UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Departamento de Electrónica e Computación

TESIS DOCTORAL

TÉCNICAS DE ANOTACIÓN SEMÁNTICA ORIENTADAS A MEJORAR EL ACCESO E INTERPRETACIÓN DE LA INFORMACIÓN CLÍNICA

Presentada por:

María Meizoso García

Dirigida por:

María Jesús Taboada Iglesias

Diego Martínez Hernández

Santiago de Compostela, Septiembre de 2015

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Dña. María Jesús Taboada Iglesias, Profesora Titular de Universidad del Área de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Santiago de Compostela

D. Diego Martínez Hernández, Profesor Titular de Universidad del Área de Física Aplicada de la Universidad de Santiago de Compostela

HACEN CONSTAR:

Que la memoria titulada **TÉCNICAS DE ANOTACIÓN SEMÁNTICA ORIENTADAS A MEJORAR EL ACCESO E INTERPRETACIÓN DE LA INFORMACIÓN CLÍNICA** ha sido realizada por **Dña. María Meizoso García** bajo nuestra dirección en el Departamento de Electrónica e Computación de la Universidad de Santiago de Compostela, y constituye la Tesis que presenta para optar al título de Doctor.

Santiago de Compostela, Septiembre de 2015

María Jesús Taboada Iglesias
Codirectora de la tesis

Diego Martínez Hernández
Codirector de la tesis

María Meizoso García
Autora de la tesis

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecimentos

Primero agradecer todo lo aprendido durante este tiempo a todos los compañeros del grupo: Jose, Diego, Charo y, sobre todo, agradecer a Chus por toda la paciencia, dedicación y compresión que ha tenido que gastar en mi, tanta que creo que ya no le puede quedar mucha más.

Expresar mi agradecimiento al apoyo económico recibido durante estos años, mencionando a los proyectos que han financiado esta investigación: *OntoNeuroPhen* (FIS2012-PI12/-00373) del Instituto de Salud Carlos III, *Gestión de Terminologías Médicas para Arquetipos* (TIN2009-14159-C05-05) del Ministerio de Economía y Competitividad e *HYGIA* (TIN2006-15453-C04-02) del Ministerio de Educación y Ciencia.

A aquellos que compartimos la convivencia, comidas, cafés, risas, agobios y más: mis compañeros de piso, laboratorio y congresos. A Juan Ángel, Cris, Miguel, Julián, Jose, Josito, Adri, Víctor, Bea, . . . , y a los que no nombro, estéis más o menos lejos.

No me puedo olvidar de mis padres, mi hermana, Marcos, la madrina y el padrino. Muchas gracias.

Santiago de Compostela, Septiembre de 2015

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Índice general

Capítulo 1

Introducción

Hoy en día, algunas de las principales prioridades de los sistemas de salud nacionales y regionales incluyen la prevención de enfermedades, el incremento de la esperanza de vida, la mejora de la calidad de vida y la reducción de las admisiones en los servicios de emergencia. Con estas prioridades en mente, los servicios sanitarios tienen como objetivo adaptar los sistemas de información actuales para que posibiliten el desarrollo e integración de herramientas informáticas avanzadas que faciliten el diagnóstico precoz, el tratamiento personalizado y la vigilancia automatizada de la salud. La fragmentación de la información del paciente en diferentes lugares y formatos dificulta enormemente su acceso y procesamiento adecuados. Por ello, los sistemas informáticos sanitarios deben de ser capaces, primero, de intercambiar datos entre todas las unidades que los integran y, segundo, de tener la habilidad para interpretar la información presente en los datos que intercambian, tanto en el contexto correcto como en un tiempo razonable. Todo ello es algo que el personal sanitario realiza de forma natural hoy en día en su trabajo diario. Sin embargo, los sistemas actuales se han diseñado e implementado sin seguir, muchas veces, el flujo de trabajo usual de dicho personal. Esto dificulta su uso, llegando a entorpecer la práctica clínica diaria y suponiendo, en muchas ocasiones, una carga de trabajo excesiva. Las técnicas semánticas pueden proporcionar las herramientas necesarias para cubrir el gap actual, por un lado, entre la información que necesita compartir el personal sanitario (es decir, el intercambio de conocimiento en la organización sanitaria) y, por otro lado, entre los sistemas informáticos (es decir, el intercambio de datos a nivel técnico).

1.1. Estado Actual

Para permitir la gestión integrada de los datos de los pacientes, los sistemas de información actuales deben dotarse de los mecanismos necesarios para intercambiar e interpretar tanto datos como conocimiento sobre la salud de los pacientes, sin que dicho intercambio interfiera en el flujo de trabajo del personal sanitario [3]. La implantación de la historia clínica electrónica (HCE) en los servicios sanitarios ofrece un amplio abanico de nuevas e innovadoras posibilidades para intercambiar información entre los diferentes servicios sanitarios. El sistema actual sanitario en Galicia (Sergas) integra en el sistema de información IANUS toda la información del paciente, incluyendo episodios codificados, texto, gráficos e imágenes digitales en formato DICOM y receta electrónica. Sin embargo, dicha información se encuentra repartida entre diferentes sistemas de información que no son interoperables y que no tienen un acceso estandarizado a ella, lo que dificulta enormemente el intercambio de información entre todos los sistemas que albergan la información de los pacientes en el sistema sanitario actual. Aunque la mayoría de los sistemas de información sanitarios, como el gallego, integran gran cantidad de datos clínicos textuales, tanto no codificados como codificados, incluyendo imágenes médicas digitales, el acceso a la información sigue siendo tedioso y requiere mucho esfuerzo manual. Esto se debe, en primer lugar, a que una parte importante de la información reside en sistemas independientes y no interoperables con la HCE, como puede ser la información sobre patología anatómica y los datos de laboratorio, entre otros.

Además, de dotar de interoperabilidad a los sistemas de información sanitarios actuales, también se hace imprescindible que durante el intercambio de conocimiento y datos, éstos no pierdan su significado original para que la información se pueda procesar automáticamente, sin necesidad de la intervención humana. A esta habilidad informática se la conoce como interoperabilidad semántica ([99, 2]).

La reducción de la carga de trabajo del personal médico también juega un papel importante para incrementar el uso de los actuales sistemas de información, que se han diseñado e implementado sin seguir, muchas veces, el flujo de trabajo usual de dicho personal.

Veamos un ejemplo de dificultad de acceso a la información clínica a través de un servicio actual de vigilancia de salud, cuyo objetivo es identificar pacientes que se adaptan a un perfil clínico específico. El proceso de acceso a la información en este supuesto incluye las siguientes etapas.

El clínico primero debe escribir una lista de términos especializados que ayuden a localizar los pacientes que se adaptan al perfil (Figura 1.1). Para asegurarse el mayor éxito posible

Figura 1.1: La dificultad de acceso a la información clínica.

en la búsqueda, la lista de términos debe ser tan exhaustiva como sea factible. Segundo, el clínico debe precisar en qué recursos informáticos se encuentra dicha información. Sin embargo, el clínico no tiene claro dónde está localizada dicha información, ya que obviamente carece del conocimiento sobre la estructura informática de su organización. Tercero, el servicio informático de vigilancia realiza la búsqueda de los términos especificados por el clínico en los recursos informáticos seleccionados. Cuarto, el servicio devuelve la información tal y como se encuentra almacenada en los recursos. En resumen, los servicios actuales son laboriosos y no están adaptados al flujo de trabajo del personal clínico.

En segundo lugar, aunque la HCE está diseñada como una colección integrada de imágenes médicas y datos clínicos alfanuméricos, el personal clínico está forzado a acceder a la información útil mediante la revisión en tres lugares diferentes de la HCE, que no son interoperables entre sí: la parte de los datos estructurados, la de los datos textuales y la de imagen digital.

Teniendo esto en cuenta, el proceso de búsqueda de la información clínica en la propia HCE y en las demás colecciones no integradas en la HCE sigue un procedimiento convencional (Figura 1.2). En primer lugar, el clínico se plantea una duda, sea o no por necesidad, y cuando averigua dónde buscar, elige la documentación apropiada. A continuación, organiza la información recopilada tras la exploración de la documentación. Finalmente, identifica la mejor solución al problema a resolver y, en caso de que no quede completamente determinado, volvería a comenzar el ciclo. Por tanto, el proceso de acceso e interpretación de la información clínica incluye saber dónde están almacenados los datos y comprender qué significan los datos. En organizaciones con infraestructuras de información inmensas y complejas, como la sanitaria, la estructura de la información no puede estar almacenada en la mente del personal médico. Indexar la información de forma adecuada, tal y como se ha hecho ya con

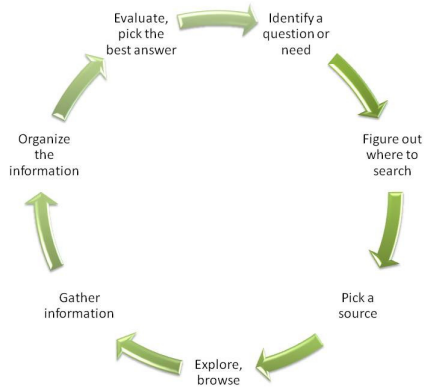


Figura 1.2: Proceso de búsqueda de información por parte del personal clínico[1].

éxito en otros entornos como el bibliotecario o la web, puede resultar adecuado. El uso de terminologías clínicas estándar, ya utilizadas por los sistemas de información para intercambiar información, puede ser la opción más destacable y de menor riesgo hoy en día para indexar los diversos recursos de almacenamiento clínico.

Actualmente, SNOMED CT se perfila como la terminología más prometedora para codificar la HCE [156]. Aún así, existen otras terminologías en uso como LOINC para codificación de los datos de laboratorio o ICD para codificación de los episodios clínicos de urgencias en España.

Uno de los principales obstáculos que impiden el acceso a la información completa del paciente en los actuales sistemas electrónicos de los sistemas sanitarios es el uso de modelos propietarios para representar la información clínica. Los estándares sobre interoperabilidad de la HCE a nivel internacional, tales como la arquitectura de información OpenEHR [119] y la norma ISO EN 13606 [34], establecen las arquitecturas para la comunicación y el intercambio de los datos de los HCE. Ambas organizaciones usan un modelo dual para separar,

– Por una parte, la información, la cual es perdurable en el tiempo y representada a través de los elementos base del modelo de referencia.

[1] `https://handsinautism.iupui.edu/inform.html`

– Por otra parte, el conocimiento, que es el contenido que evoluciona con el tiempo. Reflejado en los arquetipos que definen los conceptos clínicos a través del modelo de referencia.

Por tanto, los modelos clínicos, como los arquetipos OpenEHR, definen estructuras estandarizadas de datos con el fin de asegurar que la información clínica que se intercambie sea correcta y precisa. Además, con el uso de estos modelos clínicos la información puede estar almacenada en cualquier formato propietario. Esta independencia se consigue gracias a la vinculación existente de los elementos del arquetipo con terminologías médicas estándar. Sin embargo, solo existe un número poco significativo de conjuntos de arquetipos de acceso público que, a pesar de estar bajo actualización y mantenimiento a cargo de organismos oficiales, no poseen suficientes referencias a términos de vocabularios como para garantizar la interoperabilidad. Por este motivo, en esta tesis se realiza el esfuerzo de realizar una propuesta para la anotación de arquetipos, ya que el proceso manual equivalente requiere un gran esfuerzo por parte de especialistas clínicos.

No obstante, los servicios sanitarios no sólo demandan el acceso a la información del paciente. También requieren acceso al conocimiento sobre las recomendaciones más actuales acerca de procedimientos diagnósticos y terapéuticos. Por lo tanto, es crucial para el personal clínico el poder disponer de este tipo de información en formato estructurado y computarizado para garantizar su acceso rápido y consulta eficiente. Hay que destacar también que uno de los documentos más relevantes que recopilan este tipo de conocimiento son las guías de práctica clínica (GPC) que, a pesar de que se han intentado diseñar lenguajes formales para definirlas, la mayoría están expresadas en lenguaje natural (LN). Por estos motivos, en este trabajo hemos creado una propuesta para anotar los documentos de GPC con conceptos de terminologías estándar. Para ello, y con el objetivo de identificar evidencias en las GPC se han aplicado secuencialmente diversas técnicas de procesamiento del lenguaje natural (PLN), reconocimiento de entidades nominales (Named Entity Recognition, NER) y se han adaptando herramientas de acceso libre para su uso genérico.

1.2. Objetivos

Como ya hemos comentado, el proceso de búsqueda de la información clínica en la propia HCE, y en las demás colecciones no integradas en la HCE, sigue un procedimiento convencional que consume recursos humanos y no está adaptado al flujo de trabajo del personal clínico.

Para agilizar este proceso de búsqueda, en esta tesis se propone anotar semánticamente las diferentes colecciones de información clínica, usando las terminologías más apropiadas. De esta manera, la etapa de localización del recurso que contiene la información (Figura 1.2) sería un proceso completamente automatizado, y la búsqueda de la información en el recurso especificado vendría apoyada por la anotación semántica de dicho recurso y el uso de las ontologías o terminologías empleadas para dicha anotación. Nuestra propuesta, por tanto, está orientada a gestionar de forma eficiente la gran cantidad de información existente hoy en día en los sistemas de información clínica.

Dado que en un trabajo de tesis es imposible validar nuestra propuesta para cada uno de los recursos disponibles hoy en día en un sistema sanitario, en esta tesis nos centraremos en dos recursos. El primero es la HCE del paciente, que hoy en día se ha convertido en el núcleo central de los sistemas de información sanitaria, ya que se considera una pieza clave para la prestación eficiente y de calidad de los servicios sanitarios. En particular, haremos hincapié en la información descriptiva del paciente, dejando para un futuro la incorporación de la imagen digital. Como ya hemos comentado previamente, los arquetipos clínicos son modelos que guían la introducción y la presentación de la información clínica, por lo que se considera que son un activo de conocimiento para la recogida estandarizada de la información de la HCE. En 2011, se constituyó una iniciativa internacional llamada CIMI (Clinical Information Modeling Initiative) para potenciar un formato común para el intercambio de modelos de contenido clínico. El objetivo de dicha iniciativa es crear un conjunto de modelos clínicos validados (curated) que pueda ser implementable en cualquier sistema de información sanitario [1]. Inicialmente se ha acordado utilizar la especificación de arquetipos de ISO EN 13606 y OpenEHR como formato de representación de los modelos clínicos. Dado que los arquetipos clínicos ISO EN 13606 y OpenEHR son muy similares, hemos optado por centrar nuestros esfuerzos, que son limitados, en un solo formato. En particular, hemos seleccionado OpenEHR, puesto que, a diferencia del resto, persigue una representación completa de la HCE.

El segundo recurso estudiado en esta tesis doctoral son la guías de práctica clínica (GPC), porque constituyen una fuente sustancial e importante de conocimiento sobre las recomendaciones diagnósticas y terapéuticas basadas en la evidencia. El acceso a este conocimiento en el punto de atención al paciente es crucial para mejorar la calidad de la asistencia sanitaria y reduce costes innecesarios ([46, 139]).

El objetivo general de esta tesis es proporcionar técnicas semánticas orientadas a dotar de interoperabilidad a los sistemas de información sanitarios actuales y facilitar el proceso

de búsqueda semántica de la información, reduciendo de esta manera la carga de trabajo del personal médico. Este objetivo general puede desglosarse en dos objetivos específicos:

1. **Diseño, desarrollo e implementación de un conjunto técnicas para la anotación semántica automatizada de los modelos de datos clínicos** orientados a formalizar, normalizar y compartir la recopilación de la información del paciente que se almacena en la HCE. En el momento de desarrollo de la tesis no existían métodos automatizados para anotar arquetipos con SNOMED CT, aunque se habían desarrollado ya algunos enfoques en el marco de diferentes proyectos de investigación. La mayoría de las propuestas suponían puntos de vista parciales al problema de la anotación semántica automatizada. Por ello, el primer objetivo de esta tesis doctoral fue desarrollar un conjunto de técnicas que conjuntamente permitiesen la anotación semántica automatizada de modelos de datos clínicos con ciertas garantías de validez. Para llevar a cabo este objetivo, hemos planteado una aproximación basada en la combinación de dos métodos básicos de equiparación o mapping (conocidos como métodos léxicos y basados en contexto) con el fin de producir la anotación semántica, en primer lugar, y posteriormente validarla. Las técnicas léxicas encuentran anotaciones utilizando la sinonimia de la ontología utilizada, mientras que las técnicas basadas en el contexto identifican similitudes semánticas entre la estructura del arquetipo y las relaciones en la ontología.

2. **Diseño, desarrollo e implementación de un conjunto técnicas para la anotación semántica automatizada de guías de práctica clínica**. Es frecuente el uso de técnicas de Procesamiento de Lenguaje Natural (PLN) para analizar de forma automática los textos referentes a registros e informes de pacientes. Sin embargo, la mayoría de los métodos utilizados han sido desarrollados para sistemas específicos, y su adaptación a una nueva aplicación requiere una gran inversión en recursos humanos. Por lo tanto, se necesita una nueva investigación para determinar si tales métodos se pueden reorientar hacia nuevas aplicaciones y metas, obteniendo el mismo rendimiento. Con el fin de proporcionar acceso al conocimiento de las GPC basadas en la evidencia sobre procedimientos diagnósticos y terapéuticos, en el punto de atención al paciente, en esta tesis doctoral nos propusimos anotar semánticamente y de manera automatizada GPCs. La idea principal era enriquecer estos documentos con una ontología, para hacerlos interpretables computacionalmente. El enfoque propuesto en esta tesis doctoral consiste en elaborar y aplicar una combinación secuencial de varios métodos utilizados tradicionalmente en

PLN, para anotar gradualmente las frases relevantes de la GPC con conceptos de una ontología.

1.3. Estructura de la Memoria

En esta sección, se hará un resumen por capítulos del contenido de esta memoria.

1.3.1. Capítulo 2 Fuentes de Información

Para conseguir la interoperabilidad en todos los sistemas clínicos, es necesario realizar un análisis de los tipos de fuentes de información utilizadas por el personal clínico desde un punto de vista que implique la posibilidad de ser estandarizados y procesados por una computadora. Por este motivo, examinamos los recursos de información médica en función de su organización y establecemos tres categorías con respecto a la estructura del recurso clínico: sin estructura, semiestructurado y estructurado.

Consideramos fuente no estructurada a aquella que está expresada en LN. Lo que implica como ventaja una gran expresividad, ya que se permite emplear siglas y abreviaturas no estándar, oraciones incompletas, la escritura a mano, ... No obstante, y debido a esa expresividad, es la que precisa del procesado más complejo para la extracción del conocimiento. Además, se trata del recurso más abundante. Ejemplos de estas fuentes son los informes creados por los médicos o las GPC.

Cuando la información está organizada de alguna forma que no implique una relación semántica entre sí pero se facilita el acceso a ella, decimos que esa fuente es semiestructurada. En este nivel organizativo, la localización de la información y el procesado es más sencillo. Los datos pueden estar simplemente ordenados según algún criterio como el alfabético o agrupados en base a algún principio. Ejemplos de estos recursos son los arquetipos o los vocabularios especializados en un dominio concreto.

Finalmente, estarían las fuentes de información estructurada. En ellas cada concepto constituyente es tomado como unidad de conocimiento y está definido en función a unas características propias y a la forma en la que se relaciona con otros conceptos. Las ontologías son los modelos más significativo para este tipo de recurso. Existen gran variedad de ellas centradas en áreas específicas, lo que implica que puede haber contenido que se solape pero que, sin embargo, esté definida desde puntos de vista diferentes y con distintas características y/o granularidades.

Realizamos esta clasificación con el objetivo de conseguir la computabilidad y interoperabilidad de los recursos documentales y ser integrados con la mayor facilidad posible en los sistemas clínicos. Por ello, es necesario extraer el conocimiento de los recursos menos estructurados, gracias a la anotación con elementos de las fuentes completamente estructuradas, estándar y de uso más generalizado.

1.3.2. Capítulo 3 Técnicas de Reconocimiento de Entidades y Relaciones

En función de las características de estructuración de las fuentes de información descritas en el capítulo 2, es necesario aplicar la técnica de extracción de conocimiento adecuada o una combinación de ellas. Por este motivo, en este capítulo se han estudiado todas las tácticas para identificar los conceptos y las relaciones.

Primero, debido a que el procesado completamente automático del LN es complicado y a que los recursos estructurados necesitan pasar por un proceso inicial de organización durante su creación, es necesario aplicar técnicas que impliquen la interacción con el personal experto, es decir, las manuales. Estas técnicas permiten mejorar y no arrastrar errores en pasos posteriores del procesamiento del LN, así como, garantizar la coherencia y la precisión de los recursos.

En segundo lugar se comentan las técnicas no manuales clasificadas en elementales y relacionales. Las primeras son las tácticas aplicadas sobre los términos y entidades que constituyen la fuente de información de forma aislada. Estas técnicas se destinan a las características individuales de cada elemento, como por ejemplo, la nomenclatura, la clasificación, tipo de datos, atributos concretos ... Las relacionales son aquellos métodos que tienen en cuenta cómo esas entidades se relacionan con las otras dentro de la fuente: entidades más generales y específicas, elementos constituyentes, relaciones lógicas, el contexto en el que se enmarca el término ....

Finalmente, se describen brevemente una serie de estrategias para poder escoger el candidato más apropiado dentro del conjunto resultante de aplicar las técnicas anteriores (desambiguación) y ejemplos de diferentes combinaciones eficientes de esas tácticas para obtener el mejor resultado.

La segunda parte de este capítulo se centra en comentar las herramientas más eficientes y utilizadas en el proceso de reconocimiento de entidades. Primero se describen aquellas centradas en el análisis sintáctico que es el primer paso en el proceso de reconocimiento de

entidades y nos permite configurar la identificación de información desde el nivel más bajo. El otro tipo de herramientas son aquellas que resuelven el problema completo, bien sea porque de forma independiente son capaces de recuperar las entidades, así como las relaciones, incluidas en un texto, o bien, porque una combinación de ellas lo permiten, es decir, conforman distintas capas en la resolución del problema.

En resumen, este capítulo nos proporciona la documentación suficiente para poder resolver el problema de reconocimiento de entidades y relaciones.

1.3.3. Capítulo 4 Una Propuesta para la Anotación Semántica de Modelos de Datos Clínicos

En este capítulo comienza la primera parte del trabajo desarrollado para esta tesis para conseguir la interoperabilidad entre sistemas clínicos. Para ello se han escogido los arquetipos como un ejemplo de fuente semiestructurada de modelos clínicos que almacenan información sanitaria sobre el paciente (historia, pruebas, medicación, ...). Este tipo de modelo permite la estandarización hacia otros modelos o sistemas través del uso de terminologías médicas, es decir, asocia cada elemento del contenido del arquetipo con conceptos de un vocabulario estándar. SNOMED CT es el vocabulario que hemos escogido como destino en las anotaciones por ser el que mejor se ajusta a la HCE, tratándose además de una fuente estructurada, estándar y ampliamente utilizada.

Debido al esfuerzo que precisa un procedimiento de anotación manual, se ha realizado una propuesta para su realización de forma automática. Esto se ha conseguido gracias a utilizar las técnicas y herramientas de reconocimiento de entidades y relaciones explicadas en el capítulo 3. También se explican las condiciones bajo las que se realiza el experimento: cómo se realiza la selección del conjunto de arquetipos, las premisas sobre las que se basa, tecnologías que se utilizan para llevarlos a cabo, validación de resultados, análisis de estos ...

1.3.4. Capítulo 5 Una Propuesta para la Anotación Semántica de Guías de Práctica Clínica

Para seguir estudiando la interoperabilidad, después del trabajo para la anotación de fuentes de modelos de datos clínicos semiestructurados, solamente falta un recurso por analizar para extraer la información contenida en él. Este último tipo de recursos son los expresados en LN, es decir, aquellos que no poseen ningún tipo de estructuración. Como fuentes clínicas

expresadas en LN hemos escogido las GPC, las cuales marcan las pautas al personal sanitario a la hora de realizar un diagnóstico, establecer las pruebas a realizar, asignar un tratamiento ...

En el trabajo expuesto en este capítulo, se ha vinculado el contenido de las GPC sobre recomendaciones diagnósticas y terapéuticas con conceptos de la terminología UMLS Metathesaurus y con las relaciones existentes en la Semantic Network de UMLS. Para conseguir estos vínculos, se han utilizado herramientas especializadas en el reconocimiento de entidades de acceso libre y en las técnicas comentadas en el capítulo 3. Estas tecnologías se han combinado para obtener los mejores resultados y los más precisos. También se incluye un estudio sobre la evolución de las GPC en el tiempo, en cuando a la influencia que puede tener sobre la información relevante y que es objeto de procesado. Por lo tanto, en esta parte de la memoria está dedicada a explicar esa adaptación y acoplamiento de tecnologías, los problemas encontrados y resultados obtenidos.

1.3.5. Capítulo 6 Conclusiones y Líneas Futuras

El último capítulo está dedicado a realizar una reflexión sobre los trabajos realizados en esta tesis doctoral, sus limitaciones y aportaciones, mejoras que se podrían realizar, así como posibles investigaciones posteriores.

1.4. Publicaciones

A continuación, se realiza una enumeración de los artículos publicados durante el transcurso de este trabajo:

Publicaciones en revistas JCR

- María Taboada, María Meizoso, Diego Martínez, David Riaño y Albert Alonso. "Combining open-source natural language processing tools to parse clinical practice guidelines". *Expert Systems* 30, no. 1: 3-11, 2013. Categoría: *Computer science. Theory & methods*. Índice de impacto JCR 2013: 0,75. Posición 55/102 (Cuartil 3)

- María Meizoso, Jose L. Allones, Diego Martínez y María Taboada. "Semantic similarity-based alignment between clinical archetypes and SNOMED CT: an application to observations". *International Journal of Medical Informatics* 81, no. 8: 566-578,

2012. Categoría: *Computer Science. Information Systems*. Índice de impacto JCR 2012: 2,411. Posición 18/132 (Cuartil 1)

Publicaciones en congresos

- Jose L. Allones, María Meizoso, María Taboada, Diego Martínez y Serafín Tellado. "Combining lexical and structure-based methods to align clinical archetypes to SNOMED CT". En *16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*. Disponible en Advances in Smart Systems Research, Workshop Papers from KES Conferences, Vol. 2 No. 1, págs. 27-32. Future Technology Publications, 2012.

- Jose L. Allones, María Taboada, María Meizoso, Diego Martínez y Serafín Tellado. "Combining mapping methods to align clinical archetypes to SNOMED CT". En *10th Terminology and Knowledge Engineering Conference (TKE 2012)*. Madrid, España, 2012.

- María Meizoso, Jose L. Allones, María Taboada, Diego Martínez y Serafín Tellado. "Automated mapping of observation archetypes to SNOMED CT concepts". En *4th International Conference on Interplay between Natural and Artificial Computation*. Disponible en Foundations on Natural and Artificial Computation, págs. 550-561. Springer Berlin Heidelberg, 2011.

- María Taboada, María Meizoso, Diego Martínez y Serafín Tellado. "A Study of Extracting Knowledge from Guideline Documents". En *12th International Conference on Computer Aided Systems Theory*. Disponible en Computer Aided Systems Theory - EUROCAST 2009, págs. 195-202. Springer Berlin Heidelberg, 2009.

- María Taboada, María Meizoso, Diego Martínez y José J. Des. "Using Lexical, Terminological and Ontological Resources for Entity Recognition Tasks in the Medical Domain". En *3rd International Conference on Interplay between Natural and Artificial Computation*. Disponible en Knowledge Management for Health Care Procedures, págs. 21-31. Springer Berlin Heidelberg, 2008.

Capítulo 2

Fuentes de Información

Dentro de este capítulo analizaremos las diferentes fuentes lingüísticas existentes en el ámbito clínico. Estos recursos se clasificarán en función de su organización interna de la información, es decir, de cómo están estructurados.

2.1. Fuentes No Estructuradas

Consideramos fuentes no estructuradas aquellas en las que la información está expresada en lenguaje natural (LN), es decir, en donde el procesamiento computacional directo no es posible sin realizar algún tipo de preprocesado previo.

2.1.1. La Historia Clínica Expresada en Lenguaje Natural

Los registros electrónicos médicos contienen principalmente los datos clínicos del paciente que incluyen la historia personal y familiar, el estado clínico, los tratamientos dispensados, y otra información relevante sobre los diagnósticos y los resultados de pruebas. El uso de LN para describir esta información proporciona un nivel completo de expresividad, pero dificulta el procesamiento computacional, lo que interfiere con uno de los principales retos de la sanidad electrónica [56], concretamente con la interoperabilidad semántica de los sistemas de salud.

