva similar a «BSD 2-Clause License», pero con una tercera cláusula que prohíbe utilizar el nombre del proyecto o sus contribuidores a promover productos similares sin consentimiento escrito.

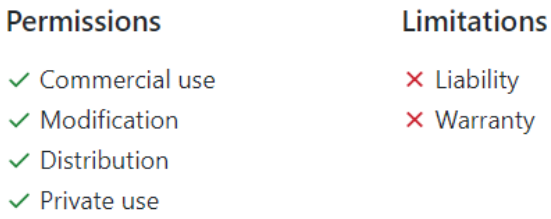

FIG. 6.2. PERMISOS DE LA LICENCIA DE GOOGLE COLABORATOY

En la Figura 6.3 [72] se muestran los permisos y prohibiciones de la licencia de uso de SpaCy. «MIT License» es una licencia simple y corta que solo requiere preservar el copyright. Se distribuye trabajos y modificaciones con licencia sin el código fuente.

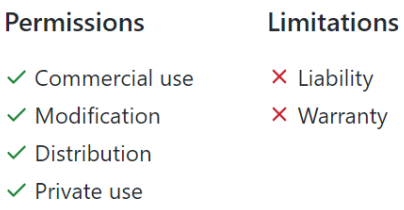

FIG. 6.3. PERMISOS DE LA LICENCIA DE SPACY

En el presente trabajo se ha utilizado el corpus propuesto por la organización CLEF. Se puede utilizar con la debida mención. Sin embargo, hay que tener en cuenta la ley de protección de datos. Los datos clínicos son de especial sensibilidad y se debe asegurar la anonimización, eliminando datos que puedan hacer referencia a personas físicas para garantizar una mayor confidencialidad y seguridad.

Protección de datos

La protección de las personas físicas en relación con el tratamiento de datos personales es un derecho fundamental, quedando regulado en el artículo 8, apartado 1, de la Carta de Derechos Fundamentales de la Unión Europea[73] donde se establece que los ciudadanos de la Unión tienen derecho a que se protejan sus datos personales.

Para poder garantizar estos derechos se aprobó el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo[74], de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de esos datos.

Con el fin de adaptar la legislación española a la normativa europea[75] y en cumplimiento del artículo 18.4 de la Constitución Española[76] donde se establece que «la Ley limitará el uso de la informática para garantizar el honor y la intimidad personal y la familiar de los ciudadanos y el pleno ejercicio de sus derechos» fue aprobada la Ley Orgánica 3/2018[77], de 5 de diciembre, de Protección de Datos y Garantía de Derechos Digitales (LOPDGDD). El objetivo de esta ley es hacer que las empresas y las organizaciones tengan un compromiso mayor con el tratamiento de los datos y archivos personales y regular la protección de datos.

La LOPDGDD solo efectúa referencias concretas a la sanidad en su disposición adicional decimoséptima y la disposición final novena.

El apartado 2 de la mencionada disposición adicional establece los criterios por los que se rige el tratamiento de datos en la investigación en salud. Fija con carácter general la necesidad de contar con el consentimiento de los interesados, con algunas excepciones como la realización de estudios científicos por las instituciones públicas con competencias en vigilancia de la salud pública en situaciones de excepcional relevancia y gravedad para la salud pública.

Se considera lícito el uso de datos personales seudonimizados con fines de investigación en salud y, en particular, biomédica.

Cuando se lleve a cabo un tratamiento con fines de investigación en salud pública y, en particular, biomédica se procederá a: 1.º Realizar una evaluación de impacto que determine los riesgos que se deriven del tratamiento. 2.º Someter la investigación científica a las normas de calidad sobre buena práctica clínica. 3.º Adoptar medidas dirigidas a evitar el acceso por los investigadores a datos de identificación. 4.º Designar un representante legal.

La disposición final novena modifica el apartado 3 del artículo 16 de la Ley 41/2002[78], de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica, estableciéndose que el acceso a la historia clínica con fines judiciales, epidemiológicos, de salud pública, de investigación, o de docencia obliga a preservar los datos de identificación personal del paciente separados de los de carácter clínico asistencial, de tal forma que se asegure el anonimato con carácter general, excepto que el paciente haya prestado el consentimiento para no separarlos.

La LOPDGDD establece en su artículo 5.1 que «Los responsables y encargados del tratamiento de datos así como todas las personas que intervengan en cualquier fase de este estarán sujetas al deber de confidencialidad (...)». A continuación, señala en su punto 2

que «La obligación general (…) será complementaría de los deberes de secreto profesional». Por tanto, los datos de carácter sanitario quedarían incluidos dentro de este deber de confidencialidad.

7. EVALUACIÓN Y DISCUSIÓN DE LOS RESULTADOS

Se ha realizado la predicción en español e inglés sobre las 3001 historias del conjunto test, que incluye background y goldstandard. Sin embargo, solo se consideran las predicciones del conjunto goldstandard con 250 historias para la evaluación.

Tras filtrar y obtener la predicción de los códigos ICD-10 de los 250 documentos contenidos en el conjunto test, es necesario comparar las predicciones con las anotaciones manuales asignadas por especialistas codificadores.

