UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

Detección de eventos adversos en historiales clínicos mediante Procesamiento del Lenguaje Natural

DANIEL MARTOS LÓPEZ

Trabajo Fin de Máster en Ingeniería y Ciencia de Datos

Curso 2020-2021 - Septiembre 2021

Directores:

Rafael Pastor Vargas
Carlos Luis Sánchez Bocanegra

Autorización de difusión

Daniel Martos López

Septiembre de 2021

El abajo firmante, matriculado en el Máster en Ingeniería y Ciencia de Datos de la Escuela Técnica Superior de Ingeniería Informática, autoriza a la Universidad Nacional de Educación a Distancia (UNED) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "Detección de eventos adversos en historiales clínicos mediante Procesamiento del Lenguaje Natural", realizado durante el curso académico 2020-2021 bajo la dirección de Rafael Pastor Vargas y con la colaboración externa de dirección de Carlos Luis Sánchez Bocanegra en el Departamento de Sistemas de Comunicación y Control, y a la Biblioteca de la UNED a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Resumen

El objetivo de este Trabajo Fin de Máster (TFM) es poner en práctica las competencias adquiridas en las enseñanzas del Máster en Ingeniería y Ciencia de Datos, más concretamente en el área del Procesamiento del Lenguaje Natural aplicado a textos clínicos. Mediante este trabajo se pretende diseñar un método que ayude al profesional sanitario a predecir eventos adversos, los cuales se definen como el daño físico no intencionado que es causado por los cuidados sanitarios más que por la enfermedad subyacente del paciente, a través de un catálogo de triggers ya definido, buscar patrones y agrupar en virtud de éstos obteniendo como resultado una mejora en la Seguridad del Paciente.

La metodología seguida ha sido la evaluación de diferentes herramientas cuya finalidad es la identificación de términos o conceptos clínicos en español o inglés dentro del vocabulario SNOMED-CT y su integración con el lenguaje de programación Python para la construcción de nuestra propia herramienta detectora de eventos adversos.

El resultado ha sido un método capaz de detectar posibles eventos adversos en texto clínico en español.

Palabras clave

Texto clínico, evento adverso, trigger, lenguaje natural, API REST, MetaMapLite, UMLS, SNOMED-CT.

Índice general

Índice I

Índice de figuras III

Índice de tablas IV

1. Introducción 1
- 1.1. Motivación . . . 1
- 1.2. Objetivos . . . 3
- 1.3. Metodología . . . 3
- 1.4. Estructura de la memoria . . . 3

2. Estado del arte 5
- 2.1. Minería de textos . . . 5
  - 2.1.1. Procesamiento del Lenguaje Natural . . . 6
  - 2.1.2. Extracción de Información . . . 7
  - 2.1.3. Aprendizaje Automático . . . 9
- 2.2. UMLS . . . 11
  - 2.2.1. Metatesauro . . . 12
  - 2.2.2. Red Semántica . . . 13
  - 2.2.3. Lexicón Especializado . . . 14
  - 2.2.4. Acceso a UMLS . . . 14
  - 2.2.5. SCTSPA . . . 15
- 2.3. SNOMED-CT . . . 16
- 2.4. Análisis de aplicaciones . . . 17
  - 2.4.1. cTakes . . . 17
  - 2.4.2. FreeLing . . . 19
  - 2.4.3. CUTEXT . . . 21
  - 2.4.4. SnowStorm . . . 22
  - 2.4.5. MetaMap . . . 23
  - 2.4.6. Stanza . . . 27
  - 2.4.7. Resumen . . . 29

3. Diseño del método 30
- 3.1. Fuentes de datos . . . 30
  - 3.1.1. Texto clínico . . . 30
  - 3.1.2. MetaMapLite . . . 30

3.1.3. Google Cloud Translation API . . . 31
3.1.4. UMLS REST API . . . 31
3.1.5. Abreviaturas UMLS . . . 31
3.1.6. Catálogo de triggers . . . 32
3.2. Diseño e implementación . . . 32
3.2.1. API RESTful . . . 32
3.2.2. Worker . . . 33
3.2.3. Trigger . . . 34
3.3. Despliegue . . . 36
3.4. Resultados . . . 37
3.4.1. TrigerApp . . . 38

4. Conclusiones y trabajo futuro **43**
4.1. Conclusiones . . . 43
4.2. Trabajo futuro . . . 43

Bibliografía **48**

Glosario **49**

A. Catálogo de triggers **50**

B. Ficheros Docker **54**
B.1. docker-compose.yml . . . 54
B.2. Dockerfile-trigger . . . 55
B.3. Dockerfile-api . . . 56

C. Ficheros de ejemplo **58**
C.1. Fichero de entrada . . . 58
C.2. Fichero de salida . . . 60

Índice de figuras

2.1. Técnicas de preprocesamiento de texto. Fuente: (Lourdusamy & Abraham, 2018) . . . 6
2.2. Fuentes de conocimiento UMLS. Fuente: UMLS . . . 12
2.3. Modelo lógico SNOMED-CT. Fuente: (IHTSDO, 2018) . . . 16
2.4. Ejecución del CVD de cTakes . . . 18
2.5. Servicios de análisis disponibles para cada lengua. Fuente: (Padró et al. 2011) 19
2.6. Depuración API Python de FreeLing . . . 20
2.7. Ejemplo de salida de fichero FreeLing . . . 20
2.8. Ejemplo de salida de fichero CUTEXT . . . 21
2.9. Resultado de la consulta del término "colesterol" a SnowStorm . . . 23
2.10. Diagrama del sistema MetaMap. Fuente: (Aronson & Lang, 2010) . . . 25
2.11. Ejemplo de salida de fichero MetaMap . . . 27
2.12. Ejemplo de salida de fichero Stanza . . . 28

3.1. Estructura del proyecto . . . 33
3.2. Diseño del servicio API RESTful basado en estados de una tarea de análisis 34
3.3. Documentación en formato Swagger del servicio API RESTful . . . 35
3.4. Trigger "B5. Laxante o enema" a partir de los ancestros del término detectado "sulfato de magnesio" . . . 36
3.5. Arquitectura completa del método Trigger . . . 37
3.6. Fragmento de texto clínico del documento S0212-16112007000800011-1 procedente del corpus SPACCC (Intxaurrondo, 2018) . . . 38
3.7. Pantalla principal de TriggerApp . . . 39
3.8. Pantalla "Nueva tarea" de TriggerApp . . . 40
3.9. Pantalla "Editar" de TriggerApp . . . 40
3.10. Pantalla "Eliminar" de TriggerApp . . . 41
3.11. Pantalla "Consulta" de TriggerApp . . . 41
3.12. Pantalla de información en "Consulta" de TriggerApp . . . 42
3.13. Pantalla "Catálogo de Trigger" de TriggerApp . . . 42

Índice de tablas

2.1. Resumen de las aplicaciones analizadas 29

A.1. Catálogo de triggers codificados en SNOMED-CT asociados a cuidados generales 51
A.2. Catálogo de triggers codificados en SNOMED-CT asociados a medicamentos o tratamientos 52
A.3. Catálogo de triggers codificados en SNOMED-CT asociados a resultados de laboratorio o microbiología 53
A.4. Catálogo de triggers codificados en SNOMED-CT asociados a pruebas diagnósticas 53

Capítulo 1

Introducción

1.1. Motivación

Según la Organización Mundial de la Salud (OMS) la **seguridad del paciente** (WHO, 2019) es una disciplina de la atención de la salud que surgió con la evolución de la complejidad de los sistemas de atención de la salud y el consiguiente aumento de los daños a los pacientes en los centros sanitarios. Su objetivo es prevenir y reducir los riesgos, errores y daños que sufren los pacientes durante la prestación de la asistencia sanitaria. Una piedra angular de la disciplina es la mejora continua basada en el aprendizaje a partir de los errores y eventos adversos. Dichos **eventos adversos** (EA) son incidentes que causan daño al paciente, y se considera daño cualquier alteración estructural o funcional del organismo o todo efecto perjudicial derivado de ella. Los daños comprenden las enfermedades, las lesiones, los sufrimientos, las discapacidades y la muerte, y pueden ser físicos, sociales o psicológicos (WHO, 2009).

Cuando un paciente padece un daño accidental en el proceso de atención, la confianza en el profesional se deteriora. Es una experiencia traumática y dolorosa pero no solo para el paciente y su familia, sino también para los profesionales sanitarios que se ven involucrados y que se convierten así en segundas víctimas de dicho EA. La tercera víctima es la organización de salud que sufre a consecuencia de un EA una importante pérdida de su reputación entre los ciudadanos y pacientes que llegan a desconfiar de los servicios que presta (Torijano-Casalengua et al. 2016). Una forma de comprobar si el compromiso por mejorar la seguridad de los pacientes alcanza sus objetivos es medir la frecuencia de EA. Recientes estudios han puesto de manifiesto que, en estos años, no se ha logrado una reducción significativa de su número. Esto puede ser debido, en parte, a que ahora medimos con mayor exhaustividad que hace unos años, a que somos más conscientes del problema y, también, a que comprendemos mejor la naturaleza de los EA, incluidos los que tienen causas más complejas de detectar (Carrillo et al. 2020). Actualmente la detección de estos EA se realiza mediante estudios retrospectivos de cohortes que constan de una **revisión manual de una selección aleatoria de la historia clínica electrónica (HCE) por parte de los profesionales sanitarios**, con el coste en tiempo y recursos que ello conlleva.

La **HCE** comprende el conjunto de los documentos relativos a los procesos asistenciales de cada paciente en formato electrónico. Según la «Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica» el contenido mínimo de la historia clínica será el siguiente:

- La documentación relativa a la hoja clínico estadística.
- La autorización de ingreso.
- El informe de urgencia.
- La anamnesis y la exploración física.
- La evolución.
- Las órdenes médicas.
- La hoja de interconsulta.
- Los informes de exploraciones complementarias.
- El consentimiento informado.
- El informe de anestesia.
- El informe de quirófano o de registro del parto.
- El informe de anatomía patológica.
- La evolución y planificación de cuidados de enfermería.
- La aplicación terapéutica de enfermería.
- El gráfico de constantes.
- El informe clínico de alta.

En la mayoría de los documentos enumerados anteriormente existe una gran cantidad de campos en formato texto no estructurado, que unido a que la implantación de dicha HCE en las instituciones sanitarias está en un nivel lo suficientemente maduro, puede resultar de interés la aplicación de técnicas de Inteligencia Artificial (IA) y más concretamente del área del Procesamiento del Lenguaje Natural (PLN) para la extracción de información y conocimiento.

En (Jick, 1974) se introdujo el concepto de **triggers** (gatillos, disparadores, pistas) para detectar EA en las historias clínicas. La metodología consiste en revisiones retrospectivas de un número limitado de historias clínicas seleccionadas al azar buscando en ellas pistas o indicios (triggers) que lleven a identificar posibles eventos adversos. Actualmente no se

dispone de un gold standard para la detección de EA en pacientes hospitalizados, aunque existen distintos métodos para ello, siendo los principales los siguientes: notificación de incidentes, revisiones sistemáticas y protocolizadas de historias clínicas, observación directa, uso de sistemas electrónicos y la herramienta, promovida principalmente por el Institute for Healthcare Improvement (IHI), mediante la metodología Global Trigger Tool (GTT). (Guzmán-Ruiz et al. 2015)

1.2. Objetivos

El **objetivo principal** es diseñar un método que ayude al profesional sanitario en la detección de eventos adversos a partir de texto no estructurado procedente de la documentación sanitaria de la historia clínica electrónica mediante técnicas de Minería de Textos y Procesamiento del Lenguaje Natural para la prevención y mejora en la seguridad del paciente.

Los **objetivos secundarios** son:

1. Evaluar las distintas soluciones disponibles para el análisis de texto clínico no estructurado.

2. Diseñar un servicio fácil de usar, accesible y escalable.

3. Permitir recuperar los textos analizados para una futura evaluación por parte de expertos.

1.3. Metodología

Es mucha, y muy diversa, la bibliografía relativa al Procesamiento del Lenguaje Natural en análisis de textos clínicos no estructurados en inglés pero muy escasa para el caso de textos en español (Santamaría & Krallinger, 2018). La primera tarea a la que nos enfrentamos es identificar cuáles son las particularidades en el análisis de este tipo de textos, así como de su terminología, y a continuación evaluar las distintas aplicaciones que nos puedan servir de apoyo para la consecución de los objetivos propuestos. Finalmente, diseñaremos un método que permita al personal sanitario la detección de eventos adversos en textos clínicos no estructurados en español.

1.4. Estructura de la memoria

Este documento se ha estructurado entorno a cuatro capítulos principales:

1. **Introducción**: en este capítulo se hace una breve justificación sobre la motivación del trabajo propuesto así como una definición de los objetivos propuestos con la realización del mismo.

2. **Estado del arte**: en este capítulo haremos una pequeña introducción a las tareas que abarcan la minería de textos, una descripción del sistema UMLS, la teminología SNOMED-CT y las pruebas realizadas con algunas aplicaciones de análisis de textos necesarias para la consecución de nuestros objetivos.

3. **Diseño del método**: en este capítulo describiremos las fuentes de datos utilizadas y el diseño del método propuesto, así como la implementación del servicio, despliegue y una demo de utilización del mismo.

4. **Conclusiones y trabajos futuros**: en este último capítulo se presentan las conclusiones extraídas con este trabajo, así como aquellas líneas de trabajo futuro que puedan mejorar el sistema propuesto.

Capítulo 2

Estado del arte

En este capítulo haremos una pequeña introducción a las tareas que abarcan la minería de textos (Luque Guzmán, 2020), una descripción del sistema UMLS (Bravo et al. 2018b), la teminología SNOMED-CT y las pruebas realizadas con algunas aplicaciones de análisis de textos necesarias para la consecución de nuestro objetivo.

2.1. Minería de textos

La minería de textos (MT) es un área orientada a la extracción de conocimiento a partir de información de contenido textual basada en un conjunto de técnicas avanzadas que permiten descubrir eventos de interés, novedosos y que explícitamente eran inexistentes en las colecciones textuales de las que se partían (Weiss et al. 2015). Los sistemas de MT permiten extraer información relevante desde distintas fuentes textuales para generar un nuevo conocimiento.

Generalmente, la mayoría de los sistemas basados en MT siguen una serie de procesos para obtener conocimiento (relevante, nuevo y desconocido a priori) partiendo de colecciones de datos textuales. Se parte como entrada de una información con contenido textual que debe ser sometida a una fase de preprocesamiento (Lourdusamy & Abraham, 2018) donde los datos no estructurados se "preparan" mediante diferentes técnicas (tokenización, stemming, etc). A continuación, le sigue una segunda fase donde la información se somete a un modelo de representación adecuado donde es transformada para poder ser interpretada. En la última fase, llamada fase de descubrimiento, se extrae el verdadero valor y conocimiento del texto de partida mediante la aplicación de ciertos métodos y técnicas (clasificación, agrupación, etc).

Existen múltiples disciplinas que nutren e influencian a la MT, algunas de las más importantes son el Procesamiento del Lenguaje Natural (PLN), Extracción de Información (EI) y Aprendizaje Automático (AA). Estas áreas son la base sobre las que se construyen la mayoría de los sistemas de MT y sus técnicas son imprescindibles para llevar a cabo el análisis y procesamiento de grandes colecciones textuales.

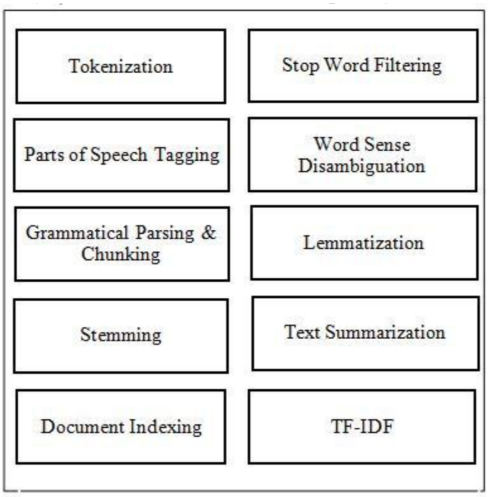

Figura 2.1: Técnicas de preprocesamiento de texto. Fuente: (Lourdusamy & Abraham, 2018)

2.1.1. Procesamiento del Lenguaje Natural

La disciplina del PLN (Chowdhury, 2003) nace en la década de los 60 como un área independiente de la Inteligencia Artificial (IA) y la Lingüística Computacional. El objetivo original de esta disciplina era estudiar los problemas derivados de la comprensión automática del lenguaje natural.

Algunos autores definen el PLN como una parte esencial de la IA que es capaz de formular mecanismos computacionales que facilitan la interrelación hombre-máquina, otros autores la definen como una disciplina que entiende la habilidad de la "máquina" para procesar y entender la información comunicada.

Uno de los primeros sistemas basados en PLN fue el denominado BASEBALL (Green Jr et al. 1961), desarrollado en la década de los 60, un interfaz orientado a la comprensión del lenguaje natural. En los años 70 se amplía el área de acción hacia otros campos como la enseñanza asistida por ordenador, comprensión del lenguaje, automatización de tareas, etc. Gracias al avance de la IA se pudo desarrollar el primer sistema de pregunta- respuesta basado en lenguaje natural.

Algunas de las principales aplicaciones del PLN son: traducción automática, interfaces humano-computadora, educación asistida por ordenador, tutores inteligentes, sistemas de búsqueda de respuestas, síntesis de voz, reconocimiento del habla, análisis de sentimientos, minería de opiniones, etc.

El lenguaje natural posee una serie de complejas características que pueden hacer, si no se tratan convenientemente, que en muchas ocasiones disminuya la efectividad de los sistemas basados en PLN. Algunos de los principales problemas a los que se enfrentan la disciplina del PLN son (Meystre et al. 2008):

- La **anáfora** se puede definir brevemente como el término empleado para hacer referencia a algo que anteriormente ya fue mencionado. Aplicando técnicas de PLN se puede paliar el problema de la resolución de las anáforas.

- Otro de los escollos que tienen que salvar los sistemas PLN es la resolución de la **ambigüedad**. En el lenguaje natural nos encontramos muy frecuentemente expresiones que pueden tener varios significados diferentes según el contexto en el que lo estemos utilizando. Cuando un ordenador tiene que seleccionar una única interpretación de entre varias, para desambiguar se requiere aplicar varias estrategias que sin la ayuda del PLN no se conseguiría.

- La detección de la **negación** es un problema de especial relevancia en algunos entornos como, por ejemplo, el jurídico o el médico. Es habitual que se utilicen expresiones negadas con expresiones positivas negadas o al revés, un ejemplo de ello lo podemos ver en la siguiente oración "no complaints of radicular pain". Es un fenómeno lingüístico difícil de detectar automáticamente, pero gracias al PLN se ha conseguido avanzar con buenos resultados en la detección de las negaciones.

- La utilización de **acrónimos** está cada día más extendida. Los acrónimos se definen como las palabras formadas por las iniciales de otras palabras que constituyen la denominación de algo. Su resolución también es uno de los problemas del ámbito del PLN. Esta proliferación de acrónimos hace que los procesos de recuperación y extracción de información puedan verse mermados si no se emplean las técnicas y procedimientos adecuados. En el ámbito de la biomedicina, algunos autores afirman que por cada cinco artículos surge una nueva sigla que llega a coincidir con un gran número de siglas preexistentes y que el uso de los acrónimos hace aumentar la polisemia y la sinonimia léxica, dos fenómenos semánticos que nos llevan de nuevo al problema de la ambigüedad. (Benavent et al. 2001)

Las fases más frecuentemente utilizadas en los sistemas PLN son (Sosa, 1997):

- **Análisis morfológico-léxico**, consiste básicamente en la obtención de palabras a partir de un texto, también se realiza la asignación de etiquetas morfológicas a las palabras de un texto (Part-Of-Speech tagging).

- **Análisis sintáctico**: da a conocer las categorías gramaticales de cada palabra y como se combinan los tokens para formar oraciones y textos.

- **Análisis semántico**: esta fase se refiere a la comprensión del lenguaje, se analiza el significado de las palabras y la resolución de ambigüedades léxicas.

- **Análisis pragmático**, se analiza cómo las oraciones se usan en un determinado contexto y cómo afecta al significado de las oraciones.

2.1.2. Extracción de Información

El avance y desarrollo de los sistemas PLN, junto con la Recuperación de Información (RI), dio origen a otra disciplina dependiente de esta denominada Extracción de Información (EI). La EI, tiene como principal objetivo encontrar y seleccionar información relevante para

el estudio de un dominio particular, denominado dominio de extracción. Pero quizás una definición más cercana a lo que es hoy en día la EI podría ser la que define esta disciplina como una ciencia que trata de identificar, clasificar y reestructurar información específica existente en fuentes desestructuradas, como por ejemplo los textos, para poder realizar su posterior procesamiento automático (Vilares, 2006).