Aunque una parte importante de los registros médicos almacenan información estructurada, una proporción significante existe en formato texto escritos en LN, es decir, como información no estructurada. El LN contiene abundantes particularidades, es muy dependiente

del contexto y posee expresiones ambiguas, especialmente las abreviaturas y los acrónimos del ámbito clínico. Es frecuente encontrarnos con errores ortográficos, la falta de aplicación o aplicación incorrecta de las reglas de la gramática, o directamente con oraciones incompletas. Esta diversidad en el estilo de escritura, en la colocación de los signos de puntuación y en la exactitud con que se describe un elemento es un problema a la hora de extraer y estandarizar el contenido de un documento.

Uno de los modos más comunes de recoger información no estructurada clínica son los exámenes médicos escritos a mano. Apkon y Singhaviranon [4] realizaron un estudio comparativo entre la recogida de información médica de forma no estructurada (informes escritos a mano) o de forma estructurada (informatizada). Este estudio demostró que los documentos electrónicos no solo reducen el tiempo de recogida de información y de la composición de notas, si no que también permiten recoger una mayor cantidad de datos. Estas observaciones se confirman sobre diferentes campos del ámbito clínico como el servicio de urgencias [10] en el cual el acceso a la información del paciente en un pequeño lapso de tiempo es imprescindible. Además, este trabajo incluye un análisis en el que participa el personal clínico sobre los pros y contras de los registros electrónicos y en papel. Otro ámbito clínico que se mejora gracias a la recolección de información electrónicamente son los registros dentales, para los que Schleyer et al. [143] además realiza un análisis de la estructuración de los datos que se deben recoger. Los anteriores autores coinciden con Roukema et al. [140], los cuales afirman que la recogida de información en papel (no estructurada) implica escritura ilegible, información ambigua e incompleta, fragmentación de datos, poca disponibilidad de estos y la demanda de tiempo en caso de revisión. También defienden la idea de que registros electrónicos estructurados permiten recoger una mayor cantidad de información. Sin embargo, debido a que la codificación limita la expresividad, consideran que se debería mantener parte de la información en lenguaje natural. En el trabajo de Macedo et al. [88] podemos observar una comparación exhaustiva, más allá de la estructuración, entre una historia clínica electrónica (HCE) y una historia clínica tradicional A.1.

Historia Clínica Electrónica

La ISO [63] define la HCE como:

> Un repositorio de información sobre el estado de salud de un sujeto de atención que es procesable computacionalmente. Son almacenados y transmitidos de forma segura, y accesibles por múltiples usuarios autorizados.

Tiene un modelo de información lógica estándar o de común acuerdo, que es independiente de los sistemas de HCE.
Su objetivo principal es dar soporte a una atención sanitaria integrada, continuada, eficiente y de calidad que refleja información pasada, presente y prospectiva.

Después de comentar los trabajos anteriores, deducimos que algunos de los motivos por los que se comienzan a utilizar las HCE son los siguientes:

- Gestionar la información de los cuidados de la salud que cada vez es más compleja.
- Conectar localizaciones de atención de pacientes separadas geográficamente.
- Permitir compartir información al equipo de atención clínica.
- Proporcionar cuidados basados en evidencias.
- Reforzar la salud de la población y la investigación.
- Mejorar la seguridad y la integridad de la información (datos duplicados, datos erróneos, . . .).
- Reducir el coste de los servicios sanitarios y gestionar los recursos de forma más efectiva.
- Respaldar la salud de la población y la investigación.
- Involucrar a los ciudadanos.
- Proteger la privacidad de los ciudadanos y hacerlos partícipes.

En los comienzos de la HCE, las primeras versiones solían almacenar la siguiente información generalizada:

- Registros de información clínica almacenados de forma anidada.
- Autores de las mediciones/observaciones realizadas al paciente.
- Fechas y tiempos asociados a los detalles de los eventos ocurridos.
- Identificadores y gestión de las versiones de los documentos.
- Unidades (tipos de datos) que se utilizan en las mediciones.
- Control de acceso (roles) dependiendo del tipo de información.

2.1.2. Guías de Práctica Clínica

Uno de los documentos clínicos no estructurados más importantes son las guías de práctica clínica (GPC). Las GPC constituyen una recurso notable de contenidos sobre recomendaciones diagnósticas y terapéuticas basadas en la evidencia y sobre el conjunto de acciones a llevar a cabo ante una situación particular; así como, sobre los datos requeridos del paciente, decisiones a tomar, restricciones entre tareas, ... Según Turner [167] de 1993 a 2006 los artículos relacionados con las guías clínicas incrementaron 11 veces su número. Field y Lohr [68] dan una definición estándar de las GPC:

> Las GPC son afirmaciones (desarrolladas sistemáticamente) para asistir a los médicos en las decisiones sobre el cuidado apropiado de pacientes bajo circunstancias clínicas específicas.

Teniendo en cuenta esto, Woolf et al [173] enuncia las siguientes afirmaciones a tener en cuenta para el desarrollo y uso de las GPC:

- Son una opción para mejorar la calidad de la atención al paciente ya que cada vez son más recurridas durante la práctica clínica.
- Tienen beneficios y perjuicios potenciales. Por lo que es necesario realizar un desarrollo riguroso que garantice un número mínimo de daños.
- Las guías pueden reducir el número de cuidados inapropiados a pacientes e introducir nuevo conocimiento a la práctica clínica puesto que incluyen las mejores prácticas actualizadas.
- Sin embargo, no solo es necesario que la GPC esté bien definida, si no que se pueda llevar a cabo. Esto se consigue añadiendo una base científica y, en la medida de los posible, la ejecución de la guía en un entorno real que garantice su efectividad.

Shiffman [136] define la guía como un documento vivo, es decir, que está constantemente ajustando las bases para la evaluación de los resultados de los pacientes así como los criterios para determinar el rendimiento.

El personal clínico podría sacar partido de los sistemas informáticos si se informatizara el conocimiento contenido en las GPC. Tanto Grimshaw y Russell [46] como Rosser [139] han reconocido indiscutiblemente que el acceso a este conocimiento en el punto de atención

al paciente mejoraría la calidad de la asistencia sanitaria y reduciría costes innecesarios. Sin embargo, el formato no estructurado de estas las convierten en poco apropiado para la representación formal. Las guías no solo presentan los problemas típicos del procesado del lenguaje natural (PLN) que pueden influir directamente en su propio tratamiento, si no que también nos podemos encontrar problemas de contenido que afectan a la calidad de la información que se extrae, por ejemplo, el incumplimiento de los estándares de desarrollo y formato de guías, fallos en la identificación y el resumen de las evidencias, así como en la formulación de las recomendaciones médicas.

Verificación de Guías de Práctica Clínica

Como comentamos, la otra parte de la investigación sobre GPC está destinada a fijar unos criterios para la evaluación de su contenido, la cual, no siendo el objetivo directo de este trabajo, se debe tener en cuenta porque influye en el procesado de las guías la forma en la que estén enunciadas y organizadas. El libro de Lohr y Field [24] recoge recomendaciones para consolidar el proceso de desarrollo y el uso de las guías, ejemplos de guías, pautas para la evaluación de la robustez de las mismas y seis casos de estudio de profesionales que utilizan las GPC.

Trabajos como los de Grol [47] y de Graham [45] establecen indicadores sobre la validez del contenido de las guías. En el caso de Graham, marca ocho atributos: validez, fiabilidad y reproducibilidad, aplicabilidad clínica, flexibilidad clínica, claridad, documentación, desarrollo por parte de un proceso multidisciplinario y los planes para su revisión. Más tarde, la Conference on Guideline Standardization (COGS[1]) intenta definir los elementos clave que debe incluir una GPC. Según Shiffman [136], se trata de definir un estándar para las guías que potencie su calidad y que facilite la implementación. Además, Shiffman elaboró una lista de dieciocho contenidos necesarios: visión general, foco, meta, usuarios y marco, características de la población sobre la que se aplica, datos del desarrollador, financiamiento, conjunto de evidencias médicas, criterios de clasificación de recomendación, métodos para sintetizar evidencias, revisión previa a la publicación, plan de actualización, definiciones, sugerencias y justificaciones, daños y beneficios potenciales, preferencias de pacientes, algoritmo e implementación del estudio. Vlayen [169] realiza una comparativa entre diferentes herramientas que valoran guías clínicas incluyendo el proyecto AGREE (Appraisal of Guideline Research and Evaluation) [165]. Otro estudio similar es el de Turner [167] que compara seis manuales

[1] gem.med.yale.edu/cogs/

para el desarrollo de GPC de los que extrae 14 elementos a tener en cuenta durante el proceso de desarrollo. Simera [154] propone un catálogo[2] de recursos, educación y entrenamiento para realizar buenos informes de investigación sobre la salud y para asistir al desarrollo, divulgación e implementación de guías de informes robustas. La necesidad de desarrollar y evaluar las guías de forma correcta sigue siendo una necesidad a día de hoy como se muestra en el trabajo de Sierig [151].

Hasta ahora, los trabajos que hemos comentado exponen la necesidad de incluir una base científica en el desarrollo de las GPC. Sin embargo, el trabajo de Shiffman, Michel et al [150] comenta 5 problemas que influyen en la aplicación de las GPC:

– defectos en la calidad de la guía debido a la falta de documentación,

– lenguaje impreciso, puesto que en ocasiones no es sencillo la identificación de las recomendaciones expuestas en la guía. Las recomendaciones deben ser decidibles (las condiciones para llevar a cabo una recomendación deben ser claras) y ejecutables (las acciones de la recomendación deben estar especificadas con determinación). Shiffman especifica situaciones que pueden llevar a ambigüedad como el uso de pasivas o la palabra *consider*. Estas consideraciones son importantes para el PLN y la extracción del conocimiento.

– Deficiencias en la extracción de conocimiento de una recomendación: cuando se propone la ejecución de una recomendación los beneficios, riesgos, costes y daños que se pueden producir deben estar manifiestos.

– Implementación ineficaz y difícil comprensión.

– Problemas a la hora de la formalización de las guías. Patel [124] y Ohno-Machado [118] comentan su experiencia en la representación de guías.

El Esfuerzo de Crear Guías de Práctica Clínica Computables

La conversión de las GPC a formato electrónico ofrece ventajas importantes a la hora de su procesado. Ejemplos de ello son, por ejemplo, la facilidad de consulta sobre el contenido de las guías, la rápida distribución electrónica o la ayuda que ofrecen a los médicos a la hora de toma de decisiones.

[2] `http://www.equator-network.org`

Debido a los problemas que acarrea la falta de estructuración Schiffman, Michel et al [150] han propuesto una solución que traduce las recomendaciones del médico expresadas en LN a un lenguaje natural controlado. Esta conversión tiene por fin implementar sistemas de soporte a la decisión[3] utilizando lógica de primer orden. El lenguaje controlado ACE (Attempto Controlled English) es un un subconjunto de la lengua inglesa con restricciones en vocabulario y gramática. Estas restricciones mejoran la consistencia terminológica, reducen la ambigüedad, manejan un vocabulario congruente, las frases se crean en base a plantillas, y simplifican las estructuras de oración y del texto. Además, existe la posibilidad de traducir textos escritos en ACE a lógica de primer orden obteniendo así un resultado más computable y apto para el razonamiento automático. El principal problema de ACE es la limitación de su léxico y de las reglas de la gramática. Esta acotación hace que no se tengan en cuenta vocabulario específico del ámbito médico o que no se puedan expresar correctamente los diferentes niveles de obligaciones impuestas en las recomendaciones médicas. En este trabajo no incluye ningún mecanismo de control del contenido de las guías como tener en cuenta los puntos de verificación fijados por la COGS o AGREE.

Anteriormente a ACE, los ingenieros del conocimiento, con la asistencia de expertos clínicos, realizaron la tarea de crear otros lenguajes formales que tenían por objetivo la traducción de texto médico para poder representar GPC de forma electrónica. Tal es el caso de los lenguajes comentados por Clercq y otros [23]: Asbru, EON, PROforma, Guideline Interchange Format (GLIF), ... Sobre estos lenguajes se crearon una serie de herramientas para el procesado de GPC como las analizadas en Isern y Moreno [62]: AsbruView ([80]), Arezzo ([55]), Tallis ([157]) o Protege ([44]), entre otros. Sin embargo, a pesar de la asistencia proporcionada por estos sistemas, la complejidad y el esfuerzo que supone la obtención manual del conocimiento contenido en las GPC es notable. Por ello, es indiscutible la necesidad de técnicas que ayuden a la automatización de adquisición del conocimiento de los textos de las GPC.

Shiffman [150] también tiene en cuenta el problema de la ambigüedad en la implementación clínica de las guías. En ocasiones, los autores de las guías, a pesar de intentar reflejar evidencias con una precisión científica, introducen ambigüedad en las recomendaciones intencionadamente con el objetivo de reflejar incertidumbre. En concreto Shiffman habla de términos de las guías que no están definidos de una forma clara, lo que provoca que sea más difícil integrarlas en sistemas electrónicos y que esta incertidumbre no influya negativamente

[3] Sistema que compara información médica del paciente con una base de conocimiento para guiar al personal clínico ofreciendo sugerencias focalizadas en la situación específica del paciente.

en la toma de decisión del médico. Es tarea de los desarrolladores hacer un esfuerzo para resolver esa indeterminación. De esta reflexión podemos concluir la dificultad y necesidad de realizar un alineamiento correcto del contenido de la guía con sistemas terminológicos.

El último esfuerzo por la estructuración de las GPC es GEM (Guideline Elements Model) [149] el cual proporciona una metodología para tal fin. Estamos hablando de un modelo para documentos de práctica clínica basado en XML (eXtensible Markup Language). GEM es una estándar ASTM[4] que almacena y organiza toda la información heterogénea contenida en los documentos de las GPC. El objetivo de este modelo es facilitar la traducción de documentos en LN a un formato estándar entendible por una computadora. Además, no es necesario tener conocimientos de programación para poder trabajar con GEM. Un punto fuerte de este estándar es que permite almacenar información sobre las guía y su proceso de desarrollo de forma estructurada a cualquier nivel de abstracción, así como codificar las recomendaciones médicas. GEM está diseñado para ser usado durante todo el ciclo de vida de las GPC: desarrollo, divulgación, implementación y mantenimiento; y para reflejar el propósito de los desarrolladores. El editor GEM Cutter[5] nos permite trabajar con este lenguaje.

2.2. Fuentes Semiestructuradas

Consideramos fuentes semiestructuradas a aquellas en las que la información está contenida en elementos sin relaciones que los vinculen entre sí pero que, sin embargo, están organizados de alguna forma que facilite el acceso o gestión de la fuente.

2.2.1. Vocabulario

El primer ejemplo de recurso semiestructurado son los vocabularios. Los diccionarios tradicionales tal y como los conocemos están pensados para ser leídos por personas, no por máquinas. Por este motivo es necesario crear vocabularios que combinen información lexicográfica y computación moderna.

Un vocabulario se define como un conjunto de pares $W(f,s)$, donde f es un conjunto de caracteres de un alfabeto finito, y s es un elemento dentro de un conjunto de significados. A f se le conoce como palabra de un lenguaje y, por lo tanto, un diccionario es una lista de palabras ordenadas alfabéticamente.

[4] American Society for Testing Materials: `http://www.astm.org/`
[5] `http://ycmi.med.yale.edu/GEM/`

Un vocabulario electrónico puede crearse bien, de forma automática escaneando las páginas de un diccionario en papel y procesando esas imágenes, o bien, de forma manual ([37]). El método automático permite la creación de un vocabulario más amplio y de una forma más rápida. Sin embargo, crear un vocabulario desde cero y a mano es un proceso lento, laborioso, caro y obtendremos un diccionario con menos entradas. La gran ventaja de hacerlo a mano es que se puede este se puede ampliar con contenido que puedan requerir diferentes aplicaciones.

2.2.2. Lexicón Computacional

Comencemos por la definición de base de datos léxica. Una base de datos léxica es un sistema de almacenamiento de información lingüística organizada según un determinado modelo de datos que posibilita el almacenamiento, recuperación y modificación de los mismos. El modelo de datos puede tener una estructura jerárquica, de red o relacional, siendo esta última la de más amplia difusión. Por lo tanto, sería un recurso más estructurado que un simple vocabulario. Una base de datos léxica tiene la función de responder a consultas sobre los datos que contiene, ya sea desde prestaciones propias o a partir de aplicaciones externas, permitiendo la reutilización de la información contenida. El comportamiento de una base de datos léxica es pasivo, pues las operaciones sobre sus datos son realizadas por aplicaciones que deben ser iniciadas explícitamente.

Formalmente las bases de datos no están planeadas para guardar información compleja sino grandes volúmenes de datos, con lo cual su aplicación en la representación de la información léxica dificulta la visión de conjunto debido a la tendencia a la atomicidad de los datos almacenados. El concepto de atomicidad implica que un elemento sea atómico (o escalar), cuando no puede fragmentarse en partes más pequeñas. Por ejemplo, en la codificación de un número telefónico, si se opta por codificar la información en tres valores separados (prefijo internacional, prefijo interprovincial y número del abonado), se posibilita una gestión impensable en caso de que todo el número se tomara como un único valor. Si el número telefónico es segmentado según los distintos prefijos que lo componen, un programa de comunicaciones que marque el número y establezca la conexión en forma automática puede utilizarlo. Por tanto, el concepto de atomicidad no es consecuente con las características de la información léxica.

Podemos decir que WordNet es la base de datos léxica de referencia. WordNet define las formas f como un conjunto de caracteres ASCII y los significados s como un conjunto de sinónimos que denotan el mismo concepto (*synset*). Por lo tanto, además de incluir las ca-

racterísticas de vocabulario que comentamos anteriormente, agrupa por categorías sintácticas (sustantivos, verbos, adjetivos, adverbios y elementos funcionales) y las variantes morfológicas para cada categoría sintáctica. WordNet tiene en cuenta varias relaciones semánticas entre palabras y significados pero siempre enfocado desde un punto de vista de herencia léxica basada en las relaciones de hiponimia[6] y hiperonimia[7].

Las bases de datos léxicas son utilizadas en la Lingüística como fuentes de información léxica a reutilizar por otros recursos, por ejemplo un lexicón computacional o una base de datos terminológica [5].

Un **lexicón computacional** almacena y caracteriza formalmente el conocimiento lingüístico a través de reglas en cada uno de sus niveles de análisis (fonológico, morfológico, sintáctico, semántico y pragmático) que le permiten realizar inferencias. Moreno Ortiz [122] define el lexicón computacional como:

> Repositorios de información léxica elaborados con el objeto de servir de soporte representacional a diversas aplicaciones en el ámbito de las tecnologías del lenguaje humano.

La finalidad de este recurso es ofrecer información léxica para usuarios no humanos; sus usuarios finales son los sistemas de PLN que adoptan un enfoque basado en conocimiento (knowledge-based) que necesitan incorporar conocimiento lingüístico explícito y un conocimiento de carácter general para realizar una tarea específica [5].

SPECIALIST Lexicon

El SPECIALIST Lexicon ([93, 97]) es uno de los tres recursos de conocimiento del proyecto Unified Medical Language System (UMLS[8]) desarrollado por la US National Library of Medicine (NLM[9]). El lexicón SPECIALIST ha sido creado para ofrecer la información léxica necesario para el sistema de PLN SPECIALIST. Cuando hablamos del lexicón SPECIALIST, estamos hablando de un lexicón completo tanto biomédico como de lengua inglesa, es decir, el SPECIALIST abarca tanto vocabulario general en lengua inglesa como vocabulario específico de biomedicina.

[6] Relación existente entre una palabra cuyo significado está incluido en los significados de otras palabras más específicas denominadas hipónimos. Por ejemplo, *disease* es hiperónimo de *diabetes, tuberculosis, type 1 diabetes*.

[7] Relación existente entre un hipónimo con otra palabra en cuyo significado se encuentra englobado el del hipónimo. Hipónimos de *symtoms* serían *weight loss, spirometric abnormality, hyperglycemia, hypotension*.

[8] `https://uts.nlm.nih.gov/home.html`

[9] `http://www.nlm.nih.gov/`

Para cada entrada o término del lexicón también se almacena la información sintáctica, morfológica y ortográfica necesaria para las tareas de PLN. Cada entrada puede estar constituida por una o varias palabras. Cada registro léxico tiene una forma base asociada a una categoría gramatical, un identificador único y opcionalmente un conjunto de variantes ortográficas. La forma base es la raíz del elemento léxico, el singular en caso de un nombre, el infinitivo en caso de un verbo y la forma positiva cuando se refiere a un adjetivo o adverbio.

La información léxica incluye categoría gramatical, variaciones morfológicas (por ejemplo, nombres en singular y plural, la conjugación de los verbos, el positivo, comparativo y superlativo de adjetivos y adverbios), y otros patrones complementarios (por ejemplo, los objetos u otros parámetros que suelen acompañar a ciertos verbos, nombres y adjetivos). El lexicón reconoce once categorías gramaticales: verbos, nombre, adjetivos, adverbios, auxiliares, modales, pronombres, preposiciones, conjunciones, complementos y determinantes. En este extracto del SPECIALIST podemos ver las características comentadas:

```
{base=anesthetic spelling_variant=anaesthetic entry=E0354094
   cat=noun variants=reg variants=uncount}
{base=anesthetic spelling_variant=anaesthetic entry=E0330019
   cat=adj variants=inv position=attrib(3) position=pred stative}
```

Los patrones básicos de la oración de un lenguaje están determinados por el número y naturaleza de los complementos que acompañan al verbo. El lexicón reconoce cinco patrones de complemento: intransitivo, transitivo, ditransitivo, conector y transitivo complejo. Las entradas verbales también recogen las formas conjugadas (las partes principales del verbo). Según la conjugación, los verbos se clasifican en regulares, regulares Greco-latinos o irregulares. Las entradas de lexicón referentes a nombres incluyen variantes ortográficas y la generación del plural. Cuando es importante, se incorpora información sobre los patrones de complementos del nombre y la información de nominalización. Además, para la flexión y codificación de complemento del nombre, los adjetivos en el lexicón tiene códigos para indicar las posiciones sintácticas en las cuales podrían aparecer. El adjetivo puede ser cualitativo, clasificador, o de color. Los adverbios en el lexicón están codificados para indicar sus propiedades de modificación.

Los elementos léxicos se generan codificando una gran cantidad de fuentes, incluyendo elementos léxicos de los registros de citas de MEDLINE[10], y un gran conjunto de ítems léxicos de diccionarios médicos o de lengua inglesa general.

[10] MEDLINE contiene citas y sinopsis de revistas de contenido biomédico a nivel mundial.

El SPECIALIST Lexicon (fichero formateado en registros que representan la unidad léxica) junto con otros ficheros relacionados se publica anualmente como uno de los recursos de conocimiento de UMLS desde 1994. Su distribución junto al UMLS está disponible como recurso *open source* sujeto a unos términos y condiciones concretas. El formato XML para la unidad léxica de registro surgió en 2003 a través de LexAccess. En 2006, se publican los API de XML schemas y JAXB (Java Architecture XML Binding). Todos los ficheros son difundidos en formato UTF-8. Es en 2009, cuando se añade un fichero ASCII, LEXICON.ascii, con el objetivo de proyectos ASCII de PLN. En 2013, todas las derivaciones en el Lexicon (incluyendo zeroD, suffixD, y prefixD) junto con información de negación se añaden a la versión anual (derivation.data).

2.2.3. Arquetipos

La siguiente fuente de información que comentaremos serán los arquetipos, de los cuales se puede decir que son una evolución de las HCE.

Hasta la creación de los arquetipos, la información generalizada que se almacenaba en las HCE consistía en un conjunto de registros de información clínica sobre los que se ejercía un control de acceso y para los que se recogían datos de tiempo y de autores que realizan la observación, control de versiones, unidades de las mediciones. Sin embargo, nunca se había explotado el conocimiento del dominio clínico.

Un *arquetipo clínico* es una especificación formal para representar una entidad clínica concreta como una observación clínica, un hallazgo, un plan o un tratamiento dentro de una HCE. Esta especificación nace de un acuerdo entre profesionales, con el objetivo adicional de garantizar la interoperabilidad.

Estas especificaciones fueron creadas y mantenidas por OpenEHR, confirmadas por CEN (EN 13606 Parte 2), aceptadas por la International Organization for Standardization (ISO) y su calidad fue ratificada por EuroRec[11]. Otras instituciones que realizaron propuestas de arquetipos han sido el Royal College of Physicians [126], NHS Localigal Record Architecture [35], UCL Chronic Disease Management [87] o *Good electronic Health Record* (GeHR) [11]. Por otro lado, hasta la fecha, existen varios organismos que han participado en el desarrollo de repositorios de arquetipos online [108, 166, 112]: OpenEHR, el *UK National Health Service Connecting for Health (NHS CFH)* [113] y la *National E-Health Transition Authority*

[11] European Institute for Health Records: `http://www.eurorec.org/`

(NEHTA) [109]. Los arquetipos de estos repositorios son mantenidos y actualizados por varios grupos de expertos que cooperan en diferentes ámbitos.

El Modelo Dual y las Características de los Arquetipos

Para una representación de los registros clínicos electrónicos que permita separar claramente la información y el conocimiento se ha propuesto una arquitectura basada en el *modelo dual*. Esta separación es necesaria puesto que la información es perdurable pero el conocimiento evoluciona con el tiempo. Con esto se consigue representar el conocimiento de un dominio particular adaptándolo a las innovaciones que surjan.

El modelo dual se estructura en base a un *modelo de referencia* compuesto por los elementos básicos para representar cualquier información de la HCE, y en base a los *arquetipos*, los cuales definen formalmente conceptos clínicos (por ejemplo medida de la presión sanguínea, informe de alta, medida de glucosa en sangre o la historia familiar) utilizando combinaciones restringidas y estructuradas de los elementos del modelo de referencia y proporcionándoles significado semántico gracias a la alineación con conceptos de terminologías médicas. La interacción del *modelo de referencia* (para almacenar los datos) y el *modelo de arquetipos* (para describir semánticamente esas estructuras de datos) proporciona una gran capacidad de evolución a los sistemas de información. El conocimiento (los arquetipos) irán evolucionando con el tiempo pero los datos permanecerán inalterados. De esta forma se consigue que el significado clínico (datos, información y conocimiento) pueda ser representado de forma consistente.

El *modelo de referencia* contiene las entidades básicas para representar cualquier información de la HCE. Además, es orientado a objetos, esto quiere decir que, por un lado, está formado por unos bloques constituyentes (*clases*) que son combinados en base a unas reglas para dar lugar a estructuras mas complejas pero sin relación semántica como comentábamos al principio y, por otro lado, por información de contexto (por ejemplo, autores, versiones ...) también representada por clases. Como en cualquier lenguaje orientado a objetos, todas estas clases utilizan una serie de tipos de datos primitivos. La estructura de un **arquetipo** está formada por tres secciones principales y que explicaremos sobre el esquema del arquetipo de *Apgar Score* generado por el Clinical Knowledge Manager de OpenEHR[12] (figura 2.1):

[12] `http://openehr.org/ckm/`

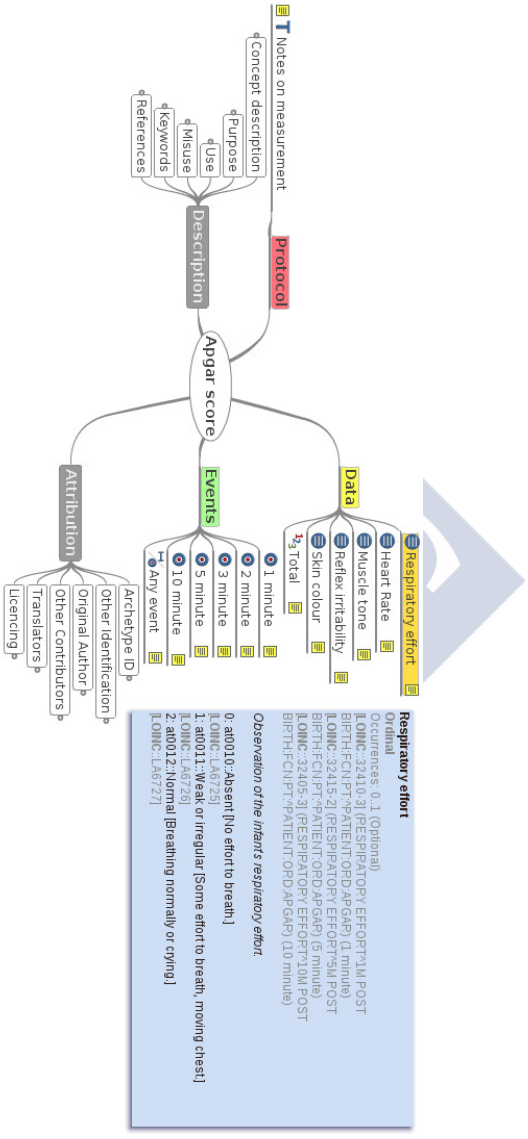


Figura 2.1: Representación del arquetipo correspondiente a la observación de Apgar.