Es importante explorar el desempeño de los sistemas de codificación y clasificación de texto clínico para determinar cuán aplicables son a la industria de la codificación[1]. Un enfoque para evaluar la precisión del código ICD es examinar las fuentes de errores que conducen a la asignación de un código de diagnóstico que no es una representación justa de la condición real del paciente. Los errores que diferencian el código ICD de la enfermedad verdadera incluyen errores de medición tanto aleatorios como sistemáticos. Al comprender estas fuentes de error, los usuarios pueden evaluar las limitaciones de las clasificaciones y tomar mejores decisiones basadas en ellas.

Para ello se utilizan diversas métricas de evaluación. Algunas de ellas son las oficiales, que tienen en cuenta todos los códigos y sus sub-categorías. Otras métricas son especiales e incluyen relajaciones de la clasificación y simplifican el problema.

7.1. Métricas de Evaluación Oficiales

Los códigos predichos tienen un valor de probabilidad asociado. Estos se ordenan en orden descendiente en una lista, ocupando el primero el ranking 1 o el quinto el ranking 5. Se han calculado las métricas más utilizadas en extracción de la información: precisión, exhaustividad y valor F. Estos valores no tienen en cuenta el orden de relevancia de los códigos de una historia clínica, sino que consideran como un conjunto de valores. Para tener en cuenta la certeza de la predicción en la evaluación se ha utilizado una métrica más: Mean Average Precision.

Precision/Accuracy (precisión)

Indica cuántos resultados son realmente positivos entre todos los resultados que han sido predichos como positivos. Por ello, es útil cuando tenemos altos costes de positivos falsos. Expresa cómo de exacto el sistema es, hallando cuántas clases predichas son realmente positivas. En la búsqueda de los códigos del conjunto de historias de caso clínico, la precisión es el número exacto de códigos correctos divididos por el número total de

códigos detectados.

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{7.1}$$

Se ha calculado tres tipos de precisión: P, P_codes y P_cat.

Esta es una de las medidas más importantes. La precisión de los códigos es buena si reflejan la enfermedad del paciente adecuadamente, y ésta impacta en la calidad de las decisiones clínicas, sobre investigación o sobre financiación [4]. Es de gran importancia para los usuarios de la codificación y afecta a cada aplicación de manera diferente.

Se debe de considerar la precisión dentro de su dominio de clasificación. Este conocimiento mejora la precisión y fortalece las decisiones basadas en esas clasificaciones. Por ejemplo, se requiere de una precisión diferente a la hora de informar sobre las tasas de letalidad de una enfermedad que a la hora de reembolsar a los hospitales por sus recursos.

Recall (exhaustividad o recuerdo)

Indica cuántos verdaderos hemos podido etiquetar entre todos los positivos reales. La exhaustividad mide la fracción de positivos verdaderos sobre los resultados predichos. Es una medida crítica cuando existe un coste elevado de obtener Falsos Negativos. En este caso mide el número de códigos correctos divididos por el número total de códigos que deberían haber sido detectados.

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{7.2}$$

Se ha calculado tres tipos de exhaustividad: R, R_codes y R_cat.

F-measure (valor F), específicamente con valor $\beta = 1$

El valor F, F1-score o F-measure es la media armónica de la precisión y la exhaustividad. Permite equilibrar la precisión y la exhaustividad. Es interesante considerarla cuando se tiene una gran cantidad de Negativos Verdaderos. Es una buena métrica cuando la distribución de clases no está balanceada, lo cual es el caso de la tarea.

$$\text{F1-score} = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \tag{7.3}$$

Se ha calculado tres tipos de valor F: F1, F1_codes y F1_cat.

MAP (Media de precisión ponderada) [79]

Es una métrica muy extendida en problemas con ranking que ha mostrado tener buena discriminación y estabilidad. En las tareas de Recuperación de la Información (IR) actuales, la métrica de exhaustividad no es muy relevante, al incluir miles de documentos

relevantes que no son de interés para cierto usuarios. Un ejemplo es que si interesan los códigos de un tipo específico, no deberían tenerse en cuenta en la evaluación el resto de códigos. Además, la precisión y la exhaustividad se miden respecto a una lista de códigos sin tener en cuenta el orden de relevancia de la predicción. A la hora de valorar la importancia del orden en los códigos descritos, se debe de considerar la Media de Precisión Ponderada. Se han calculado MAP y MAP_codes.

Dada una lista de códigos predichos en una historia clínica, ordenados desde el que parece más fiable (top 1) hasta el último elemento, MAP es la media de las puntuaciones de precisión media de cada elemento de la lista.

Las puntuaciones de precisión media de cada elemento de la lista son calculadas para cada elemento de la lista y sus anteriores elementos. Por ejemplo, si se disponen de 5 códigos, MAP es la media de las precisiones medias del primer, segundo, tercer, cuarto y quinto código. Siendo la precisión media del primero calculada solo con el primer código ordenado y siendo la precisión media del segundo igual a la precisión considerando los primeros dos códigos, etc.

$$\text{MAP} = \frac{\sum_{q=1}^{\text{number of codes per history}} \text{AveragePrecision(q)}}{\text{number of codes per history}} \tag{7.4}$$

siendo q la posición del código, desde el primero al último.