Los primeros estudios relacionadas con la EI se ubican a mediados de los años 60's, pero es a finales de los años 80 cuando esta tecnología comienza a tener auge, lo cual se debe principalmente a tres factores. En primer lugar, el poder computacional que ya empezó a estar disponible con bastante potencia en dicha época; segundo, el exceso de información textual existente en formato electrónico; y por último , la intervención de la Agencia de Defensa de los Estados Unidos (DARPA), que promocionaron durante los años de 1987 a 1998 las siete conferencias de entendimiento de mensajes (MUC) y activaron durante los años de 1990 a 1998 el programa TIPSTER (programa de investigación sobre recuperación y extracción de información del gobierno de EEUU) donde las MUC's fueron incluidas. Las MUC's fueron las que inicialmente fomentaron las competencias entre distintos grupos de investigación. Las cuales se llevaron a cabo con el objetivo de desarrollar sistemas de EI. Muchos sistemas de extracción de información (SEI) surgieron gracias a los MUCs, por ejemplo FASTUS, CRYSTAL, Autoslog, etc. Quizás uno de los sistemas de EI más conocido sea FASTUS (Finite State Automaton Text Understanding System), un sistema capaz de extraer información desde texto libre en inglés, japonés y otros lenguajes. Se aplicó inicialmente a la tarea de extraer información de artículos sobre el terrorismo en América Latina para la conferencia MUC-4. Otro foro que ha contribuido históricamente en el ámbito de los sistemas de RI y EI, son las conferencias TREC (Text Retrieval Conference), patrocinadas por el NIST (National Institute of Standards and Technology) y el Departamento de Defensa de los Estados Unidos, que comenzaron en 1992. En este mismo ámbito de la RI y EI, otro foro que ha aportado gran conocimiento en estas disciplinas son las conferencias Cross Language Evaluation Forum (CLEF). CLEF es un foro de evaluación que apoya el uso y desarrollo de aplicaciones para la gestión y manejo de librerías digitales. Para ello, desarrollan infraestructuras de prueba, mejora y evaluación de sistemas de recuperación de información multimodal y multilingüe. CLEF nació en enero de 2000, como una evolución de una línea de estudio que se había formado en TREC junto con un grupo de voluntarios europeos, entre 1997 y 1999, para el estudio de los lenguajes multilingües europeos. Todas estas conferencias iniciadas en los años 80 y 90 supusieron un gran avance y un gran marco de referencia para los futuros investigadores en las áreas de la RI y EI.

Algunas de las tareas más importantes llevadas a cabo gracias a los SEI son, entre otras, el reconocimiento de entidades nombradas, resolución de correferencias, reconocimiento de relación entre entidades, reconocimiento de expresiones temporales, etc.

El objetivo final de un SEI es partir de un texto no estructurado y llegar a conseguir, a través de una cascada de módulos que van aportando estructuración al documento, un conjunto de información estructurada y relevante, gracias al filtrado de información a través de la aplicación de determinadas reglas.

2.1.3. Aprendizaje Automático

La disciplina del Aprendizaje Automático (AA) (Jordan & Mitchell, 2015), considerada como una rama de la IA, tiene como objetivo la construcción de sistemas capaces de adquirir conocimiento y aprender automáticamente en base a un conjunto de datos de entrenamiento. El AA puede considerarse un proceso de inducción del conocimiento. El auge de la investigación y workshops dedicados expresamente a la disciplina del AA tuvieron lugar al inicio de los años 80 (aunque fue en los años 60 cuando surgen los primeros artículos relacionados con AA). Los primeros investigadores en convertir al AA en una de las subáreas de la IA de mayor importancia y a los que debemos el crecimiento actual de esta disciplina fueron, entre otros, Michalski, Carbonell, Mitchell y Dietterich.

Una de las primeras definiciones de la palabra "aprendizaje" se recoge en (Simon, 1983) donde se detalla como el aprendizaje marca cambios adaptativos en el sistema que pueden permitir que se realicen las mismas tareas con mayor eficacia cada vez, por tanto, el propósito del aprendizaje es mejorar el rendimiento de algunas clases de tareas. Quizás una de las definiciones más citadas sea la propuesta en (Mitchell, 2006), donde se afirma que "un programa aprende de la experiencia respecto a una clase de tareas y una medida de la eficiencia si su eficiencia en las tareas se incrementa con la experiencia". Para Dietterich, el problema de definir aprendizaje se reducía a definir conocimiento.

Otros autores más actuales definen el AA como la disciplina de la IA dedicada al diseño de algoritmos para identificar regularidades, patrones o reglas sobre un conjunto de datos o el estudio de algoritmos que pueden aprender relaciones complejas o patrones a partir de datos empíricos y tomar decisiones precisas en base a ellos.

Las aplicaciones del AA han sido múltiples: análisis de mercado, análisis de riesgos crediticios, detección de fraudes, clasificación de secuencias de ADN, soporte al diagnóstico y pronóstico médico, reconocimiento de patrones, problemas de clasificación, reconocimiento de imágenes, reconocimiento de spam, sistemas de recomendación, soporte a motores de búsqueda, análisis de tendencias, etc. La aplicación de estas técnicas, a lo largo de los últimos 20 años, ha contribuido a mejorar nuestro día a día en prácticamente todos los sectores de la sociedad actual, demostrándose que estas técnicas tienen un alto grado de eficacia y fiabilidad.

En líneas generales, podemos realizar una clasificación de los tipos de AA más comúnmente utilizados según el mecanismo y los métodos que usan para aprender (Jordan & Mitchell, 2015):

- Aprendizaje **supervisado**: quizás uno de los más empleados en AA, se caracteriza porque el proceso de aprendizaje es controlado gracias a un conjunto de datos de entrenamiento previamente etiquetados, de ahí que reciba el nombre de supervisado. El principal objetivo del aprendizaje supervisado es generar una función capaz de predecir nuevos datos de entrada en base al modelo aprendido que se apoya en un conjunto inicial de datos de entrenamiento.

  Las modalidades más importantes dentro del aprendizaje supervisado son la clasificación y la regresión. El objetivo de la clasificación es predecir la clase (o clases) a la

que pertenece una instancia en base a patrones de entrada previamente etiquetados. Así, a partir de un conjunto entrenado se infiere un modelo (denominado clasificador) que será utilizado para categorizar nuevas instancias no etiquetadas. El objetivo de la regresión es similar al de la clasificación, predecir un valor de salida en base a un patrón de entrada etiquetada, pero se infiere un valor continuo en lugar de categórico. Algunos de los algoritmos de aprendizaje supervisados de uso más extendido son las Máquinas de Vector Soporte (SVM), vecinos más cercanos (KNN), Naïves Bayes, árboles de decisión y redes neuronales. Son muchos los trabajos donde se han aplicado las técnicas de aprendizaje supervisado en áreas como la categorización de documentos, clasificación de imágenes, análisis de sentimientos, detección spam, detección de enfermedades, etc.

- Aprendizaje **no supervisado**: al contrario de lo que ocurre con el aprendizaje supervisado, no parte a priori de ningún conjunto de datos etiquetados. El objetivo de este tipo de aprendizaje es realizar agrupaciones de conjuntos de datos similares en base a las características que comparten. Este tipo de aprendizaje es imprescindible cuando se dispone de conjuntos de datos no etiquetados y cuando no se puede asumir el coste necesario para categorizar una gran colección de información.

  Una de las tareas más utilizadas dentro del aprendizaje no supervisado es el clustering, siendo los algoritmos K-means y DBSCAN los de uso más extendido. Las técnicas de clustering han sido aplicadas en múltiples sectores como segmentación de clientes, documentación médica, apoyo al diagnóstico médico, etc.

Aprendizaje profundo (Deep Learning)

Durante la última década han ido surgiendo numerosas soluciones a problemas de PLN basadas en Redes Neuronales gracias a la irrupción de los embeddings (representaciones matriciales estáticas del texto, en las que cada palabra del vocabulario está codificada en un vector) por la publicación en 2013 de word2vec (Mikolov et al. 2013) y la aparición de modelos tan avanzados como BERT (Devlin et al. 2019) y GPT-3 (Brown et al. 2020), basados en Transformers que partiendo de los embeddings, se aplican varias capas, denominadas de autoatención, que mezclan los vectores de representación de las palabras hasta conseguir otros tantos contextualizados, es decir, estos nuevos vectores tendrán información del resto del texto, no solo de la palabra, sino también de cómo se usa en un documento y qué otras la acompañan (Vaswani et al. 2017).

En (Peterson & Liu, 2020) podemos encontrar una posible solución al problema de relacionar conceptos médicos en texto no estructurado en SNOMED-CT usando Bidirectional Long Short-Term Memory (BiLSTM) (Schuster & Paliwal, 1997) y Clinical BERT (Alsentzer et al. 2019). También se puede ver en (Campillos-Llanos et al. 2021) el proceso anotación de entidades UMLS usando este tipo de enfoques o la creación de un espacio vectorial de conceptos SNOMED-CT con word2vec como en (Soriano et al. 2019).

2.2. UMLS

El UMLS[1] (Unified Medical Language System) es un sistema formado por un conjunto de archivos y software que integra y unifica vocabularios biomédicos. La versión actual del UMLS (2021AA) contiene 220 recursos terminológicos para 25 lenguas diferentes, de los cuales 146 son vocabularios ingleses, 9 españoles y 1 para el vasco.

El UMLS es un sistema que garantiza referencias cruzadas entre vocabularios y ontologías gracias al análisis léxico de los términos (Bodenreider, 2004). Se emplean técnicas de procesamiento que están basadas en unidades léxicas, de forma que los términos se comparan en función de lo que parecen significar (en el contexto de una frase). Se ha de señalar que el sistema UMLS no está diseñado para la consulta humana, sino que tiene como objetivo ayudar a los desarrolladores de software en una mejor implementación de sus sistemas relacionados por ejemplo en:

- Extracción de Información (EI).
- Construcción de Corpora.
- Procesamiento del Lenguaje Natural (PLN).
- Indexación Automática.
- Soporte al desarrollo de Historias Clínicas Electrónicas (HCE).

Como hemos comentado, el UMLS integra las ontologías, terminológicas y vocabularios más importantes en el campo biomédico. Obviamente, la gran mayoría de fuentes están en inglés, pero también incluye terminologías en otros idiomas como el español (CPTSP, CPCSPA, LNC-ES-AR, LNC-ES-CH, LNC-ES-ES, MDRSPA, MSHSPA, SCTSPA y WHOSPA) y el vasco (ICPCBAQ).

El sistema UMLS está formado principalmente por tres fuentes de conocimiento:

1. El Metatesauro (Metathesaurus): Términos y códigos unificados de diferentes vocabularios y ontologías biomédicos.
2. La Red Semántica (Semantic Network): mantiene una organización entre los conceptos del Metatesauro mediante categorías generales (tipos semánticos) y sus relaciones (relaciones semánticas).
3. El Léxico Especializado (SPECIALIST Lexicon): proporciona la información léxica necesaria para el desarrollo de sistemas de PLN. Incluye la información sintáctica, morfológica y ortográfica de palabras del inglés y términos biomédicos.

Específicamente, el sistema UMLS ha utilizado la Red Semántica y el Léxico Especializado para producir el Metatesauro. Aunque se utilicen ambas herramientas para la producción del Metatesauro, puede accederse a ellas por separado o en cualquier combinación dependiendo las necesidades. La producción del Metatesauro implica:

[1] http://www.nlm.nih.gov/research/umls/

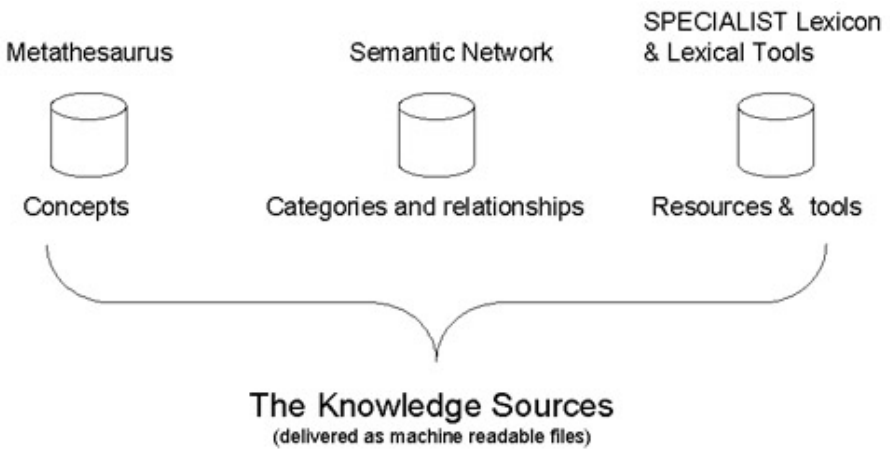

Figura 2.2: Fuentes de conocimiento UMLS. Fuente: UMLS

- Procesar los términos y códigos que utilizan las herramientas de léxico.
- Agrupar términos sinónimos en conceptos.
- Categorizar conceptos por tipos semánticos de la Red Semántica.
- Incorporar las relaciones y los atributos proporcionados por vocabularios.
- Liberar los datos en un formato común

2.2.1. Metatesauro

El Metatesauro es el componente más poderoso del sistema del UMLS. Está formado por una gran base de datos la cual contiene información sobre conceptos relacionados con el dominio biomédico, integrados de muchas fuentes de conocimiento e incluyendo toda su variedad terminológica y relaciones entre ellos. Actualmente contiene más de 100 vocabularios de diferentes recursos de los cuales un 67 % son en Inglés.

Básicamente el Metatesauro está organizado en **Conceptos** (los sinónimos son agrupados en un único concepto) y **Relaciones** (los conceptos son relacionados entre ellos). Cada Concepto abarca una unidad de significado y a su vez, está compuesto por diferentes partes organizadas de manera jerárquica:

- Concepto (CUI): Conjunto de todos los sinónimos haciendo referencia al mismo concepto.
- Término (LUI): Agrupa las variantes léxicas de cada concepto.
- String (SUI): Cada término (o String) diferente en cada idioma es identificado como un nuevo String. Cualquier variación morfológica (mayúsculas, acrónimos, signos de puntuación) entre los términos será asociado a un nuevo SUI.

- Atom (AUI): Es el concepto más básico del Metatesauro. Para cada String se diferencia su procedencia dependiendo del recurso del cual procede. Es decir, cada ocurrencia de un String en cada vocabulario (recurso) será identificado con un AUI.

El Metatesauro consiste en unos 40 ficheros de datos (conceptos, atributos y relaciones), metadatos e índices. Los ficheros son listados a continuación indicando también su contenido:

- Terminología: Conceptos, nombres, sinónimos y sus fuentes.
- Atributos: Los atributos son añadidos durante la construcción del Metatesauro y aplica a todos los términos de un concepto.
- Relaciones: El Metatesauro también incluye relaciones entre conceptos. La mayoría de estas relaciones vienen de fuentes de vocabulario individuales, y solo unas pocas son añadidas por la NLM durante la construcción.
- Metadatos: Los ficheros que contienen información de metadatos describen i) características de la versión actual del Metatesauro; ii) cambios entre la versión actual y la anterior; y iii) la historia de los identificadores de concepto (CUIs) desde 1991 hasta el día de hoy.
- Índices: Para facilitar a los programadores el desarrollo de aplicaciones que recuperen todos los nombres de conceptos que incluyen palabras específicas o grupos de palabras, se proporcionan tres índices a los nombres de conceptos: un Word Index, un Normalized Word Index (solo palabras en inglés), y un Normalized String Index (solo palabras en inglés).

2.2.2. Red Semántica

La Red Semántica fue creada en un esfuerzo para proporcionar una organización semántica al sistema UMLS y sus vocabularios integrados. Al contrario del tamaño del Metatesauro, la Red Semántica es una pequeña estructura compuesta por 135 tipos semánticos, los cuales garantizan una categorización consistente de todos los conceptos representados en el Metatesauro. Los 135 tipos semánticos están organizados en dos categorías, cada una con un árbol jerárquico: Entity y Event.

Cada concepto en el Metatesauro está asignado a uno o más **tipos semánticos**. Los tipos semánticos están vinculados entre sí a través de **relaciones semánticas**. El UMLS proporciona 54 relaciones semánticas para sus 135 tipos semánticos. El vínculo principal es el enlace "IS_A", el cual establece la jerarquía de la Red Semántica. Además, se han definido 5 categorías principales para el resto de los 53 enlaces no jerárquicos o asociativos: physically_related_to, spatially_related_to, temporally_related_to, functionally_related_to y conceptually_related_to.

Para una mayor organización del sistema UMLS, se han definido 15 colecciones diferentes de tipos semánticos, llamados **grupos semánticos** (Semantic Groups). Estos grupos

facilitan la clasificación de los conceptos de Metatesauro en un número menor de grupos semánticamente consistentes. Los principales grupos semánticos son: organismos, estructuras anatómicas, funciones biológicas, productos químicos, eventos, objetos físicos y conceptos o ideas.

Las relaciones se asientan entre los nodos del nivel más alto de la Red siempre que se puede y, generalmente, se heredan, gracias al enlace "IS_A", por todos los hijos de dichos nodos.

2.2.3. Lexicón Especializado

El objetivo del Lexicón Especializado (conocido como SPECIALIST Lexicon) es proporcionar la información léxica necesaria para sistemas de PLN. Mayoritariamente incluye entradas de palabras en inglés y vocabulario biomédico.

Para cada entrada de una palabra o término (formado por varios elementos léxicos) se incluye la información sintáctica, morfológica y ortográfica. La forma base no tiene flexión, es decir, en el caso de los verbos se usa el infinitivo, el singular para sustantivos y la forma positiva en el caso de adjetivos y adverbios. La información léxica incluye la categoría sintáctica, variación de la inflexión (singular o plural para los sustantivos, conjugación de los verbos, el comparativo, superlativo para los adjetivos y adverbios) y posibles patrones de complementación (objetos y otros argumentos que los verbos nombres y adjetivos pueden regir). El lexicón reconoce múltiples categorías sintácticas o partes del discurso: verbos, nombres, adjetivos, adverbios, auxiliares, modales, pronombres, preposiciones, conjunciones y determinantes.

Los patrones básicos de la oración se determinan por el número y la naturaleza de los complementos que rigen los verbos. El lexicón reconoce cinco tipos generales de complementación: intransitiva, transitiva, ditransitiva, linking y transitiva-compleja. Las entradas verbales contemplan las formas flexivas del verbo, si son regulares o irregulares. En cuanto a los sustantivos, se recogen patrones de pluralización y de nominalización.

2.2.4. Acceso a UMLS

Los Servicios de Terminología UMLS (UTS) proporcionan acceso vía Internet a las fuentes que forman el conocimiento del UMLS. El propósito del UTS es hacer el UMLS más accesible a los usuarios y en particular a los desarrolladores de software.

El acceso al UTS está solo disponible a quien ha firmado el acuerdo de la Licencia del Metatesauro del UMLs y ha activado su cuenta UTS. Para poder acceder al servicio, la primera vez, los usuarios deben hacer clic en "Sign Up" en la web principal[2] del UTS para enviar la petición de licencia y empezar con el proceso de activación de la cuenta.

No hay cargo por licenciar el UMLS de la NLM. NLM es miembro de la IHTSDO (dueño de SNOMED-CT), y no hay ningún cargo por el uso de SNOMED-CT en los Estados Unidos y en otros países miembros, en los cuales entra España. Algunos usos de las UMLS pueden

[2] https://uts.nlm.nih.gov/uts/

requerir acuerdos adicionales con proveedores de terminologías particulares. La cuenta UTS le permite navegar, descargar y consultar el UMLS.

El acceso al sistema UMLS por UTS está disponible a través de:

- Navegadores Web:
  - Navegador Metathesaurus: permite recuperar información de conceptos UMLS, incluyendo CUIS, tipos semánticos y términos sinónimos.
  - Semantic Network Browser: permite ver los nombres, definiciones y la estructura jerárquica de la Red Semántica.
- Instalación local: permite instalar los UMLS en su ordenador y descargar los archivos a través de las UTS. La herramienta Metamorphosys, incluida con los archivos descargados, permite personalizar los UMLS de acuerdo a sus necesidades. A continuación, puede cargar los datos personalizados en su propio sistema de base de datos, como MySQL u Oracle, o puede buscar sus datos mediante el explorador RRF Metamorphosys.
- Web Services API: puede usar las interfaces de programación de aplicaciones (API) para consultar los datos UMLS dentro de su propia aplicación.