– **header**: que contiene metainformación como, por ejemplo, el identificador del arquetipo, el autor o la versión. En la figura 2.1 que representa el contenido del arquetipo de *Apgar Score*, estaría representada por las ramas grises de *Description* y *Attribution*.

– **body**: está organizada jerárquicamente, refleja la estructura y las restricciones que deben cumplir los conceptos necesarios para representar una exposición clínica en particular. Cada uno de los nodos que forman la definición del arquetipo representan conceptos clínicos mediante el uso de las clases del modelo de referencia y son referidos unívocamente mediante identificadores. En la imagen 2.1 el contenido del cuerpo del arquetipo está representado en las ramas *Data*, *Events* y *Protocol*.La definición de cada nodo debe permitir una asociación consistente y determinista a los datos del HCE original. Además, cualquier nodo del arquetipo debe poder asignarse a términos adicionales que ofrezcan un significado equivalente a su nombre haciéndolo independiente de las particularidades del LN y del idioma. Para completar la definición de los nodos, se especifican restricciones como pueden ser: tipo de datos, rango de valores que pueden tomar, si son multivaluados, frecuencia,...

  Por lo tanto, el objetivo es que estos nodos sean replicados globalmente y unívocamente cada vez que se utilicen en un arquetipo bajo el mismo contexto. Y para ello, se utilizan terminologías públicas e internacionales como por ejemplo los sistemas terminológicos SNOMED CT, LOINC[13], UMLS, ... para garantizar la semántica y, de esta forma, alcanzar la interoperabilidad entre las HCE.

– **ontology**: es la sección que describe los términos y los asocia con las terminologías puesto que los arquetipos son neutros en cuánto a lenguaje se refiere. En el recuadro azul de la figura 2.1, vemos unas referencias de los identificadores del arquetipo (*atxxxx*) y la terminología de LOINC.

Representación de los arquetipos

El **Archetype Definition Language (ADL)** es un lenguaje formal para expresar los arquetipos. Estamos hablando, por lo tanto, de un lenguaje de descripción del conocimiento. Podemos observar un ejemplo en la figura 2.1. ADL está compuesto por tres sintaxis: dADL, para la definición de datos; cADL, para la definición de las restricciones de los términos; y

[13] `https://loinc.org/`

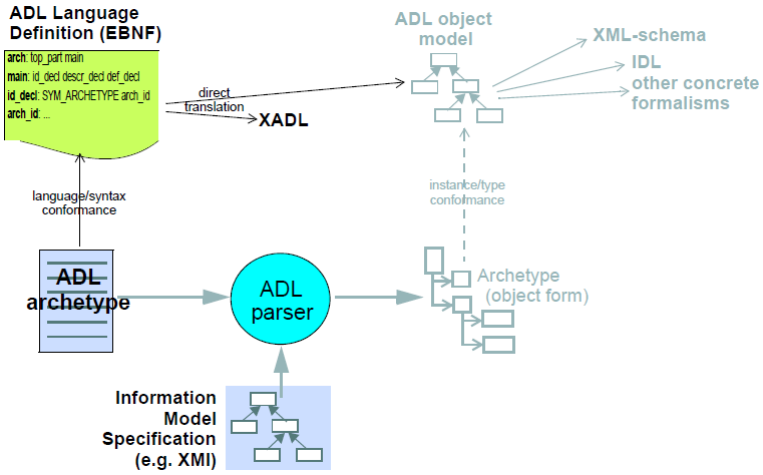


Figura 2.2: Procesado arborescente de los arquetipos [120].

FOPL, una versión de lógica de predicados de primer orden. La sintaxis de la sección de *body* viene establecida por cADL que nos permitirá aplicar las técnicas estructurales. De la misma forma, la sección *ontology* está regulada por dADL para poder extraer los nombres de los términos del arquetipo y aplicar técnicas de anotación léxica y de cadenas de caracteres.

Por ejemplo, un ADL perteneciente a OpenEHR [120] está formalmente estructurado según el modelo de referencia Archetype Object Model (AOM) [121] expresado en UML[14]. Por lo tanto un ADL puede ser parseado como un árbol de objetos tal y como se puede ver en la figura 2.2 perteneciente a los documentos de especificaciones del lenguaje ADL.

OpenEHR ha desarrollado una implementación Java para leer los arquetipos escritos en ADL y recuperar un árbol que cumple el AOM. Esta estructura arborescente es la que utilizamos en este trabajo: en la que cada nodo representa un elemento clínico asociado con un término de la sección *definition* del arquetipo y el texto correspondiente de la sección *ontology*.

[14] Unified Modeling Language: `www.uml.org`

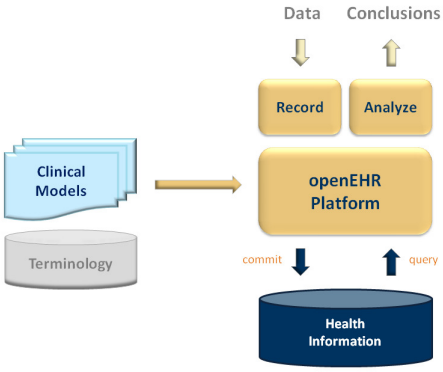


Figura 2.3: OpenEHR: arquitectura general[16].

2.2.4. Tecnologías de HCE relacionadas con el desarrollo de arquetipos

Dentro de todas las instituciones que participan en el desarrollo de la HCE directamente relacionadas con arquetipos, nos centraremos en las 3 más importantes: OpenEHR [119], la norma CEN/ISO EN 13606 [34] y HL7 [57].

OpenEHR

OpenEHR[119] es una comunidad virtual que trabaja para la interoperabilidad y la computabilidad de la información relacionada con la salud. Esta comunidad también utiliza sistemas terminológicos como se observa en la figura 2.3. Se centra en el desarrollo la de HCE y los sistemas implicados para su gestión. Es decir, el objetivo diferenciador de OpenEHR es obtener una representación completa de HCE[15]. Para conseguir esto, OpenEHR ha utilizado también el modelo dual comentado anteriormente.

Esta organización ha publicado un conjunto de especificaciones que definen un modelo de referencia de información clínica para el desarrollo de los arquetipos (modelos clínicos) que son independientes del software y del lenguaje de consultas (figura 2.4). Con esto se consigue

[15] http://www.openehr.org/programs/specification/releases/currentbaseline
[16] http://www.openehr.org/es/what_is_openehr

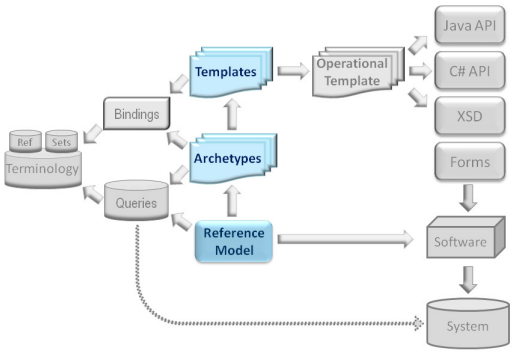


Figura 2.4: OpenEHR: arquitectura multicapa[17].

que el sistema clínico no necesite conocer los datos clínicos a priori para que puedan ser procesados, se hace posible la adaptación de los modelos clínicos sin forzar una modificación del software, y se facilite el desarrollo de software gracias al uso de plantillas que se adaptan progresivamente a los nuevos requisitos.

Con el objetivo de facilitar su difusión, los componentes de OpenEHR son abiertos (documentación, modelos y API de desarrollo).

En OpenEHR existen 4 tipos diferentes de arquetipos: observation, evaluation, instruction y action. Los arquetipos **observation** recogen información referente a cualquier fenómeno o estado de interés sobre el paciente, incluyendo resultados patológicos, tensión sanguínea, historia familiar o cualquier circunstancia social que el paciente pueda proporcionar al doctor durante un examen físico o mediante un cuestionario de seguimiento. En un arquetipo observation no solo se recopilan los datos clínicos a tener en cuenta, si no que se incluye el estado en el que se encontraba el paciente mientras se tomaban esos datos (tumbado, sentado, de pie, en descanso, embarazada) y el protocolo seguido para la recolección (clínico o referente al instrumento de medida: tensiómetro de mercurio, aire, eléctrico o holter). También es importante conocer cuándo suceden estas observaciones.

Los datos que permiten que los médicos puedan determinar o gestionar un tratamiento también están reflejados en los arquetipos. Para que un médico pueda hacer una evaluación,

[17] `http://www.openehr.org/es/what_is_openehr`

necesita información sobre el problema (signos, síntomas, cuándo tuvieron lugar, cómo evoluciona...), riesgos que puedan surgir, si se alcanza el objetivo del tratamiento, recomendaciones al paciente... El arquetipo que recoge esta información se denomina **evaluation**.

Una evaluación del médico conlleva a una decisión sobre un tratamiento (*oral corticosteroids are indicated at a peak flow of 200 l/m*). Sin embargo, una instrucción debe ser más específica y concretar el medicamento, la dosis, la vía de administración, la frecuencia... Cuando una instrucción es ejecutada por cualquier profesional del hospital (médico, enfermero, analista,...) pasa a ser una acción. Una instrucción puede estar compuesta por varias acciones. **Instruction** y **action** son, por consiguiente, los dos últimos tipos de arquetipos.

CEN/ISO EN13606 EHR Comunication Standar

La norma CEN/ISO EN 13606 [34] es una norma del Comité Europeo de Normalización (CEN[18]) que también ha sido aprobada como norma ISO. Está diseñada para lograr la interoperabilidad semántica en la comunicación de la HCE.

Las meta de la norma EN 13606 es definir una estructura de información estable y rigurosa para comunicar partes de la HCE de un único sujeto de atención (paciente). Esto es, para soportar la interoperabilidad de sistemas y componentes que necesitan comunicar (acceder, transferir, modificar o añadir) datos de HCE vía mensajes electrónicos o como objetos distribuidos:

– Conservando el significado clínico original pretendido por el autor y,

– Reflejando la confidencialidad de los datos a la que aspiran tanto autor como paciente.

Este estándar define como unidad básica el **extracto de HCE**. Este extracto es un contenedor de alto nivel de parte o de la totalidad de la historia clínica de un paciente. Permite la comunicación entre el proveedor de la historia y el receptor sin la pérdida de información. CEN garantiza la interpretación de un único extracto, sin la necesidad de depender de otra información de la historia. Se trataría de algo más simple que la arquitectura propuesta por OpenEHR, puesto que OpenEHR cumple las especificaciones de ISO/EN 13606 y las amplía.

Un arquetipo se ajustará a los requerimientos establecidos en la sección 6 de la norma ISO/EN 13606 Parte 2 [65]. Podemos ver un ejemplo en el apéndice A.2.2. De la misma forma, la información contenida en un arquetipo debe ser representada usando el modelo de

[18] `https://www.cen.eu/`

información especificado en la sección 7 de la norma ISO/EN 13606 Parte 2. En caso de ser usada dentro de la comunidad ISO EN 13606, el modelo de información del HCE seguiría el documento CEN/ISO 13606 Parte 1 [64, 67] o cualquiera versión actualizada de este. La asociación ISO EN 13606 trabaja en el desarrollo del modelado de arquetipos. Ellos establecen los componentes básicos que debe tener un HCE para proporcionar un razonamiento computacional: gramática (mediante el modelo de referencia), palabras (código de un sistema de codificación), diccionario (el sistema de codificación en sí), frases o patrones (por ejemplo, los DCM[19], arquetipos o plantillas) y un conocimiento del dominio implícito o explícito (mediante una ontología).

De la misma forma que OpenEHR, ISO EN 13606 define dos modelos conceptuales: el modelo de referencia (contiene la información clínica) y el modelo de arquetipos (contiene el conocimiento clínico). En el trabajo de [90] se realiza una equiparación entre la representación de OpenEHR e ISO EN 13606. En él se comenta que a pesar de que a nivel de sintaxis son prácticamente idénticos y teniendo en cuenta de que OpenEHR es más amplio que ISO EN 13606, es necesario introducir información complementaria en los arquetipos CEN 13606 para poder representar toda la semántica del arquetipo OpenEHR equivalente. Por ejemplo, en un arquetipo CEN 13606 para cada componente ELEMENT contiene un apartado "meaning" que especifica el tipo de elemento que es (OBSERVATION, HISTORY, POINT_EVENT, ITEM_LIST) y un apartado "parts" que define el contenido de ese componente. Se puede observar el resultado de este estudio a través del conversor de OpenEHR a ISO EN 13606 y viceversa que han desarrollado[20].

Health Level 7 (HL7)

HL7 [57] del American National Standards Institute (ANSI) desarrolla estándares para garantizar una interoperabilidad que mejore la asistencia sanitaria, mejore el flujo de trabajo, reduzca la ambigüedad y aumente la transferencia de conocimiento entre los usuarios. Así como para CEN 13606 la unidad básica de procesado es el extracto, para HL7, el **Clinical Document Architecture (CDA)** [32] es la piedra angular. El CDA está basado en el HL7 Reference Information Model (RIM) y usa los tipos de datos de HL7 V3. El contenido de estos documentos contienen una parte obligatoria textual (lo que provoca una interpretación humana del contenido) y partes opcionales estructuradas aptas para el procesamiento

[19] Detailed Clinical Models

[20] `http://sele.inf.um.es:9080/PoseacleConverter/`

computacional. Las partes estructuradas se basan en un sistema de codificación que asigna al contenido médico identificadores de terminologías como SNOMED o LOINC para la representación de conceptos. Según estas características HL7 podría clasificarse como una fuente no estructurada debido a la parte textual expresada en lenguaje natural y a la forma de agrupamiento de la información computacional. Sin embargo, debido al a influencia de OpenEHR e ISO EN 13606, HL7 evoluciona hacia el modelo dual y hacia la representación de la información clínica de forma estructurada, y por lo tanto, clasificamos HL7 como una fuente semiestructurada.

El CDA R2 [32] es una evolución de CDA R1 de cara a una representación semántica de eventos clínicos con el objetivo de intercambio de información. Un documento CDA es un objeto de información definida y completa que puede contener texto, imágenes, sonidos y otro contenido multimedia. Esta información puede ser transferida dentro de un mensaje pero puede existir independientemente del mensaje. Los documentos CDA están codificados en XML, y el RIM junto con la terminología, es el encargado de darle un significado computacionalmente procesable. El modelo CDA R2 permite tanto una gran expresividad, como una representación formal de los hechos clínicos (como por ejemplo observaciones, administración de medicamentos o efectos secundarios) para que puedan ser aplicadas e interpretadas por una computadora. La estructura del cuerpo de un documento CDA está fuertemente influenciada por los modelos CEN ENV 13606, openEHR y DICOM[21]. El CDA R2 Clinical Statement Model especifica cómo las afirmaciones clínicas se anidan dentro de una sección del documento. La semiestructuración de HL7 lo podemos observar en la constitución del CDA: el documento está organizado en secciones (*section*) dónde cada sección supone una agrupación de entradas (*entry*). La representación semántica la completa el uso de terminología. Este vínculo se hace mediante etiquetas XML (*code*) el cual especifica el sistema terminológico aplicado, el identificador del concepto al que representa y el nombre que se asigna a esa afirmación clínica. Este tipo de representación permite la postcoordinación de conceptos para conseguir una representación completa del significado de la representación clínica, gracias a las relaciones del sistema terminológico. Por ejemplo, la observación clínica *osteoarthritis of the right knee* está asociada a los conceptos de SNOMED CT *osteoarthritis* y *right knee* pero además, estos dos conceptos están vinculados mediante la relación *finding site*. Para poder afrontar el problema de la evolución hacia esta capacidad expresiva, se crearon tres ámbitos: HL7 Template, HL7 Clinical Statement Model y HL7 TermInfo.

[21] `http://dicom.nema.org/`

HL7 Template constituye un formato estándar para representar las prácticas clínicas más adecuadas, es decir, podría especificar qué secciones contiene un documento CDA R2 y qué observaciones deben ir en esas secciones. *HL7 Clinical Statement Model* está centrado en conciliar los requerimientos de afirmaciones clínicas en un modelo sencillo que pueda ser usado tanto en especificaciones V3 como en CDA, garantizando así una reusabilidad a lo largo de HL7 en la representación de medicaciones, historia familiar, signos y síntomas... La parte de la terminología está cubierta por *HL7 TermInfo*, encargada del desarrollo de una implementación de guías para el uso de SNOMED CT en el HL7 Clinical Statement Model.

La estructura básica de un documento CDA está formada por dos partes: encabezado *header* y cuerpo *component*. El *header*, al igual que en los arquetipos, contiene metainformación como identificación, título... e información referente al documento como su autor, entidad que custodia el expediente, con qué herramientas se tomaron los datos (por ejemplo, datos del escáner)... En el siguiente ejemplo de la parte *component*, vemos que el documento CDA incluye información sobre cómo se muestra la información en *html* dentro de la etiqueta *text*, en este caso utiliza una tabla y codificación *css*.

```
...
   <component>
      <section>
         <code code="003" codeSystem="7BA9BFFD-D25F-44e8-A7B0-0DF214D6845B"
            codeSystemName="e-MS Document Section Codes"
            displayName="Examination Measurements"/>
               <title>Examination Measurements</title>
               <text>
                  <table border="1">
                     <tbody>
                        <tr>
                           <th/>
                           <th>Blood Pressure</th>
                           <th>Pulse</th>
                        </tr>
                        <tr>
                           <td>
                              <content styleCode="bold">January 13, 2000
                              </content>
                           </td>
                           <td>130/80</td>
                           <td>80 BPM</td>
                        </tr>
                        ...
                     </tbody>
                  </table>
               </text>
```

...

Pero también podemos encontrar la información clínica más estructurada. Dentro de la etiqueta *entry* de la continuación del ejemplo, vemos que especifica que se trata un evento (*EVN*) de una observación (*OBS*). A su vez, la observación de la presión sanguínea está compuesta por las observaciones de presión sanguínea sistólica y diastólica. Cada una de ellas están asociadas con un término creado por el equipo e-MC (electronic Medical Summary [72]) puesto que consideran que no existe ningún elemento de la terminología utilizada. En este ejemplo, la terminología usada es LOINC como se puede observar en el observable *pulse* asociado al código de la terminología *11328-2*.

```
...
   <entry typeCode="COMP">
      <observation classCode="OBS" moodCode="EVN">
         <code code="11328-2" codeSystem="2.16.840.1.113883.6.1"
            codeSystemName="LOINC" displayName="Pulse"/>
         <effectiveTime value="20000113"/>
         <value xsi:type="PQ" value="80" unit="bpm"/>
      </observation>
   </entry>
   <entry typeCode="COMP">
      <observation classCode="OBS" moodCode="EVN">
         <code code="004"
            codeSystem="7BA9BFFD-D25F-44e8-A7B0-0DF214D6845B"
            codeSystemName="e-MS Data Element Code"
            displayName="Blood Pressure"/>
         <effectiveTime value="20000113"/>
         <entryRelationship typeCode="COMP">
            <observation classCode="OBS" moodCode="EVN">
               <code code="005"
                  codeSystem="7BA9BFFD-D25F-44e8-A7B0-0DF214D6845B"
                  codeSystemName="e-MS Data Element Code"
                  displayName="Systolic Blood Pressure"/>
               <value xsi:type="PQ" value="130" unit="mm[Hg]"/>
            </observation>
         </entryRelationship>
         <entryRelationship typeCode="COMP">
            <observation classCode="OBS" moodCode="EVN">
               <code code="006"
                  codeSystem="7BA9BFFD-D25F-44e8-A7B0-0DF214D6845B"
                  codeSystemName="e-MS Data Element Code"
                  displayName="Diastolic Blood Pressure"/>
               <value xsi:type="PQ" value="80" unit="mm[Hg]"/>
            </observation>
         </entryRelationship>
      </observation>
   </entry>
```

...

OpenEHR vs CEN/ISO EN 13606 vs HL7

Según [144], la interoperabilidad entre HCE es imprescindible para:

- compartir la información clínica de pacientes entre profesionales de la salud en un entorno multidisciplinar y compartido.
- la interoperabilidad entre organizaciones dentro de una empresa, sistemas regionales o nacionales de salud o, en el futuro, en un ámbito internacional.
- permitir la comunicación entre software de distintos fabricantes.

Trabajos como Schoeffel et al. [144] y Macedo et al.[88] establecen una comparativa entre OpenEHR, ISO EN 13606 y HL7.

OpenEHR es una especificación abierta, detallada y probada para una plataforma informática completa que garantiza una interoperabilidad tanto funcional como semántica. El objetivo es la administración de información clínica tal como la HCE y otros servicios a mayores como la terminología. OpenEHR ha tenido una gran influencia en las tres organizaciones internacionales más importantes dedicadas al desarrollo de estándares para la salud: CEN (European Committee for Standardization), HL7 (Health Level 7), e ISO. De hecho, ISO EN 13606 es un subconjunto de la especificación completa de OpenEHR.

ISO EN 13606 utiliza al igual que OpenEHR una aproximación basada en un modelado de dos niveles conocido como la metodología de arquetipo, por lo tanto, incorpora también el modelo de referencia de OpenEHR en este estándar. ISO EN 13606 es una especificación para el intercambio de extractos de HCE, no para un sistema HCE completo, por ejemplo, no es capaz de gestionar versiones. Sin embargo, OpenEHR no solo provee una especificación para la comunicación de extractos de HCE con varios niveles de complejidad si no que también proporciona una especificación completa para la creación, almacenamiento, mantenimiento y consulta de HCE. Los extractos de HCE fueron creados por CEN, un HL7 CDA puede ser considerado como un único extracto. Por este motivo, OpenEHR actualmente trabaja en la definición de un interfaz común entre los tres estándares.

Tanto OpenEHR como CEN 13606 están basado en el modelo dual y utilizan el lenguaje ADL en su especificación de arquetipos. Sin embargo, no existe una correspondencia directa

entre ambos partiendo de la idea de que están diseñados con objetivos diferentes. Por lo tanto, comparten el mismo modelo sintáctico para la definición de arquetipos pero no el mismo modelo de referencia. Existen estudios que buscan una equivalencia entre los arquetipos OpenEHR e ISO EN 13606 (por ejemplo, [90, 66]). Es sabido que OpenEHR ofrece estructuras de datos y tipos de datos más potentes. Por lo tanto, trabajos como el de Martínez-Costa et al. [90] aplica generalizaciones en el proceso de asociación e intentar mantener la semántica utilizando las propiedades las entidades y los tipos de dato. Otro esfuerzo por encontrar una equiparación entre ambas técnicas es el estándar ISO 21090 ([66]) que identifica los elementos comunes, la correspondencia entre secciones y estructuras de entidades, e incompatibilidades de ISO EN 13606. Define también una serie de tipos de datos que derivan de los de HL7 y una asociación entre tipos de OpenEHR y ISO EN 13606.

El HL7 CDA es la principal estrategia de HL7 para la interoperabilidad de HCE. El HL7 tiene la ventaja de que es una solución probada para interoperabilidad semántica entre sistemas y de que está en uso en varios países, incluyendo el sistema sanitario gallego. Sin embargo, hasta la fecha, este estándar todavía no utiliza arquetipos pero se está invirtiendo esfuerzo en ello. HL7 CDA es aproximadamente un subconjunto de los extractos 13606 con algunas pequeñas diferencias.

La relación entre los tres formatos la podemos observar en la figura 2.5.

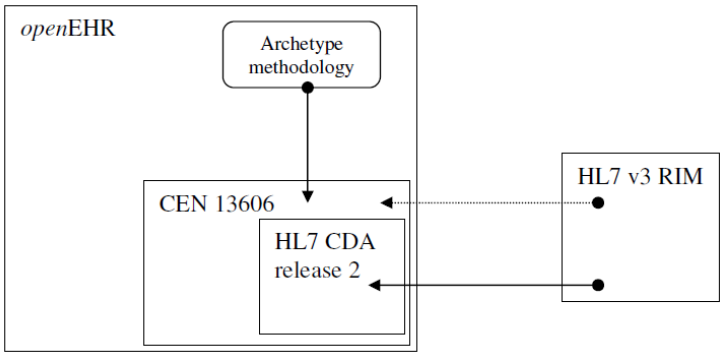


Figura 2.5: Relación semántica entre OpenEHR, CEN 13606 y HL7 CDA [144].

2.3. Fuentes Estructuradas

2.3.1. Terminología

Una terminología puede ser un vocabulario controlado, una clasificación, un tesauro o un lexicón de una disciplina o un ámbito de conocimiento. Por tanto, agrupa palabras y frases que representan las entidades y relaciones que caracterizan el conocimiento dentro de ese dominio determinado. Para denominar el concepto, se usa el término que se considera más representativo, llamado término preferido, y los demás se clasifican como sinónimos de él.

Para definir el concepto, puede haber diferentes atributos como definición, información sobre cambios, notas, ... También ayudan a definirlo relaciones con otros conceptos, que se agrupan en 3 tipos:

Relaciones de equivalencia como la ya comentada entre el término preferido y los sinónimos, que hacen referencia al mismo concepto. Otro ejemplo de relaciones de este tipo son las variantes léxicas (y abreviaturas) y los sinónimos cercanos (términos generalmente diferentes pero en el dominio son tratados como equivalentes).

Relaciones jerárquicas que están basadas en grados o niveles de superioridad y subordinación, donde el término de orden superior representa una clase o un todo y los términos subordinados se refieren a miembros o partes. Los dos tipos básicos son: BT (Broader Term), que es el término de orden superior, y NT (Narrower Term), que es el subordinado. Estas relaciones son las que diferencian un tesauro o terminología de una lista de palabras. La relación jerárquica cubre 3 situaciones diferentes y mutuamente excluyentes:

- la relación genérica (*is a*) establece el enlace entre una clase y sus miembros. Un ejemplo sencillo es *[narrower term] is a [broader term]*.
- la relación de instancia establece el enlace entre una categoría general de cosas o eventos y una instancia individual de esa categoría, a menudo un nombre propio.
- la relación todo-parte cubre situaciones en que un concepto es inherentemente incluido en otro con independencia del contexto.

Cuando un concepto está en más de una categoría se dice que posee relaciones polijerárquicas.

Relaciones asociativas (RT) permiten crear asociaciones entre términos que no son de equivalencia ni jerárquicas, pero en donde los términos están relacionados semántica o conceptualmente y es necesario reflejar esa relación en la terminología. Las más comunes son simétricas, aunque también hay algunas terminologías con relaciones asimétricas. Son las más difíciles de definir, ya que hay que hacerlo de forma explícita para evitar juicios subjetivos que provoquen inconsistencias.

Este tipo de fuentes estructuradas pueden hacer posible la integración y el agregado de recursos automáticamente no sólo al proporcionar un conjunto de términos de muchos vocabularios diferentes, sino también por la incorporación de conocimientos que suponen el alineamiento de vocabularios de contenidos específicos diferentes.

Las terminologías, por su papel de puente entre el lenguaje, la medicina y el software, se han convertido en elemento clave de los sistemas informáticos. En las décadas de los 80 y 90 del pasado siglo, se propusieron una serie de propiedades para asegurar su usabilidad a lo largo del tiempo y su interoperabilidad con otras terminologías [22, 138]:

- **Contenido**: La terminología ha de incluir todos los términos necesarios para el desarrollo de la actividad (*concept coverage*). Estos han de ser exactos (*term accuracy*) y expresivos (*term expressivity*) y, además, es deseable que tengan consistencia sintáctica.

- **Orientación al concepto**: Cada concepto de la terminología debe corresponderse con uno y sólo un significado del dominio. Todos los términos con el mismo significado se agrupan como sinónimos (es decir, se relacionan entre sí a través de una relación de sinonimia).

- **Permanencia**: El significado de un concepto una vez creado es inalterable. El nombre preferido puede evolucionar o puede ser marcado como inactivo o arcaico, pero su significado debe permanecer.

- **Identificador único no semántico**: Cada concepto del vocabulario ha de tener asociado un identificador único. Si un concepto tiene varios nombres, uno de ellos se escoge como preferido y los demás como sinónimos.