MAP@30 (Media de precisión ponderada en 30 elementos)

MAP@30 es MAP que tiene en cuenta los primeros 30 códigos y no tiene en cuenta las posiciones posteriores al código situado en la posición 30. Se ha calculado MAP@30 y MAP@30_codes.

$$\text{MAP@30} = \frac{\sum_{q=1}^{30} \text{AveragePrecision(q)}}{30} \tag{7.5}$$

siendo q la posición del código, desde el primero al situado en la posición 30.

7.2. Métricas de Evaluación Especiales

Además de la Evaluación Oficial con todos los códigos completos que incluye P, R, F1, MAP y MAP@30, se han considerado dos casos extra que aportan mayor información:

Evaluación con menos códigos

Este tipo de evaluación reduce el número de códigos y elimina los códigos del conjunto test que no habían sido utilizados en los conjuntos anteriores. Por lo tanto, sólo se

tienen en cuenta los 1767 códigos presentes en train. Las métricas calculadas de este tipo de evaluación incluyen P_codes, R_codes, F1_codes y MAP_codes.

Evaluación menos granular

En este caso se reduce la granularidad de los códigos ICD-10. Solo se tiene en cuenta si la categoría específica del código es correcta. La categoría de la etiqueta son los primeros tres primeros caracteres (una letra y dos dígitos) de los códigos. Por ejemplo, los códigos P96.5 y P96.89 pertenecen a la categoría específica P96. Por lo tanto, P96.5 se considera correcto aunque el verdadero código sea P96.89, ya que la categoría especifica es correcta. Las métricas calculadas de este tipo de evaluación incluyen P_cat, R_cat, F1_cat.

7.3. Resultados obtenidos

En este apartado se incluye el rendimiento del sistema codificador sobre el conjunto test. Para la evaluación solo se tienen en cuenta los códigos predichos de las historias clínicas del conjunto goldstandard. Las historias del conjunto background que no tienen etiquetas manuales son ignoradas. En primer lugar, el rendimiento del enfoque basado en diccionario y en segundo lugar, el rendimiento de otros sistemas participantes en la tarea.

Resultados del enfoque basado en diccionario

Las métricas obtenidas por el enfoque basado en diccionario sobre el corpus goldstandard se incluyen a continuación.

En la Tabla 7.1 se muestra precision, recall, F1-score, MAP y MAP@30 del enfoque basado en diccionario. En negrita se muestra el mejor resultado y subrayado el segundo mejor.

Nivel	Precision	Recall	F-measure	MAP	MAP@30
Oficial	0.866	0.066	0.123	0.115	0.115
Menos códigos	**0.935**	0.071	0.132	0.138	0.138
Menos granular	<u>0.889</u>	0.074	0.137	-	-

TABLA 7.1. PRECISION, RECALL, F1-SCORE Y MAP DEL ENFOQUE BASADO EN DICCIONARIO

Resultados de todos los participantes

A efectos de comparación, se han tenido en cuenta las métricas obtenidas por todos los equipos que participaron en CodiEsp Diagnósticos. En total se registraron 78 ejecuciones, ya que cada uno de los 22 equipo participantes presentó entre una y cinco predicciones diferentes. En la Tabla 7.2 se muestra precision, recall, F1-score y MAP del mejor enfoque de cada equipo. Se tienen en cuenta la evaluación oficial (todos los códigos y las subcategorías ICD-10-CM). En negrita los mejores resultados, subrayados los segundos mejores.

Nombre de equipo	P	R	F1	MAP
TeamX	0.123	0.858	0.192	0.299
SWAP	0.295	0.442	0.308	0.202
LIIR	0.124	0.055	0.76	0.44
FLE	0.74	0.633	0.679	0.519
IAM	0.817	0.592	**0.687**	0.521
BCGS	0.547	0.287	0.337	0.259
SINAI	0.45	0.544	0.488	0.314
DCIC - UNS	0.482	0.261	0.187	0.097
IMS	0.373	0.709	0.474	0.449
SSN-NLP	0.025	0.049	0.033	0.007
MEDIA	0.735	0.63	0.629	0.488
Hulat-PDPQ	**0.866**	0.066	0.123	0.115
NLP-UNED	0.542	0.089	0.153	0.1
UDC-UA	0.727	0.605	0.546	0.368
CodeICD@IITH	0.462	0.281	0.35	0.192
The Mental Strokers	0.759	0.638	0.591	0.517
ICB-UMA	0.004	**0.897**	0.009	0.482
ExeterChiefs	0.117	0.201	0.144	0.082
LSI-UNED	0.253	0.688	0.37	0.517
nlp4life	0.014	0.038	0.02	0.004
Anuj	0.741	0.621	0.676	0.505
IXA-AAA	0.004	0.858	0.009	**0.593**

TABLA 7.2. MEJOR PRECISION, RECALL, F1-SCORE Y MAP DE LOS PARTICIPANTES DE CODIESP 2020

7.4. Análisis de los resultados

En esta sección se analizan los resultados obtenidos y se comparan con el resto de participantes de CodiEsp. Esto permite comparar la efectividad de los clasificadores y estudiar las diferencias con un análisis de errores.