2.2.5. SCTSPA

Desde el punto de vista de las 9 fuentes que componen el UMLS en español, la edición en español de SNOMED-CT, SNOMED-CT Spanish Edition (SCTSPA), es la más completa incluyendo 364.897 conceptos, los cuales componen más del 76 % de todos los conceptos médicos en español. Muchos han sido los esfuerzos para poder ampliar y extender esta fuente terminológica en español (SCTSPA) y los números presentados lo demuestran. Si comparamos esta fuente terminológica con su homóloga inglesa notamos que presentan una distancia mínima. SNOMED-CT US está compuesto por 382.092 conceptos de los cuales solo 19.674 (5 %) no están incluidos en SCTSPA. Dicho esto, también hay que tener en cuenta la calidad de los conceptos representados en SCTSPA respecto a los que están en SNOMED-CT US desde el punto de vista de la variedad terminológica. En este sentido, (Perez-de-Viñaspre & Oronoz, 2015) llevaron a cabo un estudio cuantitativo sobre el número de términos completamente especificados en el contenido terminológico de las versiones en español y en inglés de SNOMED-CT. Su conclusión fue que, aunque el número de conceptos es parecido, casi 16.000 conceptos en español no presentaban "términos preferidos" o sinónimos en la versión española. (Bravo et al. 2018a)

Aunque el número de conceptos reportando esta poca variabilidad terminológica es bajo, hay que tener en cuenta este tipo de estudio a la hora de valorar también, la diversidad lingüística de una futura extensión del UMLS.

En nuestro estudio sobre el UMLS, se puede concluir que SNOMED-CT es el recurso lingüístico médico en español más importante y utilizado. Lamentablemente, esta tendencia no es la habitual en el resto de fuentes terminológicas del UMLS. Si se observa el UMLS no

a través de los recursos integrados, sino a través de los **tipos semánticos** (TUIs) que lo componen, se aprecia fácilmente un desequilibrio entre los conceptos integrados en inglés y en español.

2.3. SNOMED-CT

Las **terminologías clínicas** pueden definirse como listas de términos empleados en el ámbito médico. Normalmente, tienen algunas características ontológicas para describir formalmente los términos y sus relaciones. Las terminologías clínicas han surgido para ser usadas por los sistemas de información para capturar, procesar y transferir los datos clínicos de una forma consistente y estandarizada. Las terminologías además son claves en varios escenarios: en la integración de diversos sistemas de información, en la conexión de la HCE con los entornos de soporte a la decisión y en la reutilización de la información clínica (generada durante el proceso asistencial de los pacientes) para otros fines, como puede ser la investigación, la gestión hospitalaria o la evaluación de la calidad (Teresa Romá-Ferri & Palomar, 2008). Algunas de las terminologías más utilizadas son SNOMED-CT, MESH, LOINC, RXNORM y UMLS.

SNOMED-CT o Systematized Nomenclature of Medicine Clinical Terms (IHTSDO, 2018) es una extensa terminología clínica de atención médica, disponible en varios idiomas y usada en la actualidad por más de cincuenta países. Nace de la fusión entre SNOMED RT desarrollada por el College of American Pathologists (CAP) y el Clinical Terms Version 3 (CTV3) desarrollada por el National Health Service (NHS) de Reino Unido.

En 2007 los derechos de propiedad intelectual fueron transferidos a la International Health Terminology Standards Development Organisation (IHTSDO) quien se encarga de su mantenimiento y distribución hoy en día.

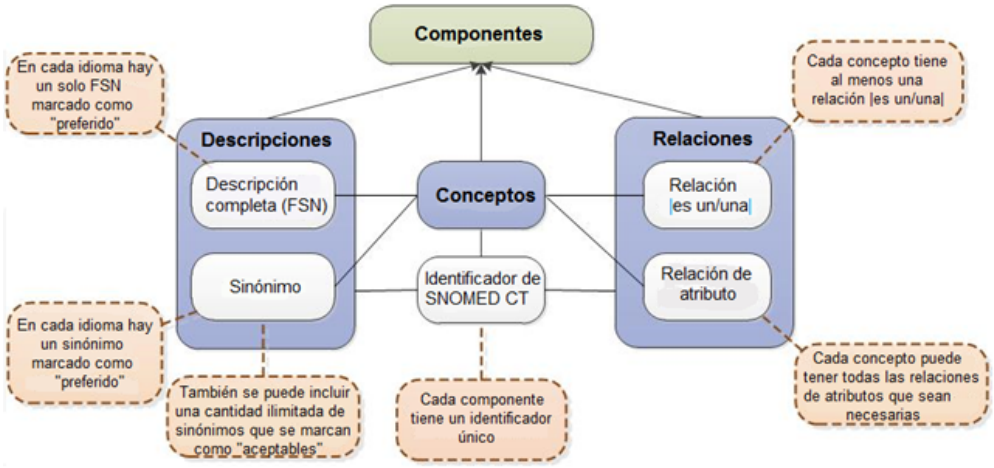

Figura 2.3: Modelo lógico SNOMED-CT. Fuente: (IHTSDO, 2018)

Los principales componentes de SNOMED-CT son:

- **Conceptos**: Los conceptos representan como su propio nombre indica conceptos o ideas clínicas. Cada uno tiene un identificador numérico único (ConceptID). Están organizados en *jerarquías* que van de lo general a lo específico a medida que se desciende en ellas.
- **Descripciones**: Un concepto puede tener más de una descripción asociada, cada una representando un sinónimo que describe la misma idea.
- **Relaciones**: Las relaciones permiten asociar a un concepto otros conceptos cuyo significado está relacionado. La relación más importante es la llamada "es un" que define la mencionada jerarquía entre conceptos. Otros ejemplos son "agente causal", "sitio de hallazgo" o "morfología asociada". Existen cuatro tipos de relaciones: definitorias, calificadoras, históricas y adicionales.

2.4. Análisis de aplicaciones

A continuación pasaremos a detallar las diferentes aplicaciones de anotación y proveedoras de terminología evaluadas para poder llevar a cabo nuestro objetivo. Para la evaluación de cada aplicación tendremos en cuenta los siguientes puntos:

- Identificación de términos médicos: concepto clínico, campo semántico, identificador unívoco UMLS, ...
- Integración con Python, por ser el principal lenguaje de programación utilizado durante el Máster.
- Posibilidad de analizar textos en español.
- Velocidad de procesamiento.

2.4.1. cTakes

cTakes[3] (clinical Text Analysis and Knowledge Extraction System), introducido en (Savova et al. 2010), es una herramienta muy conocida para la anotación semántica de documentos biomédicos en general, y particularmente para textos de investigación clínica. Está desarrollado mediante dos frameworks para PLN muy consolidados: UIMA[4] y OpenNLP[5]. Está desarrollado modularmente, formado por un conjunto de componentes de procesado de texto que aplican técnicas basadas en reglas y aprendizaje automático. cTakes reconoce conceptos biomédicos en textos y los relaciona con su identificador UMLS. Una de las carencias de cTakes es que no implementa ningún tipo de método para desambiguar entidades, pero, al ser una herramienta modular, permite la fácil integración del componente YTEX,

[3] http://ctakes.apache.org/
[4] https://uima.apache.org/
[5] https://opennlp.apache.org/

el cual está centrado en el análisis y procesamiento de texto biomédico con la capacidad de realizar la desambiguación entre conceptos UMLS.

La última versión disponible en el momento de la realización de las pruebas es la 4.0.0.1, dicha versión permite la integración con UMLS mediante API KEY (en lugar de usuario y contraseña como en versiones anteriores). cTakes está desarrollado en Java, por lo que para su ejecución resulta necesario la instalación de la máquina virtual[6]. Una vez descargado, instalado y definidas las variables de entorno necesarias se pueden comenzar las pruebas ejecutando el CAS Visual Debugger (CVD) y seleccionando el Analysis Engine (AE) "AggregatePlaintextFastUMLSProcessor.xml":

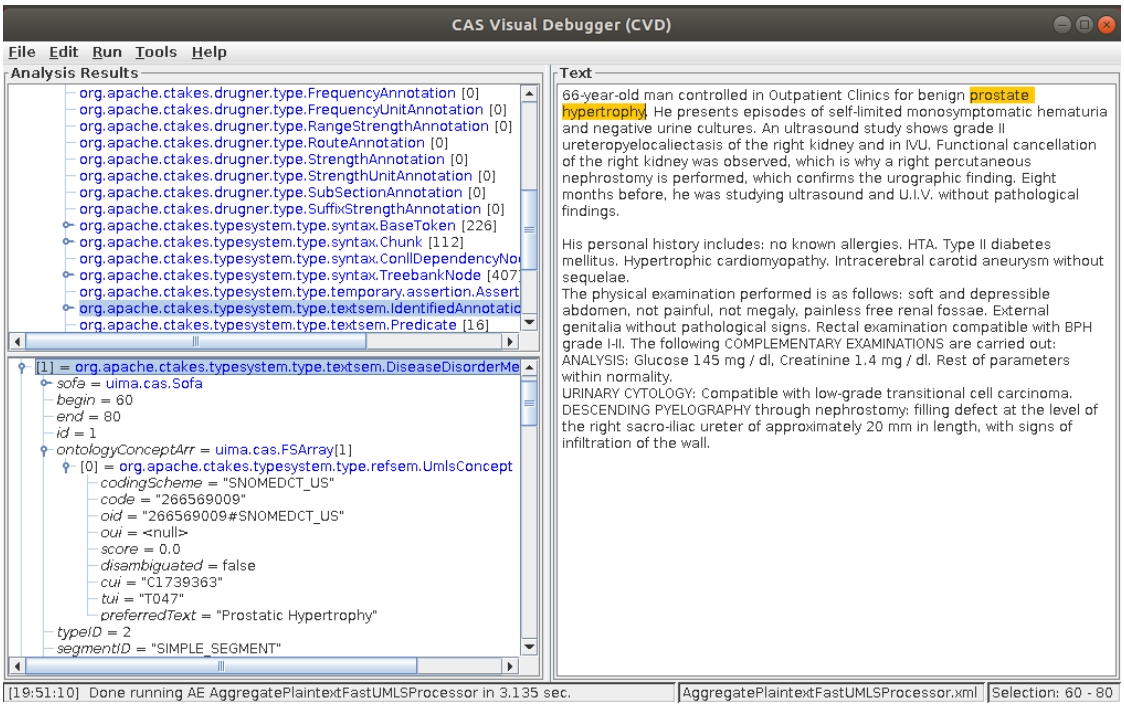

Figura 2.4: Ejecución del CVD de cTakes

Como se puede observar en 2.4 cTakes permite identificar de manera unívoca los conceptos SNOMED-CT US detectados en el texto, proporcionando además su Concept Unique Identifier (CUI). Respecto a su integración con Python no se ha encontrado ninguna librería disponible que permitiera parsear directamente los objetos resultantes para su manipulación, por lo que habría que desarrollar un pipeline que lance tareas a la herramienta y mapee posteriormente los ficheros resultantes. Respecto al idioma, cTakes no permite analizar de forma nativa los textos biomédicos en español aunque se han realizado algunas aproximaciones como en (Costumero et al. 2014). La velocidad de procesamiento es aceptable teniendo en cuenta el nivel de detalle del análisis obtenido como resultado.

[6] https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+4.0+User+Install+Guide

2.4.2. FreeLing

FreeLing[7] es una librería de código abierto para el procesamiento multilingüe automático, que proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas. FreeLing ofrece a los desarrolladores de aplicaciones de PLN, funciones de análisis y anotación lingüística de textos. FreeLing es altamente configurable, es decir, se pueden utilizar los recursos lingüísticos por defecto (diccionarios, lexicones, gramáticas, etc) o ampliarlos/adaptarlos a dominios particulares, o incluso desarrollar otros nuevos para idiomas específicos o necesidades especiales de las aplicaciones.

	as	ca	cy	en	es	gl	it	pt	ru
Tokenization	X	X	X	X	X	X	X	X	X
Sentence splitting	X	X	X	X	X	X	X	X	X
Number detection		X		X	X	X	X	X	X
Date detection		X		X	X	X		X	X
Morphological dictionary	X	X	X	X	X	X	X	X	X
Affix rules	X	X	X	X	X	X	X	X	
Multiword detection	X	X	X	X	X	X	X	X	
Basic named entity detection	X	X	X	X	X	X	X	X	X
B-I-O named entity detection				X	X	X			
Named Entity Classification				X	X				
Quantity detection		X		X	X	X		X	X
PoS tagging	X	X	X	X	X	X	X	X	X
WN sense annotation		X		X	X				
UKB sense disambiguation		X		X	X				
Shallow parsing	X	X		X	X	X		X	
Full/dependency parsing	X	X		X	X	X			
Coreference resolution					X				

Figura 2.5: Servicios de análisis disponibles para cada lengua. Fuente: (Padró et al. 2011)

FreeLing proporciona una API nativa para Python (Figura 2.6) la cual nos permite acceder a los servicios de análisis descritos en la tabla anterior para textos en español. Respecto a la velocidad de procesamiento los tiempos de respuesta son bastante bajos. El principal inconveniente detectado es que es una herramienta generalista, por lo que no detecta conceptos médicos ni tampoco los identifica de manera unívoca por lo que sería necesario la utilización de algún servidor de terminología adicional que realice dicha tarea, como por ejemplo SnowStorm.

[7] http://nlp.lsi.upc.edu/freeling

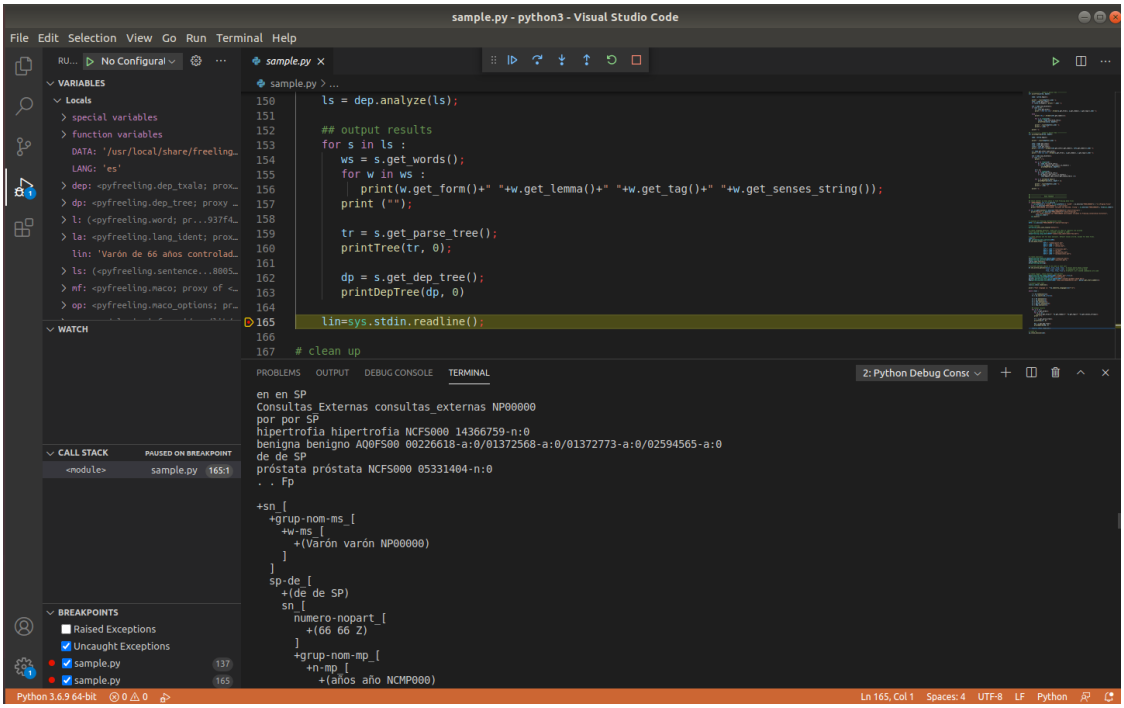

Figura 2.6: Depuración API Python de FreeLing

```
Varón varón NCMS000 1
de de SP 0.999961
66 66 Z 1
años año NCMP000 1
controlado controlar VMP00SM 1
en en SP 1
Consultas_Externas consultas_externas NP00G00 1
por por SP 1
hipertrofia hipertrofia NCFS000 0.874633
benigna benigno AQ0FS00 1
de de SP 0.999961
próstata próstata NCFS000 1
. . Fp 1

Presenta presentar VMIP3S0 0.995868
episodios episodio NCMP000 1
de de SP 0.999961
hematuria hematuria NCFS000 1
monosintomática monosintomático AQ0FS00 1
autolimitada autolimitar VMP00SF 1
y y CC 0.999989
urinocultivos urinocultivos NCMP000 1
negativos negativo AQ0MP00 0.983871
. . Fp 1

En en SP 1
estudio estudio NCMS000 0.97043
ecográfico eco_gráfico NCMS000 1
se se P00CN00 0.494509
observa observar VMIP3S0 0.989241
ureteropielocaliectasia ureteropielocaliectasia AQ0FS00 0.249948
grado grado NCMS000 0.994792
II ii NP00V00 1
de de SP 0.999961
riñón riñón NCMS000 1
derecho derecho AQ0MS00 0.117834
y y CC 0.999989
en en SP 1
U.I.V. u.i.v. NP00V00 1
se se P00CN00 0.494509
objetiva objetivar VMIP3S0 0.188737
anulación anulación NCFS000 1
funcional funcional AQ0CS00 1
```

Figura 2.7: Ejemplo de salida de fichero FreeLing

2.4.3. CUTEXT

CUTEXT[8] (Cvalue Used To EXtract Terms) es una aplicación de extracción de términos médicos multilingüe (Santamaría & Krallinger, 2018). Permite extraer términos en textos redactados en inglés, castellano, gallego y catalán.

Las principales características de CUTEXT son las siguientes:

- Está implementado en Java, por lo que es multiplataforma.
- Es multilingüe: ha sido probado en inglés, español, catalán y gallego, y se puede adaptar fácilmente a otros idiomas simplemente cambiando la configuración del archivo de texto de la etiqueta léxica.
- Los documentos de entrada pueden estar en texto plano o en pdf.
- Se puede ejecutar en modo gráfico o por consola (línea de comandos).
- Soporta numerosos parámetros de configuración, entre los más importantes: el idioma, el etiquetador, los umbrales de frecuencia y valor c y la entrada de documentos.
- La salida se proporciona en texto plano, en formato JSON[9] y/o en BioC[10].


```
{
	"terms_json":
	[
		{
			"term": "aorta",
			"frecuency": "1",
			"c-value": "1.0"
		}
		{
			"term": "aneurisma de aorta",
			"frecuency": "1",
			"c-value": "2.584962500721156"
		}
		{
			"term": "pieza",
			"frecuency": "1",
			"c-value": "1.0"
		}
		{
			"term": "signos",
			"frecuency": "2",
			"c-value": "2.0"
		}
		{
			"term": "riñón derecho",
			"frecuency": "3",
			"c-value": "6.0"
		}
		{
			"term": "diabetes",
			"frecuency": "1",
			"c-value": "1.0"
		}
```


Figura 2.8: Ejemplo de salida de fichero CUTEXT

[8] https://github.com/PlanTL-SANIDAD/CUTEXT
[9] https://www.json.org/
[10] http://bioc.sourceforge.net/

Señalar que para la ejecución de CUTEXT es necesario la instalación de la aplicación de anotación de texto TreeTagger[11] si se desea etiquetar textos en español. A diferencia de FreeLing, CUTEXT sí identifica conceptos médicos, pero no los identifica unívocamente por lo que sigue siendo necesario la utilización de un servidor de terminología adicional. Respecto al rendimiento se obtienen los resultados casi de manera inmediata y dado que el resultado es un fichero en formato JSON puede ser fácilmente integrable con Python.

2.4.4. SnowStorm

SnowStorm[12] es un servidor de terminología de SNOMED-CT construido sobre Elasticsearch[13], enfocado en el rendimiento y la escalabilidad. Proporciona una API del servidor de terminología para SNOMED International Browser, incluida la Edición Internacional y alrededor de catorce extensiones nacionales. Snowstorm se puede utilizar en implementaciones locales para consultar SNOMED CT con las siguientes características:

- Alojar múltiples extensiones junto con la Edición Internacional de SNOMED-CT.
- Búsqueda multilingüe y recuperación de contenido.
- Totalmente compatible con ECL v1.3
- Historial completo (depende de la importación RF2 completa).
- API FHIR de solo lectura.

Una vez desplegada la infraestructura Docker necesaria para las pruebas del servidor e importada la terminología SNOMED-CT ES se pueden realizar consultas de términos en español mediante su API REST, devolviendo una lista de conceptos identificados por su "conceptId".