- **Polijerarquía**: La terminología debe disponer de mecanismos para expresar la jerarquía de los conceptos. Esta propiedad es muy útil para localizar conceptos.

- **Definiciones formales**: Se expresan como una colección de relaciones con otros conceptos del vocabulario.
- **Múltiples granularidades**: Un mismo vocabulario puede adaptarse a diferentes propósitos, si presenta un nivel de granularidad[22] adecuado.
- **Múltiples vistas consistentes**: Podrán ser usadas por diferentes aplicaciones.
- **Representación del contexto**: Este se puede representar a través de información explícita sobre cómo se usan los conceptos (con restricciones).
- **Evolución controlada**: A través de descripciones claras y detalladas de los cambios ocurridos y del por qué, se pueden incorporar adecuadamente la evolución del dominio.
- **Reconocimiento de la redundancia**: La sinonimia es un tipo de redundancia deseable ya que aumenta la usabilidad de la terminología, para algunos entornos puede ser deseable que la terminología incorpore un sistema para detectarla.

Las terminologías han de dar soporte a diferentes sistemas, como almacenamiento de datos del paciente, sistemas de soporte a decisiones y sistemas de recuperación de información. La terminología ha de modelar cuatro grandes tipos de funciones [131]:

- **Conceptual**: definición formal, clasificación y composición de conceptos.
- **Lingüística**: generación y comprensión de unidades lingüísticas más complejas que etiquetas simples, incluyendo las dificultades de sinonimia, metonimia[23], alusión,...
- **Inferencial**: obtención de conclusiones sobre el mundo representado por los conceptos.
- **Pragmática**: interacción con los conceptos, hechos y lenguaje por parte de humanos a través de diálogos para realizar las tareas.

Cada función requiere diferentes formas de conocimiento, que se agrupan en 3 niveles [38]:

[22] Nivel de descomposición o grado en que pueden ser divididos los contenidos de la terminología.

[23] Figura consistente en designar una cosa con el nombre de otra con la que guarda una relación de causa a efecto, autor a sus obras...

– El **nivel léxico**, que incluye las palabras y frases que expresan los conceptos, consistente en una tipología semántica de las palabras, las reglas para enumerar posibles relaciones entre conceptos y las reglas para componer frases complejas. Por ejemplo, en SNOMED CT, el concepto identificado por el código 22298006 tiene asociadas varias descripciones: *Myocardial infarction (disorder)*, *Cardiac infarction*, *Heart attack* o *Infarction of heart*.

– El **nivel conceptual**, que representa la información contextual, la estructura de los conceptos y la nomenclatura de conceptos y sinónimos. También especifica los tipos de significado y relaciones entre conceptos. Siguiendo con el ejemplo anterior, el concepto *Myocardial infarction* tiene relación *is a* con *Injury of anatomical site (disorder)*, *Myocardial disease (disorder)* y *Structural disorder of heart (disorder)*.

– El **nivel de codificación**, que especifica cómo se enlazan las expresiones lingüísticas a los conceptos.

A continuación, en primer lugar, se comentarán dos ejemplos de las terminologías más relevantes: UMLS Metathesaurus y MeSH. En segundo lugar, se explicarán tecnologías creadas para facilitar el desarrollo y el procesado de terminologías. Por una parte, ISO-25964 es un estándar creado inicialmente con el objetivo de definición de terminologías y, por otra parte hay estándares como SKOS, que complementa tecnologías como RDF y OWL con el fin de regular el alineamiento entre terminologías.

UMLS Metathesaurus

El Unified Medical Language System[24] [14] es un repositorio de vocabularios biomédicos desarrollados por el US National Library of Medicine. UMLS es una herramienta para la integración de información a través de la unificación de terminología. US NLM crea el Metathesaurus como

> un esfuerzo para superar las dos barreras importantes en la recuperación de información por parte de las máquinas

refiriéndose a los problemas de heterogeneidad de nombres que usados para expresar el mismo concepto y la ausencia de un formato estándar para distribuir las terminologías. UMLS

[24] `https://uts.nlm.nih.gov/home.html`

Metathesaurus integra sobre 2 millones de nombres para 900.000 conceptos de más de 60 familias de vocabularios biomédicos, así como 12 millones de relaciones entre esos conceptos. Ejemplos de los vocabularios que incluye UMLS Metathesaurus son NCBI taxonomy[25], Gene Ontology, MeSH, OMIM, ... Los conceptos de UMLS no solo están relacionados entre sí, si no que también están vinculados con recursos externos como GenBank[26]. A día de hoy, estaríamos hablando de alrededor de 150 vocabularios[27]. En la siguiente figura 2.6 podemos ver un ejemplo de la integración de UMLS.

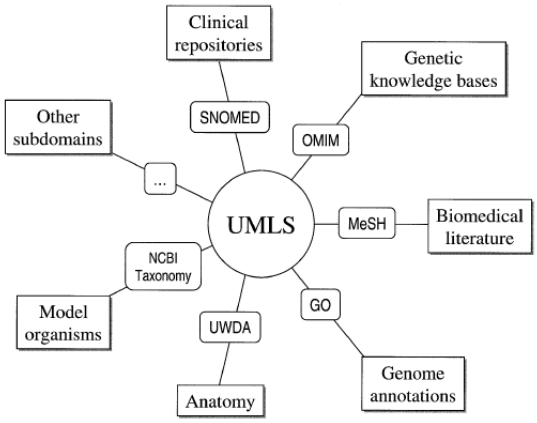


Figura 2.6: Integración de categorías en UMLS [14].

Además de la información, UMLS incluye herramientas para personalizar el Metathesaurus (MetamorphoSys), para generar variantes léxicas de los nombres de los conceptos (lvg) y para extraer conceptos UMLS de texto (MetaMap). Los recursos de conocimiento de UMLS se actualizan trimestralmente. Todos los vocabularios están disponibles para fines de investigación.

El mayor componente de UMLS es el Metathesaurus, un repositorio de conceptos biomédicos interrelacionados. Las otras dos herramientas de conocimiento en el sistema son Semantic Network, que provee de categorías de alto nivel asociadas a cada concepto del Metathe-

[25] http://www.ncbi.nlm.nih.gov/taxonomy
[26] http://www.ncbi.nlm.nih.gov/genbank/
[27] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/source_vocabularies.html

saurus y recursos léxicos como el ya comentado SPECIALIST Lexicon que genera variantes léxicas.

Con respecto a la estructuración de la información en UMLS [115], podemos decir que el conocimiento está organizado en *conceptos*, es decir, en significados. Se utilizan unos identificadores únicos (CUI) para poder referirnos a cada concepto unívocamente, pero para hacer más inteligible el concepto se utiliza un nombre preferido que puede ser configurable en función de las necesidades de nuestro sistema usuario de UMLS Metathesaurus. Los sinónimos se agrupan para formar un concepto. Un objetivo clave de Metathesaurus es interpretar el significado de un concepto en la terminología origen y vincular todos los nombres de los vocabularios origen que significan lo mismo (los sinónimos). Por este motivo, para cada sinónimo se mantiene una referencia al identificador en el vocabulario origen, así como al propio vocabulario. De la misma forma, cada nombre o *string* en UMLS tiene un identificador propio (SUI) aunque entre ellos exista una pequeña variación en la escritura (signos de puntuación, mayúsculas o minúsculas . . .) o en el idioma. Si un *string* tiene más de un significado, este será vinculado con varios conceptos (Figura 2.8). De esta manera, cada *string* de un vocabulario origen está considerada como la unidad atómica de Metathesaurus que diferencia a través del código AUI. Se puede decir que cada AUI es una ocurrencia de un *string* en cada vocabulario fuente y, por lo tanto, un AUI está relacionado con un SUI. Así mismo, un *string* en lengua inglesa puede sufrir variaciones léxicas, lo que genera un nuevo elemento de Metathesaurus identificado por un LUI y se conoce como *término*. Del igual modo que un *string*, un término puede estar relacionado con más de un concepto. También, cada término puede estar vinculado con más de un *string*, sin embargo, un *string* o un átomo están vinculados con un solo término. Un ejemplo de esta organización lo podemos ver en la figura 2.7 para el concepto *Atrial Fibrillation*.

Los conceptos están vinculados con otros a través del significado de varios tipos de relaciones, no solo por los vínculos de sinonimia entre nombres. Estas relaciones se crean a través de los vocabularios origen o a través los editores del Metathesaurus. Un concepto puede estar relacionado con otro de una forma jerárquica (*is a*: hiponimia, *part of*: meronimia) o asociativa, también llamada lógica, (*location of*, *caused by*. . .). Integra también las relaciones estadísticas entre conceptos del vocabulario de MeSH. Y finalmente, añade a cada concepto de Metathesaurus una categoría que representa el tipo semántico asignado por los editores de Metathesaurus, lo cual provee una orientación semántica en Metathesaurus. Además de los vínculos internos de Metathesaurus, no hay que olvidar las referencias cruzadas entre entida-

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**L0004238** Atrial Fibrillation (preferred) Atrial Fibrillations	**S0016668** Atrial Fibrillation (preferred)	**A0027665** Atrial Fibrillation (from MSH) **A0027667** Atrial Fibrillation (from PSY)
		S0016669 (plural variant) Atrial Fibrillations	**A0027668** Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation (preferred)	**A0027930** Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	**A0027932** Auricular Fibrillations (from MSH)

Figura 2.7: Conceptos, términos, átomos y *strings* en UMLS [115].

des pertenecientes a distintos vocabularios y que sirven de puente entre las terminologías de UMLS y recursos externos, por ejemplo, una enfermedad de OMIM puede estar relacionada con proteínas de MeSH. Estaríamos hablando de alineaciones/mappings entre recursos y UMLS.

Tanto conceptos, átomos y relaciones pueden completarse con información extra con lo que llaman *atributos* que pueden ser definiciones, tipos semánticos ...

UMLS Metathesaurus incluye una serie de índices para facilitar la recuperación de conceptos en función de las palabras o grupo de ellas por las que se realiza la búsqueda:

– *Word Index*: Este índice conecta cada palabra individual de un *string* de Metathesaurus con los identificadores de *strings*, términos y conceptos con los que está relacionado. El sistema define una palabra como un *token* formado por uno o más caracteres alfanuméricos. Existe un índice para cada idioma.

– *Normalized Word Index*: Para cada palabra en lengua inglesa normalizada mantiene una referencia con los identificadores de *strings*, términos y conceptos relacionados. Palabra normalizada se refiere a su forma raíz y en minúsculas; también, se eliminan las denominadas *stop words* (preposiciones, conjunciones ...).

Concepts (CUIs)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF only
C0009264 **Cold Temperature**	**L0215040** cold temperature	**S7669511** Cold Temperature	**A15594156** Cold Temperature (from MTH)
	L0009264 cold	**S0026353** Cold	**A0040709** Cold (from LCH)
			A4711382 Cold (from SNOMEDCT)
C0009443 **Common Cold**	**L0009443** cold common	**S0026747** Common Cold	**A0041261** Common Cold (from MSH)
	L0009264 cold	**S0026353** Cold	**A0040708** Cold (from COSTAR)
			A2880095 Cold (from SNOMEDCT)
C0024117 **Chronic Obstructive Airway Disease**	**L0498186** airway chronic disease obstructive	**S0837575** Chronic Obstructive Airway Disease	**A0896021** Chronic Obstructive Airway Disease (from MSH)
	L0008703 chronic disease lung obstructive	**S0837576** Chronic Obstructive Lung Disease	**A0896023** Chronic Obstructive Lung Disease (from MSH)
	L0009264 cold	**S0474508** COLD	**A10765219** COLD (from NCI)
			A0539536 COLD (from SNMI)

Figura 2.8: UMLS tiene en cuenta la ambigüedad [115].

– *Normalized String Index*: En este caso, los índices están formados por los *strings* en lengua inglesa normalizados del Metathesaurus. Para la normalización se usa el SPECIALIST Lexicon y, en caso de que no aparezca reflejado, se resuelve algorítmicamente.

MeSH

Medical Subject Heading (MeSH)[96] es un tesauro de vocabulario controlado de la U.S National Library of Medicine (NLM). Consiste en un conjunto de términos denominados *descriptores* organizados en una estructura jerárquica y que permite la búsqueda en varios niveles de especificidad.

Los descriptores de MeSH están estructurados alfabéticamente y jerárquicamente. En el nivel superior de la estructura jerárquica se encuentran encabezados de significado general como *Anatomy* o *Mental Disorders*; y cuanto más se desciende en la jerarquía de doce niveles, más aumenta la especificidad, por ejemplo *Ankle* y *Conduct Disorder*. En 2015, hay 27.455 descriptores en MeSH. Hay también sobre 218.000 términos que sirven de ayuda a la hora de encontrar el encabezado más apropiado, por ejemplo, *Vitamin C* es un término para *Ascorbid Acid*. Además de estos, existen más de 214.000 encabezados en un tesauro separado llamados *registros suplementarios de concepto*.

El vocabulario MeSH está bajo revisión continua. Cuentan con especialistas en áreas de las ciencias de la salud que aportan su experiencia y conocimiento. Además, es el encargado de recoger nuevos términos que van apareciendo en la literatura científica o en áreas de investigación emergentes. Estos términos se integran en el vocabulario existente y se sugiere su incorporación a MeSH. A mayores, se consulta a profesionales de diferentes disciplinas en relación con cambios de organización y se mantiene la coordinación con varios vocabularios especializados.

Este tesauro se utiliza para indexar artículos de MEDLINE/PubMed[28,29]. También es usada por la base de datos de NLM para la catalogación de libros, documentos y recursos audiovisuales pertenecientes NLM. La terminología MeSH ofrece una forma consistente de recuperar información que permite usar diferente terminología para los mismos conceptos. Cada referencia bibliográfica está asociada con un conjunto de términos MeSH que describen los

[28] PubMed ofrece acceso gratis a MEDLINE y enlaza a los artículos cuando es posible.
[29] `http://www.ncbi.nlm.nih.gov/pubmed`

contenidos del ítem. De la misma forma, para encontrar los contenidos de un tema concreto, se usa el vocabulario MeSH en las consultas de búsqueda.

MeSH, en un formato legible por la computadora, ofrece un acceso gratuito bien, a modo de iniciación a través de un buscador online[30], o bien en formato electrónico desde su sitio Web[31].

Estándar ISO-25964

El desarrollo de normas para la representación de tesauros es una tarea objetivo para la ISO. El estándar *ISO 25964 - the international standard for thesauri and interoperability with other vocabularies* está formado por dos partes:

1. La primera fue publicada en 2011 y cubre todos los aspectos de desarrollo de un tesauro monolingüe o plurilingüe reemplazando a los estándares ISO 2788 y ISO 5964. Para favorecer la interoperabilidad ha definido un modelo de datos orientado a objetos y un un esquema XML para el intercambio de información.

2. La segunda parte del estándar, publicada en 2013, tiene como objetivo principal propiciar una recuperación de información con fiabilidad a través de recursos conectados en red que han sido indexados con diferentes vocabularios. Explica cómo realizar alineamientos entre los conceptos de cada vocabulario y otras formas de uso complementario.

Podemos ver un ejemplo de la primera parte de la norma en el apéndice A.3.1.

Este estándar también busca la interoperabilidad con otras tecnologías, especialmente SKOS con el cual es completamente compatible. La diferencia principal con SKOS es que ISO-25964-1 representa conceptos, términos y las relaciones para construir un tesauro. Sin embargo, SKOS busca una asociación entre conceptos de la web. ISO-25964-2 cubre la tarea de alineamiento pero para todo tipo de tesauros.

SKOS

SKOS Simple Knowledge Organization System (Sistema Simple de Organización del Conocimiento) [100] es un proyecto desarrollado por el W3C cuya primera versión fue lanzada

[30] `http://www.ncbi.nlm.nih.gov/mesh`
[31] `http://www.nlm.nih.gov/mesh`

en el 2003. SKOS es un modelo de datos común en XML para compartir y enlazar sistemas de organización del conocimiento vía web. Podemos ver un ejemplo en A.3.2.

SKOS proporciona un modelo para la representación de la estructura básica y el contenido de esquemas de conceptos como tesauros, esquemas de clasificación, listas de encabezamientos de materia, taxonomías[32], folcsonomías[33] y otros vocabularios controlados similares. Al tratarse de una aplicación de Resource Description Framework (RDF), SKOS permite la creación y publicación de conceptos en la Web, así como vincularlos con datos en este mismo medio e incluso integrarlos en otros esquemas de conceptos. Puede ser usado de forma aislada o en combinación con un lenguaje de representación del conocimiento como Web Ontology Language (OWL).

La meta de SKOS es proveer un puente entre la práctica de diferentes comunidades del área de la biblioteconomía y ciencias de la información que trabajan en el diseño y en la aplicación de sistemas de organización del conocimiento. Además, SKOS ofrece un enlace entre estas comunidades y la Web Semántica, mediante la transmisión de los modelos existentes de la organización del conocimiento al contexto de la tecnología de la Web Semántica, y proporcionando una migración rápida de los sistemas de organización de la información a RDF.

SKOS ocupa una posición entre la explotación y el análisis de información desestructurada, la información informal con influencia social a gran escala, y la representación formal del conocimiento.

2.3.2. Ontología

Una ontología desde el punto de vista de ingeniería del conocimiento (IC) es un cuerpo estructurado del conocimiento aplicable en la gestión terminológica. Una ontología es *un conjunto de conceptos organizados jerárquicamente, representados en algún sistema informático cuya utilidad es la servir de soporte a diversas aplicaciones que requieren de conocimiento*

[32]Taxonomía: Clasificación u ordenación en grupos de cosas que tienen unas características comunes. Es un tipo de vocabulario controlado en que todos los términos están conectados mediante algún modelo estructural (jerárquico, arbóreo, facetado, etc.) y especialmente orientado a los sistemas de navegación, organización y búsqueda de contenidos de los sitios web. No exige que sus componentes estén definidos, conectados mediante un tipo específico de relaciones, es decir, simplemente requiere que sus componentes estén organizados.

[33]Folcsonomía: Clasificación colaborativa por medio de etiquetas simples en un espacio de nombres llano, sin jerarquías ni relaciones de parentesco predeterminadas. La clasificación no se realiza a través de una serie de categorías fijas y jerárquicas, como tradicionalmente se ha hecho, sino a través de lo que se denominan tags o etiquetas que son añadidas y administradas libremente por las personas que usan los sistemas. No presentan relaciones jerárquicas ni de otro tipo, pero pueden establecerse relaciones de forma natural.

específico sobre la materia que la ontología representa. Una ontología representa una vista de un dominio de aplicación común, reusable y compartible. Provee de significado a las estructuras de información que intercambian los sistemas de información. De una forma más informal, podríamos decir que una ontología define vocabularios que las máquinas pueden entender y que son especificados con la suficiente precisión como para permitir diferenciar términos y referenciarlos de manera precisa.

Las ontologías son un recurso independiente de la lengua en el que se representan conceptos organizados en una jerarquía de relaciones y características heredadas. Estos conceptos además están interconectados mediante un sistema de relaciones semánticas definidas entre los conceptos. Todo esto ayuda a la resolución de ambigüedades semánticas y en la interpretación del lenguaje, realizando inferencias basadas en la topología de la ontología para medir la afinidad semántica entre significados.

Las ontologías son la piedra angular en el desarrollo de la Web Semántica ya que permite añadir significado explícito a la información. Esto hace más fácil para las computadores procesar automáticamente e integrar la información disponible en la web.

Los avances en la Web Semántica hacen de la ontología la tecnología candidata para dar soporte a las tareas de conocimiento relacionadas con los arquetipos y las HCE. Además, las ontologías han sido nombradas por el proyecto de Semantic Health[34] como una de las tecnologías básicas para conseguir la interoperabilidad semántica de los sistemas de información de la salud. El uso de ontologías para representar conocimiento biomédico no es nuevo, ya que han sido usadas ampliamente en este campo con propósitos diferentes.

Los arquetipos clínicos se representan normalmente con el lenguaje ADL. Sin embargo, este lenguaje tiene problemas importantes a la hora de conseguir la interoperabilidad semántica. Consecuentemente, la formalización de los procesos de intercambio y transformación es más difícil que usar modelos orientados semánticamente como los ontológicos.

Por lo tanto, las ontologías proveen de un modelo semántico formal estructurado para representar los arquetipos clínicos.

SNOMED CT, Gene Ontology y Human Phenotype Ontology son uno de los ejemplos más importantes de ontologías. Además, se puede considerar la UMLS Semantic Network como una ontología de nivel superior.

[34] `http://www.semantichealthnet.eu/`

UMLS Semantic Network

La UMLS Semantic Network (SN)[35] es uno de los tres componentes de UMLS. Consiste en:

- Un amplio conjunto de categorías temáticas, o tipos semánticos, que aportan a todos los conceptos representados en el UMLS Metathesaurus una categorización consistente y,
- un conjunto significante y útil de relaciones, o relaciones semánticas, que existen entre los tipos semánticos.

La SN se define como ontología de nivel superior [92] que integra terminologías y ontologías que durante su fase de creación se enfocaron en un ámbito específico y diferente entre sí. Por lo tanto, el papel de la SN es proporcionar un entorno de alto nivel en el cual todos los conceptos poseen una representación semánticamente coherente y correcta.

La SN está formada por 133 tipos semánticos (TS) y 54 relaciones. Está representada a través de dos jerarquías de herencia simple, una que representa las entidades y otra los eventos. Las relaciones *is a* permite a los nodos (es decir, tipos semánticos) heredar propiedades de nodos de niveles superiores. Además, hay 5 categorías de relaciones asociativas que conectan los tipos semánticos. Una relación asociativa puede ser física (por ejemplo, *connected to*), funcional (*causes*), espacial (como *traverses*), temporal (*co-occurs with*) o conceptual (*degree of*).

El agrupar los tipos semánticos (grupos semánticos, GS) puede ser útil desde varios puntos de vista como por ejemplo, la visualización del conocimiento de un dominio particular; procesado del lenguaje natural donde las categorías más altas en la jerarquía, en ocasiones, son suficientes para el procesado semántico; y para evaluar si los conceptos y las relaciones que representan un dominio, son válidos. Estas agrupaciones se crearon con el objetivo de cumplir una serie de principios generales, incluyendo, validez semántica (los grupos deben ser semánticamente coherente); frugalidad (el número de grupos debe ser tan pequeño como sea posible); completitud (los grupos deben cubrir el dominio completo); exclusividad (cada concepto en el dominio debe pertenecer a un solo grupo); naturalidad (los grupos deben caracterizar el dominio de una manera que sea aceptable para un experto de dominio); y la utilidad (los grupos deben ser útiles para algún propósito).

[35] `http://semanticnetwork.nlm.nih.gov/`

Con el objetivo de alcanzar la coherencia, para la creación de la SN, se analizaron las relaciones en las que participan los GS. Estos incluyen no sólo la relación jerárquica *is a*, sino también las muchas relaciones asociativas observadas en el ámbito biomédico (por ejemplo, *treats*, *location of*, *measures*). Buscaron que para cada miembro de un grupo, cada relación fuera relevante. Y que además, exista consistencia en las relaciones que se obtienen a lo largo de los grupos.

Las relaciones de la SN se representan como triplas *(TS_1, rel, TS2)*, donde *rel* es la relación entre el tipo semántico *TS1* y el tipo semántico *TS2*. El UMLS representa un total de 558 tripletas de este tipo. En la SN, las relaciones se establecen en el nivel más alto posible de la jerarquía, y a menos que se especifique lo contrario, este tipo de relaciones se heredan a lo largo de la jerarquía. Como cada relación *rel* se relaciona con dos tipos semánticos y, cada tipo semántico pertenece a un único grupo semántico, se puede decir que dos grupos semánticos están conectados a través de una relación *(GS1, rel, GS2)*.

A partir de la SN se han realizado trabajos que la toman como base para crear nuevas ontologías de dominios específicos [107]. De la misma forma, se ha utilizado como contexto para la representación de GPC [82].

SNOMED CT

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) [156] es una amplia y completa terminología multilíngüe gestionada por el International Health Terminology Standards Development Organisation (IHTSDO[36]). Se trata de un estándar para la información clínica puesto que usa en más de 50 países y se mapea con otros estándares internacionales. Hablamos de un vocabulario clínicamente validado, semánticamente rico y controlado que permite una gran expresividad. SNOMED CT es una terminología global pero que permite ser adaptada a las necesidades de un país o una región, mediante mecanismos como los *reference sets*. Esta terminología está centrada en los conceptos. También, está traducido a cinco idiomas y en vías de traducción a otras lenguas.

SNOMED CT [53] ofrece una forma estandarizada de representar la información clínica creada por el médico permitiendo así una interpretación automatizada. SNOMED CT contribuye a la mejora del cuidado del paciente mediante el desarrollo de HCE que permita la recuperación de información clínica de una forma basada en significado, sirve de ayuda en procesos como el diagnóstico, el desarrollo de informes estadísticos consistentes, en el aná-

[36] http://www.ihtsdo.org/

lisis de costes, vigilancia de la salud pública... Los pacientes se benefician del uso de SNOMED CT porque mejora la información de la historia clínica y facilita la comunicación. Por lo tanto, SNOMED CT se usa en la recolección de gran variedad de información clínica, la conexión de bases de conocimiento, recuperación de la información, y agregación, análisis e intercambio de información entre otras funciones.

Actualmente, SNOMED CT contiene aproximadamente 311.000 conceptos activos unívocos, cada uno descrito por un término preferido y uno o más términos adicionales llamados sinónimos. Cada concepto, con significado único, está descrito lógicamente a través de sus relaciones con otros conceptos y organizado jerárquicamente.

Si un concepto en SNOMED CT posee las suficientes características como para diferenciarlo de otros conceptos similares, se dice que es *fully-defined*; en caso de no estarlo se diría que es *primitive*. Podemos ver en la figura 2.9 las características de un concepto. Los componentes principales de SNOMED CT que permiten definir las características de un concepto son tres:

1. **Conceptos**: Representan ideas clínicas, desde una *acumulación de pus en una cavidad del cuerpo (absceso)* hasta *cigoto*. Cada concepto tiene un identificador único llamado *concept identifier*. Estos componentes están organizados en jerarquías que van desde lo general a lo específico, permitiendo que se almacene información referente a detalles clínicos para posteriormente agregarla en un nivel más general. Es decir, las relaciones *is a* posibilitan representar la correspondencia lógica de inclusión jerárquica (subsumir).

2. **Descripciones**: Son las encargadas de asociar textos compresibles con los conceptos, proporcionan la parte inteligible especificando el significado de los conceptos de SNOMED CT y así, permite diferenciar unos conceptos de otros dentro de la jerarquía. Existen principalmente dos tipos de descripciones: el *Fully Specified Name (FSN)* que intenta ser un término diferenciador del concepto; y los sinónimos de un concepto que proporcionan descripciones alternativas para referirnos a este.

   En la versión internacional de SNOMED CT existen casi un millón de términos. Cada traducción de SNOMED CT añade un conjunto adicional de descripciones, las cuales enlazan términos de otros idiomas a conceptos SNOMED CT. Cada descripción tiene un identificador numérico.

3. **Relaciones**: Son vínculos que poseen un significado y que enlazan cada concepto a otros conceptos con significado vinculado. Estas relaciones denotan definiciones forma-

les y otras características del concepto, es decir, lo definen lógicamente. Cada relación tiene un identificador único.

La relación más relevante es la *is a* (*es un*) que relaciona un concepto con un concepto más general. Por ejemplo, *neumonía viral* tiene una relación de *es un* con el concepto general de *neumonía* (concepto padre). Esta relación *is a* define la jerarquía de los conceptos de SNOMED CT. También hay que tener en cuenta, que un concepto puede tener más de un concepto padre (supertipos) y que los conceptos más genéricos subsumen a los descendientes.

Otros tipos de relaciones representan otras características de un concepto. Por ejemplo, el concepto *neumonía viral* tiene una relación de *causative agent* (*agente causal*) con el concepto *virus* y está vinculado al concepto *pulmón* mediante la relación *finding site* (*localización*). En SNOMED CT existen sobre un millón de relaciones. El conjunto de conceptos que pueden formar parte del origen de este tipo de relaciones se denomina dominio y aquellos que pueden formar parte del destino de la relación se llaman rango.

En la figura 2.9 podemos ver la organización de SNOMED CT de forma gráfica.