Análisis del enfoque basado en diccionario

Se ha alcanzado un valor de precisión del 86.6 % y un 6.6 % en exhaustividad.

En primer lugar, se comenta el valor de precisión. Se puede observar que 0.866 es muy bueno. Por lo tanto, existen pocos Falsos Positivos y la mayoría de las entidades detectadas son correctas. Muestra que la exactitud es alta y el número de códigos correctos detectados se aproxima al número total de códigos detectados. Pocos códigos detectados son erróneos. Además, esta métrica es fundamental para no tomar decisiones erróneas sobre las historias clínicas. Prima la calidad de los códigos detectados sobre la cantidad. El valor utilizando sólo los códigos presentes en train es el mejor valor de todos: 0.935. Esto es debido a que se tienen en cuenta menos códigos. Además, el valor obtenido utilizando sólo las categorías específicas es mejor que la precisión oficial pero no supera la precisión con sólo códigos en train (1767 códigos). Se puede indicar que a menor número de códigos, la precisión aumenta.

En segundo lugar se presenta la exhaustividad. Este resultado de 0.066 es bastante mejorable. Indica que el número de códigos correctos es muy inferior al número de códigos que deberían haber sido detectados. Por tanto, se aprecia una alta presencia de falsos negativos. La consecuencia de ésto es que no se han detectado la mayoría de códigos y un codificador manual sí sería capaz de codificar enfermedades que este sistema no reconoce. Utilizando sólo los códigos de train y relajando los códigos a su categoría específica mejora levemente los resultados a 0.071 y 0.074 respectivamente.

En tercer lugar se obtiene el valor de F, que es la media armónica entre la precisión y la exhaustividad. Estos resultados de 0.123, 0.132 y 0.137 son los esperables dados los buenos resultados de precisión y débiles valores de exhaustividad. Es una buena métrica cuando existe una gran cantidad de negativos verdaderos y la distribución de las clases no está balanceada, lo cuál es el caso de la tarea.

En cuarto lugar, el valor de Media de Precisiones Medias ha sido más débil que la precisión. Esto indica que el hecho de ponderar con más peso las precisiones de los códigos con más probabilidad ha tenido un efecto contraproducente en la métrica. Además, los valores de MAP para más de 30 códigos por historia clínica son los mismos que el MAP, indicando que la mayoría de las historias clínicas se han etiquetado con menos de 30 códigos.

Análisis comparativo con otros participantes

A pesar de que el corpus de historias está en español, los 22 equipos participantes provienen de países diversos: Argentina, India, Italia, Alemania, España, Estados Unidos, Japón, Francia, Bélgica, Turquía, Reino Unido,

Los valores obtenidos de precisión, exhaustividad, valor F y MAP más elevados fueron 0.866, 0.897, 0.687 y 0.593, respectivamente[68]. Cada uno de los valores máximos fue obtenido con un equipo diferente.

Los resultados del enfoque basado en diccionario muestran los valores más elevados de precisión de todos los equipos participantes. La clara ventaja es el menor número de falsos negativos y la mayoría de las entidades detectadas son correctas.

Análisis de errores

En el proceso de asignación de código se pueden presentar una amplia gama de errores. Por ello, es importante especificar el proceso de código, los tipos de errores y las inconsistencias de codificación para comprender cuáles son más importantes o comunes y prevenirlos en un futuro.

La aproximación basada en diccionario no fue capaz de detectar la mayoría de los códigos debido a una falta de flexibilidad en el reconocimiento. Esto puede ser debido a la variabilidad léxica que el sistema no es capaz de capturar. Una limitación es la unión de entidades cuando los adjetivos están por separado. Por ejemplo, en la frase «metástasis hepática y renal» se detecta metástasis hepática pero no se reconoce metástasis renal.

Esto supone un problema esencial para la aproximación basada en diccionario. Dada la enorme dificultad de la tarea, se deben incorporar mejoras para aumentar la exhaustividad.

8. DISCUSIÓN FINAL

8.1. Conclusiones

Las estudios de caso clínico realizados a través del análisis de historias clínicas son una fuente de conocimiento que es necesario de estructurar y procesar para extraer información relevante para los profesionales del sector sanitario. Actualmente queda patente el enorme efecto que la gestión sanitaria ocupa en la sociedad y cómo la informática ocupa un papel fundamental en su desarrollo futuro.

Los investigadores y profesionales sanitarios requieren de herramientas expertas en el dominio biomédico para procesar textos clínicos. Este proceso es muy cansado y mecánico, lo que aumenta la situación de estrés laboral y la propensión a errores.

El objetivo del presente trabajo es codificar historias de caso clínico en español con códigos ICD-10.

Se ha construido un sistema que reconoce entidades de Diagnóstico utilizando la línea de procesos de reconocimiento de entidades nombradas de Spacy y codificando con la implementación de la distancia de Levenshtein presente en la librería Fuzzy Wuzzy. Para ello, se ha utilizado un diccionario que mapea los códigos con sus descripciones en español e inglés.