[11] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[12] https://github.com/IHTSDO/snowstorm
[13] https://www.elastic.co/

```
{
  "items" : [ {
    "term" : "colesterol",
    "active" : true,
    "languageCode" : "es",
    "module" : "450829007",
    "concept" : {
      "conceptId" : "84698008",
      "active" : true,
      "definitionStatus" : "PRIMITIVE",
      "moduleId" : "900000000000207008",
      "fsn" : {
        "term" : "colesterol (sustancia)",
        "lang" : "es"
      },
      "pt" : {
        "term" : "colesterol",
        "lang" : "es"
      },
      "id" : "84698008"
    },
    "branchPath" : "MAIN/SNOMEDCT-ES"
  }, {
    "term" : "colesterol LDI",
    "active" : true,
    "languageCode" : "es",
    "module" : "450829007",
    "concept" : {
      "conceptId" : "259569005",
      "active" : true,
      "definitionStatus" : "PRIMITIVE",
      "moduleId" : "900000000000207008",
      "fsn" : {
        "term" : "colesterol de las lipoproteínas de densidad intermedia (sustancia)",
        "lang" : "es"
      },
```


Figura 2.9: Resultado de la consulta del término "colesterol" a SnowStorm

Dicha API es fácilmente accesible desde Python. El inconveniente detectado es el rendimiento, ya que las pruebas han sido realizadas en local desde un PC de uso doméstico.

2.4.5. MetaMap

MetaMap[14] (Aronson, 2001) es un software desarrollado por el Dr. Alan Aronson en la National Library of Medicine (NLM) de Estados Unidos que cuenta con múltiples opciones de configuración y permite mapear o asociar términos que aparecen en un texto biomédico. Los términos se asocian con conceptos del Metatesauro UMLS.

Se utiliza un enfoque basado en el PLN y en técnicas lingüísticas. Además de ser aplicado tanto para recuperación de información (IR) y aplicaciones de minería de datos es uno de

[14] https://metamap.nlm.nih.gov/

los fundamentos del Medical Text Indexer (MTI) del NLM el cual se utiliza para indexar de manera semiautomática o automática literatura del NLM.

MetaMap ofrece diversas funcionalidades, de las cuales, las más importantes son (Pastore Burgos & Díaz Esteban, 2015):

- Desambiguación: Uno de los problemas principales del PLN es la ambigüedad del lenguaje, y una de las mayores debilidades de MetaMap es su incapacidad para resolver la ambigüedad del Metatesauro en las situaciones en la que dos o más conceptos comparten un sinónimo. Para solucionar este problema se incluyó un sistema de desambiguación (Word Sense disambiguation (WSD)) activable mediante la *opción -y*. De esta forma se favorecen las alternativas que tienen un tipo semántico más probable en función del contexto.

- Detección de negaciones: MetaMap es capaz de detectar cuando un concepto está negado mediante el uso de una versión extendida del algoritmo NegEx. Para que sea legible (human- readable), es necesario usar la *opción -negex*.

- Detección de acrónimos y abreviaturas definidos por el autor: En los documentos técnicos aparecen con frecuencia acrónimos y abreviaturas (Acronyms and Abbreviations (AA)) y suelen ir acompañados de definiciones o extensiones. Interesa que después de que un acrónimo o sigla haya sido definido, se asigne la misma definición en futuras apariciones. El algoritmo trata de asociar la expansión del AA, con la sigla o acrónimo, que deberá estar escrito entre paréntesis y situado después de la expansión. Existen ciertas reglas que es necesario cumplir para evitar errores:

  - Los AA no pueden contener más de 20 caracteres.
  - Las expansiones deben ser mayores que los AA correspondientes.
  - Una expansión no puede contener texto entre paréntesis.
  - Cada palabra considerada como AA debe contener como máximo 12 caracteres.
  - Los AA no pueden comenzar por “such”, “also” o “including”.

- Conceptos: Pero por supuesto la parte esencial de MetaMap es aquella que define el mapeo de términos de un texto biomédico a conceptos del Metatesauro UMLS. El algoritmo que sigue para llevar a cabo el proceso de mapeo de términos de un texto biomédico a conceptos del Metatesauro UMLS engloba diferentes **fases** que son detalladas por el creador, el Dr. Aronson:

  1. Parsing: Haciendo uso del Léxico Especializado realiza un primer análisis sintáctico siendo capaz de detectar diferentes elementos textuales, como la palabra principal de una frase.
  2. Variant Generator: Para cada frase se genera una variante utilizando de nuevo el Léxico Especializado y en esta ocasión, de manera complementaria, una base de datos de sinónimos. Dicha variante consistirá en un sintagma nominal junto con variantes ortográficas, sinónimos, acrónimos o abreviaturas entre otros.

3. Candidate Retrieval: Formar el conjunto candidato de todas las cadenas del Metatesauro que contienen al menos una de las variantes.
4. Candidate Evaluation: Asignar a cada candidato un sintagma nominal, evaluarlo o puntuarlo y ordenar los candidatos por puntuación.
5. Mapping Construction: Combinar los candidatos que están involucrados en partes disjuntas del sintagma nominal. Calcular de nuevo la puntuación y seleccionar en base a dicha puntuación.

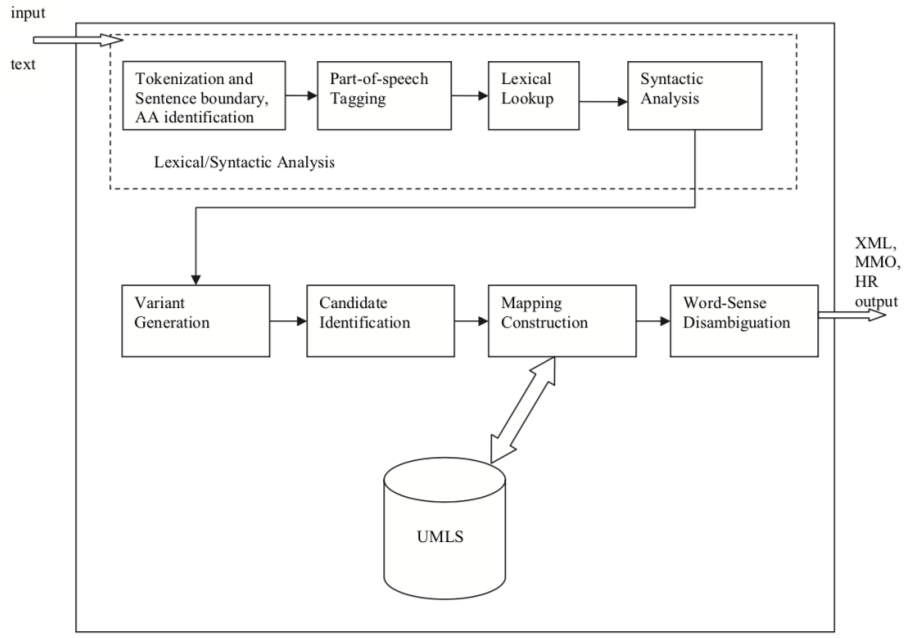

Figura 2.10: Diagrama del sistema MetaMap. Fuente: (Aronson & Lang, 2010)

MetaMap puede generar archivos con diferentes formatos de salida:

- Human-Readable: formato de salida por defecto. Muestra para cada frase del texto de entrada la propia frase, una lista de conceptos candidatos del Metatesauro asociados a cada parte de la frase, el mapeo de realizar las combinaciones de candidatos asociados a partes disjuntas y datos complementarios como la puntuación otorgada, u otros opcionales como el CUI del concepto (*-I*), los tipos semánticos (*-s*), o las fuentes (*-G*).
- MetaMap Matching Output (MMO): incluye un super conjunto de la información de salida Human-Readable y tiene formato de términos Prolog[15]. Permite realizar un post procesamiento por parte de aplicaciones Prolog. *Opción -q.*
- XML: también se puede obtener la salida MMO en formato XML. Para esto se puede usar una las siguientes *opciones*: *-XMLf*, *-XMLf1*, *-XMLn* o *-XMLn1*. La *f* indica

[15] https://en.wikipedia.org/wiki/Prolog_syntax_and_semantics

que el .xml tendrá formato y la *n* que no. El *1* indica que se generará un .xml para un archivo de entrada, o, si no se incluye, uno por cada entrada citada. Tiene el inconveniente del peso añadido de los archivos en el disco.

- Colorized MetaMap Output (MetaMap 3D): este formato proporciona información mediante colores para los conceptos mapeados por MetaMap en un texto.
- Fielded MetaMap Indexing (MMI) Output: Mismo contenido que la salida MMO, pero representado en la opción múltiples líneas que contienen campos delimitados por tabulaciones. *Opción -f*. Este tipo de salida es utilizado principalmente por el Medical Text Indexer (MTI).

Según la documentación[16] de MetaMap: "Todas las versiones de datos de MetaMap (Base, USAbase y NLM) incluyen sólo las cadenas en inglés de las fuentes UMLS. Los usuarios que necesiten incluir fuentes en idiomas distintos del inglés para procesar texto no inglés deben crear su propia base de datos tipo UMLS utilizando nuestra suite Data File Builder[17]". Hemos intentado cargar la edición en español de SNOMED-CT (SCTSPA) siguiendo el procedimiento descrito en (Rogers & Gay, 2015) para que detectara términos en español pero el proceso fallaba antes de finalizar, suponemos que por tratarse de una versión sin cambios desde el 2016.

MetaMapLite

En lo que refiere a la velocidad de ejecución y procesamiento se ha evaluado una versión más ligera de MetaMap denominada MetaMapLite, cuyo principal objetivo es proporcionar un **reconocedor de entidades con nombre** (NER) casi en tiempo real que no es tan riguroso como MetaMap pero que es mucho más rápido (Demner-Fushman et al. 2017). MetaMapLite utiliza algunas de las tablas desarrolladas originalmente para MetaMap. Actualmente, MetaMapLite no soporta la generación dinámica de variantes. Las entidades nombradas se encuentran utilizando la coincidencia más larga. La restricción por fuente UMLS y tipo semántico es opcional. El etiquetado de parte de la palabra (POS), que mejora la precisión en una pequeña cantidad (a costa de la velocidad), también es opcional. La detección de la negación está disponible utilizando el contexto de Wendy Chapman (Chapman et al. 2001) o un algoritmo nativo de detección de la negación basado en NegEx de Wendy Chapman, que es algo menos eficaz, pero más rápido. Respecto a la integración con Python se ha probado satisfactoriamente la librería *pymetamp*[18] y permite interactuar tanto con MetaMap, como con MetaMapLite.

[16] https://metamap.nlm.nih.gov/Docs/FAQ/DataVersions.pdf
[17] https://metamap.nlm.nih.gov/DataFileBuilder.shtml
[18] https://github.com/AnthonyMRios/pymetamap

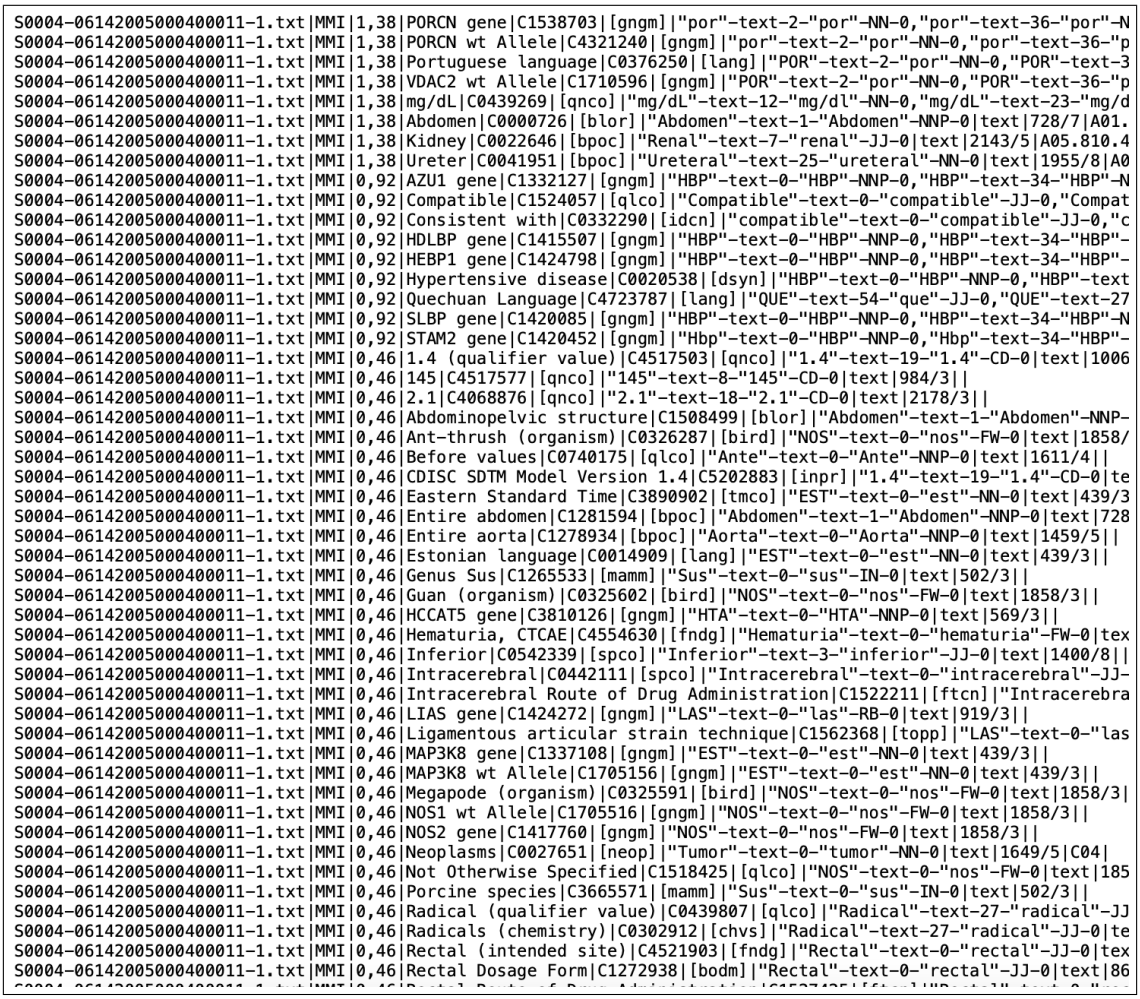

Figura 2.11: Ejemplo de salida de fichero MetaMap

2.4.6. Stanza

Stanza[19] es un paquete de análisis del lenguaje natural en Python (Qi et al. 2020). Contiene herramientas, que se pueden usar en una tubería, para convertir una cadena que contiene texto en lenguaje humano en listas de oraciones y palabras, para generar formas base de esas palabras (lematización), sus partes del discurso (POS) y características morfológicas, dar una estructura para el análisis de dependencia sintáctica y reconocer entidades nombradas (NER). El conjunto de herramientas está diseñado para ser paralelo entre más de 70 idiomas, utilizando el formalismo de dependencias universales[20].

Stanza está construido con componentes de redes neuronales de alta precisión que también permiten su entrenamiento y evaluación eficientes con sus propios datos anotados. Los

[19] https://stanfordnlp.github.io/stanza/
[20] https://universaldependencies.org/

módulos están construidos sobre la biblioteca PyTorch[21], permitiendo la ejecución sobre GPU.

Además, Stanza incluye una interfaz Python para el paquete Java CoreNLP[22] y hereda la funcionalidad adicional de éste, como el análisis de constituyentes, resolución de correferencia y coincidencia de patrones lingüísticos.

Stanza también cuenta con un analizador sintáctico y NER para textos clínicos y biomédicos, pero actualmente sólo esta disponible para textos en inglés (Zhang et al. 2021).

```
benign prostate hypertrophy       PROBLEM
self-limited monosymptomatic hematuria     PROBLEM
urine cultures    TEST
An ultrasound study       TEST
grade II ureteropyelocaliectasis of the right kidney       PROBLEM
IVU      TEST
a right percutaneous nephrostomy  TREATMENT
studying ultrasound       TEST
pathological findings     PROBLEM
known allergies  PROBLEM
Type II diabetes mellitus PROBLEM
Hypertrophic cardiomyopathy       PROBLEM
Intracerebral carotid aneurysm    PROBLEM
sequelae PROBLEM
The physical examination  TEST
soft and depressible abdomen      PROBLEM
painful  PROBLEM
megaly   PROBLEM
painless free renal fossae        PROBLEM
pathological signs        PROBLEM
Rectal examination        TEST
BPH      PROBLEM
Glucose  TEST
Creatinine       TEST
URINARY CYTOLOGY TEST
low-grade transitional cell carcinoma      PROBLEM
DESCENDING PYELOGRAPHY    TEST
Nephrostomy      TREATMENT
filling defect   PROBLEM
infiltration of the wall  PROBLEM
Creatinine clearance      TEST
Infrarenal Aortic aneurysm        PROBLEM
Hydronephrotic atrophy of the right kidney PROBLEM
visualizing intraureteral endoluminal images       TEST
Doubtful paravesical lymphadenopathy       PROBLEM
a tumor of the right ureteric tract        PROBLEM
surgery  TREATMENT
a right radical nephroureterectomy         TREATMENT
the specimen     TEST
chronic pyelonephritis    PROBLEM
changes in arteriosclerosis       PROBLEM
an inflammatory ureteral lesion   PROBLEM
infiltrate of lymphocytes in muscular layers       PROBLEM
erosion of the urothelium PROBLEM
Actinomyces      PROBLEM
serum creatinine TEST
```

Figura 2.12: Ejemplo de salida de fichero Stanza

[21] https://pytorch.org/
[22] https://stanfordnlp.github.io/CoreNLP

2.4.7. Resumen

A modo de resumen se presenta la tabla 2.1 con las características que nos interesan de cada una de las aplicaciones analizadas.

Aplicación	Texto en español	Identificación de términos médicos	Integración con Python
cTakes	No	CUIs	No
FreeLing	Sí	No	Sí, librería nativa
CUTEXT	Sí	Conceptos	Sí, mediante JSON
SnowStorm	Sí	CUIs a partir de conceptos	Sí, mediante API
MetaMapLite	No	CUIs	Sí, librería pymetamap
Stanza	No	Conceptos	Sí, librería nativa

Tabla 2.1: Resumen de las aplicaciones analizadas

Capítulo 3

Diseño del método

3.1. Fuentes de datos

3.1.1. Texto clínico

Las fuentes de datos de texto clínico utilizadas son dos:

- Spanish Clinical Case Corpus (SPACCC)[1], una colección de 1.000 casos clínicos españoles de SciELO (Intxaurrondo, 2018).
- Dataset de prueba generado en el Centro de Medicina Integral del Comahue de Neuquén, Argentina. Este dataset ha sido proporcionado en formato XLS y se ha tenido que realizar una tarea de preprocesamiento previa en Python para su conversión a formato JSON.

3.1.2. MetaMapLite

Tras el estudio realizado en el Capítulo 2 se ha decidido utilizar MetaMapLite[2] (2.4.5) como analizador de texto en lenguaje natural para la extracción de conceptos médicos por su eficacia y velocidad. La última versión disponible del software en el momento de la realización de este trabajo es la 3.6.2rc6 y la versión del Dataset UMLS utilizado es "2020AA UMLS Level 0+4+9 Dataset".

Respecto a la librería para el acceso desde Python, la versión utilizada de *pymetamap*[3] es la 0.2.

[1] http://doi.org/10.5281/zenodo.2560316
[2] https://metamap.nlm.nih.gov/MetaMapLite.shtml
[3] https://github.com/AnthonyMRios/pymetamap

3.1.3. Google Cloud Translation API

Debido a que MetaMapLite analiza textos en inglés se evaluaron las herramientas para la traducción de texto del español al inglés Google Cloud Translation API[4],basado en Redes Neuronales Recurrentes (RNN), y DeepL API[5], basado en Redes Neuronales de Convolución (CNN), eligiendo la primera opción por su alta precisión en la traducción de textos médicos, siempre y cuando no se cometan errores ortográficos o gramaticales (Khoong et al. 2019), además de por su facilidad de integración con Python mediante la librería *google-cloud-translate* en su versión 3.1.0.

3.1.4. UMLS REST API

Para la obtención de la información relativa a SNOMED-CT US y SCTSPA a partir de los conceptos UMLS CUI devueltos por MetaMapLite se ha utilizado la UMLS REST API[6] por lo que ha sido necesario registrarse para el acceso a dicho servicio e implementar los métodos necesarios para la consulta de la API:

- /content/{version}/CUI/{CUI}/atoms: Recupera átomos e información sobre átomos para un CUI conocido. Nos permite conocer los ID de SNOMED-CT US asociados a cada UMLS CUI devuelto por MetaMapLite.
- /content/{version}/source/{source}/{id}: Recupera información sobre un identificador conocido. Nos permite obtener información del vocabulario SCTSPA a partir del ID (coincide con el ID de SNOMED-CT US).
- /content/{version}/source/{source}/{id}/ancestors: Recupera todos los antepasados de un identificador conocido. Nos permite obtener la jerarquía del ID que nos servirá para la detección de eventos adversos.