En la guía de usuario de SNOMED CT [59] podemos conocer su composición y estructuración. El concepto raíz del que descienden el resto es *SNOMED CT concept*. Los descendientes directos del raíz se llaman *top level concepts* y forman las ramas principales de SNOMED CT de diferentes granularidades:

- **Clinical Finding**: Son el resultado de una observación, un seguimiento o una decisión clínica, ya sea normal (*Clear sputum (finding)*) o no (*Abnormal breath sounds (finding)*). Esta jerarquía tiene como descendiente a todos los conceptos que son enfermedades (*disease* o *disorder*)

- **Procedure**: Representan las actividades que se llevan a cabo durante el proceso clínico, por ejemplo, los invasivos, administración de medicamentos o procedimientos de imagen o de educación (*Appendectomy (procedure)*).

- **Situation with explicit context**: Engloba a los conceptos que incluyen el contexto clínico en su definición. Incluye condiciones presentes o ausentes, hallazgos clínicos actuales o pasados o relacionados con otros sujetos (*family history of glaucoma*).

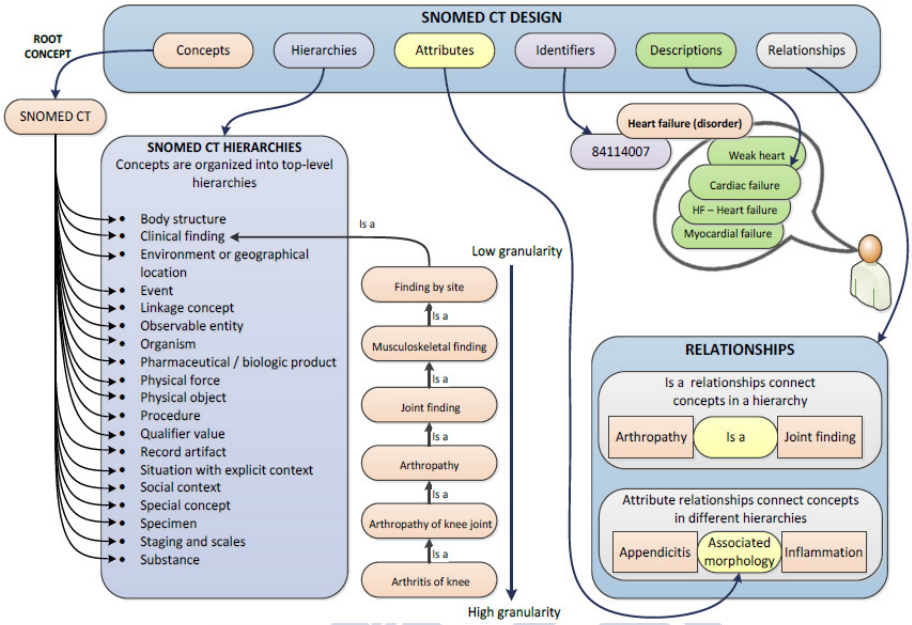


Figura 2.9: Diseño de SNOMED CT [53].

– **Observable entity**: Son el resultado de un procedimiento o la respuesta a una pregunta. Por ejemplo, *gender* sería un observable, sin embargo, *female gender* sería un *clinical finding*.

– **Body Structure**: Incluyen estructuras anatómicas normales (*Mitral valve structure (body structure)*) y no normales (*Polyp (morphologic abnormality)*).

– **Organism**: Hace referencia a organismos tanto humanos como del mundo animal, especialmente aquellos que son causa de enfermedades recogidas en SNOMED CT (*Streptococcus pyogenes (organism)*).

– **Substance**: Engloba componentes químicos activos de medicamentos, alimentos y productos químicos alérgenos, causante de reacciones adversas, tóxicos o venenos (*Insulin (substance)*).

- **Pharmaceutical/biologic product**: Su objetivo es diferenciar de una forma más evidente los productos farmacéuticos de sus componentes químicos (*substance*).
- **Specimen**: Son los conceptos que representan entes que generalmente son extraídos de un paciente tras una examen o un análisis. Estaría relacionado con la parte del cuerpo y el procedimiento de obtención entre otras (*Calculus specimen (specimen)*).
- **Special concept**: engloba a todos los conceptos inactivos que han sido retirados pero guardan alguna relación con algún concepto activo de SNOMED CT y que pueden ser útiles (*Ambiguous concept (inactive concept)*).
- **Physical object**: abarca objetos naturales y fabricados (*Artificial kidney, device (physical object)* o *Latex rubber gloves (physical object)*).
- **Physical force**: enfocada a representar aquello que puede suponer un mecanismo de lesión (*Spontaneous combustion (physical force)*).
- **Event**: Cualquier suceso excluyendo a los procedimientos y a las intervenciones clínicas (*Earthquake (event)*).
- **Environments and geographics locations**: comprende entornos y localizaciones como países, estados o regiones (*Intensive care unit (environment)*).
- **Social context**: incluye características como el status familiar o económico, etnia y religión, estilo de vida y profesión (*Caregiver (person)*).
- **Staging and scales**: referente, por ejemplo, a escalas de evaluación o fases de un tumor (*Glasgow coma scale (assessment scale)*).
- **Qualifier value**: abarcan los valores de los atributos (*linkage concepts*) de SNOMED CT cuando no son subtipos de las otras jerarquías principales de SNOMED CT. Por ejemplo, *Left (qualifier value)* es un calificador de *Laterality (attribute)*. Hay que tener en cuenta que también se pueden encontrar valores en otras jerarquías.
- **Record artifact**: se refiere a los registros de información clínica o informe realizado sobre un estado o un evento (*Clinical statement entry (record artifact)*).
- **Linkage concept**: son los conceptos utilizados para establecer relaciones entre conceptos de SNOMED CT. Es el propio concepto el que especifica el tipo de relación.

Existen dos tipos de conceptos de vinculación: aquellos que refuerzan la definición del concepto con el que están relacionados (*Laterality (attribute)*, *Finding site (attribute)* ...) y, aquellos que se refieren a relaciones históricas entre conceptos (*REPLACED BY (attribute)*).

– **SNOMED CT model component**: metainformación sobre la versión de SNOMED CT.

Recordemos que existen relaciones que sirven para definir características de los conceptos (atributos). Este tipo de relaciones complementan la definición de los conceptos pertenecientes a algunas de las jerarquías principales que son las siguientes: *clinical finding*, *procedure*, *evaluation procedure*, *specimen*, *body structure*, *pharmaceutical/biologic product*, *situation with explicit context*, *event* y *physical objects*.

SNOMED CT utiliza la expresiones precoordinadas y postcoordinadas para representar frases clínicas a cualquier nivel de detalle. Esta técnica le permite a la terminología no tener que almacenar un concepto específico para cada combinación de ideas.

La precoordinación permite separar un concepto detallado SNOMED CT en otros más específicos. Por ejemplo, *laparoscopic emergency appendectomy (174041007)* se puede separar el concepto que refleja la extracción del apéndice *appendectomy (80146002)*, el que especifica la herramienta *laparoscope (86174004)* y la que especifica la prioridad *emergency (25876001)*.

Una expresión postcoordinada es aquella que incluye dos o más conceptos SNOMED CT. Estas expresiones consiguen un mayor nivel de detalle que cualquier concepto de SNOMED CT. Por ejemplo, la expresión *Laparoscopic removal of device from abdomen* que no existe como un concepto independiente, se puede formar con la composición de los conceptos *removal of device from abdomen (68526006)* y *laparoscope (6174004)*. Para la construcción de estas composiciones SNOMED CT define una gramática

Gene Ontology

The Gene Ontology (GO) [9] es un proyecto cuyo objetivo es producir un vocabulario dinámico y controlado que se pueda aplicar a todas las células a pesar de lo reciente o cambiante que pueda ser el conocimiento celular génico y proteico.

El proyecto de GO es un esfuerzo colaborativo que trata de ofrecer descripciones consistentes de productos génicos a lo largo de distintas bases de datos. El proyecto comienza en 1998 como una colaboración entre tres modelos de bases de datos de organismos: Fly-

Base (Drosophila[37]), la base de datos Saccharomyces Genome (SGD[38]) y la Mouse Genome (MGD[39]). Desde entonces, el Consorcio GO ha crecido para incluir más bases de datos, incluyendo varias de los mayores repositorios de genomas de plantas, animales y microorganismos.

El proyecto de GO trabaja en el desarrollo de tres vocabularios estructurados controlados (ontologías) que describen productos génicos. dominios:

- Componentes celulares: partes de una célula y su entono extracelular
- La función molecular: las actividades básicas de un producto génico a nivel molecular, como la unión o la catálisis.
- Procesos biológicos: operaciones o conjuntos de eventos moleculares con un inicio y fin definidos, en lo que concierne al funcionamiento integrado de unidades vivas (células, tejidos, órganos y organismos).

La ontología GO está estructurada en base a un gráfico dirigido acíclico, y cada termino tiene definidas relaciones con uno o más términos pertenecientes al mismo dominio, y a veces a dominios diferentes. El vocabulario GO está diseñado para que sea independiente de la especie e incluye términos aplicables a células procariotas y eucariotas y a organismos unicelulares y multicelulares.

Para conseguir la correcta descripción de los productos génicos, se ha tenido que hacer un esfuerzo desde tres puntos de vista: primero, el desarrollo y mantenimiento de las ontologías en sí mismas; segundo, la anotación de los productos génicos lo que implica hacer asociaciones entre las ontologías y los genes y productos génicos de las bases de datos colaboradoras; y tercero, el desarrollo de herramientas que faciliten la creación, mantenimiento y uso de las ontologías.

El uso de términos GO por parte de las bases de datos que colaboran en el proyecto facilitan las consultas a través de ellas. Los vocabularios controlados están estructurados para que puedan ser consultados desde diferentes niveles: por ejemplo, puedes usar GO para encontrar todo lo relacionado con el genoma del ratón vinculado a la transducción de señales, o permite profundizar en materia de receptores de tirosina quinasa. Esta estructura permite a los anotadores asignar propiedades a genes o productos genéticos a diferentes niveles dependiendo del detalle de conocimiento que necesitemos sobre esa entidad.

[37] http://flybase.org/
[38] http://www.yeastgenome.org/
[39] http://www.informatics.jax.org/

Human Phenotype Ontology

El objetivo de Human Phenotype Ontology (HPO) [83] es ofrecer un vocabulario estandarizado de las anormalidades fenotípicas encontradas en la enfermedad humana. La HPO describe anormalidades fenotípicas tales como un defecto septal atrial. La HPO nace de Online Mendelian Inheritance in Man (OMIM[40]), un importante recurso de datos más allá del campo de la genética humana. La HPO actualmente está formada por información de OMIM y de literatura médica que da lugar a aproximadamente 10.000 términos. Sobre 50.000 anotaciones de enfermedades hereditarias están disponibles para su descarga y acceso usando la herramienta PhenExplorer [41]. Actualmente, la HPO se desarrolla en colaboración con miembros de la OBO Foundry[42] (Open Biological and Biomedical Ontologies), y las definiciones lógicas de los términos de HPO se desarrollan usando PATO (ontología de características fenotípicas[43]) y junto con otras ontologías como FMA (Foundational Model of Anatomy Ontology[44]), GO (Gene Ontology[45]), ChEBI (Chemical Entities of Biological Interest[46]), y MPATH (Mammalian Pathology Ontology[47]). La HPO puede usarse para diagnósticos clínicos en genética humana (Phenomizer), investigación bioinformática centrada en las relaciones entre las anormalidades del fenotipo humano y las redes bioquímicas y celulares, y como vocabulario estándar de bases de datos clínicas, entre muchas otras posibilidades ...

RDF y OWL

De la misma forma que ocurría con las terminologías, también se crearon tecnologías específicas para la definición de ontologías. Veamos el ejemplo de la organización de W3C que creó los estándares más importantes en este ámbito.

El trabajo con la Web Semántica del W3C ha estimulado un campo nuevo de desarrollo de tecnología e investigación integradora de los campos de los sistemas de base de datos, lógica formal y la World Wide Web. Este trabajo ha llevado al desarrollo de unos estándares base para la creación de la Web Semántica. El estándar Resource Description Framework (RDF) [102] proporciona una abstracción de datos común y una sintaxis para la Web. El RDF Vocabulary

[40] http://www.omim.org/
[41] http://www.human-phenotype-ontology.org/contao/index.php/hpo_browse.html
[42] http://obofoundry.org/
[43] http://obofoundry.org/wiki/index.php/PATO:Main_Page
[44] http://sig.biostr.washington.edu/projects/fm/
[45] http://www.geneontology.org/
[46] http://www.ebi.ac.uk/chebi/
[47] http://code.google.com/p/mpath/

Description (RDFS) [91] junto con el lenguaje Ontology Web Language (OWL) [95] [171] forma un lenguaje de modelado de información general para los datos en la Web (ejemplo en apéndice A.3.3). El protocolo y lenguaje de consulta SPARQL [127] es el estándar para interactuar con los datos de la Web.

Estas tecnologías ha sido empleadas a través de diferentes aplicaciones debido a que necesitan un entorno de trabajo común para publicar, compartir, intercambiar e integrar información de diferentes fuentes. La capacidad para relacionar información entre diferentes fuentes es la motivación de muchos proyectos, puesto que diferentes comunidades buscan explotar valores ocultos en los datos que están distribuidos en fuentes aisladas.

Una faceta de la Web Semántica es la de organizar mejor la enorme cantidad de información no estructurada y entendible por el humano que hay en la Web facilitando nuevas rutas para descubrir y compartir información. RDFS y OWL son lenguajes de representación formal de conocimiento, que facilitan formas de expresar un significado desde un punto de vista computacional, siendo este significado el que complementa y da estructura a la información ya presente en la Web. Sin embargo, para aplicar realmente estas tecnologías sobre volúmenes de información considerables se requiere la construcción de mapas detallados sobre dominios de conocimiento particulares, además para una descripción (es decir, anotación o clasificación) precisa de las fuentes de información a gran escala, generalmente no se puede realizar mediante un proceso automatizado. La experiencia acumulada y las buenas prácticas del área de biblioteconomía y de ciencias de la información con respecto a la organización de la información y el conocimiento son complementarias y aplicables a esta visión, puesto que se han desarrollado muchos sistemas de organización del conocimiento y todavía continúan en uso.

En resumen, podemos afirmar que OWL es un lenguaje de Ontologías Web. Con anterioridad, se utilizaron lenguajes para desarrollar herramientas y ontologías destinadas a comunidades específicas (especialmente para ciencias y aplicaciones específicas de comercio electrónico) pero que no fueron definidos para ser compatibles con la arquitectura de la World Wide Web en general, y la Web Semántica en particular. Además, OWL lo reforma proporcionando un lenguaje que utiliza la conexión proporcionada por RDF para añadir las siguientes capacidades a las ontologías:

– Capacidad de ser distribuidas a través de varios sistemas.

– Escalable a las necesidades de la Web.

– Compatible con los estándares Web de accesibilidad e internacionalización.

– Abierto y extensible.

OWL extiende RDF para permitir la expresión de relaciones complejas entre diferentes clases RDF, y mayor precisión en las restricciones de clases y propiedades específicas. Esto incluye, por ejemplo, los recursos para:

– limitar las propiedades de clases con respecto a número y tipo,

– inferir qué elementos que tienen varias propiedades son miembros de una clase en particular,

– determinar si todos los miembros de una clase tendrán una propiedad en particular, o si puede ser que sólo algunos la tengan,

– distinguir entre relaciones uno-a-uno, varios-a-uno o uno-a-varios, permitiendo que las claves foráneas de las bases de datos puedan representarse en una ontología,

– expresar relaciones entre clases definidas en diferentes documentos en la Web,

– construir nuevas clases a partir de uniones, intersecciones y complementos de otras, y

– restringir rangos y dominios para especificar combinaciones de clases y propiedades.

La Guía OWL [170] proporciona ejemplos de todo lo anterior en el área de la descripción de comida y vino.

2.3.3. Extractos Ontológicos

Las ontologías, a pesar de estar centradas en un dominio específico, tienen tendencia a la ampliación e integración de otras terminologías, vocabularios u ontologías. Esto provoca que sean más difíciles de manejar, procesar, mantener y entender. Por este motivo, cobra gran importancia la localización de segmentos de la ontología que contengan la información relevante en ese momento. Además, proporcionan beneficios como la optimización de acceso y consultas, facilita el proceso de comparación y alineación de ontologías, es posible anotarla una ontología de propósito general en secciones más específicas, también adaptarlas a aplicaciones específicas.

La estrategia más empleada para segmentar una ontología es utilizar las relaciones lógicas. El particionado comienza con uno o varios conceptos objetivo en torno a los que gira el fragmento.

Un ejemplo de creación de extractos ontológicos es el de Seidenberg [145] capaz de sacar de la ontología médica GALEN[48] segmentos que conforman una ontología por sí mismos y que se ajusta a necesidades particulares. En este caso, con respecto a los conceptos objetivo, el algoritmo extrae todos sus conceptos superiores e inferiores en la relación taxonómica, así como, los relacionados lógicamente. Para estos últimos extraídos también se tienen en cuenta aquellos conceptos relacionados jerárquicamente. Sin embargo, para evitar que el segmento se expanda demasiado, se excluyen los conceptos hermanos y aplica una serie de restricciones.

En el trabajo desarrollado en esta tesis ([99]) también se parte de un nodo de la ontología de SNOMED CT previamente alineado y se extrae el contexto en el que está encuadrado. Este los constituyen los conceptos relacionados jerárquicamente y relaciones lógicas concretas como *interprets*, tal y como veremos en el capítulo 4.

[48] Repositorio GALEN: `http://www.opengalen.org/`

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Capítulo 3

Técnicas de Reconocimiento de Entidades y Relaciones

La información digital es, a día de hoy, uno de los principales activos del sistema sanitario. Por ello, la mejora de la calidad asistencial está íntimamente ligada a la gestión y uso eficiente de la información, incluyendo la generada por la propia organización. Es en este contexto donde la extracción de la información cobra un sentido crucial.

En contraste con la recuperación de la información, la extracción de la información permite reconocer y acceder directamente a aquellos elementos de información relevantes en un contexto clínico determinado, disminuyendo drásticamente la cantidad de información que necesitamos leer. Dentro de la extracción de información, el reconocimiento de entidades (*Named Entity Recognition*, NER) está orientado a la identificación de semánticas relevantes en un texto. A partir del reconocimiento de estas entidades, es posible identificar relaciones, escenarios o contextos. Una de las múltiples aplicaciones directas del reconocimiento de entidades es la anotación semántica automatizada, en la que nos centramos en esta tesis doctoral.

La anotación semántica está orientada a identificar formalmente conceptos y relaciones en documentos de textos. Por ejemplo, usando una ontología sobre fármacos, la anotación semántica relacionaría el término *propranolol* con una instancia de *betabloqueante* y con la instancia *insuficiencia cardíaca* de la clase *enfermedad* o *síndrome*, evitando así la ambigüedad que pudiera surgir sobre ese término. Esta anotación puede realizarse de manera manual o puede automatizarse para conseguir procesos escalables ante grandes colecciones.

Este capítulo presenta el conjunto de técnicas aplicadas a lo largo de este trabajo de tesis, las cuales han sido desarrolladas para reconocer entidades relevantes de una ontología o terminología dada una descripción clínica en lenguaje natural. Abarca varios tipos de tácticas empezando por las manuales, las mejores cuando se parte de cero en la creación de un recurso o como apoyo a la anotación; las aplicadas sobre recursos de información que poseen algún tipo de estructura, bien esté aplicada a cada elemento básico e individual del recursos o a las uniones entre ellos y; para completar la documentación del proceso de anotación, las estrategias que implican una combinación de las tácticas anteriores para conseguir la mayor eficiencia posible y los mejores resultados.

Finalmente, debido a que existen gran cantidad de herramientas que ayudan en el proceso de reconocimiento de entidades y relaciones, se realiza una breve descripción de las más relevantes y utilizadas conformando una muestra que abarca todas las etapas de este procedimiento de identificación.

En la figura 3.1, vemos un esquema arborescente general de clasificación de las técnicas desarrolladas y aplicadas en este trabajo. Dichas técnicas han sido desarrolladas específicamente para anotar descripciones clínicas en lenguaje natural (LN) con conceptos relevantes de una ontología como SNOMED CT o una terminología como UMLS. En la figura 3.1 puede distinguirse entre un primer nivel de técnicas de equiparación: manuales, no manuales y las que son una combinación de ambas (metaequiparación).

La primera rama representa las tácticas de anotación manuales, las cuales son necesarias para ayudar a deducir el procedimiento que seguirá el proceso automático. Son las tácticas principalmente aplicadas durante el proceso de creación, ampliación y mantenimiento de los recursos de información, además de usarse como etapa inicial de asistencia al proceso de extracción del conocimiento.

Como ejemplo dentro del trabajo realizado, en el capítulo 5 la anotación de GPC fue realizada por integrantes del grupo con conocimientos clínicos. Ellos seleccionaron la información destacada de la guía sobre los procedimientos relacionados con el diagnóstico y terapéuticos sobre los que se experimentará la extracción del conocimiento. Este marcado, también nos ayuda a identificar en versiones posteriores de la guía cuál es la información relevante, por ejemplo, la administración de un cierto medicamento o la educación y seguimiento del paciente.

La segunda rama, refleja las técnicas que se utilizan en las fuentes semiestructuradas o estructuradas y que no suponen un proceso manual, es decir, que pueden ser computables.

Hacia abajo a la izquierda, están las que se aplican sobre cada componente atómico de la fuente de forma individualizada y, hacia la derecha, las que estudian las ligazones del recurso de información.

Las tácticas más utilizadas dentro de las individuales son aquellas que tienen en cuenta similitudes entre las cadenas de caracteres que se usan para nombrar los elementos (términos), las cuales se usan repetidas veces a lo largo de este trabajo. El primer paso que aplicamos en este procesado es la eliminación de caracteres especiales y números, conversión a minúsculas, . . . , es decir, lo que llamamos normalización. Una vez que se tokeniza el término, es decir, se divide en palabras, utilizamos una equiparación parcial o total. Equiparación parcial implica que solo una parte del término origen coincide con la de destino y, total es aquella en la que todos los caracteres coinciden.

Para comparaciones de caracteres más complejas utilizamos las ventajas que proporciona un procesado léxico. Este tipo de procesado implica también una normalización pero en este caso centrada en extraer la raíz de la palabra o generar variantes de la palabra y hacer la comparación, es decir, elimina sufijos que indican número o forma verbal, procesan acrónimos, ... Para ello, utilizamos directamente herramientas ya desarrolladas para tal fin como el lvg[1] de las Lexical Tools de UMLS, o indirectamente a través de MetaMap o el servicio terminológico de UMLS (UTS[2]). Este servicio, que utilizamos vía web o, principalmente, a través de un API Java, nos permite además integrar una nueva técnica individual que incluye el uso de terminologías, vocabularios u ontologías, incluso permite manejar diferentes idiomas. En este caso, hemos creado un método que nos ayude a anotar un texto con SNOMED CT utilizando los sinónimos de otras terminologías puente, es decir, el término origen y el de SNOMED CT destino no tienen por qué parecerse, simplemente es suficiente con que el término origen cumpla las condiciones de semejanza con alguno de los conceptos de las terminologías que integra el Metathesaurus y que tengan como sinónimo a alguno de SNOMED CT.

También se comentan dentro de las técnicas individuales, las centradas en otras características de los elementos como podría ser los tipos de datos. Es decir, podríamos establecer una condición sobre el tipo de datos (texto, numérico, . . .) al que pertenece el elemento. Pero la restricción que se suele aplicar en este tipo de trabajos es el filtrado por el tipo semántico o clase semántica a la que pertenece el elemento, de esta forma, nosotros podemos elegir dentro del conjunto obtenido como resultado los conceptos SNOMED CT que pertenezcan al tipo

[1] `http://lsg3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/web/index.html`
[2] `https://uts.nlm.nih.gov/home.html`

semántico de *observable* o, para Metathesaurus, los términos que sean procedimientos *Therapeutic or Preventive Procedure*, dependiendo del problema que estemos resolviendo en ese momento.

Después de comprender las tácticas individuales, las siguientes a comentar y que están al mismo nivel en el árbol del diagrama son las técnicas relacionales. Estas buscan principalmente semejanzas o establecen reglas y condiciones entre los vínculos jerárquicos, contextuales y lógicos de la fuente de información.

Las relaciones taxonómicas son las más importantes a la hora de trabajar la semántica de un texto o de un modelo de información. Circunstancia que se da también en estos trabajos. Estos enlaces nos permiten crear un contexto, más general o más específico, bajo el que enfocar la semántica concreta sobre la que queremos trabajar en un momento dado. Cuando las combinamos con las técnicas individuales de caracteres podemos jugar con la composición y/o concatenación de los términos del contexto y generar nuevas cadenas y así, mejoramos los resultados de equiparación.

Las relaciones contextuales, se refieren a los vínculos parte-todo (mereológicas) o a la agrupación de elementos. Recordemos que en el capítulo 2 decíamos que un arquetipo es una fuente de información semiestructurada porque agrupaba elementos clínicos unos dentro de otros y esto no implicaba que el concepto que engloba sea un hiperónimo del englobado. Pues bien, son este tipo de relaciones otras que también nos proporcionan un contexto dentro del arquetipo. De las misma forma que hacíamos con las taxonómicas, podemos aplicar sus métodos a este tipo de relaciones.

Las últimas relaciones que nos quedan por comentar son las lógicas. Recapitulando, SNOMED CT estaba constituido por este tipo de relaciones, entre otras. En este trabajo, gracias a la creación del gold estándar manual (anotación manual), se ha encontrado una equivalencia entre la asociación de elementos *ELEMENT* y *VALUE* en el arquetipo y la asociación *interprets* de SNOMED CT. Esto nos ha permitido limitar el ámbito de búsqueda de anotación. También tratamos una relación lógica como una taxonómica a la hora de aplicar métodos de extracción de entidades.

En general, las relaciones nos permiten extraer un contexto enfocado de búsqueda en la anotación, similar a lo comentado en 2.3.3. Este contexto se presupone semánticamente más acertado y permite aplicar técnicas dentro de él que obtienen resultados diferentes, por ejemplo, una equiparación parcial, siendo esta menos restrictiva que una total, mejora los resultados en cuanto a ambigüedad se refiere.

Además, podemos establecer patrones sobre sobre las relaciones. Por ejemplo, como comentamos, el árbol resultante de un análisis sintáctico está formado por agrupaciones (relaciones mereológicas): una oración, está formada por una frase nominal y otra verbal, a su vez, la verbal puede incluir una nominal o preposicional. Pues bien, son este tipo de condiciones o patrones los que utilizamos para extraer las relaciones semánticas dentro de una oración, puesto que dos frases nominales unidas por un verbo puede identificar una relación perteneciente a la red semántica de UMLS. Muy a grandes rasgos, este es el criterio sobre el que funciona la herramienta SemRep que también utilizamos para extraer conocimiento y, en el caso de que no se obtengan resultados aplicamos una combinación de los métodos que hemos desarrollado para localizar la relación.

Todos estos métodos no consiguen una eficacia aceptable si no se combinan de la forma adecuada, por eso, en la última rama de la figura 3.1 se incluyen estrategias de mezcla y de selección resultados óptimos de eficiencia demostrada.

Prácticamente, cualquier condición o regla establecida sobre casi cualquiera de los métodos citados nos permite establecer un criterio para la selección del resultado más representativo tal y como muestra la parte del árbol 3.1 correspondiente a las estrategias de desambiguación (filtrado por tipo semántico, similitud máxima entre cadenas de caracteres, prioridades a la hora de aplicar las diferentes técnicas, . . .). Pero es que además, en este trabajo utiliza la semántica proporcionada por las relaciones lógicas y contextuales para, además de desambiguar, validar la anotación que se lleva a cabo.

Si prestamos atención, a lo largo de la explicación de la figura 3.1, en muchos de los casos no nos es posible ejemplificar las técnicas sobre nuestro trabajo de forma aislada, generalmente siempre las hemos comentado como una combinación de métodos del árbol 3.1 descritos previamente. Pues bien, estas estrategias aplicadas en la tesis a su vez son conjuntadas de forma secuencial o paralela para crear estrategias más complejas todavía. El realizar la interpretación del árbol 3.1 hacia la derecha tiene por objetivo mostrar las técnicas aplicadas desde las más sencillas a las más complejas y que precisan de los métodos de más a la izquierda para poder aplicarse de manera eficiente. Por eso, comentamos las estrategias individuales antes que las relacionales, comentamos la poda (*pruning*) de la fuente de información entera o sobre unos resultados parciales aplicando condiciones sobre características particulares de los elementos del recurso o hablamos de la extracción de un contexto en base a una anotación previa (*anchor*).