En el análisis de resultados se ha obtenido un 86 % de precisión, siendo el valor más elevado de todos los enfoques de la competición Codiesp. Sin embargo la exhaustividad obtuvo un valor en la media del 6.6 % y lejos de la capacidad humana y de Modelos de Lenguaje de detectar enfermedades. Estos valores destacan que el enfoque no fue suficiente para detectar la mayoría de códigos presentes en las historias debido a la variabilidad léxica de los términos.

Se ha comprobado que los resultados mejoran al relajar los códigos completos en categorías específicas.

Para llegar a implementar este sistema de codificación en los centros sanitarios, se requiere comercializar el software y verificar el producto, y cumplir la Ley de Protección de datos a nivel nacional y europeo.

8.2. Discusión

Respecto a los resultados, son considerados positivos para ser la primera participación en una tarea de una competición internacional que lleva realizándose durante varios años, aumentando el nivel de dificultad progresivamente. El sistema NER presentado es sencillo ya que utiliza un modelo con capas convolucionales de la librería de Spacy, y una

coincidencia difusa de cadenas basado en la distancia de Levenshtein. Se ha comprobado que este enfoque funciona bien en el lenguaje español y obtiene los mejores resultados en precisión de todos los equipos participantes.

Sin embargo, existen una serie de limitaciones para utilizar enfoques de Aprendizaje Profundo debido a la alta complejidad de la tarea. Esto es debido a los pocos ejemplo de cada clase en el entrenamiento, las clases desbalanceadas, un alto número de códigos y clases a codificar y mucha variabilidad léxica en las historias clínicas.

Para poder utilizar modelos de Aprendizaje Profundo sin que haya insuficiencias en el corpus, se deben de incluir técnicas en el futuro como las siguientes:

- Reducir el número de códigos irrelevantes, que apenas aparecen en las historias clínicas y que pueden mejorar las métricas sin un grave impacto en la precisión.
- Incorporar recursos lingüísticos y terminológicos adicionales. Es posible incluir otros corpus con historias clínicas en español que hayan sido codificados con la terminología ICD-10-CM. Se puede utilizar para mejorar el sistema. Por ejemplo, los corpus Licacs[80] e Ibecs [81].
- Utilizar modelos para inferir categorías, como mecanismos de atención o modelos de auto-aprendizaje como BERT.
- Incluir más lenguajes. Utilizando modelos multi-lenguaje y otros recursos lingüísticos. Esto permite implementar un sistema que combine diferentes lenguajes además del español y el inglés.
- Añadir detección de abreviaturas. Se puede introducir un tratamiento de abreviaturas para capturar expresiones que hacen referencia a enfermedades pero que no han sido tenidas en cuenta al incluir abreviaturas.
- Detección de errores tipográficos. Los errores de escritura o tipográficos dificultan el procesamiento del texto y disminuyen la exhaustividad. Existen mecanismos para paliar sus consecuencias.
- Incluir reglas semánticas que aumenten el número de características que sirven para clasificar.
- Mejorar el filtrado tras la fase de detección de entidades, para aumentar la precisión.

BIBLIOGRAFÍA

[1] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders y W. R. Hersh, "A systematic literature review of automated clinical coding and classification systems," DOI: 10.1136/jamia.2009.001024. [En línea]. Disponible en: www.jamia.org.

[2] *CodiEsp*. [En línea]. Disponible en: https://temu.bsc.es/codiesp/.

[3] *Métodos de investigación / N.J. Salkind.* [En línea]. Disponible en: https://www.researchgate.net/publication/31749735_Metodos_de_investigacion_NJ_Salkind.

[4] K. J. O'Malley et al., *Measuring diagnoses: ICD code accuracy*, oct. de 2005. DOI: 10.1111/j.1475-6773.2005.00444.x. [En línea]. Disponible en: /pmc/articles/PMC1361216/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1361216/.

[5] *Delving into computer-assisted coding - PubMed*. [En línea]. Disponible en: https://pubmed.ncbi.nlm.nih.gov/15559840/.

[6] *ICD-10-GM-2020 Code Suche*. [En línea]. Disponible en: https://www.icd-code.de/.

[7] *ICD-Scout: Volltextsuche im systematischen und alphabetischen Verzeichnis der ICD-10-GM*. [En línea]. Disponible en: http://www.icdscout.de/.

[8] *ICD-10 Training Tool*. [En línea]. Disponible en: https://apps.who.int/classifications/apps/icd/icd10training/.

[9] *ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification*. [En línea]. Disponible en: https://www.cdc.gov/nchs/icd/icd9cm.htm.

[10] *eCIE-Maps - Documentación*. [En línea]. Disponible en: https://eciemaps.mscbs.gob.es/ecieMaps/documentation/documentation.html.

[11] "Unidad Técnica de Codificación CIE-10-ES Ministerio de Sanidad, Servicios Sociales e Igualdad," inf. téc.

[12] A. Geissbuhler y C. Kulikowski, "Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research," inf. téc. 1, 2008, pp. 128-172.