Además de los métodos descritos anteriormente también se han implementado los métodos necesarios para la autenticación:

- /cas/v1/api-key: Recupera un Ticket Granting Ticket (TGT), válido durante 8h., a partir de una API KEY asociada a la cuenta de registro.
- /cas/v1/tickets/{TGT}: Recupera un ticket de servicio de un solo uso para cada operación.

3.1.5. Abreviaturas UMLS

Para aportar más información al resultado devuelto por nuestra herramienta, se ha decodificado las abreviaturas que utiliza UMLS relativas a los tipos de términos[7] y tipos

[4] https://cloud.google.com/translate
[5] https://www.deepl.com/docs-api
[6] https://documentation.uts.nlm.nih.gov/rest/home.html
[7] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html

semánticos[8].

3.1.6. Catálogo de triggers

Como indicamos en el Capítulo 1 utilizaremos el concepto de trigger para detectar eventos adversos en el texto analizado. Para ello necesitamos codificar en SNOMED-CT de forma manual el catálogo de triggers presentados en la "Tabla 1 Listado de triggers" de (Guzmán-Ruiz et al. 2015) haciendo uso del SNOMED-CT Browser[9] en su edición en Español para obtener su SCTID cuyo resultado podemos ver en las Tablas A.1, A.2, A.3 y A.4 del Apéndice A.

3.2. Diseño e implementación

La herramienta diseñada consta de los siguientes módulos desarrollados en Python:

- *api.py*: implementación de un servicio API RESTful utilizando el framework FastAPI[10] para poder lanzar tareas de detección de eventos adversos a la herramienta desde cualquier entorno web.
- *worker.py*: implementación de un gestor de tareas de análisis encargado de procesar las peticiones procedentes de *api.py*, preprocesarlas y enviarlas al detector.
- *trigger*: librería principal con el código del detector de eventos adversos.

3.2.1. API RESTful

El servicio API RESTful se ha diseñado teniendo en cuenta un sistema de colas de los tres posibles estados (Figura 3.2) por los que puede ir pasando una tarea de análisis de texto (solución propuesta ante la imposibilidad de poder realizar la detección de manera síncrona debido al tiempo requerido para realizar dicho proceso):

1. **Estado pendiente (pending)**: estado inicial de una tarea recién creada.
2. **Estado procesando (processing)**: estado intermedio durante el cual el detector esta ejecutando el análisis sobre la tarea.
3. **Estado completado (completed)**: estado final con el resultado de la tarea de detección de eventos adversos.

Para ello se han implementado los métodos HTTP:

[8] https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml
[9] https://browser.ihtsdotools.org/
[10] https://fastapi.tiangolo.com/

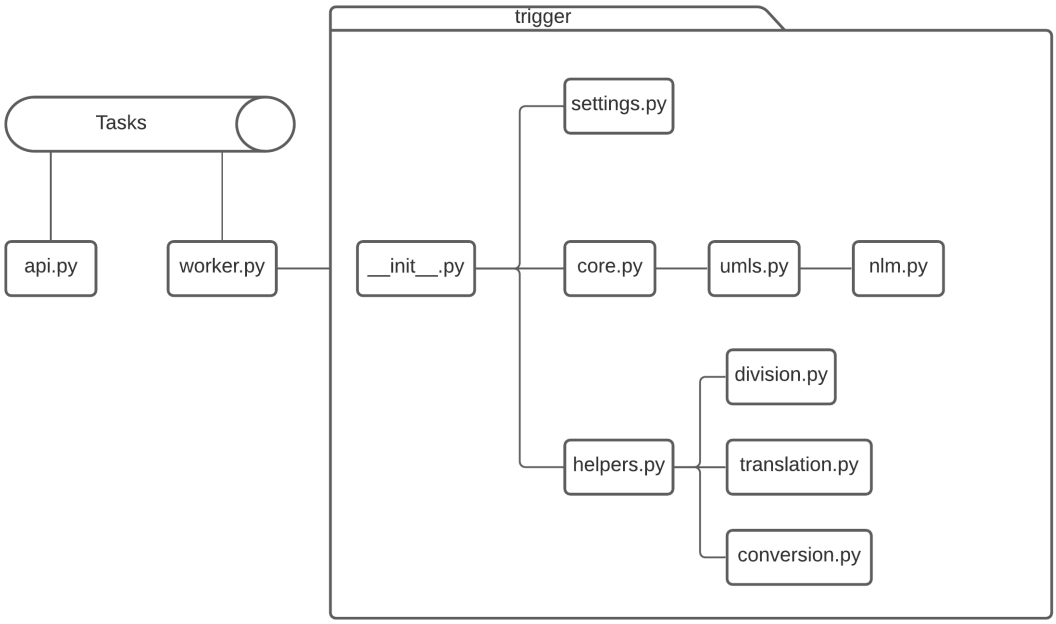

Figura 3.1: Estructura del proyecto

- POST: crea una nueva tarea de detección y lo añade a la cola de tareas pendientes.
- GET: devuelve una lista de todas las tareas en cualquier estado.
- DELETE: elimina todas la tareas en cualquier estado.
- PUT {id}: modifica la tarea con identificador {id} en estado pendiente.
- GET {id}: devuelve la información relativa a la tarea con identificador {id}.
- DELETE {id}: elimina la tarea con identificador {id}.

La documentación de la API ha sido generada de manera automática en formato Swagger[11] (Figura 3.3) y ReDoc gracias al uso del framework FastAPI, la cual nos permite ejecutar los métodos implementados directamente sobre un navegador web.

3.2.2. Worker

El Worker se encarga de mover las tareas entre las distintas colas de estados, realiza un preprocesamiento previo del texto (limpieza de espacios en blanco innecesarios) y la división de los párrafos en subtareas que pasa al detector de eventos adversos que, junto al total de palabras del texto, nos permite obtener un porcentaje del avance del proceso completo.

[11] https://swagger.io/

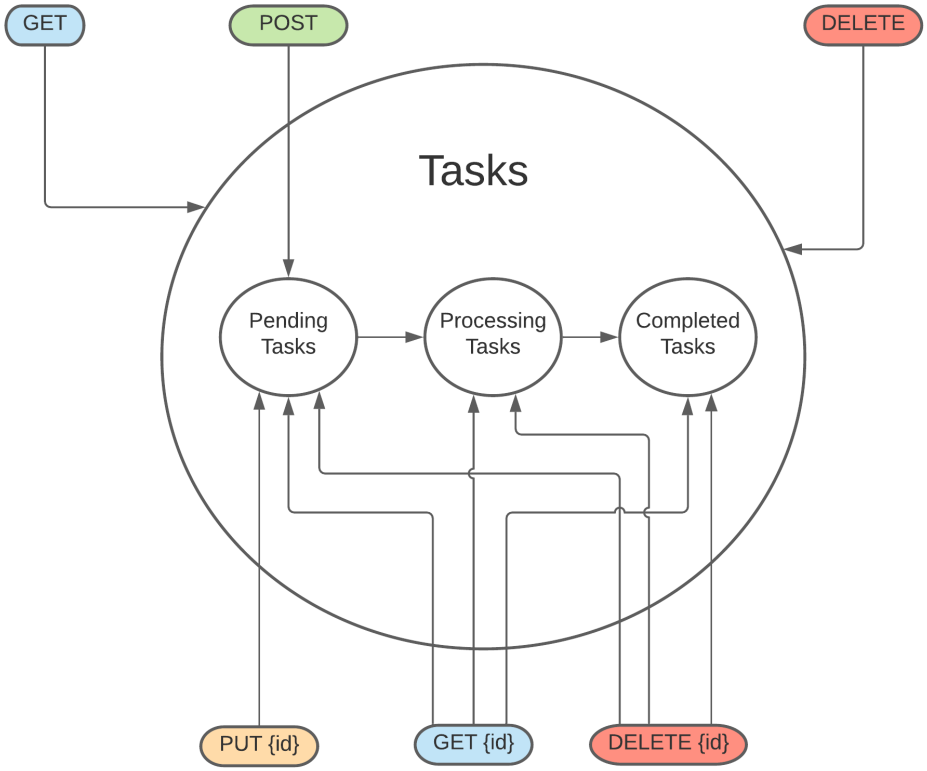

Figura 3.2: Diseño del servicio API RESTful basado en estados de una tarea de análisis

3.2.3. Trigger

Como muestra la Figura 3.1, la librería Trigger se subdivide en los módulos:

- *settings.py*: en este módulo se cargan las variables de entorno necesarias por la herramienta: rutas locales de ficheros, URIS de acceso a servicios externos, API KEY de servicios externos, etc.
- *core.py*: en este módulo se realizan los pasos necesarios de los que consta el proceso de análisis:
  1. División del texto de entrada en oraciones (módulo *division.py*), usando para ello la librería de NLP *stanza*[12] (Qi et al. 2020).
  2. Traducción de las oraciones del español al inglés (módulo *translation.py*), usando para ello la librería *google-cloud-translate*.
  3. Envío de las oraciones traducidas al detector (módulo *umls.py*). El módulo *umls.py* contiene el **algoritmo principal** de la herramienta, el cual obtiene todos los

[12] https://stanfordnlp.github.io/stanza/

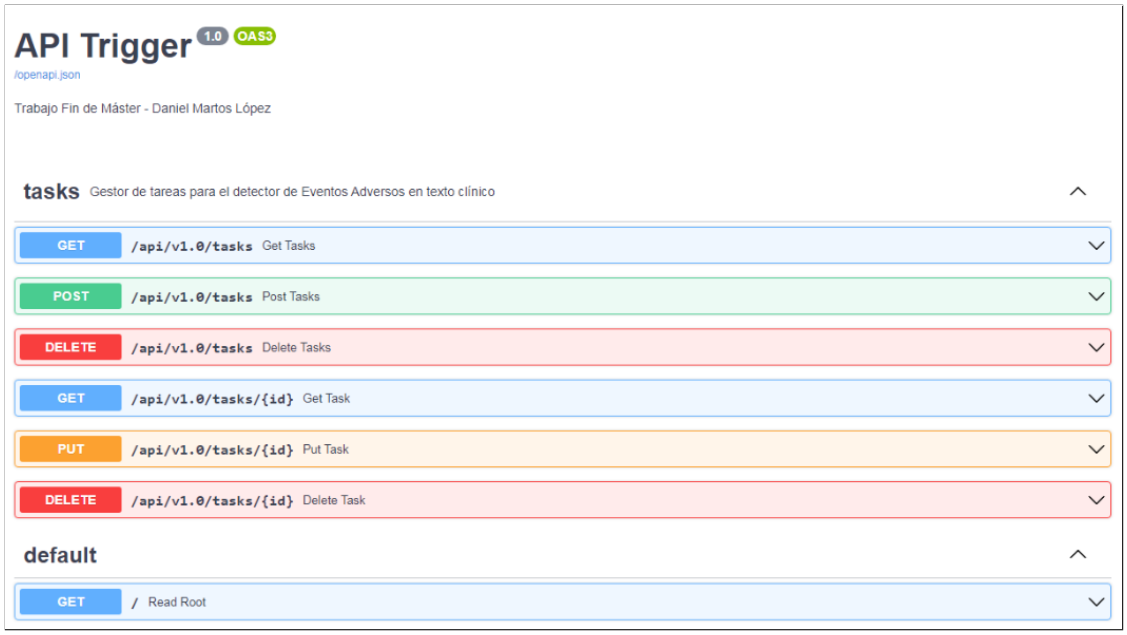

Figura 3.3: Documentación en formato Swagger del servicio API RESTful

conceptos UMLS del texto traducido usando para ello la librería *pymetamap* y la herramienta MetaMapLite, para cada concepto obtiene sus ancestros en la jerarquía de SNOMED-CT (módulo *nlm.py*), usando para ello la UMLS REST API, y los cruza con la lista de triggers previamente definida; en caso de existir alguna coincidencia se considera que se ha detectado un posible evento adverso (Figura 3.4).

- *helpers.py*: este módulo contiene otros módulos secundarios utilizados por el módulo principal (*core.py*) cuya funcionalidad ya se comentó anteriormente (*division.py* y *translation.py*), además de un módulo *conversion.py* que se utilizó para convertir el dataset de prueba generado en el Centro de Medicina Integral del Comahue de Neuquén de XLS a formato JSON, usando para ello la librería *xlrd*.

Optimización

Para reducir el número de consultas a la UMLS REST API se decidió cachear los siguientes objetos:

- CUI: lista de códigos de SNOMED-CT asociados a un concepto de UMLS.
- UI: información en español SCTSPA de un código SNOMED-CT.
- ANCESTORS: lista de ancestros asociada a un código SNOMED-CT.

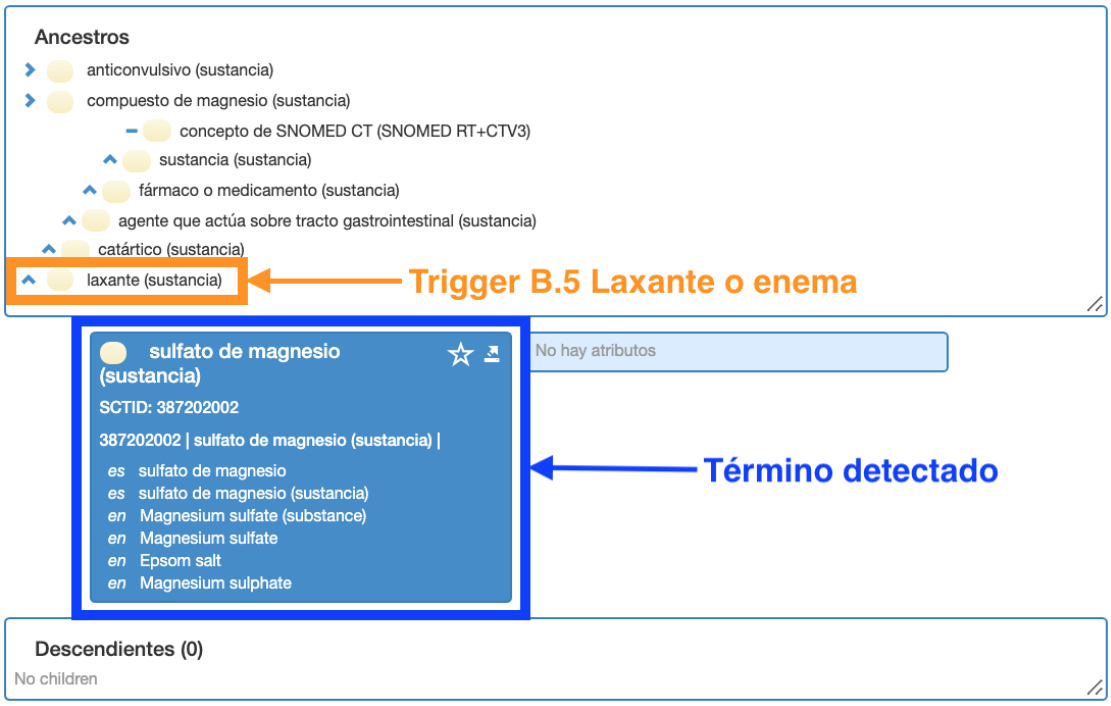

Figura 3.4: Trigger "B5. Laxante o enema" a partir de los ancestros del término detectado "sulfato de magnesio"

3.3. Despliegue

Para simplificar y automatizar el despliegue de nuestro método Trigger hemos utilizado contenedores de software basados en la tecnología Docker[13] que proporciona una capa adicional de abstracción y automatización de virtualización, además de facilitar el despliegue en la nube. Se han creado dos contenedores compartidos a través de la herramienta Docker Compose[14] (Apéndice B):

- api: a partir de la imagen base *tiangolo/uvicorn-gunicorn-fastapi:python3.7* se despliega el servicio API RESTful comentado en la Sección 3.2.1.
- trigger: a partir de la imagen base *ubuntu:bionic* se realiza toda la instalación y configuración necesaria para el despliegue del módulo Worker (Sección 3.2.2) y la librería principal Trigger (Sección 3.2.3).

[13] https://docs.docker.com/reference/
[14] https://docs.docker.com/compose/

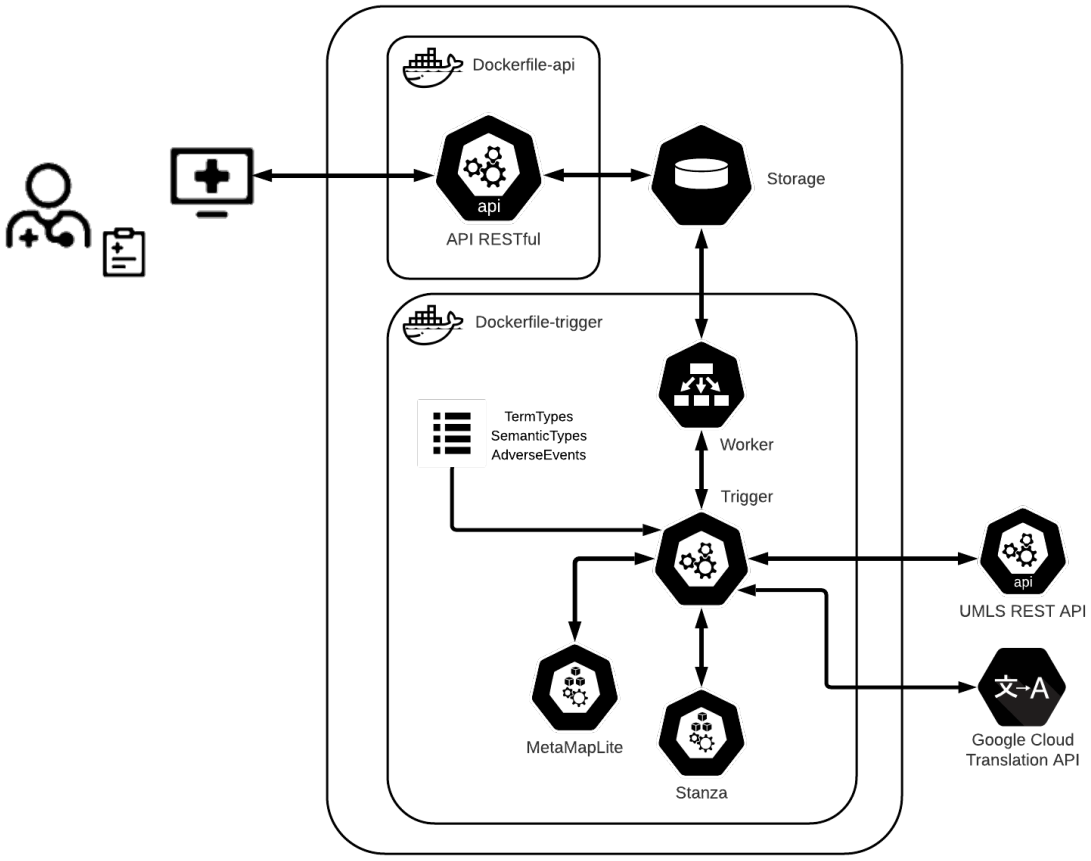

Figura 3.5: Arquitectura completa del método Trigger

3.4. Resultados

Se puede consultar en el Apéndice C el resultado de procesar el texto clínico de la Figura 3.6 con nuestro método. En el texto de ejemplo se ha detectado, entre otros, el efecto adverso "Estreñimiento por inmovilidad" a partir del trigger "B.5 Laxante o enema" derivado del concepto "sulfato de magnesio" (Figura 3.1).

Como no se ha encontrado un gold standard de casos clínicos en español anotados con eventos adversos necesario para obtener métricas de evaluación y comparación con otros sistemas (Dalianis, 2018) se ha desplegado la herramienta [15] para su evaluación de forma manual por parte de expertos del Centro de Medicina Integral del Comahue de Neuquén (Argentina).