El árbol de la figura 3.1 tiene en cuenta también tácticas relacionadas con la teoría de grafos o conjunto y probabilísticas entre otras. Algunas de ellas son usadas indirectamente a través de las herramientas explotadas en estos trabajos. Y, a pesar de que estas técnicas no se implementan en este trabajo, sin embargo igualmente forman parte del proyecto y se exponen en otras tesis.

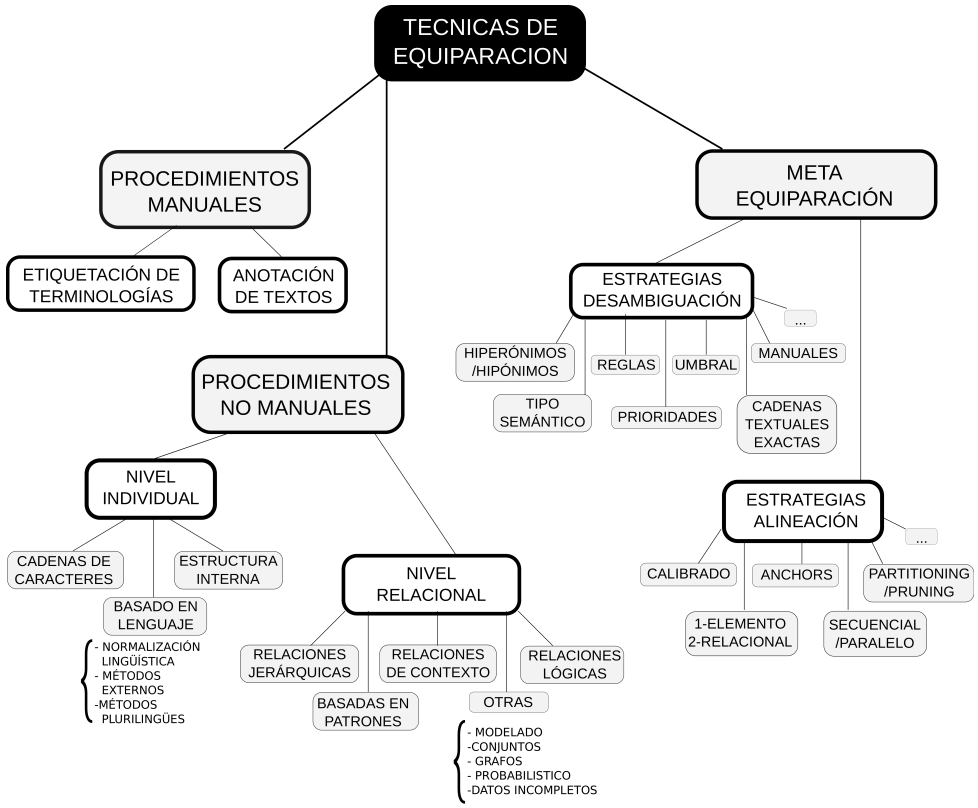
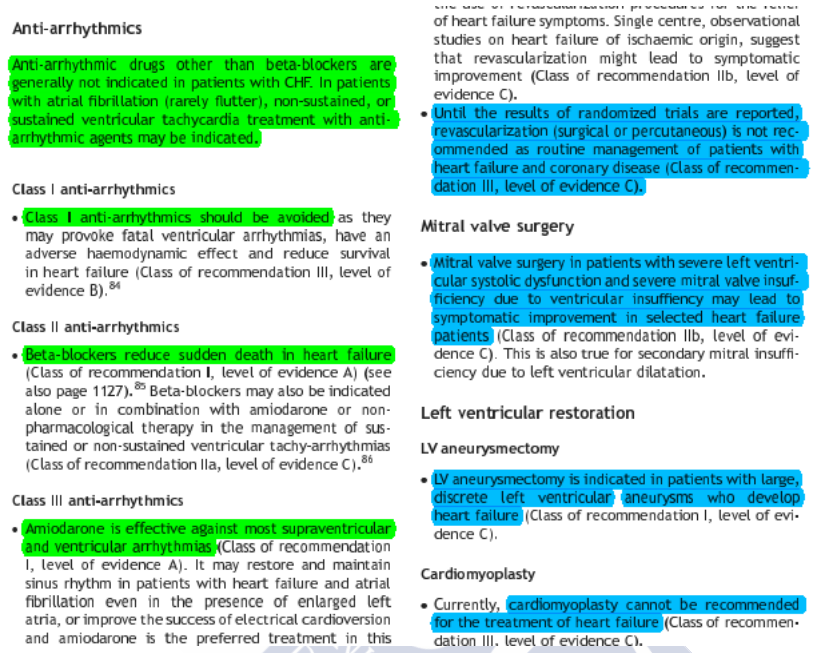


Figura 3.1: Esquema de las técnicas utilizadas.

3.1. Procedimientos manuales

Como procedimientos manuales entendemos aquellos en los que entran en juego uno o varios expertos humanos para realizar identificación de información relevante, etiquetado de



Anti-arrhythmics

Anti-arrhythmic drugs other than beta-blockers are generally not indicated in patients with CHF. In patients with atrial fibrillation (rarely flutter), non-sustained, or sustained ventricular tachycardia treatment with anti-arrhythmic agents may be indicated.

Class I anti-arrhythmics

- Class I anti-arrhythmics should be avoided as they may provoke fatal ventricular arrhythmias, have an adverse haemodynamic effect and reduce survival in heart failure (Class of recommendation III, level of evidence B).[84]

Class II anti-arrhythmics

- Beta-blockers reduce sudden death in heart failure (Class of recommendation I, level of evidence A) (see also page 1127).[85] Beta-blockers may also be indicated alone or in combination with amiodarone or non-pharmacological therapy in the management of sustained or non-sustained ventricular tachy-arrhythmias (Class of recommendation IIa, level of evidence C).[86]

Class III anti-arrhythmics

- Amiodarone is effective against most supraventricular and ventricular arrhythmias (Class of recommendation I, level of evidence A). It may restore and maintain sinus rhythm in patients with heart failure and atrial fibrillation even in the presence of enlarged left atria, or improve the success of electrical cardioversion and amiodarone is the preferred treatment in this

of heart failure symptoms. Single centre, observational studies on heart failure of ischaemic origin, suggest that revascularization might lead to symptomatic improvement (Class of recommendation IIb, level of evidence C).

- Until the results of randomized trials are reported, revascularization (surgical or percutaneous) is not recommended as routine management of patients with heart failure and coronary disease (Class of recommendation III, level of evidence C).

Mitral valve surgery

- Mitral valve surgery in patients with severe left ventricular systolic dysfunction and severe mitral valve insufficiency due to ventricular insuffiency may lead to symptomatic improvement in selected heart failure patients (Class of recommendation IIb, level of evidence C). This is also true for secondary mitral insufficiency due to left ventricular dilatation.

Left ventricular restoration

LV aneurysmectomy

- LV aneurysmectomy is indicated in patients with large, discrete left ventricular aneurysms who develop heart failure (Class of recommendation I, level of evidence C).

Cardiomyoplasty

- Currently, cardiomyoplasty cannot be recommended for the treatment of heart failure (Class of recommendation III, level of evidence C).

Figura 3.2: Marcado en colores de los procedimientos terapéuticos y diagnósticos en la GPC para el diagnóstico y tratamiento del infarto de corazón crónico publicada por la Sociedad Europea de Cardiología.

términos o alineación de terminologías. En la imagen 3.2 podemos observar un ejemplo de marcación de la GPC por parte de expertos clínicos. Esta marcación es parte de la anotación realizada por los expertos de nuestro grupo a partir de la cual se extrajo automáticamente la información útil sobre procedimientos diagnóstico y terapéuticos que se detallan en el capítulo 5. Este proceso es tedioso, lento e imperfecto (ya que la capacidad humana es limitada) y depende del operador (personas diferentes identifican correspondencias diferentes). Vemos un resumen de estos procedimientos en la figura 3.3

3.1.1. Anotación de Textos

Generalmente los enfoques centrados en el campo de extracción de conocimiento necesitan documentos anotados por los humanos puesto que un análisis de esta técnica ayuda a centrar la especificación de requisitos. La anotación manual también provee información para el sistema que se va a desarrollar, como por ejemplo, la extracción de reglas de forma auto-

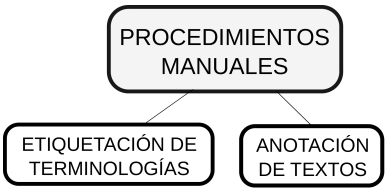


Figura 3.3: Resumen técnicas de alineamiento manuales.

mática o a mano. Y, evidentemente, permite la obtención de un *gold standard* contra el que evaluar los resultados.

Sin embargo, el proceso de anotación manual de documentos no es una práctica común debido a varios motivos:

– Se necesitan expertos que supervisen el proceso o directamente son ellos los que realizan la anotación. Estos expertos no solo deben tener conocimiento del área clínica sino que también deben conocer el recurso con el que están anotando. Encontrar expertos con esta motivación es complicado porque es un procedimiento monótono y poco recompensado.

– En ocasiones, las anotaciones no se realizan utilizando conceptos de ontologías o de terminologías, recursos que se consideran indispensables para la integración de datos, la interoperabilidad y la extracción de conclusiones sobre los documentos. No obstante, esta integración se ve dificultada por la existencia de un importante número de ontologías disponibles que se solapan entre ellas y que están representadas en diferentes formatos, lo que dificulta además el acceso automático a estas. Sin embargo, existe algún recurso como el UMLS Metathesaurus que está orientado integrar todos estos recursos terminológicos.

Por este motivo, a día de hoy existen diferentes herramientas que nos ayudan a anotar texto de forma automática tal y como veremos en la sección 3.5.2.

Alineamiento de Ontologías

Como comentábamos, existe un problema a la hora de la elección del recurso que utilizamos para realizar la notación. Esto es debido al gran número de recursos terminológicos y

ontologías existentes con diferente granularidad, distinto formato, ... Por este motivo se han hecho esfuerzos para integrar y equiparar los recursos terminológicos. Un ejemplo de recurso ya conocido en esta tesis es el UMLS Metathesaurus sobre el que se apoyan otros trabajos como el de Taboada et al. [162] en el que se integran fuentes nuevas en el sistemas Metathesaurus.

La Ontology Alignment Evaluation Initiative (OAEI[3]) es un consenso para evaluar las técnicas de mapeo de ontologías. OAEI ofrece al desarrollador de alineamientos un conjunto de datos para poder evaluar su propia técnica. Lo que implica que se podrían tomar como el *gold standard* necesario. Todas las ediciones de OAEI tienen algún ejercicio que trata conocimiento médico. Estos ejercicios consisten en tomar dos ontologías como entrada y producir como salida una alineación, es decir, un conjunto de correspondencias entre las entidades de esas ontologías relacionadas semánticamente.

Uso de Terminologías en la Anotación de Textos

El ejemplo más importante de procedimiento manual de etiquetado de texto empleando terminologías es la creación de una cita de PubMed[4] [94]. En la que tanto su título como el abstract se indexan con términos de MeSH[5] [96], terminología ya conocida de la sección 2.3.1. Otro caso sería el corpus GENIA [78] que utiliza una ontología para anotar las sinopsis (abstract) de artículos investigación almacenados en la base de datos de MEDLINE. Sin embargo, a pesar de las anotaciones con Gene Ontology (GO) [9], la indexación de PubMed y el corpus de GENIA, la mayoría de los datos médicos están no estructurados y son raramente descritos con conceptos ontológicos.

Como se comenta en [130], los primeros trabajos en anotación han sido con nombres de genes y proteínas. Sobre este campo existen documentos ya validados que se pueden tomar como *gold standard*. Sin embargo, existe una gran cantidad de recursos clínicos cuyo objetivo es extraer toda la información relacionada con el paciente o enfermedades contenida en registros médicos electrónicos, informes de alta, guías clínicas y resúmenes de ensayos clínicos. El problema principal de este tipo de anotaciones es que los resultados no están publicados para poder evaluar y comparar los sistemas. Un caso similar es el de los repositorios abiertos de arquetipos clínicos: varios grupos de expertos que cooperan en diferentes ámbitos son los

[3] `http://oaei.ontologymatching.org/`
[4] Standard literature database of biomedicine
[5] Medical Subject Headings es un vocabulario controlado biomédico creado por el US NLM

encargados de mantener y actualizar los arquetipos, sin embargo, con excepción de un pequeño número de casos, son poco frecuentes las asociaciones entre vocabularios estándar y arquetipos.

3.1.2. Etiquetado de Terminologías

Investigadores en Biomedicina han decidido usar las ontologías y las terminologías para describir la información en este campo y así, convertirlo en conocimiento formal y estructurado. Por ejemplo, frecuentemente se usa la GO para describir funciones moleculares, células y procesos biológicos de productos génicos. Es por este motivo que los responsables de GO establecen protocolos para incitar a la comunidad médica o génica a la colaboración en el proyecto con respecto a la ontología y a anotaciones clínicas.

Veamos ejemplos concretos en los que fue necesario realizar un proceso de etiquetado de ontologías: UMLS Semantic Network, Gene Ontology, Human Phenotype Ontology y SNOMED CT.

UMLS Semantic Network

Tal y como vimos en la sección 2.3.2, gracias a la red semántica de UMLS (SN), todos los conceptos de la terminología UMLS Metathesaurus están vinculados a un tipo semántico (TS). Por lo tanto, a parte del evidente esfuerzo necesario para asociar cada concepto del Metathesaurus con un tipo semántico de la SN, primero fue necesaria la elección de los tipos semánticos y de las relaciones, jerárquicas y asociativas, que iban a formar parte de la SN [15].

HPO Annotation Guide

La ontología Human Phenotype Ontology (HPO), comentada en el capítulo anterior, ha desarrollado una guía de anotación de los términos constituyentes de la ontología[6].

Estas anotaciones están reflejadas en un fichero de texto plano donde cada línea refleja la asociación entre una enfermedad y una propiedad clínica de esa enfermedad. De la misma forma, los datos dentro de cada línea están separados por tabulaciones. Entre la información

[6] `http://www.human-phenotype-ontology.org/contao/index.php/annotation-guide.html`

que se almacena para cada característica habría: cualificadores de grado, la referencia de dónde se tomaron los datos de la asociación, frecuencia, coocurrencia con otras propiedades, sinónimos, cuándo fue realizada la asociación ...

Anotación de GO

Según el proyecto GO la anotación es la práctica que permite identificar actividades así como localizar información relacionada con genes. Ello se realiza acorde con dos principios: se crean referencias entre los elementos de GO y las bases de datos origen y se indican las evidencias de esa asociación. Es decir, realiza una alineación manual entre los elementos de GO o con otras bases de datos independientes que participan en el proyecto. Estas referencias cruzadas, maximizan la utilidad de la ontología minimizando así la redundancia. Por ejemplo, mediante la combinación de la ontología de procesos de GO con una segunda ontología que describe la estructura anatómica de la drosophila[7], se podría crear una ontología basada en la mosca.

A la hora de la anotación, no es fácil establecer una diferencia entre el nombre de un producto génico y su función molecular. Por este motivo, muchas funciones moleculares de GO llevan la palabra *activity* al final. Para tratar estas trabas y otras relacionados con el proceso de anotación, el proyecto GO ofrece una guía de anotación[8,9]. Los grupos encargados de abastecer el repositorio de GO deben tener un conjunto de anotación asociado, que puede coincidir o no. Estos grupos de anotación también están encargados de evitar la redundancia. Los miembros del consorcio del proyecto GO publican esta información como parte del acceso a datos de AmiGO 2[10], buscador GO y herramienta de búsqueda. Con el objetivo de facilitar el proceso de anotación, GO crea unas versiones de las ontologías reducidas para trabajar con genomas o conjuntos de genes.

SNOMED CT Release Formats

De la misma forma que las otras herramientas, SNOMED CT crea una serie de documentos que describen las diferentes formas en las que se representan los componentes, los derivados que dependen de las necesidades de funcionalidad y las expresiones. Estas representaciones incluyen los ficheros en los que está distribuido SNOMED CT, así como posible

[7] Coloquialmente, mosca de la fruta

[8] `http://www.geneontology.org/GO.doc.shtml#annotation`

[9] `http://geneontology.org/page/annotation-tools-downloads-contributing-go`

[10] `http://amigo.geneontology.org/amigo/software_list`

representaciones que pueden ser usadas en la asistencia para la implementación u optimización de funciones particulares.

El contenido de SNOMED CT se distribuye bajo licencia como un conjunto de ficheros. La especificación de los acuerdos de nomenclatura usados en los ficheros de SNOMED CT se encuentra en la sección de Release File Specifications del documento oficial [60]. Actualmente hay dos formatos disponibles:

– Release Format 1 (RF1): Se trata de la especificación creada para el primer lanzamiento de SNOMED CT en el año 2002, que a mayores contiene pequeñas correcciones.

– Release Format 2 (RF2): Es un borrador de especificaciones que añade un número considerable de mejoras.

Con respecto al RF2, esta está dividida en dos partes:

– The Core Component Guide: encargado de la representación de conceptos, descripciones y relaciones contenidas en la primera versión de SNOMED CT.

– The Reference Sets Guide: que describe un patrón común para añadir información adicional a los elementos que forman el núcleo de SNOMED CT. También detalla las formas en las que se usan los patrones para representar funcionalidades básicas (como la concreción del lenguaje, control de cambios y asociaciones) y funcionalidades opcionales (subconjuntos, alineamiento y jerarquías alternativas para la navegación).

En 2012 RF2 se convirtió en el formato de publicación principal de SNOMED CT mientras que RF1 será retirado en un futuro. Sin embargo, para casos especiales, el IHTSDO permite recuperar la información en formato RF1 a través de una conversión que parte de RF2.

3.2. Alineamiento a Nivel de Individual

Este tipo de alineamiento considera cada elemento de texto que se procesa de forma individual. Puede aplicarse a una fuente con alguna forma de estructuración, como por ejemplo una terminología, o sin estructurar como un informe médico. En este tipo de alineación, para el procesado de la porción de texto nunca se tendrán en cuenta las relaciones con otros elementos textuales de la misma fuente. En general, las técnicas que se van a explicar a continuación

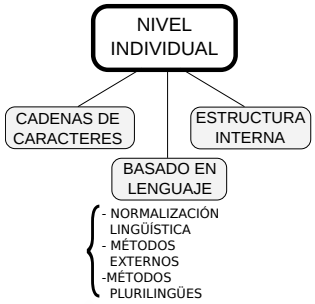


Figura 3.4: Resumen técnicas de alineamiento a nivel individual.

no se usan de forma aislada, si no que sirven como base para métodos que se aplican globalmente a la fuente de información y que son combinados para reforzar sus ventajas. Veamos un ejemplo de estas tácticas en la figura 3.4

3.2.1. Alineamiento de Cadenas de Caracteres

Estos métodos se basan en la estructura de las cadenas a comparar. Se pueden buscar combinaciones exactas de letras o palabras, combinaciones similares o subcadenas. Si las subcadenas a comparar son de tamaño fijo se habla de Q-gramas, es decir, subcadenas de tamaño Q [123]. Dada una cadena c se introducen caracteres de inicio y de final de cadena (# y $ u otros símbolos no existentes en el alfabeto utilizado) y se obtiene una lista de Q-gramas mediante el uso de una ventana de tamaño q que se deslizará a través de los caracteres de la cadena. Podemos ver un ejemplo de cómo se forman los Q-gramas en la tabla 3.1.

La comparación de caracteres nos informa de que *illness* y *sickness* poseen cierta similitud. Sin embargo, también informa que *illegal* está más relacionado con *illnes* que *disease*, para resolver este problema, necesitaríamos tener en cuenta relaciones como la sinonimia que comentaremos en el siguiente apartado. Estas técnicas son:

– **normalización**, busca reducir las cadenas a comparar a un formato común, por ejemplo, cambiar mayúsculas por minúsculas, eliminación de acentos, signos de puntuación y número y, por último, normalizar los espacios. Con estas técnicas, se consiguen disminuir las alteraciones debidas, por ejemplo, a las variantes del idioma y aumentar los sinónimos. El inconveniente sería una posible pérdida de significado aunque se au-

Cadena Original	Posición de la ventana	Q-grama Extraído	Lista de Q-gramas
Glucose	[#G]lucose$	[#G]	[#G]
	#[Gl]ucose$	[Gl]	[#G][Gl]
	#G[lu]cose$	[lu]	[#G][Gl][lu]
	#Gl[uc]ose$	[uc]	[#G][Gl][lu][uc]
	#Glu[co]se$	[co]	[#G][Gl][lu][uc][co]
	#Gluc[os]e$	[os]	[#G][Gl][lu][uc][co][os]
	#Gluco[se]$	[se]	[#G][Gl][lu][uc][co][os][se]
	#Glucos[e$]	[e$]	[#G][Gl][lu][uc][co][os][se][e$]

Tabla 3.1: Ejemplo de generación de Q-gramas para una cadena según [123].

mentan las posibilidades de encontrar resultados, como sería el caso de *angiotensin 2 blockers* que se procesaría finalmente como *angiotensin blockers*.

– **técnicas de subcadenas** con las que se mide la similitud basándose en los caracteres comunes que comparten las cadenas, como la subcadena máxima, prefijos o sufijos. Otro caso podría ser la localización exacta de una cadena de texto concreta (patrón).

– **medición de distancias** que evalúan si una cadena puede ser una versión errónea de otra. Para ello, se mide el número mínimo de cambios para obtener una cadena desde otra. Se usa para determinar la similitud entre cadenas con diferencias de deletreo. Por ejemplo, la distancia Levenshtein calcula el número mínimo de inserciones, sustituciones y eliminaciones de caracteres requeridos para transformar una cadena en otra.

– **medidas estadísticas** que calculan la importancia de una palabra en una cadena utilizando medidas como TFIDF (term frequency-inverse document frequency[11]). Este enfoque funciona bien aplicado a textos largos.

– **comparaciones de camino** que compara no solo las cadenas que se están procesando, si no que también la secuencia de cadenas de los elementos que nos llevan a esa cadena. Ejemplos de esto son los espacios de nombres; en una estructura arborescente, las etiquetas de los nodos que van desde la raíz hasta el nodo que se procesa o; la sección en la que se encuentra el texto dentro del índice de un documento.

[11] A menudo se usa como un factor de ponderación en minería de datos y recuperación de información. Su valor aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada con la frecuencia de la palabra en la colección de documentos, ayudando así a controlar el hecho de que algunas palabras son más comunes que otras.

3.2.2. Alineamiento Basado en el Lenguaje

A diferencia del alineamiento anterior que solamente tenía en cuenta los conjuntos de caracteres, en esta aproximación se tienen en cuenta las cadenas consideradas como textos que pueden ser separados en una o varias palabras/términos. Estas palabras no se agrupan en un conjunto como se hace en recuperación de información, si no que se encuadran dentro de una estructura gramatical. Hay que tener en cuenta que los **términos** son frases que identifican conceptos, por lo tanto, estas técnicas nos serán de gran utilidad para reconocerlos rodeados de otra información. Estas estrategias dependen del procesamiento del lenguaje natural (PLN) para extraer los términos significantes de un texto como, por ejemplo, las propiedades:

– ortográfica: variantes autorizadas en la escritura de un mismo término.

– morfológica: variaciones en la forma y la función. Por ejemplo, el singular y plural de un nombre, conjugación de verbos, formas comparativas y superlativas de los adjetivos ...

– cierta información sintáctica: puesto que se conoce la categoría gramatical a la que pertenece una palabra, por consiguiente, se pueden tener en cuenta ciertos complementos que suelen acompañar al término.

Normalización lingüística

Consiste en transformar un término en una forma estandarizada que sea fácilmente reconocible. Se basa en las propiedades gramaticales anteriormente citadas podemos ver un ejemplo en la siguiente tabla 3.2.

Se puede decir que los pasos generales para el procesado basado en lenguaje de texto son los siguientes:

– **Tokenización**: consiste en la segmentación de cadenas de caracteres en secuencias de tokens a través de la puntación, mayúsculas, espacios en blanco, dígitos, ...

– **Lematización**: para cada token obtenido en el paso anterior se realiza un análisis morfológico para llevar a cabo una normalización eliminando la información de género y número entre otros y obteniendo la raíz de la palabra.

– **Extracción del término**: apoyándose en la repetición de morfología similar en las frases y en el uso de patrones.

Tipo	Subtipo	Ejemplo
Morfológico	Flexión	*daily dose*
	Derivación	*daily dosage*
	Flexión y Derivación	*daily dosages*
Sintáctico	Inserción	*current daily dose*
	Permutación	*dose by day*
	Coordinación	*initial and daily dose*
Morfo-sintáctico	Derivación y Coordinación	*initial and daily dosage*
	Flexión y Permutación	*doses by day*
Semántico		*daily prescription*
Plurilingüe	Español	*dosis diaria*

Tabla 3.2: Ejemplo de los tipos de normalización lingüística.

– **Eliminación de stop words**: Son tokens que se identifican como artículos, preposiciones, conjunciones, ... Estos son eliminados puesto que no contienen significado a la hora de la alineación.

Después de aplicar estos procedimientos, lo siguiente que se procesará serán términos, no palabras.

Métodos Externos y Plurilingües

Estos métodos utilizan recursos lingüísticos a mayores de los que se quiere equiparar con el objetivo de encontrar similitudes entre términos. Ejemplos de estos recursos son lexicones, tesauros o terminologías ya explicados en el capítulo 2. Se trata de material que puede ser especializado en un dominio y que contiene información que no existe en el lenguaje cotidiano, nombres específicos y abreviaturas centradas en el dominio. Poseen la gran ventaja de la sinonimia y la desventaja de la homonimia. Por otra parte, independientemente del nivel de estructuración de la fuente extra o la granularidad en el dominio, estos recursos proveen al proceso de alineamiento de un entorno común entre la información a mapear.

Los recursos con los que se trabaja pueden ser plurilingües permitiendo alineaciones sobre más de un idioma. El empleo de ellos nos permiten hacer diferentes tipos de alineaciones: monolingües, plurilingües (recursos que están representados en diferentes idiomas) y de cruce de idiomas (alineamientos de diferentes idiomas).

3.2.3. Alineamiento Basado en La Estructura Interna de las Entidades

Si hablamos de fuentes semiestructuradas o estructurada, este alineamiento estaría fundamentado en las restricciones o procesados que se aplican sobre las definiciones de los elementos, los tipos de datos, la cardinalidad de los atributos o las claves, es decir, de la estructura interna de las entidades sin hacer referencia a las relaciones con otras. Se usan principalmente para eliminar incompatibilidades. Sin embargo, este tipo de técnicas no ofrecen mucha información sobre los elementos a comparar; concretando sobre tipos de datos: elementos diferentes podrían tener el mismo tipo de datos y distintos modelados de un mismo concepto podrían tener tipos incompatibles. Debido a estas incompatibilidades, se usan en conjunto con otras técnicas.

Ejemplo de esto podría ser, dentro de una ontología, que los términos que se desean alinear pertenezcan a un determinado tipo semántico. Por ejemplo, los elementos principales pertenecientes a un arquetipo de tipo observable, solo puedan estar asociados con las entidades de SNOMED CT que sean descendientes de la entidad *observable entity*.

Puesto que se trata de la estructura interna de los elementos, otra situación que podríamos incluir aquí sería el alineamiento de definiciones. Muchos de los recursos que se utilizan en el proceso de asociación nos proveen de una definición de conceptos. Esta definición, escrita en lenguaje natural, es utilizada para la extracción de conocimiento o para alineamientos, así lo establece Noy [116] en un estudio sobre integración semántica.

Knight y Luk [79] realizaron uno de los primeros trabajos utilizando las definiciones de conceptos. Ellos definieron un algoritmo de alineamiento de definiciones que se basa en la idea de que los significados de dos palabras deberían mapear si sus dos definiciones comparten palabras en común. Además, en la asociación de definiciones utilizan también los sinónimos del concepto. Un ejemplo es Ehler et al. [33] que utilizan las definiciones de los términos de Gene Ontology (GO) para etiquetar artículos con estos términos. Otra empleo es el de Murata et al. [104] los cuales utilizan las definiciones para extraer sinónimos en lengua japonesa.