[13] F. Jiang et al., *Artificial intelligence in healthcare: Past, present and future*, dic. de 2017. DOI: 10.1136/svn-2017-000101. [En línea]. Disponible en: http://svn.bmj.com/.

[14] J. Ramón Rabuñal Dopico, A. Pazos Sierra y N. York, "Encyclopedia of Artificial Intelligence," inf. téc.

[15] *Google*. [En línea]. Disponible en: https://www.google.es/.

[16] *PubMed*. [En línea]. Disponible en: https://pubmed.ncbi.nlm.nih.gov/.

[17] M. A. Hearst y M. A. Hearst, "Untangling Text Data Mining," *UNIVERSITY OF MARYLAND*, pp. 3-10, 1999. [En línea]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1393.

[18] R. M. R. Zavala, "Modelo híbrido para el reconocimiento de entidades biomédicas en textos biomédicos," Tesis doct., UC3M, 2018.

[19] L. S. Larkey y W. B. Croft, "Automatic Assignment of ICD9 Codes To Discharge Summaries," inf. téc.

[20] I. Goldstein, A. Arzrumtsyan y O. Uzuner, "Three approaches to automatic assignment of ICD-9-CM codes to radiology reports.," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2007, pp. 279-283, 2007. [En línea]. Disponible en: /pmc/articles/PMC2655861/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655861/.

[21] R. Farkas y G. Szarvas, "Automatic construction of rule-based ICD-9-CM coding systems," en *BMC Bioinformatics*, vol. 9, BMC Bioinformatics, abr. de 2008. DOI: 10.1186/1471-2105-9-S3-S10. [En línea]. Disponible en: https://pubmed.ncbi.nlm.nih.gov/18426545/.

[22] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean y R. A. Dudley, "N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit," *Journal of the American Medical Informatics Association*, vol. 21, n.º 5, pp. 871-875, 2014. DOI: 10.1136/amiajnl-2014-002694. [En línea]. Disponible en: https://pubmed.ncbi.nlm.nih.gov/24786209/.

[23] R. Kavuluru, A. Rios e Y. Lu, "An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records," *Artificial Intelligence in Medicine*, vol. 65, n.º 2, pp. 155-166, oct. de 2015. DOI: 10.1016/j.artmed.2015.04.007.

[24] E. Scheurwegs, B. Cule, K. Luyckx, L. Luyten y W. Daelemans, "Selecting relevant features from the electronic health record for clinical code prediction," *Journal of Biomedical Informatics*, vol. 74, pp. 92-103, oct. de 2017. DOI: 10.1016/j.jbi.2017.09.004.

[25] B. Shickel, P. J. Tighe, A. Bihorac y P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, n.º 5, pp. 1589-1604, sep. de 2018. DOI: 10.1109/JBHI.2017.2767063.

[26] T. Baumel et al., "Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment," inf. téc. [En línea]. Disponible en: www.aaai.org.

[27] W. Zhang, J. Yan, X. Wang y H. Zha, "Deep Extreme Multi-label Learning," 2018. DOI: 10.1145/3206025.3206030. [En línea]. Disponible en: https://doi.org/10.1145/3206025.3206030.

[28] S. Pakhomov, J. Buntrock y P. Duffy, "High throughput modularized NLP system for clinical text," en *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2005, pp. 25-28. DOI: 10.3115/1225753.1225760. [En línea]. Disponible en: http://umlslex.nlm.nih.gov.

[29] C. Kolářik, M. Hofmann-Apitius, M. Zimmermann y J. Fluck, "Identification of new drug classification terms in textual resources," en *Bioinformatics*, vol. 23, Oxford Academic, jul. de 2007, pp. 264-272. DOI: 10.1093/bioinformatics/btm196. [En línea]. Disponible en: http://www.nlm.nih.gov/research/umls/meta4.html.

[30] "NATURAL LANGUAGE PROCESSING SECOND EDITION," inf. téc.

[31] K. M. Hettne et al., "A dictionary to identify small molecules and drugs in free text," *Bioinformatics*, vol. 25, n.º 22, pp. 2983-2991, nov. de 2009. DOI: 10.1093/bioinformatics/btp535. [En línea]. Disponible en: http://www.biosemantics.org/chemlist..

[32] L. Zhou et al., "Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes.," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2011, pp. 1639-1648, 2011. [En línea]. Disponible en: /pmc/articles/PMC3243163/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243163/.

[33] T. C. Rindflesch, L. Tanabe, J. N. Weinstein y L. Hunter, "EDGAR: extraction of drugs, genes and relations from the biomedical literature.," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 517-528, 2000. DOI: 10.1142/9789814447331{_}0049. [En línea]. Disponible en: /pmc/articles/PMC2709525/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2709525/.

[34] M. A. Levin, M. Krol, A. M. Doshi y D. L. Reich, "Extraction and mapping of drug names from free text to a standardized nomenclature.," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2007, pp. 438-442, 2007. [En línea]. Disponible en: /pmc/articles/PMC2655777/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655777/.