[15] http://cmic.grupocemico.com.ar:8194

*Varón de 44 años con enfermedad de Crohn diagnosticado 15 años antes, que ingresa en el servicio de Cirugía General por un cuadro de dolor abdominal y diarrea de una semana de evolución, compatible con un brote de su enfermedad. En tratamiento domiciliario con corticoides y mesalazina, el paciente es portador de una ileostomía desde hace 2 años tras realizarle una colectomía subtotal, con muñón rectal cerrado a raíz de un brote de su enfermedad. Le es realizada una bioquímica general en la que destaca un calcio corregido por la albúmina de 8,42 mg/dl (rn: 8,7-10,6), K 2,6 mg/dl (rn: 3,6-4,9), albúmina 2.5 mg/dl (rn 4-5,2), glucosa y sodio dentro de los valores de referencia. Posteriormente se le realiza un **TAC** craneal en el que no se aprecian anormalidades. Presenta un electrocardiograma del día del ingreso con un ritmo sinusal normal a 79 lpm, sin alargamiento del espacio QT ni PR. A los dos días presenta de nuevo una convulsión tónico-clónica, generalizada de un minuto de duración. Tras ser valorado por el Servicio de Neurología se le realiza un electroencefalograma que resulta ser un trazado sin hallazgos patológicos y se inicia tratamiento con carbamacepina. Ante la persistencia de las convulsiones de las mismas características a pesar del tratamiento médico, se le realiza un registro electroencefalográfico de 24 horas de duración en el que no se evidencian alteraciones. Dado que el paciente presenta datos de deshidratación y desnutrición, así como disminución de la ingesta y astenia, realizan una interconsulta a la Unidad de Nutrición, para valoración de soporte nutricional en caso necesario.*

*Ante la existencia de deshidratación con balances hídricos negativos debido a un elevado débito de la ileostomía, y cifras bajas de electrolitos en sangre se inicia tratamiento monitorizado por vía parenteral, para reposición de volumen, electrolitos, y otros micronutrientes, entre ellos fósforo y magnesio. Previamente se realizó una extracción de sangre para la determinación de dichos micronutrientes ante la sospecha de un posible déficit. Ese mismo día el paciente sufre una convulsión similar a las previas. En la analítica realizada presenta un magnesio de 0,76 mg/dl (rn: 2,40-5,40) con un fósforo, calcio, potasio y sodio dentro de los valores de referencia. Se pauta una perfusión con altas dosis de **sulfato de magnesio** con la progresiva normalización de sus niveles en sangre, siendo suficiente para el tratamiento de mantenimiento el aporte de lactato de magnesio en altas dosis por vía oral, como tratamiento de mantenimiento. Tras estabilizar las cifras con aporte vía oral, se le retira la medicación antiepiléptica, no apareciendo más episodios de convulsiones.*

Figura 3.6: Fragmento de texto clínico del documento S0212-16112007000800011-1 procedente del corpus SPACCC (Intxaurrondo, 2018)

3.4.1. TrigerApp

Para complementar todo el trabajo realizado, se ha desarrollado una herramienta web básica bautizada con el nombre de **TriggerApp** para que los profesionales sanitarios puedan evaluar nuestro método realizando pruebas y consultando los resultados de manera fácil y sencilla. Desde la pantalla principal (Figura 3.7) se muestra información de las tareas lanzadas, el estado de cada una de ellas y el número de posibles eventos adversos detectados en el texto analizado en caso de que existan. La herramienta contiene las siguientes funcionalidades:

- Nueva tarea: esta opción (Figura 3.8) permite al profesional lanzar nuevas tareas de análisis.
- Editar: permite modificar (Figura 3.9) el texto clínico mientras la tarea esté pendiente

de analizar.

- Eliminar: elimina (Figura 3.10) una tarea de la pantalla principal.
- Consulta: pulsando sobre el título de la tarea podremos ver el resultado devuelto por nuestro método (Figura 3.11), resaltando el texto de otro color en caso de que se haya detectado algún posible evento adverso; pulsando sobre dicho texto nos mostrará más información sobre éste (Figura 3.12), así como el término que lo desencadenó.

Además, como información adicional, se ha incorporado el "Catálogo de Trigger" (Figura 3.13) del Apéndice A con enlaces de cada trigger a su definición en la página oficial de SNOMED CT Browser[16] para facilitar a los profesionales la validación del método.

Desde el punto de vista más técnico la herramienta TriggerApp ha sido desarrollada utilizando los frameworks Flask[17] (backend) y Bootstrap[18] (frontend). Para el despliegue en Docker ha sido necesario la adición de un nuevo servicio de proxy inverso con Nginx[19] para mantener el servicio web de TriggerApp y el servicio API RESTful funcionando simultáneamente.

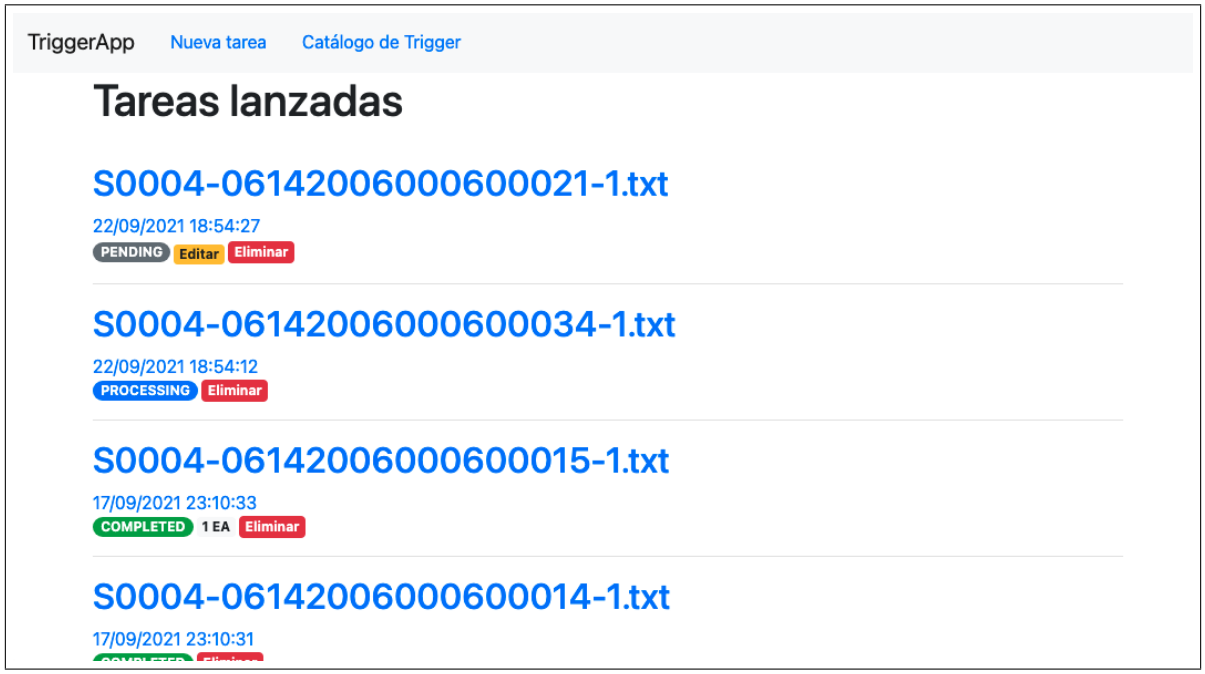

Figura 3.7: Pantalla principal de TriggerApp

[16] https://browser.ihtsdotools.org
[17] https://flask.palletsprojects.com
[18] https://getbootstrap.com
[19] https://www.nginx.com

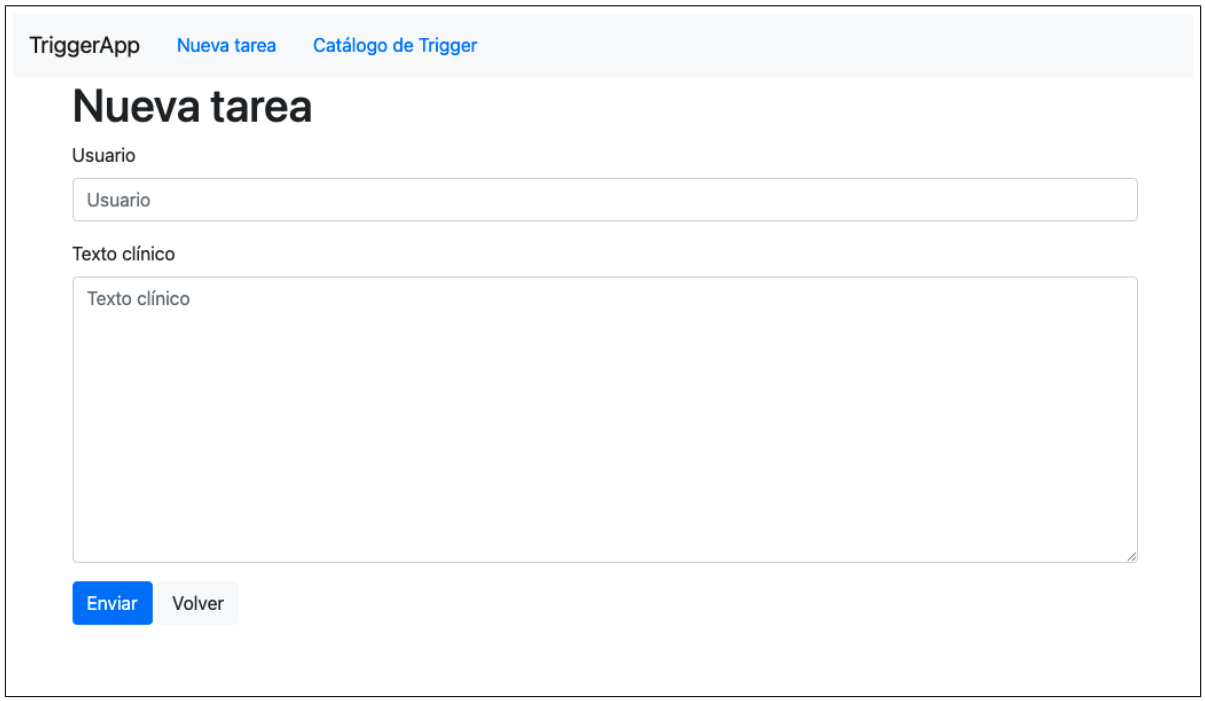

Figura 3.8: Pantalla "Nueva tarea" de TriggerApp

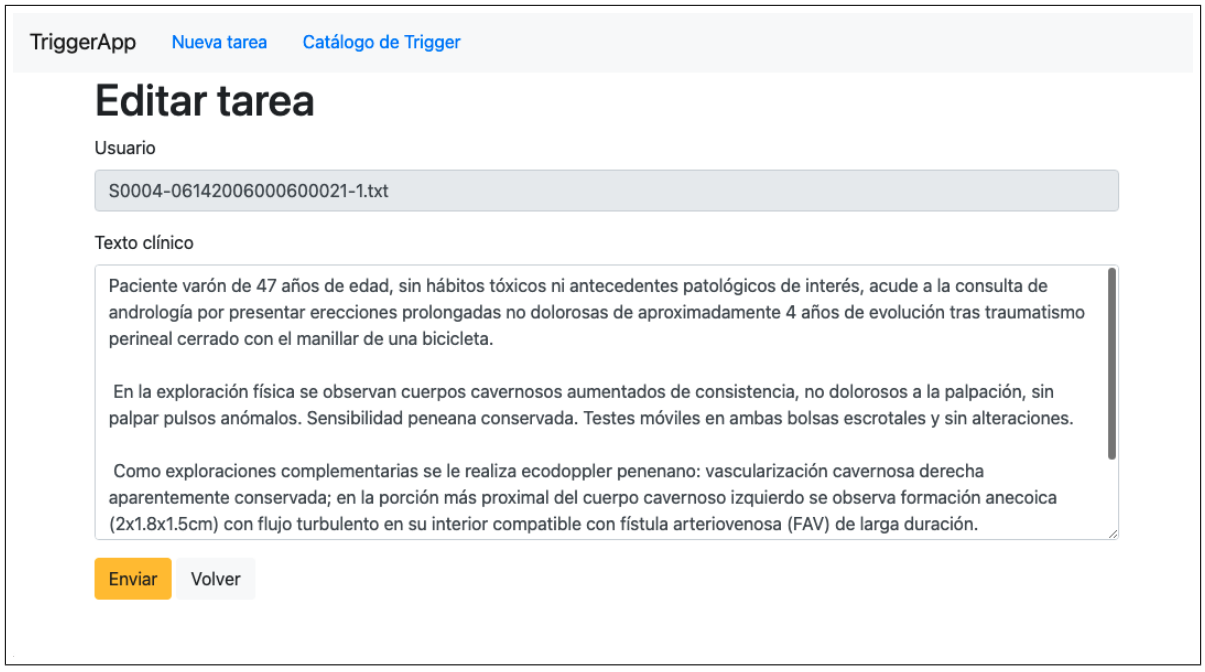

Figura 3.9: Pantalla "Editar" de TriggerApp

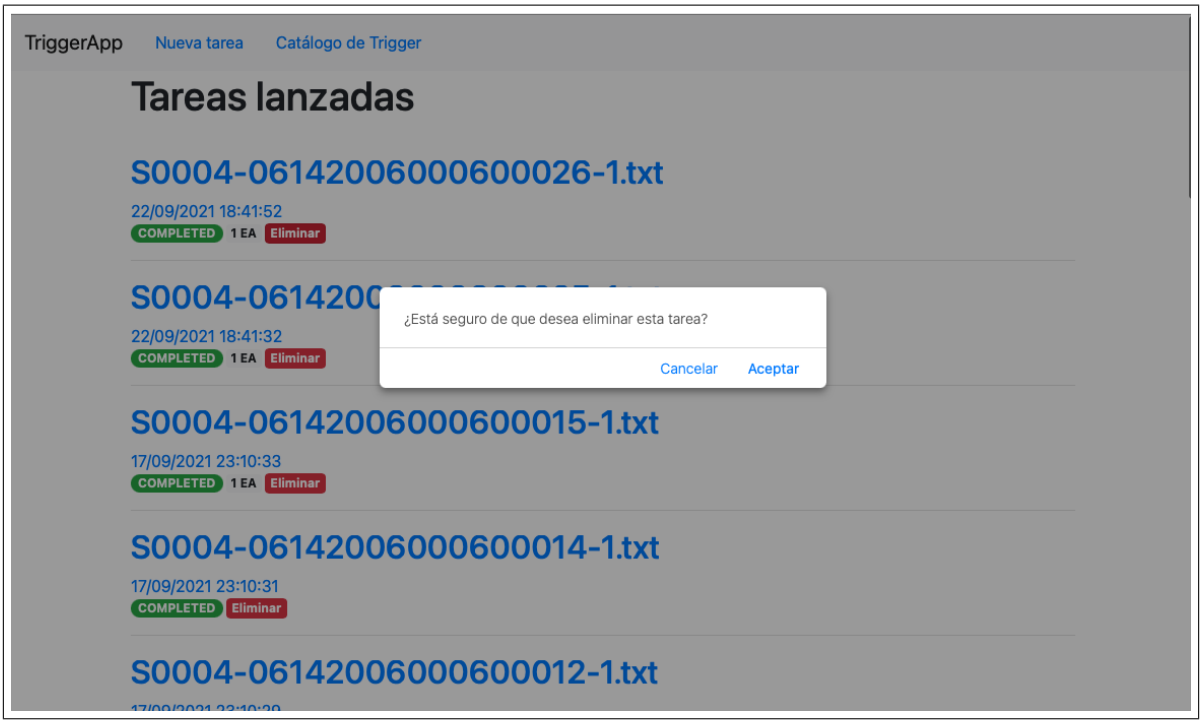

Figura 3.10: Pantalla "Eliminar" de TriggerApp

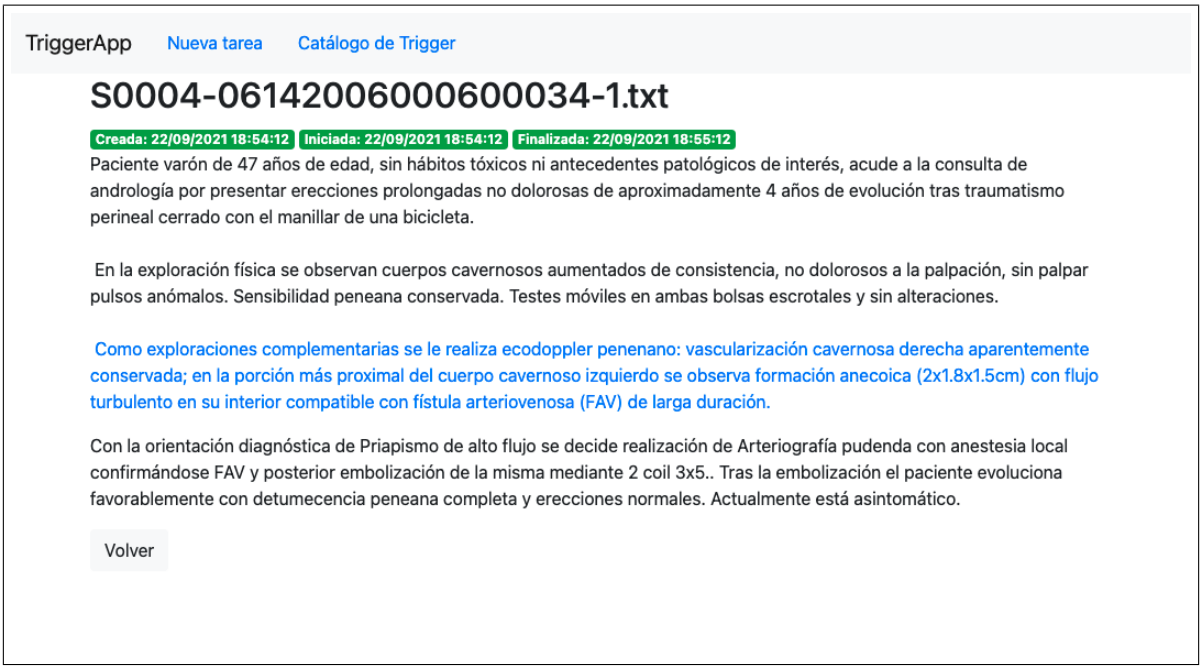

Figura 3.11: Pantalla "Consulta" de TriggerApp

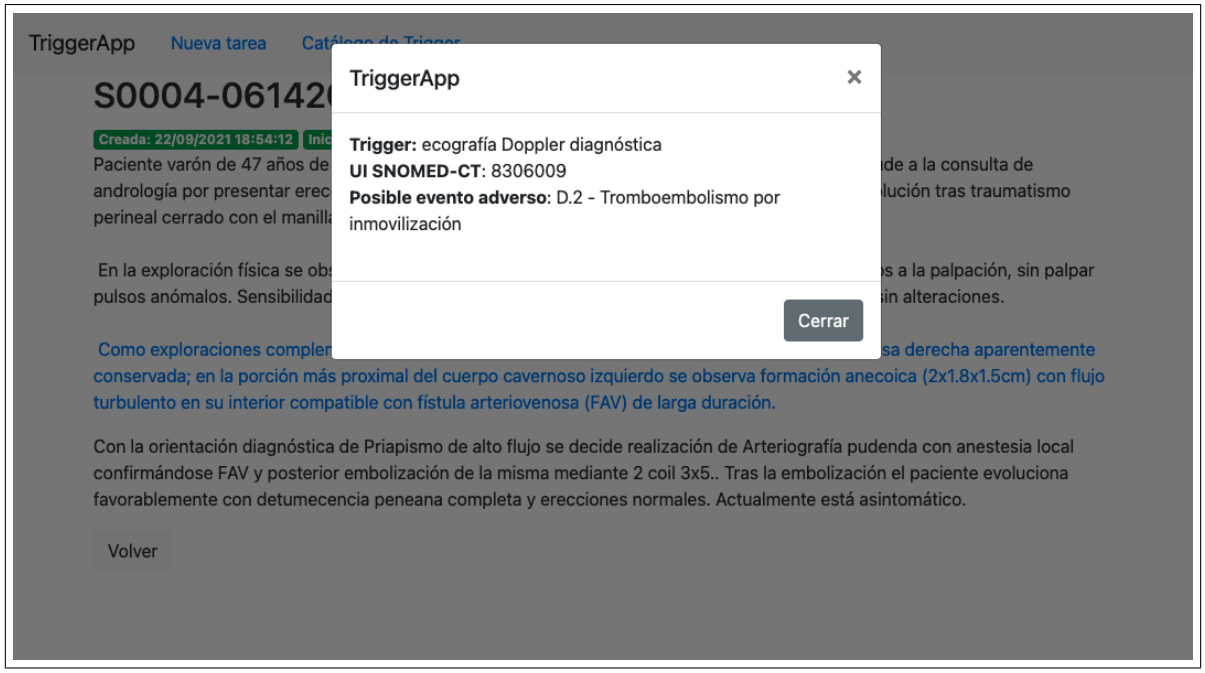

Figura 3.12: Pantalla de información en "Consulta" de TriggerApp

Número	Trigger	SCTID	Semántica	Ejemplo de posible evento adverso asociado
A.1	Úlceras por presión	420226006	Anomalía morfológica	Úlcera por presión
A.2	Caídas	217082002	Evento	Sobresedación
A.3	Rash	271807003	Trastorno	Reacción alérgica a medicamento
A.4	IAM en paciente hospitalizado	57054005	Trastorno	Procedimientos invasivos
A.5	AVC en paciente hospitalizado	266257000	Trastorno	Cardioversión de FA sin anticoagular
A.6	ETEV en paciente hospitalizado	429098002	Trastorno	Falta de profilaxis en paciente

Figura 3.13: Pantalla "Catálogo de Trigger" de TriggerApp

Capítulo 4

Conclusiones y trabajo futuro

4.1. Conclusiones

Respecto a la consecución de los objetivos planteados en el Epígrafe 1.2, con el diseño del método para detectar eventos adversos y su inclusión en una herramienta se ha alcanzado el objetivo principal, además, mediante la utilización de Docker y el servicio API RESTful como interfaz, se han alcanzado los objetivos secundarios de usabilidad, accesibilidad y escalabilidad (obj. sec. 2), ya que permite ser integrada fácilmente desde cualquier plataforma externa en producción o desarrollarla desde cero, así como de recuperar los textos analizados (obj. sec. 3). En referencia a su integración con la HCE se podrían incorporar los resultados obtenidos por nuestro método mediante el uso de estándares como HL7 (Health Level Seven) o CDA[1] (Clinical Document Architecture).