3.3. Alineamiento a Nivel Relacional

Al contrario que las técnicas que realizan alineaciones a nivel de elemento, las técnicas globales al recurso tienen en cuenta las relaciones que vinculan las entidades de las fuentes de información entre sí. Vemos un esquema en la figura 3.5

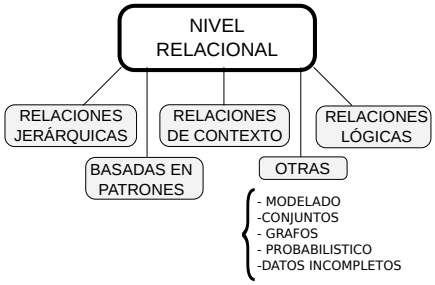


Figura 3.5: Resumen técnicas de alineamiento a nivel relacional.

El procesamiento del recurso de información en forma de grafo es el tipo de técnicas de alineamiento a nivel relacional más utilizado. Este enfoque se asienta sobre algoritmos que consideran la fuente como un grafo etiquetado. En ocasiones, basan la equiparación en la comparación de la posición de las entidades dentro del grafo, es decir, se basan en el *principio de vecindad* que especifica que si dos entidades son similares, las entidades próximas también lo serán. Los tipos más importantes de asociaciones que tienen en cuenta la estructura en forma de grafo son aquellos que toman como referencia los vínculos taxonómicos (*is a*), los mereológicos (*part of*) y otros tipos de relaciones que definen lógicamente a las entidades de la fuente de información.

3.3.1. Alineamiento Basado en Relaciones Jerárquicas o Taxonómicas

Las relaciones jerárquicas, también llamadas taxonómicas, son las *is a*. Se podría decir que son el tipo de relación más importante y el más estudiado a la hora de la asociación entre entidades.

Existen varias clases de mediciones para comparar los nodos de una taxonomía. Principalmente, lo que se compara es la distancia entre nodos dentro de la jerarquía. Por ejemplo, la medida Wu-Palmer [174] tiene en cuenta que dos nodos próximos a la raíz están muy cerca uno del otro en la jerarquía, sin embargo, pueden ser muy diferentes conceptualmente; por otro lado, dos nodos para los que uno es descendiente de otro y además existe una distancia considerable entre ellos pueden ser más similares conceptualmente comparados con los no-

dos del ejemplo anterior. La técnica anteriormente comentada tendría en cuenta la estructura jerárquica completa del recurso, sin embargo, es posible aplicarlo de forma no global:

– **Reglas de super o subclases**: Establece que dos nodos son similares si lo son sus padres o sus hijos. Reglas que no funcionan en caso de que existan varias super o subclases, para lo que hay que añadir más condiciones. También es un problema el alineamiento de un nodo debido a la fuerte dependencia de la similitud entre padres o hijos.

– **Alineamiento basado en el camino entre nodos**: Especifica que dados dos pares de nodos relacionados entre sí y bajo la misma rama jerárquica, aquellos nodos pertenecientes al camino que une un nodo con otro, son candidatos a ser similares. Para este caso, hay que tener en cuenta que tendríamos que partir de unas alineaciones ya confirmadas (anchors) y que habría que completar con otros métodos de vinculación para poder llevar a cabo el alineamiento.

3.3.2. Alineamiento Basado en Relaciones de Contexto o Mereológicas

Otro método relacional a tener en cuenta es la mereológica, es decir, las relaciones *part of*. Se basa en la idea de que si dos nodos son equivalentes, también deberían estar relacionadas la partes que lo constituyen, y viceversa. Lo cual es de gran utilidad a la hora de establecer vínculos entre nodos. El principal problema de este tipo de relaciones es que en no es sencillo identificar este tipo de relación dentro de un recurso.

Un ejemplo de este tipo de relaciones parte-todo lo vemos en el árbol sintáctico de una oración. Una oración está constituida por un conjunto de frases, por ejemplo, una nominal y otra verbal. Y de la misma, forma cada una de esas puede estar constituida por mas frases: preposicionales, nominales, adverbiales, ...

Otro ejemplo serían el ADL que especifica arquetipos OpenEHR. Normalmente, un arquetipo utiliza para la definición un *ITEM_LIST* o *ITEM_TREE* que contiene varios *ELEMENTS* anidados. De la misma forma, un *ELEMENT* contienen una lista de los posibles valores que pueden tomar.

3.3.3. Alineamiento Basado en Relaciones Lógicas

A diferencia de los apartados anteriores, este trata el problema general de alineamiento de entidades basado en otras relaciones que pueden definir lógicamente la entidad. Además,

mientras en los otros tipos de alineamiento basado en relaciones solamente se vinculaban las entidades, en este caso, también podemos deducir si dos relaciones son equivalentes ante la premisa de que los pares de entidades que relacionan lo son.

Una forma de alineación utilizando estas relaciones podrían ser: si una entidad de la fuente origen está definida lógicamente a través de relaciones que la vinculan a un grupo de entidades; de la misma forma, si este grupo de entidades de la fuente origen está alineado con otro grupo de entidades de la fuente destino; y además, el último conjunto de entidades destino poseen un vínculo común con una entidad concreta: entonces podríamos alinear esta entidad destino con la entidad origen inicial.

Desde el punto de vista de teoría de grafos, estas relaciones se pueden considerar como los arcos de un grafo y a su vez como propiedades de las entidades (nodos) que unen eses arcos. La única diferencia con las secciones anteriores es que estas relaciones pueden contener circuitos. Estos se podría abordar, por ejemplo, primero aplicando un alineamiento sobre las etiquetas de los elementos individuales, segundo intentando aplicar alineamiento taxonómico y, finalmente, seguir algún tipo de criterio para encontrar similitudes como por ejemplo los siguientes:

– Hijos: dos conceptos son estructuralmente similares si sus hijos inmediatos son altamente similares.

– Hojas: dos conceptos no hoja son similares si sus conjuntos de conceptos hoja son similares, aún cuando sus hijos inmediatos no lo sean.

3.3.4. Alineamiento Basado en Patrones

Lo primero que se necesita para poder aplicar este tipo de alineamiento es un conjunto de patrones predefinidos contra los que establecer la equiparación. En la definición de patrón no se utilizan ejemplos concretos si no que se especifican elementos más abstractos como por ejemplo sus tipos de datos, tipo semántico o categoría.

También es frecuente el uso de antipatrones que indican lo que no debe suceder. Se utilizan para encontrar inconsistencias e incoherencias.

Un ejemplo de aplicación de patrones podría ser la identificación de ciertas estructuras en el árbol sintáctico correspondiente a una oración: una frase nominal nos ayudaría a identificar una unidad textual con significado (concepto), un verbo una relación, ...

3.3.5. Otras Formas de Alineamiento Relacional

A mayores de los citados en la sección 3.3, hay que citar también otro tipo de enfoques estructurales.

La primera es la técnica basada en modelado que tiene en cuenta una interpretación semántica del contenido y, por lo tanto, se pueden aplicar métodos deductivos como técnicas proposicionales o de descripción lógica.

Otras serían las técnicas basadas en conjunto de entidades que se centran en comparar grupos de entidades, en vez de hacerlo aisladamente, y decidir si esas clases están vinculadas o no. Para ello explotan el razonamiento de teoría de conjuntos o técnicas completas de análisis de datos y estadísticas.

Existen también métodos iterativos para el cálculo de la similitud en grafos basados también en el principio de vecindad y que además nos permiten localizar ciclos en el recurso que estamos utilizando.

En el caso de que los datos sean incompletos o parciales, igualmente existen soluciones que nos permiten estimar parámetros.

Finalizamos con los métodos probabilísticos cuyo objetivo es mejorar los candidatos de alineamientos usados en combinación con otras técnicas. Podrían ser modelos de esto las redes bayesianas o las de Markov.

3.4. Metaequiparación

En las secciones previas se explicaban las técnicas de alineamiento y de equiparación más básicas y comentadas al nivel más bajo de aplicación. Sin embargo, una sola técnica de las anteriores no es capaz de proporcionar resultados satisfactorios al problema de la anotación. Por este motivo, es necesario realizar una combinación de estas tácticas básicas que garanticen un nivel de eficiencia y de precisión aceptables. A continuación, procedemos a su explicación en las dos secciones siguientes.

3.4.1. Estrategias de Alineación para la Extracción de Términos

Como comentamos, las tácticas anteriores no se utilizan de forma aislada para obtener resultados si no que hay que combinarlas como por ejemplo muestra la figura 3.6.

Figura 3.6: Resumen estrategias alineación.

Generalmente, en el caso de que el recurso sea de gran tamaño, el proceso de alineación comienza con la división de la fuente de información en porciones más pequeñas (*partioning*) o se ignora parte de esa fuente (*pruning*).

Es frecuente utilizar *anchors*, es decir, vínculos prefijados o confirmados antes de la aplicación de alguna de las técnicas.

Normalmente se aplican primero las técnicas a nivel de elemento para realizar alineaciones iniciales y, a continuación, como soporte para las tácticas relacionales.

Si se integran métodos/herramientas de alineamiento, suele ser necesario un calibrado de esta para que se adapte al problema concreto que se está resolviendo. El objetivo puede ser, bien mejorar la calidad de los resultados de mapeo (precisión, recall, F-measure) como intenta esta tesis o bien, mejorar su rendimiento a nivel de consumo de recursos. Este ajuste se puede hacer antes del alineamiento, previo análisis, o después y puede ser manual, semiautomático o automático. También podría mejorarse el alineamiento utilizando técnicas de aprendizaje máquina que nosotros no utilizaremos porque no disponemos de un conjunto de datos de control.

Es necesario hacer una composición de los resultados obtenidos a través de varias técnicas que fueron aplicadas de forma independiente, es decir, hay que agregar e integrar los resultados al resultado final aplicando algún tipo de criterio u operador. En el caso de que las técnicas se ejecuten de forma secuencial, no es necesario aplicar ningún criterio especial puesto que el resultado de una técnica será la entrada del siguiente método. En el caso contrario de que se obtengan resultados utilizando dos métodos diferentes y sobre una misma entrada habrá que seleccionar el más adecuado. Para solucionar este problema habrá que aplicar estrategias de desambiguación como las comentadas en el siguiente apartado.

Término extraído	Conceptos UMLS Recuperados	Tipos Semánticos	Tipo de Equivalencia
early morning			Ninguna
dyspnea	dyspnea	Sign or Symptom	Simple
chronic obstructive pulmonary disease	*chronic obstructive airway disease *pulmonary disease, chronic obstructive, severe early-onset	*Disease or Syndrome *Disease or Syndrome	Ambigua

Tabla 3.3: Ejemplo de términos médicos extraídos de una GPC y de sus correspondientes conceptos UMLS Metathesaurus obtenidos.

3.4.2. Estrategias de Desambiguación

La aplicación de un método de alineación no garantiza que el resultado sea único, si no que podemos obtener varios resultados y, dentro este grupo de candidatos unos serán más válidos que otros. Para ello, hay que saber discriminar dentro de este conjunto de resultados y decidir cuál o cuáles son los más adecuados, pudiendo utilizar las técnicas de alineamiento anteriormente citadas para este proceso.

En el momento que intentamos alinear un término, la técnica empleada puede devolver cero, uno o varios entidades de la fuente destino. En el primer caso, decimos que no existe un vínculo; en el segundo, la asociación es simple (uno a uno); y en el tercer caso, el mapeo es ambiguo porque un término se asocia a varias entidades. En la tabla 3.3, podemos ver un ejemplo de cada situación posible.

A continuación se explican los ejemplos de tácticas para la desambiguación mostrados en la figura 3.7.

Volviendo a la tabla 3.3 del ejemplo, se puede observar que el tercer término, *chronic obstructive pulmonary disease* es un término ambiguo ya que tiene varios conceptos asociados. Si además comprobáramos las relaciones que poseen esos conceptos con otros, verificaremos que existe una relación jerárquica entre sí, en la que el concepto *chronic obstructive airway disease* es un concepto más general que engloba a *pulmonary disease, chronic obstructive, severe early-onset* y, por consiguiente, podríamos tener en consideración sólo el primer concepto. Por el contrario, dependiendo del problema, también puede ser interesante elegir el concepto más específico en lugar del más general.

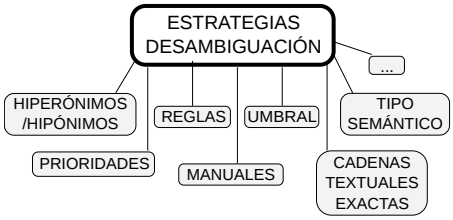


Figura 3.7: Resumen estrategias desambiguación.

Se puede decir que un candidato es más adecuado que otro cuando el número de caracteres compartidos es mayor en uno que en otro (3.2.1). Por este motivo, la tendencia es intentar localizar un candidato cuyos caracteres coincidan exactamente con el término origen y asociarlo como candidato más adecuado. En la tabla 3.3 sería *pulmonary disease, chronic obstructive*

El filtrado de resultados en función del tipo o categoría semántica es uno de los más utilizados. Esta técnica ya fue comentada en el alineamiento individual basado en la estructura interna de la entidad 3.2.3. Esta situación se daría por ejemplo cuando estamos intentando alinear un texto que se centra en el examen de un paciente, habrá cantidad de términos se estén vinculados con conceptos *observable* de SNOMED CT.

Otra posible solución sería adjudicar diferentes prioridades a las tácticas utilizando puntuaciones o reglas, es decir, una aproximación basada en semántica sería más fiable que una basada en cadenas de caracteres. Por ejemplo, supongamos que utilizamos una técnica de alineación que combina la coincidencia parcial de caracteres que componen los términos y la condición de pertenencia a un contexto lógico. Entonces, los resultados obtenidos con esta combinación de tácticas, tendrán más peso que una correspondencia exacta de caracteres. Concretándolo con un ejemplo, si tenemos los siguiente conceptos de SNOMED CT *Increased energy (finding)* y *Increased (qualifier value)* como candidatos para vincular con el término *increased* que pertenece al contexto de *energy*, el concepto que se seleccionará como válido no será el que coincida exactamente en caracteres (*Increased (qualifier value)*), si no que será el que contiene información del término y del contexto (*Increased energy (finding)*).

También se pueden establecer otro tipo de criterios para escoger/desambiguar los vínculos como hacerlo de forma manual por parte de los usuarios, utilizar umbrales para garantizar asociaciones de cierta calidad o hacerlo a través de aprendizaje automático.

3.5. Herramientas para la Equiparación

A continuación, se van a comentar una serie de herramientas que son candidatas para la resolución del problema de alineamiento y anotación.

Debido a la variedad de herramientas que existen cada una de ellas puede resolver una o varias partes del problema de extracción de conocimiento a diferentes niveles. Por este motivo, clasificamos las herramientas en aquellas especificas que ayudan al análisis sintáctico y otras más generales que se centran en el reconocimiento de entidades nominales. Las que resuelven problemas más básicos en cuanto a NLP se refiere se integran dentro de las que implican semántica, tal es el caso de, MGREP y OBA; SPECIALIST NLP tools, MetaMap y SemRep; OpenNLP y Stanbol; Stanford en GATE, ...

3.5.1. Herramientas para el Análisis Sintáctico

La segmentación de textos es un problema que aún no tiene una solución satisfactoria, sigue siendo muy complicado encontrar herramientas software de este tipo. Las etapas más básicas de la extracción de información (EI) son tres: segmentación de textos en oraciones (Sentence Splitter), etiquetado de cada palabra con una categoría gramatical (part-of-speech, POS) y desambiguación del sentido de la palabra, y son estas etapas en las que nos hemos basado para la realización del estudio de la selección de herramientas. Las dos primeras etapas son en las que participan las herramientas de análisis sintáctico.

Una serie de software de muestra que puede ayudar en el análisis sintáctico podría clasificarse en:

- Segmentación el texto, tales como JTextTile [49] o TextSeg [168] entre otras comentadas en [111].

- Etiquetado de las palabras, como LASSIE [74], SPECIALIST NLP Tools, OpenNLP o Stanford.

Si tenemos en cuenta que un analizador de lenguaje natural es un programa que trabaja sobre la estructura gramatical de las oraciones, por ejemplo, qué grupos de palabras van juntas (como frases) y qué las palabras son el sujeto u objeto de un verbo. Debido a la completitud funcional que nos ofrecen los analizadores, nos quedamos con los recursos de OpenNLP y Stanford. Otros motivos por los que las hemos seleccionado son:

- Herramienta de dominio público.
- Implementada en Java, por lo que se simplifica su integración posterior con otras herramientas de terminología médica.
- Ha sido probada con textos largos. Realizando algunas tareas de preprocesado, tales como, sustitución de abreviaturas, símbolos matemáticos (>, <, =, etc.), preparación de tablas, sustitución de abreviaturas por nombres completos; se han obtenido resultados con un 90 % de precisión.

A continuación se enumeran las particularidades encontradas en el uso de este tipo de herramientas sobre las GPC, algunos de ellos fueron posteriormente solventados mediante un preprocesado automático y manual del texto extraído:

- Las abreviaturas del tipo *e.g.* no son procesadas de la forma esperada.
- Las unidades de medida deben estar correctamente escritas, en cuanto a espacios se refiere, para que el análisis del módulo de etiquetado sea válido. Por lo tanto, texto de la forma *60mmHg*, debería escribirse *60 mmHg* para que se pueda identificar cada elemento correctamente en el análisis sintáctico.
- Expresiones matemáticas no son correctamente analizadas por la herramienta. Expresiones de este tipo serían $50\,\% \leq FEV_1 < 80\,\%$ *predicted* o $\frac{FEV_1}{FVC} < 0{,}70$
- Signos de puntuación suelen producir análisis defectuosos por parte del módulo de chunker y parsing que explicaremos a continuación. Un ejemplo de un análisis erróneo del signo / en la frase nominal *night/early morning* se analizaría night/early como adjetivo de morning, mientras que el procesado correcto implica separar *night* y *early* en dos adjetivos diferentes.
- En ocasiones, el uso de mayúsculas en el texto y versalita para títulos de secciones en el texto puede provocar que esas palabras sean etiquetadas como nombre propio cuando en realidad no lo son.
- El análisis de palabras no comunes, como por ejemplo nombres de medicamentos, puede ser equivocado. Así, palabras como *tiotropiumm* son analizadas como interjecciones.
- Enumeraciones simples en las que algún elemento contenga una frase verbal, ésta será detectada como el núcleo verbal de la enumeración.

– Cuando el analizador se encuentra con una enumeración de varias frases nominales separadas por el signo ortográfico , o las conjunciones *and* u *or* cabe la posibilidad de que este conjunto de frases se analice como un todo, cuando lo que realmente nos interesa es que analice cada frase nominal por separado para facilitar su extracción posteriormente.

– En la fase de detección de oraciones si se encuentra con frases nominales aisladas que constituyen oraciones independientes, éstas pueden ser incluidas erróneamente en un análisis conjunto con la oración consecutiva en el texto que incluya una frase verbal.

– Debido a la gran ambigüedad del lenguaje, en cualquier herramienta de análisis sintáctico podríamos encontrarnos con errores de asignación de frases, veámoslo sobre un ejemplo: si nos encontramos con la siguiente frase *follow-up and home care arragements* el módulo TreeBank Chunker de el analizador la dividirá en las dos frases siguientes *follow-up* y *home care arragements*, cuando en realidad el texto se estaba refiriendo a *follow-up arragements* y *home care arragements*. Para solucionar este tipo de errores, tanto computacionalmente como en una comunicación oral, texto escrito,... es necesaria cierta información referente al contexto de la frase.

OpenNLP

OpenNLP tools[12] es una colección de herramientas desarrolladas en java para el PLN. Este conjunto de herramientas utiliza un paquete software (Maxent) con posibilidad entrenamiento y que hace uso de modelos de máxima entropía para la resolución de ambigüedades. Hay que destacar también que OpenNLP trabaja sobre gran variedad de idiomas como español, francés, alemán, portugués, danés, holandés, chino, árabe, ... Un ejemplo de integración de OpenNLP en un conjunto de componentes dedicados a la gestión del contenido semántico es Apache Stanbol[13] para el cual constituye el entorno de PLN por defecto.

Las herramientas OpenNLP permiten la realización de un análisis sintáctico. Veamos cómo se produciría un análisis de este tipo sobre dos párrafos de la GPC usada en este trabajo:

```
Severe COPD is characterized by repeated exacerbations that almost
 always have an impact on patients' quality of life.
Cigarette smokers have a higher prevalence of respiratory symptoms
```

[12] http://opennlp.apache.org/
[13] https://stanbol.apache.org/index.html

```
and lung function abnormalities. Pipe and cigar smokers have
greater COPD morbidity, although their rates are lower than those
for cigarette smokers. Passive exposure to cigarette smoke may
also contribute to respiratory symptoms and COPD.
```

Sentence Detector Como su propio nombre indica, se encarga de la identificación de oraciones.

```
Severe COPD is characterized by repeated exacerbations that
 almost always have an impact on patients' quality of life.
Cigarette smokers have a higher prevalence of respiratory
 symptoms and lung function abnormalities.
Pipe and cigar smokers have greater COPD morbidity, although
 their rates are lower than those for cigarette smokers.
Passive exposure to cigarette smoke may also contribute to
 respiratory symptoms and COPD.
```

Tokenizador Este módulo es el encargado de dividir las oraciones en tokens separados por espacios. Estos tokens son necesarios para el análisis en módulos sucesivos. Cada palabra se corresponde con un token, pero algunas de ellas pueden dividirse en varios.

```
Severe COPD is characterized by repeated exacerbations that
 almost always have an impact on patients ' quality of life .
```

Parts-of-speech(POS) Tagger o también etiquetador de categoría gramatical. Este módulo asigna a cada token una etiqueta empleando un diccionario de etiquetas y los modelos entrenados. Este etiquetado no proporciona ninguna información de la estructura de la oración, los módulos siguientes se encargarán de ello.

```
Severe/NNP COPD/NNP is/VBZ characterized/VBN by/IN
 repeated/VBN exacerbations/NNS that/IN almost/RB always/RB
 have/VBP an/DT impact/NN on/IN patients/NNS '/POS quality/NN
 of/IN life/NN ./.
```

Chunker Treebank Es el encargado de identificar las frases nominales, verbales, adverbiales, . . . dentro de las oraciones sin un consumo excesivo de memoria.

```
[NP Severe/NNP COPD/NNP ] [VP is/VBZ [VP characterized/VBN
 [PP by/IN [NP [NP repeated/VBN exacerbations/NNS ]
 [WHNP that/IN ] [ADVP almost/RB always/RB ]  [VP have/VBP
 [NP [NP an/DT impact/NN ] [PP on/IN [NP [NP [NP patients/NNS
 '/POS ] quality/NN ] [PP of/IN [NP life/NN ] ] ]
 ] ] ] ] ] ] ] ./.
```

Treebank Parser o analizador Treebank. Este módulo etiqueta los tokens agrupando las frases en una jerarquía, construyendo árboles de oraciones en modo texto. Un ejemplo del árbol sintáctico desarrollado gráficamente puede observarse en la figura 3.8. Funciona de forma más eficiente, cuanto mayor sea el volumen de texto analizar.

```
(TOP (S (NP (NNP Severe) (NNP COPD)) (VP (VBZ is) (VP (VBN
characterized) (PP (IN by) (NP (NP (VBN repeated) (NNS
exacerbations)) (SBAR (WHNP (IN that)) (S (ADVP (RB almost)
(RB alwaUMLSys)) (VP (VBP have) (NP (NP (DT an) (NN impact))
(PP (IN on) (NP (NP (NP (NNS patients) (POS ')) (NN quality))
(PP (IN of) (NP (NN life))))))))))) (. .)))
```

Stanford

Este paquete software [20] es una implementación Java de analizadores probabilísticos de LN, optimizado para analizadores basados en gramáticas libres de contexto probabilísticas (GLCP) y lexicalizadas por separado y en conjunto. La versión original fue desarrollada por Dan Klein y Christopher Manning y, posteriormente, se amplió con la internacionalización y el modelado específico de idioma, entradas y salidas de la herramienta configurables, la compactación gramática, análisis reticular, análisis k, salida de dependencias gramaticales ”mecanografiada“, soporte a usuarios, . . .

El analizador probabilístico lexicalizado implementa un modelo de producto factorizado, es decir, posee expertos separados para la estructura de frase basada en GLCP y para la dependencia léxica. Ambos se combinan de la forma más eficiente gracias a la inferencia exacta

utilizando el algoritmo A^*. El software bien puede utilizarse simplemente como un analizador de gramática libre de contexto estocástico no lexicalizado, o bien como un sistema de análisis estadístico. También proporcionan una interfaz gráfica de usuario para ver el árbol que forma la estructura de la frase generado por el analizador.

Además de proporcionar un analizador de inglés, el analizador puede ser y ha sido adaptado para trabajar con otros idiomas. Un analizador de chino basado en la Treebank china, también se incluye un analizador alemán basado en el corpus Negra y analizadores árabe basados en la Penn Treebank árabe. El analizador también se ha utilizado para otros idiomas, como el italiano, búlgaro y portugués.

El software genera como salida árboles de dependencia universal[14], de dependencias Stanford[15] o con la estructura de la frase (relaciones gramaticales). Estos formatos están disponibles sólo para Inglés y Chino.

Para el ejemplo anterior, el resultado final de un análisis sintáctico realizado por Stanford[16] sería el siguiente en el que podemos observar que el conjunto de etiquetas es el mismo que OpenNLP y que el resultado obtenido es igual:

```
(ROOT
 (S
   (NP (NNP Severe) (NNP COPD))
   (VP (VBZ is)
     (VP (VBN characterized)
       (PP (IN by)
         (NP
           (NP (VBN repeated) (NNS exacerbations))
           (SBAR
             (WHNP (WDT that))
             (S
               (ADVP (RB almost) (RB always))
               (VP (VBP have)
                 (NP
                   (NP (DT an) (NN impact))
                   (PP (IN on)
                     (NP
```

[14] http://universaldependencies.github.io/docs/
[15] http://nlp.stanford.edu/software/stanford-dependencies.shtml
[16] http://nlp.stanford.edu:8080/parser/index.jsp

```
                    (NP
                      (NP (NNS patients) (POS '))
                      (NN quality))
                    (PP (IN of)
                      (NP (NN life))))))))))))))
(. .)))
```

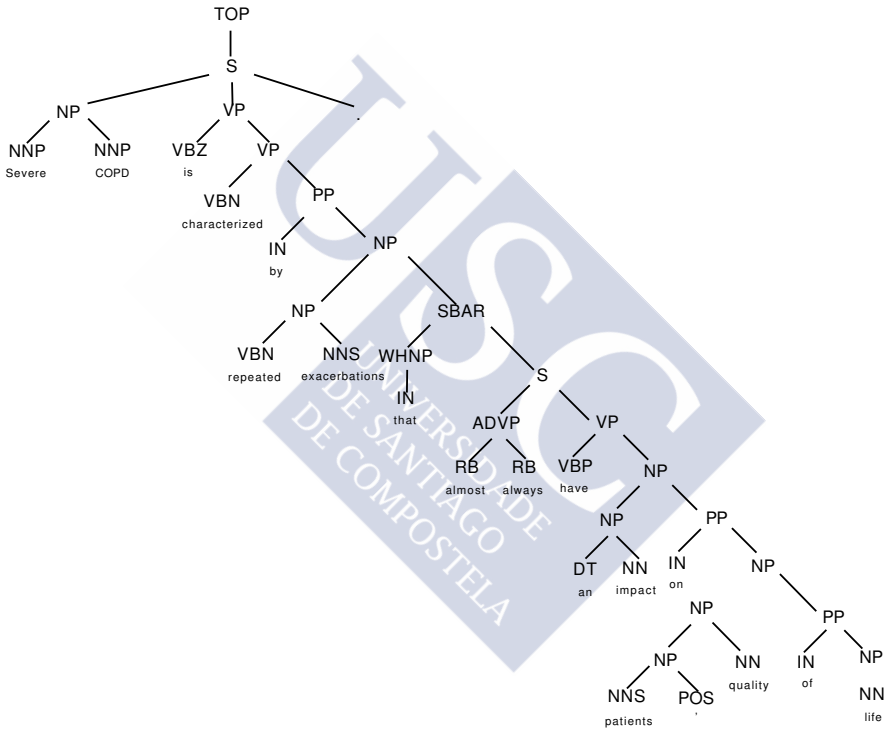


Figura 3.8: Desarrollo gráfico del árbol sintáctico creado por OpenNLP

The SPECIALIST NLP Tools

Otra herramienta importante para el análisis sintáctico es el SPECIALIST NLP Tools, a pesar de que su uso sea más frecuente de forma embebido en otras herramientas más que su aprovechamiento de forma independiente.

Las herramientas SPECIALIST Natural Language Processing[17] han sido desarrolladas por el grupo de sistemas léxicos del centro Lister Hill National de las comunicaciones biomédicas. Fue creado para investigar la contribución que las técnicas de PLN pueden hacer para mediar entre el lenguaje de los usuarios y el de los recursos biomédicos.