[35] D. Sanchez-Cisneros, P. Martínez e I. Segura-Bedmar, "Combining dictionaries and ontologies for drug name recognition in biomedical texts," en *International Conference on Information and Knowledge Management, Proceedings*, New York, New York, USA: ACM Press, 2013, pp. 27-30. DOI: 10.1145/2512089.2512100.

[En línea]. Disponible en: http://dl.acm.org/citation.cfm?doid=2512089.2512100.

[36] *The CLEF Initiative (Conference and Labs of the Evaluation Forum) - Homepage*. [En línea]. Disponible en: http://www.clef-initiative.eu/.

[37] *CLEF eHealth Lab Series*. [En línea]. Disponible en: https://clefehealth.imag.fr/.

[38] A. Névéol et al., "CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French," inf. téc. [En línea]. Disponible en: https://www.cdc.gov/.

[39] A. Névéol et al., "CLEF eHealth 2018 Multilingual Information Extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian," inf. téc. [En línea]. Disponible en: http://www.istat.it/..

[40] M. Li et al., "ECNU at 2018 eHealth Task1 Multilingual Information Extraction," inf. téc.

[41] R. Flicoteaux, "ECSTRA-APHP @ CLEF eHealth2018-task 1: ICD10 Code Extraction from Death Certificates," inf. téc. [En línea]. Disponible en: http://www.cepidc.inserm.fr/.

[42] S. Cossin, V. Jouhet, F. Mougin, G. Diallo y F. Thiessard, "IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates," inf. téc.

[43] G. Maria Di Nunzio, "Classification of ICD10 Codes with no Resources but Reproducible Code. IMS Unipd at CLEF eHealth Task 1," inf. téc. [En línea]. Disponible en: https://www.tidyverse.org.

[44] A. Atutxa et al., "IxaMed at CLEF eHealth 2018 Task 1: ICD10 Coding with a Sequence-to-Sequence approach," inf. téc.

[45] Julia, "0 Coding of French and Italian Death Certificates with Character-Level Convolutional Neural Networks," inf. téc., 2018.

[46] M. Almagro, S. Montalvo, A. Díaz De Ilarraza y A. Pérez, "MAMTRA-MED at CLEF eHealth 2018: A Combination of Information Retrieval Techniques and Neural Networks for ICD-10 Coding of Death Certificates," inf. téc.

[47] J. Gobeill y P. Ruch, "Instance-based learning for ICD10 categorization," inf. téc.

[48] P. López-, M. Carlos Díaz-Galiano, M. Teresa Martín-Valdivia y L. Alfonso Ureña-López, "Machine learning to detect ICD10 codes in causes of death," inf. téc. [En línea]. Disponible en: https://snowballstem.org/.

[49] K. Réby, S. Cossin, G. Bordea y G. Diallo, "SITIS-ISPED in CLEF eHealth 2018 Task 1 : ICD10 coding using Deep Learning," inf. téc.

[50] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé y M. Krallinger, "Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020," inf. téc. [En línea]. Disponible en: https://www.who.int/classifications/icd/icdonlineversions/en/.

[51] *guilopgar/CLEF-2020-CodiEsp: Contribution of the ICB-UMA team to the CLEF eHealth 2020 Task 1: Multilingual Information Extraction*. [En línea]. Disponible en: https://github.com/guilopgar/CLEF-2020-CodiEsp.

[52] N. García-Santa y K. Cetina, "FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding," inf. téc. [En línea]. Disponible en: http://www.fujitsu.com/emea/about/fle/.

[53] A. Blanco, A. Pérez y A. Casillas, "IXA-AAA at CLEF eHealth 2020 CodiEsp Automatic classification of medical records with Multi-label Classifiers and Similarity Match Coders," inf. téc.

[54] J. Costa, I. Lopes, A. Carreiro, D. Ribeiro y C. Soares, "Fraunhofer AICOS at CLEF eHealth 2020 Task 1: Clinical Code Extraction From Textual Data Using Fine-Tuned BERT Models," inf. téc. [En línea]. Disponible en: https://www.aicos.fraunhofer.pt/.

[55] E. Moons, A. Khanna, A. Akkasi y M.-F. Moens, "A Comparison of Deep Learning Methods for ICD Coding of Clinical Records," *Applied Sciences*, vol. 10, n.º 15, p. 5262, jul. de 2020. DOI: 10.3390/app10155262. [En línea]. Disponible en: https://www.mdpi.com/2076-3417/10/15/5262.

[56] A. Miranda-Escalada, A. Gonzalez-Agirre y M. Krallinger, "CodiEsp corpus: Spanish clinical cases coded in ICD10 (CIE10) - eHealth CLEF2020," mayo de 2020. DOI: 10.5281/ZENODO.3837305. [En línea]. Disponible en: https://zenodo.org/record/3837305.

[57] *Multi-Class and Structured Classification - ppt video online download*. [En línea]. Disponible en: https://slideplayer.com/slide/3833223/.

[58] J. Lafferty, A. Mccallum, F. C. N. Pereira y F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data Part of the Numerical Analysis and Scientific Computing Commons Recommended Citation Çonditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence DataÇonditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," inf. téc., 2001, pp. 282-289. [En línea]. Disponible en: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers.