Tras el análisis realizado de las herramientas disponibles (obj. sec. 1) se puede concluir que existen diferencias entre el estado de maduración del PLN de textos clínicos en español con respecto a textos clínicos en inglés pero que, gracias a iniciativas como la creación del Plan de Impulso de las Tecnologías del Lenguaje[2] del Ministerio de Asuntos Económicos y Transformación Digital, se fomenta y desarrolla el uso de las tecnologías del lenguaje en este ámbito.

4.2. Trabajo futuro

Algunas de las mejoras que se podrían desarrollar en el futuro son:

- Permitir la opción de utilizar otros analizadores diferentes a MetaMapLite, como por ejemplo scispaCy[3].
- Incorporar un proceso de anonimización del texto clínico durante el paso de preprocesamiento.

[1] http://www.hl7spain.org/cda/
[2] https://plantl.mineco.gob.es/sanidad/Paginas/sanidad.aspx
[3] https://allenai.github.io/scispacy/

- Mejorar la infraestructura propuesta utilizando Redis para el gestor de colas o MongoDB para el almacenamiento de documentos para su posterior consulta y evaluación.
- Añadir una capa de seguridad de acceso al servicio API RESTful mediante JSON Web Token (JWT).

Bibliografía

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T. & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. *arXiv preprint arXiv:1904.03323*, 72-78. https://doi.org/10.18653/v1/w19-1909

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 17-21.

Aronson, A. R. & Lang, F. M. (2010). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229-236. https://doi.org/10.1136/jamia.2009.002733

Benavent, A., Iscla, A., Benavent, R. A., Amador Iscla, A. & Rafael, C. (2001). Problemas del lenguaje médico actual. (II) Abreviaciones y epónimos. *Papeles Médicos*, *10*(4), 170-176.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, *32*(suppl_1), D267-D270.

Bravo, À., Saggion, H. & Accuosto, P. (2018a). Estudio de viabilidad de una versión en español del sistema UMLS. Entregable 9.

Bravo, À., Saggion, H. & Accuosto, P. (2018b). ET2: Introducción al uso de UMLS en tareas relacionadas con Tecnologías del Lenguaje.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*. http://arxiv.org/abs/2005.14165

Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A. & Moreno-Sandoval, A. (2021). A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, *21*(1), 1-19.

Carrillo, I., Mira, J. J., Astier-Peña, M. P., Pérez-Pérez, P., Caro-Mendivelso, J., Olivera, G., Silvestre, C., Mula, A., Nuin, M. Á., Aranaz-Andrés, J. M., Fernández, A., González de Dios, J., Nebot, C., Vitaller, J., Caride Miana, E., Asencio Aznar, A., Rodríguez Sempere, V., Hervella Durantez, M. I., Molina Santiago, A., ... Palacios Palomares, C. (2020). Eventos adversos evitables en atención primaria. Estudio retrospectivo de cohortes para determinar su frecuencia y gravedad. *Atención Primaria*, *52*(10), 705-711. https://doi.org/10.1016/j.aprim.2020.02.008

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, *34*(5), 301-310.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51-89. https://doi.org/10.1002/aris.1440370103

Costumero, R., García-Pedrero, Á., Gonzalo-Martín, C., Menasalvas, E. & Millan, S. (2014). Text analysis and information extraction from Spanish written documents. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8609 LNAI*, 188-197. https://doi.org/10.1007/978-3-319-09891-3_18

Dalianis, H. (2018). Evaluation metrics and evaluation. *Clinical Text Mining* (pp. 45-53). Springer.

Demner-Fushman, D., Rogers, W. J. & Aronson, A. R. (2017). MetaMap Lite: An evaluation of a new Java implementation of MetaMap. *Journal of the American Medical Informatics Association*, *24*(4), 841-844. https://doi.org/10.1093/jamia/ocw177

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171-4186.

Estado, J. & I, J. C. (2015). Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica. http://www.boe.es/buscar/pdf/2002/BOE-A-2002-22188-consolidado.pdf

Green Jr, B. F., Wolf, A. K., Chomsky, C. & Laughery, K. (1961). Baseball: an automatic question-answerer. *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, 219-224.

Guzmán-Ruiz, O., Ruiz-López, P., Gómez-Cámara, A. & Ramírez-Martín, M. (2015). Detección de eventos adversos en pacientes adultos hospitalizados mediante el método Global TriggerTool. *Revista de Calidad Asistencial*, *30*(4), 166-174. https://doi.org/10.1016/j.cali.2015.03.003

IHTSDO. (2018). *Guía de introducción a SNOMED CT*.

Intxaurrondo, A. (2018). SPACCC. https://doi.org/10.5281/ZENODO.2560316

Jick, H. (1974). Drugs—remarkably nontoxic. *New England Journal of Medicine*, *291*(16), 824-828.

Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Khoong, E. C., Steinbrook, E., Brown, C. & Fernandez, A. (2019). Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA internal medicine*, *179*(4), 580-582.

Lourdusamy, R. & Abraham, S. (2018). A Survey on Text Pre-processing Techniques and Tools. *International Journal of Computer Sciences and Engineering*, *6*(03), 148-157.

Luque Guzmán, C. (2020). Text Mining y Medicina: Una aproximación a la detección temprana de enfermedades.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, *17*(01), 128-144.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Mitchell, T. M. (2006). *The Discipline of Machine Learning* (Vol. 17). Carnegie Mellon University, School of Computer Science, Machine Learning.

Padró, L. et al. (2011). Analizadores Multilingües en FreeLing. *Linguamática*, *3*(2), 13-20.

Pastore Burgos, A. & Díaz Esteban, A. (2015). *Herramienta para búsqueda de casos médicos semejantes* (Tesis doctoral).

Perez-de-Viñaspre, O. & Oronoz, M. (2015). SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. *BMC Medical Informatics and Decision Making*, *15*(2), S5. https://doi.org/10.1186/1472-6947-15-S2-S5

Peterson, K. J. & Liu, H. (2020). Automating the Transformation of Free-Text Clinical Problems into SNOMED CT Expressions. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, *2020*, 497-506.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Rogers, W. J. & Gay, C. (2015). MetaMap Data File Builder. *Builder*.

Santamaría, J. & Krallinger, M. (2018). Construcción de recursos terminológicos médicos para el español: el sistema de extracción de términos CUTEXT y los repositorios de términos biomédicos. *Procesamiento del Lenguaje Natural*, *61*, 49-56.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507-513. https://doi.org/10.1136/jamia.2009.001560

Schuster, M. & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681. https://doi.org/10.1109/78.650093

Simon, H. A. (1983). Why Should Machines Learn? *Machine Learning* (pp. 25-37). Elsevier. https://doi.org/10.1016/b978-0-08-051054-5.50006-6

Soriano, I. M., Pena, J. L. C., Fernandez Breis, J. T., Roman, I. S., Barriuso, A. A. & Baraza, D. G. (2019). Snomed2Vec: Representation of SNOMED CT terms with Word2Vec. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, *2019-June*, 678-683. https://doi.org/10.1109/CBMS.2019.00138

Sosa, E. (1997). Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones (Parte I). *El profesional de la información*, *6*.

Teresa Romá-Ferri, M. & Palomar, M. (2008). Análisis de terminologías de salud para su utilización como ontologías computacionales en los sistemas de información clínicos. *Gaceta Sanitaria*, *22*(5), 421-433. https://doi.org/10.1157/13126923

Torijano-Casalengua, M. L., Astier-Peña, P. & Mira-Solves, J. J. (2016). El impacto que tienen los eventos adversos sobre los profesionales sanitarios de atención primaria y sus instituciones. *Atención Primaria*, *48*(3), 143-146. https://doi.org/10.1016/j.aprim.2016.01.002

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999-6009.

Vilares, J. (2006). Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español. *Procesamiento del lenguaje natural*, (36), 57-58.

Weiss, S. M., Indurkhya, N. & Zhang, T. (2015). *Fundamentals of predictive text mining*. Springer.

WHO. (2009). Más que palabras . Marco Conceptual de la Clasificación Internacional para la Seguridad del Paciente Informe Técnico Definitivo Enero de 2009. http://www.who.int/patientsafety/implementation/icps/icps_full_report_es.pdf

WHO. (2019). Seguridad del paciente. Consultado el 13 de julio de 2021, desde https://www.who.int/es/news-room/fact-sheets/detail/patient-safety

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D. & Langlotz, C. P. (2021). Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*.

Glosario

concepto entidades más relevantes de un dominio. 10

lexicon diccionario. 11

metatesauro tesauro multi-lenguaje que contiene millones de conceptos biomédicos y sanitarios, sus nombres, sinónimos y sus relaciones utilizado en las diferentes herramientas de UMLS.. 11

ontología formulación de un exhaustivo y riguroso esquema conceptual dentro de uno o varios dominios dados, con la finalidad de facilitar la comunicación y el intercambio de información entre diferentes sistemas y entidades.. 11

prolog lenguaje de programación mediante el paradigma lógico con técnicas de producción final interpretada.. 25

semántica significado de una unidad lingüística.. 11

terminología conjunto de las unidades de expresión y comunicación que permiten transferir y comunicar el pensamiento especializado. Una terminología persigue el objetivo de fijar unas unidades terminológicas como formas normalizadas y de referencia que descartan las demás variantes para denominar un mismo concepto con el fin de alcanzar una comunicación profesional precisa, moderna y unívoca.. 3

tesauro lista de palabras con significados similares, sinónimos, habitualmente acompañada por otra lista de antónimos. Normalmente está reducido a un campo concreto de la lengua como puede ser el ámbito biomédico. 49

vocabulario catálogo o lista de palabras, ordenadas con arreglo a un sistema, y con definiciones o explicaciones sucintas.. 11

Apéndice A

Catálogo de triggers

Número	Trigger	SCTID	Semántica	Ejemplo de posible evento adverso asociado
A.1	Úlceras por presión	420226006	Anomalía morfológica	Úlcera por presión
A.2	Caídas	217082002	Evento	Sobresedación
A.3	Rash	271807003	Trastorno	Reacción alérgica a medicamento
A.4	IAM en paciente hospitalizado	57054005	Trastorno	Procedimientos invasivos
A.5	AVC en paciente hospitalizado	266257000	Trastorno	Cardioversión de FA sin anticoagular
A.6	ETEV en paciente hospitalizado	429098002	Trastorno	Falta de profilaxis en paciente inmovilizado
A.7	Hemorragia	131148009	Hallazgo	Procedimientos invasivos
A.8	Agitación	24199005	Hallazgo	Uso de anticolinérgicos
A.9	Sobresedación	410011004	Procedimiento	Uso de benzodiacepinas
A.10	Hipotensión	45007003	Trastorno	Sobremedicación hipotensora
A.11	Neumonía de nuevo comienzo	425464007	Trastorno	Neumonía nosocomial
A.12	Flebitis	61599003	Trastorno	Infección de vía venosa
A.13	Neumotórax	36118008	Trastorno	Procedimientos invasivos
A.14	Sondaje	387651008	Procedimiento	Retención urinaria por fármacos
A.15	Extravío o repetición de petición	118635009	Procedimiento	Retraso en el diagnóstico
A.16	Fiebre	386661006	Hallazgo	Infección nosocomial
A.17	Reingreso en 30 días	417005	Procedimiento	Infección nosocomial
A.18	Trasferencia a nivel de cuidados más alto	305351004	Procedimiento	Complicación de procedimiento invasivo
A.19	Muerte	419620001	Evento	Complicación de procedimiento invasivo

Tabla A.1: Catálogo de triggers codificados en SNOMED-CT asociados a cuidados generales

Número	Trigger	SCTID	Semántica	Ejemplo de posible evento adverso asociado
B.1	Antieméticos	52017007	Producto	Anestesia
B.2	Antidiarreico	398836001 55946005 387040009	Sustancia	Diarrea asociada a antibióticos
B.3	Antihistamínico	373268008	Sustancia	Reacción alérgica a medicamento
B.4	Antipsicótico	372482001	Sustancia	Delirium en paciente hospitalizado
B.5	Laxante o enema	372800002 61919008	Sustancia y Procedimiento	Estreñimiento por inmovilidad
B.6	Prednisona u otro corticoide	116602009 307513002	Sustancia y Procedimiento	Reacción alérgica a medicamento
B.7	Glucagón o suero glucosado al 10 %	66603002	Sustancia	Hipoglucemia insulínica
B.8	Vitamina K	65183007	Sustancia	Hemorragia por anticoagulantes
B.9	Transfusión	33389009	Procedimiento	Hemorragia por procedimiento invasivo
B.10	Flumazenilo o naloxona	372890007	Sustancia	Sobreefecto de sedantes o narcóticos
B.11	Cese brusco de medicación	406149000	Situación	Alteración hidroelectrolítica por fármacos

Tabla A.2: Catálogo de triggers codificados en SNOMED-CT asociados a medicamentos o tratamientos

Número	Trigger	SCTID	Semántica	Ejemplo de posible evento adverso asociado
C.1	Cultivo positivo a C. difficile	122209009	Procedimiento	Diarrea asociada a antibióticos
C.2	Glucosa <50 mg/dl	51798006	Hallazgo	Hipoglucemia insulínica
C.3	INR >6	313341008	Hallazgo	Hemorragia por anticoagulantes
C.4	Caída de hemoglobina o hematocrito >25 %	165397008 165414004	Hallazgo	Hemorragia por procedimiento invasivo
C.5	Elevación de los niveles de creatinina basal ×2	166717003	Hallazgo	Insuficiencia renal por fármacos
C.6	Na <120 o >150 mEq/l	166693005	Hallazgo	Alteración hidroelectrolítica por fármacos
C.7	K <2 o >5 mEq/l	166690008 166689004	Hallazgo	Alteración hidroelectrolítica por fármacos
C.8	Hemocultivo positivo	8730001000004107	Hallazgo	Infección nosocomial
C.9	Urocultivo positivo	365697000	Hallazgo	Infección nosocomial

Tabla A.3: Catálogo de triggers codificados en SNOMED-CT asociados a resultados de laboratorio o microbiología

Número	Trigger	SCTID	Semántica	Ejemplo de posible evento adverso asociado
D.1	Gastroscopia	173822004	Procedimiento	Hemorragia por anticoagulantes
D.2	Pruebas de detección de coágulo (TAC/eco Döppler)	77477000 8306009	Procedimiento	Tromboembolismo por inmovilización

Tabla A.4: Catálogo de triggers codificados en SNOMED-CT asociados a pruebas diagnósticas

Apéndice B

Ficheros Docker

B.1. docker-compose.yml

```
version: '3'
services:
  trigger:
    build:
      context: .
      dockerfile: Dockerfile-trigger
    container_name: worker-trigger
    restart: unless-stopped
    volumes:
      - ../../trigger/src:/opt/trigger
      - ./tasks:/opt/tasks
    environment:
      - ONLY_TRIGGER=true
      - APP_DIR=data/
      - LOG_FILE=/opt/tasks/trigger.log
      - LOG_LEVEL=INFO
      - AUTHENTICATION_BASE_URI=https://utslogin.nlm.nih.gov
      - ENDPOINT_BASE_URI=https://uts-ws.nlm.nih.gov/rest
      - UMLS_API_KEY=XXX
      - ANALYZER=MML
      - MML_MIN_CONCEPT_SCORE=0.4
      - MML_DIR=/opt/public_mm_lite/
      - GOOGLE_APPLICATION_CREDENTIALS=XXX
      - PENDING_DIR=/opt/tasks/pending/
      - PROCESSING_DIR=/opt/tasks/processing/
      - COMPLETED_DIR=/opt/tasks/completed/
  api:
```

```
    build:
      context: .
      dockerfile: Dockerfile-api
    container_name: api-trigger
    restart: unless-stopped
    volumes_from:
      - trigger
    ports:
      - 80:80
    environment:
      - MODULE_NAME=api
      - ACCESS_LOG=/opt/tasks/api_access.log
      - ERROR_LOG=/opt/tasks/api_error.log
      - LOG_FILE=/opt/tasks/api.log
      - LOG_LEVEL=INFO
      - PENDING_DIR=/opt/tasks/pending/
      - PROCESSING_DIR=/opt/tasks/processing/
      - COMPLETED_DIR=/opt/tasks/completed/

```

B.2. Dockerfile-trigger

```
FROM ubuntu:bionic

ENV PYTHONDONTWRITEBYTECODE 1
ENV PYTHONUNBUFFERED 1
ENV TZ Europe/Madrid
ENV DEBIAN_FRONTEND noninteractive
ENV HOME /home/root

WORKDIR /opt

# Requisito previo:
↪ https://metamap.nlm.nih.gov/download/metamaplite/public_mm_lite_3.6.2rc6_binaryonly.zip
COPY requires/public_mm_lite_3.6.2rc6_binaryonly.zip /opt
# Requisito previo:
↪ https://metamap.nlm.nih.gov/download/metamaplite/public_mm_data_lite_usabase_2020aa.zip
COPY requires/public_mm_data_lite_usabase_2020aa.zip /opt
# Requisito previo: repo https://github.com/AnthonyMRios/pymetamap.git
ADD requires/pymetamap.tar.gz /opt
# Requisito previo: credenciales Google-Cloud
```

```
COPY requires/XXX.json /opt

RUN apt-get update && \
    apt-get install -y tzdata software-properties-common unzip default-jre libaio1
    ↪  python3-pip && \
    apt-get clean && \
    rm -rf /var/lib/apt/lists/*
RUN ln -snf /usr/share/zoneinfo/$TZ /etc/localtime && echo $TZ > /etc/timezone &&
↪  dpkg-reconfigure --frontend noninteractive tzdata
RUN unzip public_mm_lite_3.6.2rc6_binaryonly.zip && \
    unzip public_mm_data_lite_usabase_2020aa.zip && \
    rm -f public_mm_*.zip
RUN cd pymetamap && \
    python3 setup.py install && \
    cd .. && \
    rm -rf pymetamap
RUN pip3 install --upgrade pip && \
    pip install --no-cache-dir python-dotenv==0.17.1 stanza==1.2
    ↪  google-cloud-translate==3.1.0 xlrd==2.0.1 beautifulsoup4==4.9.3
RUN sed 's/CipherString = DEFAULT@SECLEVEL=2/#CipherString = DEFAULT@SECLEVEL=2/g'
↪  /etc/ssl/openssl.cnf > /etc/ssl/openssl.cnf_new && \
    mv /etc/ssl/openssl.cnf_new /etc/ssl/openssl.cnf

WORKDIR /opt/trigger

# Python 3.6.9 (default)
ENTRYPOINT ["python3"]

CMD ["worker.py"]
```

B.3. Dockerfile-api

```
FROM tiangolo/uvicorn-gunicorn-fastapi:python3.7

ENV PYTHONDONTWRITEBYTECODE 1
ENV PYTHONUNBUFFERED 1

ENV TZ=Europe/Madrid
RUN ln -snf /usr/share/zoneinfo/$TZ /etc/localtime && echo $TZ > /etc/timezone

RUN pip install --no-cache-dir python-dotenv==0.17.1
```

```
10
11 RUN sed 's/CipherString = DEFAULT@SECLEVEL=2/#CipherString = DEFAULT@SECLEVEL=2/g'
   ↪  /etc/ssl/openssl.cnf > /etc/ssl/openssl.cnf_new
12 RUN mv /etc/ssl/openssl.cnf_new /etc/ssl/openssl.cnf
13
14 WORKDIR /opt/trigger
```

Apéndice C

Ficheros de ejemplo

C.1. Fichero de entrada

```
{
    "request": {
        "user": "Foo",
```

```
    "text": "Varón de 44 años con enfermedad de Crohn diagnosticado 15 años antes, que ingresa en el servicio de Cirugía General por un cuadro de dolor abdominal y diarrea de una semana de evolución, compatible con un brote de su enfermedad. En tratamiento domiciliario con corticoides y mesalazina, el paciente es portador de una ileostomía desde hace 2 años tras realizarle una colectomía subtotal, con muñón rectal cerrado a raíz de un brote de su enfermedad. Le es realizada una bioquímica general en la que destaca un calcio corregido por la albúmina de 8,42 mg/dl (rn: 8,7-10,6), K 2,6 mg/dl (rn: 3,6-4,9), albúmina 2.5 mg/dl (rn 4-5,2), glucosa y sodio dentro de los valores de referencia. Posteriormente se le realiza un TAC craneal en el que no se aprecian anormalidades. Presenta un electrocardiograma del día del ingreso con un ritmo sinusal normal a 79 lpm, sin alargamiento del espacio QT ni PR. A los dos días presenta de nuevo una convulsión tónico-clónica, generalizada de un minuto de duración. Tras ser valorado por el Servicio de Neurología se le realiza un electroencefalograma que resulta ser un trazado sin hallazgos patológicos y se inicia tratamiento con carbamacepina. Ante la persistencia de las convulsiones de las mismas características a pesar del tratamiento médico, se le realiza un registro electroencefalográfico de 24 horas de duración en el que no se evidencian alteraciones. Dado que el paciente presenta datos de deshidratación y desnutrición, así como disminución de la ingesta y astenia, realizan una interconsulta a la Unidad de Nutrición, para valoración de soporte nutricional en caso necesario.\nAnte la existencia de deshidratación con balances hídricos negativos debido a un elevado débito de la ileostomía, y cifras bajas de electrolitos en sangre se inicia tratamiento monitorizado por vía parenteral, para reposición de volumen, electrolitos, y otros micronutrientes, entre ellos fósforo y magnesio. Previamente se realizó una extracción de sangre para la determinación de dichos micronutrientes ante la sospecha de un posible déficit. Ese mismo día el paciente sufre una convulsión similar a las previas. En la analítica realizada presenta un magnesio de 0,76 mg/dl (rn: 2,40-5,40) con un fósforo, calcio, potasio y sodio dentro de los valores de referencia. Se pauta una perfusión con altas dosis de sulfato de magnesio con la progresiva normalización de sus niveles en sangre, siendo suficiente para el tratamiento de mantenimiento el aporte de lactato de magnesio en altas dosis por vía oral, como tratamiento de mantenimiento. Tras estabilizar las cifras con aporte vía oral, se le retira la medicación antiepiléptica, no apareciendo más episodios de convulsiones.",
    "analyzer": "MML",
    "onlyTrigger": true,
    "requestTimestamp": 1627896047.415472,
    "additionalInfo": {
        "episodio": 1325621,
```

```
            "servicio": "Cirugía General",
            "fechaIngreso": "22/05/2021",
            "fechaAlta": null
        }
    },
    "result": null,
    "startTimestamp": null,
    "endTimestamp": null
}
```

C.2. Fichero de salida

```
{
    "request": {
        "user": "Foo",
```

```
    "text": "Varón de 44 años con enfermedad de Crohn diagnosticado 15 años antes, que ingresa en el servicio de Cirugía General por un cuadro de dolor abdominal y diarrea de una semana de evolución, compatible con un brote de su enfermedad. En tratamiento domiciliario con corticoides y mesalazina, el paciente es portador de una ileostomía desde hace 2 años tras realizarle una colectomía subtotal, con muñón rectal cerrado a raíz de un brote de su enfermedad. Le es realizada una bioquímica general en la que destaca un calcio corregido por la albúmina de 8,42 mg/dl (rn: 8,7-10,6), K 2,6 mg/dl (rn: 3,6-4,9), albúmina 2.5 mg/dl (rn 4-5,2), glucosa y sodio dentro de los valores de referencia. Posteriormente se le realiza un TAC craneal en el que no se aprecian anormalidades. Presenta un electrocardiograma del día del ingreso con un ritmo sinusal normal a 79 lpm, sin alargamiento del espacio QT ni PR. A los dos días presenta de nuevo una convulsión tónico-clónica, generalizada de un minuto de duración. Tras ser valorado por el Servicio de Neurología se le realiza un electroencefalograma que resulta ser un trazado sin hallazgos patológicos y se inicia tratamiento con carbamacepina. Ante la persistencia de las convulsiones de las mismas características a pesar del tratamiento médico, se le realiza un registro electroencefalográfico de 24 horas de duración en el que no se evidencian alteraciones. Dado que el paciente presenta datos de deshidratación y desnutrición, así como disminución de la ingesta y astenia, realizan una interconsulta a la Unidad de Nutrición, para valoración de soporte nutricional en caso necesario.\nAnte la existencia de deshidratación con balances hídricos negativos debido a un elevado débito de la ileostomía, y cifras bajas de electrolitos en sangre se inicia tratamiento monitorizado por vía parenteral, para reposición de volumen, electrolitos, y otros micronutrientes, entre ellos fósforo y magnesio. Previamente se realizó una extracción de sangre para la determinación de dichos micronutrientes ante la sospecha de un posible déficit. Ese mismo día el paciente sufre una convulsión similar a las previas. En la analítica realizada presenta un magnesio de 0,76 mg/dl (rn: 2,40-5,40) con un fósforo, calcio, potasio y sodio dentro de los valores de referencia. Se pauta una perfusión con altas dosis de sulfato de magnesio con la progresiva normalización de sus niveles en sangre, siendo suficiente para el tratamiento de mantenimiento el aporte de lactato de magnesio en altas dosis por vía oral, como tratamiento de mantenimiento. Tras estabilizar las cifras con aporte vía oral, se le retira la medicación antiepiléptica, no apareciendo más episodios de convulsiones.",
    "analyzer": "MML",
    "onlyTrigger": true,
    "requestTimestamp": 1627896047.415472,
    "additionalInfo": {
        "episodio": 1325621,
```

```
            "servicio": "Cirugía General",
            "fechaIngreso": "22/05/2021",
            "fechaAlta": null
        }
    },
    "result": [
        {
            "text": "Varón de 44 años con enfermedad de Crohn diagnosticado 15 años antes, que ingresa en el servicio de Cirugía General por un cuadro de dolor abdominal y diarrea de una semana de evolución, compatible con un brote de su enfermedad. En tratamiento domiciliario con corticoides y mesalazina, el paciente es portador de una ileostomía desde hace 2 años tras realizarle una colectomía subtotal, con muñón rectal cerrado a raíz de un brote de su enfermedad. Le es realizada una bioquímica general en la que destaca un calcio corregido por la albúmina de 8,42 mg/dl (rn: 8,7-10,6), K 2,6 mg/dl (rn: 3,6-4,9), albúmina 2.5 mg/dl (rn 4-5,2), glucosa y sodio dentro de los valores de referencia. Posteriormente se le realiza un TAC craneal en el que no se aprecian anormalidades. Presenta un electrocardiograma del día del ingreso con un ritmo sinusal normal a 79 lpm, sin alargamiento del espacio QT ni PR. A los dos días presenta de nuevo una convulsión tónico-clónica, generalizada de un minuto de duración. Tras ser valorado por el Servicio de Neurología se le realiza un electroencefalograma que resulta ser un trazado sin hallazgos patológicos y se inicia tratamiento con carbamacepina. Ante la persistencia de las convulsiones de las mismas características a pesar del tratamiento médico, se le realiza un registro electroencefalográfico de 24 horas de duración en el que no se evidencian alteraciones. Dado que el paciente presenta datos de deshidratación y desnutrición, así como disminución de la ingesta y astenia, realizan una interconsulta a la Unidad de Nutrición, para valoración de soporte nutricional en caso necesario.",
            "sentences": [
                {
                    "sentence": "Varón de 44 años con enfermedad de Crohn diagnosticado 15 años antes, que ingresa en el servicio de Cirugía General por un cuadro de dolor abdominal y diarrea de una semana de evolución, compatible con un brote de su enfermedad.",
                    "traduction": "A 44-year-old man with Crohn's disease diagnosed 15 years earlier, who was admitted to the General Surgery service due to a week-long history of abdominal pain and diarrhea, compatible with an outbreak of his disease."
                },
                {
```

```
            "sentence": "En tratamiento domiciliario con corticoides y mesalazina, el paciente es portador de una ileostomía desde hace 2 años tras realizarle una colectomía subtotal, con muñón rectal cerrado a raíz de un brote de su enfermedad.",
            "traduction": "Under home treatment with corticosteroids and mesalazine, the patient has had an ileostomy for 2 years after undergoing a subtotal colectomy, with a closed rectal stump as a result of an outbreak of his disease."
        },
        {
            "sentence": "Le es realizada una bioquímica general en la que destaca un calcio corregido por la albúmina de 8,42 mg/dl (rn: 8,7-10,6), K 2,6 mg/dl (rn: 3,6-4,9), albúmina 2.5 mg/dl (rn 4-5,2), glucosa y sodio dentro de los valores de referencia.",
            "traduction": "A general biochemistry is performed in which a calcium corrected by albumin stands out of 8.42 mg / dl (rn: 8.7-10.6), K 2.6 mg / dl (rn: 3.6-4 , 9), albumin 2.5 mg / dl (rn 4-5.2), glucose and sodium within the reference values."
        },
        {
            "sentence": "Posteriormente se le realiza un TAC craneal en el que no se aprecian anormalidades.",
            "traduction": "Subsequently, a cranial CT scan is performed in which no abnormalities are appreciated.",
            "concepts": [
                {
                    "name": "CT scan",
                    "posStart": 24,
                    "posEnd": 31,
                    "umls": [
                        {
                            "name": "X-Ray Computed Tomography",
                            "score": 13.81,
                            "cui": "C0040405",
                            "semantic": "Diagnostic Procedure",
                            "snomedctUs": [
                                {
                                    "name": "CAT - Computerised axial tomography",
                                    "ui": "A3105529",
                                    "sctid": "77477000",
                                    "termType": "British synonym",
                                    "sctspa": {
```

```
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "CAT - Computerized axial tomography",
            "ui": "A3105530",
            "sctid": "77477000",
            "termType": "Designated synonym",
            "sctspa": {
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "CAT scan",
            "ui": "A2896691",
            "sctid": "77477000",
            "termType": "Designated synonym",
            "sctspa": {
```

```
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "CT - Computerised tomography",
            "ui": "A3106031",
            "sctid": "77477000",
            "termType": "British synonym",
            "sctspa": {
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "CT - Computerized tomography",
            "ui": "A3106033",
            "sctid": "77477000",
            "termType": "Designated synonym",
            "sctspa": {
```

```
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "Computed axial tomography",
            "ui": "A3113341",
            "sctid": "77477000",
            "termType": "Designated synonym",
            "sctspa": {
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "Computerised axial tomography",
            "ui": "A3113367",
            "sctid": "77477000",
            "termType": "British preferred term",
            "sctspa": {
```

```
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "Computerised tomograph scan",
            "ui": "A3113371",
            "sctid": "77477000",
            "termType": "British synonym",
            "sctspa": {
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "Computerised tomography",
            "ui": "A3113372",
            "sctid": "77477000",
            "termType": "British synonym",
            "sctspa": {
```

```
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "Computerised transaxial tomography",
            "ui": "A3365896",
            "sctid": "77477000",
            "termType": "British synonym",
            "sctspa": {
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "Computerized axial tomography",
            "ui": "A3113375",
            "sctid": "77477000",
            "termType": "Designated preferred name",
            "sctspa": {
```

```
            "name": "tomografía axial computarizada",
            "ui": "77477000"
        },
        "trigger": [
            {
                "name": "tomografía axial computarizada",
                "ui": "77477000",
                "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
            }
        ]
    },
    {
        "name": "Computerized axial tomography (procedure)",
        "ui": "A3365898",
        "sctid": "77477000",
        "termType": "Full form of descriptor",
        "sctspa": {
            "name": "tomografía axial computarizada",
            "ui": "77477000"
        },
        "trigger": [
            {
                "name": "tomografía axial computarizada",
                "ui": "77477000",
                "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
            }
        ]
    },
    {
        "name": "Computerized tomograph scan",
        "ui": "A3113378",
        "sctid": "77477000",
        "termType": "Designated synonym",
        "sctspa": {
```

```
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "Computerized tomography",
            "ui": "A3113379",
            "sctid": "77477000",
            "termType": "Designated synonym",
            "sctspa": {
                "name": "tomografía axial computarizada",
                "ui": "77477000"
            },
            "trigger": [
                {
                    "name": "tomografía axial computarizada",
                    "ui": "77477000",
                    "adverseEvent": "D.2 - Tromboembolismo por inmovilización"
                }
            ]
        },
        {
            "name": "Computerized transaxial tomography",
            "ui": "A2895778",
            "sctid": "77477000",
            "termType": "Designated synonym",
            "sctspa": {
```

```
                                    "name": "tomografía axial
                                    ↪  computarizada",
                                    "ui": "77477000"
                                },
                                "trigger": [
                                    {
                                        "name": "tomografía axial
                                        ↪  computarizada",
                                        "ui": "77477000",
                                        "adverseEvent": "D.2 -
                                        ↪  Tromboembolismo por
                                        ↪  inmovilización"
                                    }
                                ]
                            }
                        ]
                    }
                ]
            }
        ]
    },
    {
        "sentence": "Presenta un electrocardiograma del día del ingreso con
        ↪  un ritmo sinusal normal a 79 lpm, sin alargamiento del espacio
        ↪  QT ni PR.",
        "traduction": "He presented an electrocardiogram on the day of
        ↪  admission with a normal sinus rhythm at 79 bpm, without QT space
        ↪  lengthening or PR."
    },
    {
        "sentence": "A los dos días presenta de nuevo una convulsión
        ↪  tónico-clónica, generalizada de un minuto de duración.",
        "traduction": "Two days later, he presented again a generalized
        ↪  tonic-clonic seizure lasting one minute."
    },
    {
        "sentence": "Tras ser valorado por el Servicio de Neurología se le
        ↪  realiza un electroencefalograma que resulta ser un trazado sin
        ↪  hallazgos patológicos y se inicia tratamiento con
        ↪  carbamacepina.",
```

```
                "traduction": "After being assessed by the Neurology Service, an
                ↪  electroencephalogram was performed, which turned out to be a
                ↪  tracing without pathological findings, and treatment with
                ↪  carbamazepine was started."
            },
            {
                "sentence": "Ante la persistencia de las convulsiones de las mismas
                ↪  características a pesar del tratamiento médico, se le realiza un
                ↪  registro electroencefalográfico de 24 horas de duración en el
                ↪  que no se evidencian alteraciones.",
                "traduction": "Given the persistence of seizures of the same
                ↪  characteristics despite medical treatment, a 24-hour
                ↪  electroencephalographic recording was performed in which no
                ↪  alterations were evidenced."
            },
            {
                "sentence": "Dado que el paciente presenta datos de deshidratación y
                ↪  desnutrición, así como disminución de la ingesta y astenia,
                ↪  realizan una interconsulta a la Unidad de Nutrición, para
                ↪  valoración de soporte nutricional en caso necesario.",
                "traduction": "Given that the patient presents data of dehydration
                ↪  and malnutrition, as well as decreased intake and asthenia, they
                ↪  consult the Nutrition Unit to assess nutritional support if
                ↪  necessary."
            }
        ]
    },
    {
```

```
330             "text": "Ante la existencia de deshidratación con balances hídricos
                ↪  negativos debido a un elevado débito de la ileostomía, y cifras bajas de
                ↪  electrolitos en sangre se inicia tratamiento monitorizado por vía
                ↪  parenteral, para reposición de volumen, electrolitos, y otros
                ↪  micronutrientes, entre ellos fósforo y magnesio. Previamente se realizó
                ↪  una extracción de sangre para la determinación de dichos micronutrientes
                ↪  ante la sospecha de un posible déficit. Ese mismo día el paciente sufre
                ↪  una convulsión similar a las previas. En la analítica realizada presenta
                ↪  un magnesio de 0,76 mg/dl (rn: 2,40-5,40) con un fósforo, calcio,
                ↪  potasio y sodio dentro de los valores de referencia. Se pauta una
                ↪  perfusión con altas dosis de sulfato de magnesio con la progresiva
                ↪  normalización de sus niveles en sangre, siendo suficiente para el
                ↪  tratamiento de mantenimiento el aporte de lactato de magnesio en altas
                ↪  dosis por vía oral, como tratamiento de mantenimiento. Tras estabilizar
                ↪  las cifras con aporte vía oral, se le retira la medicación
                ↪  antiepiléptica, no apareciendo más episodios de convulsiones.",
331             "sentences": [
332                 {
333                     "sentence": "Ante la existencia de deshidratación con balances
                        ↪  hídricos negativos debido a un elevado débito de la ileostomía,
                        ↪  y cifras bajas de electrolitos en sangre se inicia tratamiento
                        ↪  monitorizado por vía parenteral, para reposición de volumen,
                        ↪  electrolitos, y otros micronutrientes, entre ellos fósforo y
                        ↪  magnesio.",
334                     "traduction": "Given the existence of dehydration with negative
                        ↪  water balances due to a high output of the ileostomy, and low
                        ↪  levels of electrolytes in the blood, parenterally monitored
                        ↪  treatment is started to replace volume, electrolytes, and other
                        ↪  micronutrients, including phosphorus and magnesium."
335                 },
336                 {
337                     "sentence": "Previamente se realizó una extracción de sangre para la
                        ↪  determinación de dichos micronutrientes ante la sospecha de un
                        ↪  posible déficit.",
338                     "traduction": "Previously, a blood extraction was carried out to
                        ↪  determine these micronutrients due to the suspicion of a
                        ↪  possible deficit."
339                 },
340                 {
341                     "sentence": "Ese mismo día el paciente sufre una convulsión similar
                        ↪  a las previas.",
342                     "traduction": "That same day the patient suffers a seizure similar
                        ↪  to the previous ones."
```

```
        },
        {
            "sentence": "En la analítica realizada presenta un magnesio de 0,76 mg/dl (rn: 2,40-5,40) con un fósforo, calcio, potasio y sodio dentro de los valores de referencia.",
            "traduction": "In the analysis carried out, it presents a magnesium of 0.76 mg / dl (rn: 2.40-5.40) with phosphorus, calcium, potassium and sodium within the reference values."
        },
        {
            "sentence": "Se pauta una perfusión con altas dosis de sulfato de magnesio con la progresiva normalización de sus niveles en sangre, siendo suficiente para el tratamiento de mantenimiento el aporte de lactato de magnesio en altas dosis por vía oral, como tratamiento de mantenimiento.",
            "traduction": "An infusion with high doses of magnesium sulfate is prescribed with the progressive normalization of its blood levels, the intake of magnesium lactate in high doses by mouth being sufficient for maintenance treatment, as maintenance treatment.",
            "concepts": [
                {
                    "name": "magnesium sulfate",
                    "posStart": 31,
                    "posEnd": 48,
                    "umls": [
                        {
                            "name": "magnesium sulfate",
                            "score": 4.14,
                            "cui": "C0024480",
                            "semantic": "Inorganic Chemical, Pharmacologic Substance",
                            "snomedctUs": [
                                {
                                    "name": "Magnesium sulfate",
                                    "ui": "A2937229",
                                    "sctid": "387202002",
                                    "termType": "Designated preferred name",
                                    "sctspa": {
                                        "name": "sulfato de magnesio",
                                        "ui": "387202002"
                                    },
                                    "trigger": [
```

```
                {
                    "name": "laxante",
                    "ui": "372800002",
                    "adverseEvent": "B.5 - Estreñimiento por inmovilidad"
                }
            ]
        },
        {
            "name": "Magnesium sulfate (substance)",
            "ui": "A3547729",
            "sctid": "387202002",
            "termType": "Full form of descriptor",
            "sctspa": {
                "name": "sulfato de magnesio",
                "ui": "387202002"
            },
            "trigger": [
                {
                    "name": "laxante",
                    "ui": "372800002",
                    "adverseEvent": "B.5 - Estreñimiento por inmovilidad"
                }
            ]
        },
        {
            "name": "Magnesium sulphate",
            "ui": "A29539886",
            "sctid": "387202002",
            "termType": "British synonym",
            "sctspa": {
                "name": "sulfato de magnesio",
                "ui": "387202002"
            },
            "trigger": [
                {
                    "name": "laxante",
                    "ui": "372800002",
                    "adverseEvent": "B.5 - Estreñimiento por inmovilidad"
                }
            ]
```

```
                                        }
                                    ]
                                }
                            ]
                        }
                    ]
                },
                {
                    "sentence": "Tras estabilizar las cifras con aporte vía oral, se le retira la medicación antiepiléptica, no apareciendo más episodios de convulsiones.",
                    "traduction": "After stabilizing the figures with oral administration, the antiepileptic medication was withdrawn, and no more episodes of seizures appeared."
                }
            ]
        }
    ],
    "startTimestamp": 1627896075.237859,
    "endTimestamp": 1627896124.308381
}
```