[59] A. Mccallum y W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons," inf. téc.

[60] *Colaboratory – Google*. [En línea]. Disponible en: https://research.google.com/colaboratory/faq.html.

[61] *spaCy 101: Everything you need to know · spaCy Usage Documentation*. [En línea]. Disponible en: https://spacy.io/usage/spacy-101.

[62] *explosion/spacy-models: Models for the spaCy Natural Language Processing (NLP) library*. [En línea]. Disponible en: https://github.com/explosion/spacy-models.

[63] R. Haldar y D. Mukhopadhyay, "Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach," ene. de 2011. [En línea]. Disponible en: http://arxiv.org/abs/1101.1232.

[64] *Python Advanced: Recursive and Iterative Implementation of the Edit Distance*. [En línea]. Disponible en: https://www.python-course.eu/levenshtein_distance.php.

[65] *fuzzywuzzy · PyPI*. [En línea]. Disponible en: https://pypi.org/project/fuzzywuzzy/.

[66] *Submission – CodiEsp*. [En línea]. Disponible en: https://temu.bsc.es/codiesp/index.php/2020/02/06/submission/.

[67] *GitHub - TeMU-BSC/CodiEsp-Evaluation-Script: Evaluation library for CodiEsp Task*. [En línea]. Disponible en: https://github.com/TeMU-BSC/CodiEsp-Evaluation-Script.

[68] L. Goeuriot et al., "Overview of the CLEF eHealth Evaluation Lab 2020," en, Springer, Cham, sep. de 2020, pp. 255-271. DOI: 10.1007/978-3-030-58219-7{_}19. [En línea]. Disponible en: http://link.springer.com/10.1007/978-3-030-58219-7_19.

[69] J. Palotti, H. Scells y G. Zuccon, "Trectools: An open-source python library for information retrieval practitioners involved in TREC-like campaigns," en *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, jul. de 2019, pp. 1325-1328. DOI: 10.1145/3331184.3331399.

[70] *GanttProject - Free Project Management Application*. [En línea]. Disponible en: https://www.ganttproject.biz/.

[71] *colaboratory/LICENSE at master · jupyter/colaboratory*. [En línea]. Disponible en: https://github.com/jupyter/colaboratory/blob/master/LICENSE.

[72] *spaCy/LICENSE at master · explosion/spaCy*. [En línea]. Disponible en: https://github.com/explosion/spaCy/blob/master/LICENSE.

[73] *Boletín Oficial del Estado*, 2010. [En línea]. Disponible en: https://www.boe.es/doue/2010/083/Z00389-00403.pdf.

[74] "Diario Oficial de la Unión Europea.," inf. téc., 2016. [En línea]. Disponible en: https://www.boe.es/doue/2016/119/L00001-00088.pdf.

[75] *La protección de datos en la UE | Comisión Europea*. [En línea]. Disponible en: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_es.

[76] *Boletín Oficial del Estado*. 1978. [En línea]. Disponible en: https://www.boe.es/buscar/pdf/1978/BOE-A-1978-31229-consolidado.pdf.

[77] J. Del Estado, "Boletín Oficial del Estado.," inf. téc., 2018. [En línea]. Disponible en: https://www.boe.es/boe/dias/2018/12/06/pdfs/BOE-A-2018-16673.pdf.

[78] *Boletín Oficial del Estado*. 2002. [En línea]. Disponible en: https://www.boe.es/buscar/pdf/2002/BOE-A-2002-22188-consolidado.pdf.

[79] D. Parra, P. Asistente y P. Chile, "Métricas de Evaluación Métricas de Evaluación IIC 3633-Sistemas Recomendadores," inf. téc.

[80] | *LILACS*. [En línea]. Disponible en: https://lilacs.bvsalud.org/es/.

[81] *IBECS*. [En línea]. Disponible en: https://ibecs.isciii.es/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&base=IBECS&lang=e.

ANEXO I. LISTA DE ABREVIATURAS

AHIMA	American Health Information Management Association
BOW	Bolsa de palabras
CBOW	Bolsa de palabras continua
CLEF	Conference and Labs of the Evaluation Forum
CM	Modificación Clínica (en ICD)
CNN	Convolutional Neural Networks - Redes de Neuronas Convolucionales
CRF	Conditional Random Field - Campo aleatorio condicional
ECDL	European Conference for Digital Libraries - Conferencia Europea de Bibliotecas Digitales
ICD	International Classification of Diseases - Clasificación Internacional de Diagnósticos
IE	Information Extraction - Extracción de la Información
IM	Information Management - Gestión de la Información
IR	Information Retrieval - Recuperación de la Información
MAP	Mean Average Precision - Media de Preción Ponderada
MRF	Markov random Fields - Campo aleatorio de Márkov
NCHS	Centro Nacional de Estadísticas de Salud
NER	Name-Entity Recognition - Reconocimiento de Entidades Nombradas
NLP	Natural Language Processing - Procesamiento del Lenguaje Natural
OMS	Organización Mundial de la Salud
SVC	Support Vector clustering/classification
SVM	Support Vector Machine - Máquinas de Soporte Vectorial
WHO	World Health Organization