Esta herramienta está creada sobre el SPECIALIST Lexicon ya comentado en la sección 2.2.2. Entre las facilidades que ofrecen este conjunto de herramientas están aquellas que facilitan el acceso a la información del lexicón: generar variantes léxicas para acceder a UMLS o sobre textos en lenguaje natural, revisión de la ortografía, etiquetado de categoría gramatical, herramientas gráficas ...

3.5.2. Herramientas de Reconocimiento de Entidades

A día de hoy existen diferentes herramientas que nos ayudan a anotar texto. Todo proceso de anotación de textos no estructurados comienza por una identificación de conceptos o entidades en ese texto. El identificador de entidades más conocido a día de hoy es MetaMap [6], el cual es tomado como referencia y como *gold standard*. Otro identificador que últimamente está consiguiendo gran relevancia es MGrep [148]. MGrep realiza un alineamiento de cadenas de caracteres lo que obtiene una mayor precisión de resultados, sin embargo, a pesar de que Bhatia [13] no lo comenta en su artículo centrado en las diferencias entre MetaMap y MGrep, MetaMap debería tener un recall[18] mayor puesto que el número de conceptos recuperados supera al de MGrep. Otra diferencia es que MGrep, al contrario que nuestro trabajo, se usa en entornos en los que prima el tiempo de respuesta, por eso tampoco utiliza técnicas de PLN. MGrep posee la ventaja de que atiende a una gran diversidad de formato de recursos clínicos y terminologías, por la contra, MetaMap está demasiado ligado a UMLS lo que implica un preprocesado laborioso del uso de una terminología diferente.

Veamos ahora algunos ejemplos de herramientas de identificación de entidades.

METAMAP

MetaMap [7] es la herramienta para la identificación de entidades semánticas en texto natural por excelencia. El resto de sistemas que utilizan identificadores de conceptos o entidades la toman como referencia y como *gold standard*.

[17] `http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html`
[18] RECALL: fracción de conceptos recuperados relevantes frente al total de documentos relevantes

MetaMap [19] es un software configurable que alinea textos biomédicos con el UMLS Metathesaurus[20]. Esta herramienta [6] ofrece una serie de opciones para controlar tanto la salida como el procedimiento, por ejemplo, dividir o no las frases nominales, tener en cuenta las preposiciones o ignorar el orden de las palabras.

En sus comienzos MetaMap realizó el alineado de forma manual para conseguir deducir información precisa sobre cómo debería funcionar el algoritmo de mapeo que iban a utilizar.

Dependiendo de cómo se mapeara la frase, MetaMap clasificaba la salida en cuatro categorías:

– Correspondencia simple: en la que la frase de entrada es igual a la cadena de caracteres de Metathesaurus.

– Correspondencia compleja: en la que, a pesar de que no existe una correspondencia exacta con Metathesaurus, cada parte del nombre tiene su correspondencia simple.

```
Processing: external ventricular restoration
Phrase: external ventricular restoration
>>>>> Phrase
external ventricular restoration
<<<<< Phrase
>>>>> Mappings
Meta Mapping (851):
   660   External (Extrinsic) [Spatial Concept]
   660   Ventricular (Heart Ventricle) [Body Part, Organ, or
         Organ Component]
   827   Restoration (Type of restoration) [Intellectual
         Product]
Meta Mapping (851):
   660   External (Extrinsic) [Spatial Concept]
   660   Ventricular [Spatial Concept]
   827   Restoration (Type of restoration) [Intellectual
         Product]
<<<<< Mappings
```

[19] http://metamap.nlm.nih.gov/
[20] https://uts.nlm.nih.gov/home.html

– Correspondencia parcial: El término se asocia con Metathesaurus pero hay al menos una palabra de la frase o Metathesaurus que no participan en la asociación. Las alineaciones parciales tienen variantes:

  - Correspondencia parcial normal: En el siguiente ejemplo, para la palabra *thrombo* no existe correspondencia.

```
Processing: Thrombo embolism prophylaxis
Phrase: Thrombo embolism prophylaxis
>>>>> Phrase
thrombo embolism prophylaxis
<<<<< Phrase
>>>>> Mappings
Meta Mapping (790):
   660   EMBOLISM (Embolism) [Pathologic Function]
   827   Prophylaxis (Prophylactic treatment) [Therapeutic
         or Preventive Procedure]
Meta Mapping (790):
   660   EMBOLISM (Embolism) [Pathologic Function]
   827   prophylaxis (prevention & control) [Intellectual
         Product]
Meta Mapping (790):
   660   Embolism, NOS (Embolus) [Finding]
   827   Prophylaxis (Prophylactic treatment) [Therapeutic
         or Preventive Procedure]
Meta Mapping (790):
   660   Embolism, NOS (Embolus) [Finding]
   827   prophylaxis (prevention & control) [Intellectual
         Product]
<<<<< Mappings
```

  - Correspondencia parcial con intervalos: Los resultados de MetaMap añaden palabras diferentes a las de la consulta.

```
Processing: peptide measurement
Phrase: peptide measurement
>>>>> Phrase
peptide measurement
```

```
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
   827   Peptide Hormone Measurement [Laboratory Procedure]
Meta Mapping (1000):
   827   Peptide YY Measurement [Laboratory Procedure]
<<<<< Mappings
```

- Sobrecorrespondencia: El nombre buscado tiene una o más asociaciones en Metathesaurus.

```
Processing 00000000.tx.1: anaemia
Phrase: anaemia
>>>>> Phrase
anaemia
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
  1000   ANAEMIA (Anemia) [Disease or Syndrome]
Meta Mapping (1000):
  1000   Anemia (Genus Anemia) [Plant]
<<<<< Mappings
```

– Sin correspondencia: ninguna parte de la frase buscada está asociada con Metathesaurus.

Los pasos generales que aplica MetaMap de forma secuencial para realizar un mapeo automático son los siguientes:

Análisis sintáctico: herramientas del SPECIALIST (3.5.1) segmentan el texto en frases nominales y verbales produciendo un análisis sintáctico de alto nivel que también proporciona un etiquetado gramatical. En esta fase también ignora las palabras sin contenido semántico (stop words).

Generación de variantes: Para las frases extraídas, se generan las variantes utilizando la herramienta comentada en 2.2.2. Las variantes son un conjunto de frases que englo-

ban variaciones ortográficas, abreviaturas, acrónimos, sinónimos, variantes léxicas y las combinaciones de estas.

Obtención de candidatos: De la lista de variantes se obtiene una de Metacandidatos. Estos son un conjunto de conceptos Metathesaurus que se obtienen aplicando las tácticas de correspondencia sobre las variantes.

Evaluación de los candidatos: cada metacandidato se evalúa contra el texto de entrada comparando las palabras de la frase origen con las del candidato. Finalmente, se obtiene un peso aplicando un algoritmo basado en los siguientes principios:

- Contenido del núcleo de la frase nominal por parte del candidato.
- Calculo de la distancia entre las variantes y la cadena de Metathesaurus.
- Cobertura en el sentido de cuántas palabras de la frase y de Metathesaurus participan en el mapeo.
- Cohesividad, teniendo en cuenta la secuencia máxima de palabras continuas que participan en la alineación.
- Implicación de una palabra en el mapeo, evaluando qué importancia tendría su eliminación en la asociación con Metathesaurus.

Construcción de la asociación: combinando los candidatos en conjuntos disjuntos para conseguir una representación completa de la frase nominal original asignando a cada uno un peso final para posteriormente ordenarlos de forma descendiente en base a un ranking y pasando al problema de la ambigüedad. Podemos ver el resultado que genera en los ejemplos anteriores de categorías de clasificación de salida de MetaMap.

MGREP

Es una estructura de datos basado en el árbol de radix[21] que permite un mapeo rápido y eficiente de texto contra un conjunto de términos de diccionario ([148]).

Es usado por Open Biomedical Annotator (OBA[22]) en su primera fase de correspondencia de cadenas de caracteres. No está centrado en extraer conocimiento preciso de un recurso

[21] Estructura de datos arborescente de búsqueda en el que las claves son cadenas de caracteres. Las claves definen la posición dentro del array de forma que, para un nodo concreto, todos sus descendientes tienen un prefijo común. Por ejemplo, para el nodo con clave *health* los hijos serán *healthy*, *healthcare*, *healthier*, ...

[22] `http://bioportal.bioontology.org/annotator`

médico. Su objetivo principal es la escalabilidad y el tiempo de respuesta para poder gestionar la gran cantidad y diversidad de documentos clínicos existentes.

Lependu [84] utiliza esta herramienta para la detección efectos secundarios graves a causa del uso de medicamentos. Este sistema consiste básicamente en la anotación de documentos no estructurados de historiales de pacientes que hacen referencia al fármaco y a la enfermedad que se supone que produce ese fármaco. En este caso, no es necesaria una extracción del conocimiento completo incluido en el registro, si no que solamente identifica un par de señales de forma rápida en una elevada cantidad de documentos. Por este motivo no usa técnicas de NLP de base, solo lo hace exclusivamente para la detección de oraciones negadas.

SAPHIRE

SAPHIRE, Semantic and Probabilistic Heuristic Information Retrieval Environment de la Universidad de Oregon Health Sciences ([50]) es un sistema de indexación que usa una aproximación basada en la búsqueda de patrones para extraer de texto los conceptos incluidos en el vocabulario de Metathesaurus, por lo tanto, no hace uso de información sintáctica.

SAPHIRE puede ejecutarse de dos formas: con un alineamiento total o parcial. Su funcionamiento se basa en la segmentación de la entrada en palabras y en utilizar esas palabras para indexar conceptos del Metathesaurus. La lista de candidatos se crea en función de unas condiciones sobre el peso de la palabra y el número de alineamientos existentes entre las palabras de origen y los conceptos de Metathesaurus. Utiliza el concepto de *palabras frecuentes* (en función de la repetición en la entrada y en diferentes documentos) para mejorar el tiempo de respuesta de la herramienta.

Debido a que SAPHIRE proporciona un alineamiento múltiple posteriormente incluyó criterios de desambiguación basados en el tipo semántico al que pertenecen los conceptos candidatos ([132]).

INDEXFINDER

IndexFinder ([177]) es un algoritmo para el alineamiento de conceptos UMLS para aplicaciones en tiempo real.

IndexFinder usa técnicas sintácticas y semánticas para filtrar el resultado del alineamiento. Además tiene en cuenta el problema de la ambigüedad de las siglas, utiliza la normalización de UMLS, aplica desambiguación semántica y sintáctica. Sin embargo, no es una herramienta basada en PLN, lo que lo hace más rápida.

GATE

GATE [25] es una arquitectura, un marco de trabajo y un entorno de desarrollo para ingeniería del lenguaje. Está organizado en componentes donde cada uno de ellos realiza una función y están comunicados entre sí mediante tecnologías estándar abiertas lo que garantiza la interoperabilidad, la reusabilidad, reduce el solapamiento de los módulos y, además, es plurilingüe.

GATE[23] nos permite desarrollar una aplicación en el entorno gráfico solamente con especificar qué recursos de procesado deseamos utilizar y sobre qué datos los vamos a ejecutar. Los resultados se pueden mostrar en diferentes formatos XML, RTF, HTML, SGML, correo electrónico o texto plano. Debido a la flexibilidad de GATE, podemos encontrarnos multitud de herramientas desarrolladas sobre este entorno que se pueden usar de forma modular, por ejemplo, para el etiquetado gramatical de palabras, identificador de oraciones, el reconocimiento de entidades, la identificación de referencias entre elementos en un texto, etiquetador UMLS . . . Además de herramientas de evaluación como *AnnotationDiff* para realizar mediciones sobre los resultados obtenidos y para visualizarlos y herramientas para hacer seguimiento del procesado del lenguaje. En junio del 2015, se lanza la versión 8.1 de la herramienta.

OPEN BIOMEDICAL ANNOTATOR

El Open Biomedical Annotator (OBA) [71] es un servicio web[24] que nos permite etiquetar texto en lenguaje natural y etiquetarlo con conceptos de ontologías biomédicas.

Cada asociación tiene ligado un valor en función del contexto en el que fue generado. Primero, hace una equiparación de cadenas de caracteres entre el texto y todos los términos de la ontología. OBA no aplica ninguna técnica de procesado del lenguaje natural para la asociación, por ejemplo, dada la frase: *The heart has a failure* obtenemos respuesta para *heart* y para *failure*. Sin embargo, para la frase: *The heart had failed* solo se obtienen resultados para *heart*.

Existe la posibilidad de incluir como asociaciones válidas los ascendientes a los resultados del paso anterior. Es el usuario de OBA el encargado de especificar los niveles de la jerarquía que quiere incluir. De la misma forma, los desarrolladores consideran la posibilidad de validez de los conceptos que están a una distancia concreta (llamada distancia semántica) del concepto mapeado en la fase inicial. Finalmente, OBA utiliza las referencias que tienen

[23] `http://gate.ac.uk`
[24] `http://bioportal.bioontology.org/annotator`

las ontologías a otros recursos para realizar las asociaciones. Utilizando este anotador han creado el índice Open Biomedical Resources (OBR) que permite al usuario buscar todos los documentos relacionados etiquetados con un anotador concreto.

CONNAN

CONNAN [133] es un anotador de conceptos biomédicos. Usa un enfoque de filtrado incremental para reducir la lista de las frases candidatas específicas del dominio antes de decidir el mejor alineamiento con la frase origen. CONNAN mejora el tiempo de ejecución sobre otros anotadores y es fácil de usar, lo que lo hace adecuado para entornos online. Para la evaluación de las alineaciones utiliza medidas como la cobertura y la coherencia que mide gracias a dos tipos diferentes de filtros y a la semántica de las palabras de la ontología. Permite aplicar un alineamiento exacto, así como un alineamiento flexible obteniendo respectivamente una precisión del 90 % y 95 %.

SemRep

La herramienta de SemRep está encuadrada dentro del proyecto de *Semantic Knowledge Representation* (SKR[25]).

Este proyecto lleva a cabo una investigación básica en PLN haciendo uso de los medios de conocimiento proporcionados por UMLS. Un recurso básico es el programa SemRep ([135, 134, 51]), que extrae predicados semánticos del texto. SemRep fue creado originalmente para la investigación biomédica, y ahora, se está desarrollando una metodología general para extender su dominio de aplicación. En estos momentos, su dominio se centra en en la prevención ante casos de epidemia de gripe, fomento de la salud, y efectos del cambio climático en la salud.

El proyecto SKR mantiene una base de datos de 68 millones de predicados SemRep extraídos de las citas de MEDLINE. Esta base de datos es compatible con la aplicación web *Semantic MEDLINE* (ver figura 3.9), que integra la búsqueda en PubMed, predicados de SemRep, el resumen automático y visualización de datos. La aplicación está destinada a ayudar a los usuarios a administrar los resultados de búsquedas en PubMed. La salida se visualiza como un gráfico informativo con enlaces a las citas de MEDLINE originales. También se pro-

[25] `http://skr3.nlm.nih.gov/index.html`

porciona un cómodo acceso a los recursos de conocimiento adicionales, tales como Entrez Gene, Genetics Home Reference, y UMLS Metathesaurus.

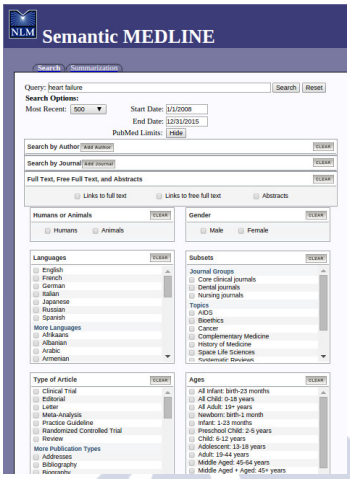


Figura 3.9: Servicio Web para el buscador de MEDLINE.

SKR se dedica al desarrollo de aplicaciones innovadoras para la gestión de información en biomedicina, así como a investigación básica. El equipo del proyecto utiliza predicados semánticos para encontrar publicaciones que responden a preguntas utilizadas durante la creación de guías de práctica clínica con el apoyo del National Heart, Lung, Blood Institute. Además, la tecnología semántica MEDLINE se adaptó para trabajar en otros ámbitos.

La investigación más importante de SKR está centrada desarrollar y aplicar el reconocimiento en literatura utilizando predicados semánticos. Uno de estos proyectos trabaja la fisiología del sueño y patologías asociadas, tales como la disminución de la calidad del sueño en hombres de edad avanzada, el síndrome de piernas inquietas y la apnea obstructiva del sueño; otro estudio utiliza los predicados y la teoría de grafos para el resumen automático de textos biomédicos. Además, el equipo SKR está colaborando en el uso de predicados semánticos para ayudar a interpretar los resultados de los experimentos de microarrays, para investigar métodos estadísticos avanzados con el objetivo de mejorar la gestión de la información, y para hacer frente a las necesidades de información de los médicos en el punto de atención.

En resumen, principalmente, SKR realiza investigación y desarrollo en informática biomédica relacionados con la salud de los consumidores, los datos clínicos, tratamiento de imagen,

y procesamiento del lenguaje natural para informar y capacitar a pacientes, profesionales de la salud, investigadores y al público en general.

Dentro del proyecto SKR, centrándonos en SemRep, como ya comentamos es un software que extrae predicados semánticos (tripletas sujeto-objeto-relación) de LN biomédico. Tanto el sujeto como el objeto de cada predicado son conceptos de UMLS Metathesaurus y la relación (en mayúsculas) es una relación de la Red Semántica UMLS. En Rindflesch et al. [135] explican el funcionamiento básico de SemRep para la identificación de relaciones de hiperonimia en textos clínicos. A grandes rasgos, esta herramienta toma las SPECIALIST NLP tools para realizar un análisis sintáctico e identificar las frases nominales. A continuación, utiliza MetaMap para realizar la asociaciones entre las frases nominales y los conceptos del Metathesaurus filtrando por los TS que son necesarios. Finalmente, intenta extraer las relaciones hiperónimas identificando una frase nominal larga observando su núcleo, el uso de verbos específicos como *be* y en frases nominales consecutivas en la oración (coordinación).

SemRep se puede ejecutar de forma interactiva o en modo batch utilizando el SKR Scheduler. SemRep también está disponible como un programa de ejecución en local para Linux. Un ejemplo de ejecución de SemRep es el siguiente:

```
IMPUT TEXT:
   aldosterone antagonists, angiotensin-converting enzyme
   inhibitors, beta-blockers and diuretics in advanced heart
   failure.

RESULTS:
   00000000.tx.1|relation|C0012798|Diuretics|phsu|phsu|||TREATS|
        C0018801|Heart failure|dsyn|dsyn||
   00000000.tx.1|relation|C0001645|Adrenergic beta-Antagonists|
        phsu|phsu|||TREATS|C0018801|Heart failure|dsyn|dsyn||
   00000000.tx.1|relation|C0003015|Angiotensin-Converting Enzyme
        Inhibitors|phsu|phsu|||TREATS|C0018801|Heart failure|dsyn|
        dsyn||
   00000000.tx.1|relation|C0002007|Aldosterone Antagonists|phsu|
        phsu|||TREATS|C0018801|Heart failure|dsyn|dsyn||
```

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Capítulo 4

Una Propuesta para la Anotación Semántica de Modelos de Datos Clínicos

Uno de los principales retos de la sanidad electrónica es la interoperabilidad semántica de los sistemas de salud. Pero, esto solo será posible si se estandariza la recopilación, la representación y el acceso a los datos del paciente. Gracias a acuerdos llevados a cabo por expertos, existen modelos de datos clínicos, como los arquetipos OpenEHR, que definen estructuras de datos que garantizan la precisión de la información sanitaria. Además, son una alternativa para normalizar los datos clínicos por medio de la anotación de los términos utilizados en la definición de los modelos con vocabularios médicos estandarizados. Sin embargo, el esfuerzo necesario para establecer la unión entre términos del arquetipo y conceptos de una terminología estándar es considerable. El propósito de este capítulo de la tesis es proporcionar un enfoque automatizado para anotar los términos de los arquetipos OpenEHR con la terminología externa SNOMED CT, incluyendo la capacidad para hacerlo a nivel semántico.

4.1. Introducción

Los registros electrónicos médicos contienen principalmente los datos clínicos del paciente que incluyen la historia personal y familiar, el estado clínico, los tratamientos dispensados, y otra información relevante sobre los diagnósticos y los resultados alcanzados. El uso de len-

guaje natural para describir esta información proporciona un nivel completo de expresividad, pero dificulta el procesamiento computacional, lo que interfiere con uno de los principales retos de la sanidad electrónica [56], concretamente con la interoperabilidad semántica de los sistemas de salud. Lograr esto proporcionaría la posibilidad de gestionar automáticamente fragmentos arbitrarios de la información del paciente desde diferentes historias clínicas electrónicas, sin la necesidad de enlaces específicos entre ellas, al mismo tiempo que aumentaría la calidad de los servicios médicos asistenciales. Pero esto sólo será posible si los sistemas sanitarios se dotan de la capacidad de compartir la información de una manera significativa, inequívoca y precisa. Por lo tanto, la estandarización de la recopilación, representación y acceso a la información detallada y completa del paciente es esencial y por este motivo se han desarrollado los arquetipos. Como ya se ha comentado previamente en el capítulo 2 de esta tesis doctoral, varias propuestas de registro de historias clínicas, como *Good electronic Health Record* (GeHR) [11], OpenEHR [119] o CEN ISO EN13606 [34], se han enfocado a apoyar la asistencia sanitaria con modelos que permiten representar la información clínica de una manera correcta, robusta, fiable y sin ningún tipo de ambigüedad. Estas propuestas establecen la normalización de los datos clínicos en una arquitectura de tres capas [56], incluyendo:

- modelos de referencia,
- definiciones de la estructura de datos clínicos y
- sistemas de terminología clínica.

El modelo de referencia proporciona las piezas constituyentes para ser reutilizadas en la especificación de aspectos particulares de la información clínica. El uso de estos bloques constituyentes, hacen que los expertos puedan acordar una definición formal de los datos clínicos para garantizar una información médica precisa. Los principales candidatos para la definición de estructuras de datos son los arquetipos OpenEHR e ISO EN 13606 y las plantillas de HL7. Los arquetipos OpenEHR, a parte de ser los únicos que representan la historia clínica de forma completa, proporcionan una opción para normalizar el contenido de los datos clínicos que recopilan la información relacionada con el paciente. Esta estandarización la hacen por medio de la anotación entre los términos empleados en la definición del modelo con los conceptos pertenecientes a vocabularios médicos estándar. Como se comentó en el capítulo de 2, existen varias instituciones que han participado en el desarrollo de repositorios de arquetipos online que son mantenidos y actualizados por varios grupos de expertos que cooperan en diferentes

ámbitos. No obstante, con la excepción de un pequeño número de casos, son poco frecuentes las anotaciones de los arquetipos disponibles públicamente con uno o varios vocabularios estándar. Sin embargo, dicha anotación es un paso crucial para conseguir la interoperabilidad semántica entre los diferentes sistemas de información sanitaria.

Cuando se desarrolla un conjunto importante de arquetipos y son estables, el esfuerzo necesario para anotar fragmentos de arquetipo a una terminología estándar es considerable. Hasta ahora, la *International Health Terminology Standards Development Organisation (IHTSDO)* [58] y la Fundación OpenEHR han mostrado su interés en colaborar con el fin de indagar cómo se puede asociar la terminología SNOMED CT con los arquetipos OpenEHR para apoyar la historia clínica electrónica [54]. Además, una vez definidas las anotaciones, es necesario interpretarlos adecuadamente y garantizar su validez, y esta tarea es costosa en tiempo y recursos humanos. Por lo tanto, se necesitan métodos automatizados con el objetivo de simplificar el proceso de anotación para, a continuación, poder interpretar y evaluar las anotaciones resultantes de una manera más eficaz que la revisión manual.

4.2. Técnicas de anotación automática

El presente capítulo de esta tesis doctoral propone un enfoque automatizado para anotar los términos que componen los arquetipos con la terminología externa SNOMED CT. Este enfoque aplica una combinación de dos métodos básicos de alineación o mapping (conocidos como métodos léxicos y basados en contexto) con el fin de producir la anotación, en primer lugar, y posteriormente validarla. Las técnicas léxicas encuentran anotaciones utilizando la sinonimia de SNOMED CT, mientras que las técnicas basadas en el contexto o relacionales identifican similitudes semánticas entre la estructura del arquetipo y las relaciones en SNOMED CT.

4.2.1. Técnicas orientadas al descubrimiento de anotaciones

Algunas disciplinas como, por ejemplo, las ciencias de la información [176, 101], las bases de datos [30, 158] o la ingeniería ontológica [75, 117, 36, 76], han estado trabajando activamente en la mejora de las técnicas orientadas a descubrir equivalencias (o mappings) entre conceptos a través de distintos recursos. Estas técnicas se pueden aplicar ahora para crear anotaciones de forma automatizada. Las principales técnicas basadas en nombres utilizan las propiedades léxicas de las palabras para encontrar correspondencias entre conceptos.

Las técnicas basadas en estructuras utilizan propiedades estructurales (tales como, las relaciones compartidas a través de los recursos) para encontrar correspondencias entre conceptos. Generalmente, estas técnicas se utilizan en combinación con las basadas en nombres, puesto que está demostrado que se aumenta el rendimiento general [42], tal y como comentábamos en la sección 3.4.1. Además, estos procedimientos también se pueden utilizar para validar correspondencias léxicas [75, 117, 36, 76, 42, 162] (sección 3.4.2). Las técnicas basadas en recursos lingüísticos (secciones 3.2.2 y 3.2.2) utilizan medios externos (por ejemplo, diccionarios, lexicones y tesauros) buscando la asociación gracias a las relaciones lingüísticas entre las palabras (por ejemplo, sinónimos, hipónimos); las de recuperación de la información (RI) recuperan datos no estructurados que satisfacen cierta información contenidos en grandes colecciones de datos [89]. Finalmente, las basadas en contexto utilizan la información sobre la proximidad estructural y semántica, que permite identificar similitudes entre los conceptos, para encontrar alineaciones entre ellos [76].

Aunque hoy en día no hay métodos automatizados disponibles para anotar arquetipos con SNOMED CT, se han desarrollado algunos enfoques en el marco de proyectos de investigación [175, 85, 128, 141]. La tabla 4.1 resume los enfoques más relevantes y previos a nuestro trabajo, incluyendo también el estudio correspondiente a esta tesis doctoral. En el sistema desarrollado por Yu et al. [175] se han aplicado técnicas de RI (Lucene[1]) para anotar los términos del arquetipo, mostrando un buen tiempo de respuesta. Sin embargo, el método RI utilizado está más centrado en documentos de texto largos que en las breves descripciones de conceptos SNOMED donde la repetición de palabras es menos frecuente. Por otra parte, en algunas ocasiones este sistema no es capaz de encontrar anotaciones que alguna de las técnicas de normalización lingüística es capaz de hacer. La investigación de Lezcano et al. [85] utiliza las herramientas UMLS para anotar los términos del arquetipo con conceptos SNOMED CT. Además, utiliza un método de contexto para reducir la ambigüedad de la anotación. Este método consiste en la concatenación de cada término con el perteneciente al nivel superior en la jerarquía del arquetipo con el fin de generar expresiones más específicas y así aumentar la cobertura de la anotación. En el trabajo de Qamar y Rector [128] se integran varios de los métodos anteriores con técnicas basadas en recursos lingüísticos y reglas de post-filtrado. Este enfoque también incluye técnicas léxicas y recursos terminológicos externos, en combinación con técnicas basadas en el contexto, donde tienen en cuenta tanto la estructura anidada de los arquetipos como la estructura de definición jerárquica y lógica de SNOMED CT, con el

[1] `http://lucene.apache.org/java/docs/index.html`

Investigación	Basado en nombres	Basado en Contexto y en Estructura			Basada en recursos lingüísticos	Recuperación de información
		Contexto Estructural	Contexto Jerárquico	Contexto Lógico		
Yu et al.						X
Lezcano et al.		X			X	
Qamar y Rector	X	X			X	X
Berges et al.	X					
Nuestro trabajo	X	X	X	X	X	

Tabla 4.1: Resumen de las técnicas usadas por otros trabajos.

objetivo de aumentar el número de asociaciones correctas. Por otra parte, las técnicas basadas en el contexto se utilizan para validar las uniones léxicas resultantes, así como para resolver los problemas que surgen con las anotaciones ambiguas. Recientemente, se ha publicado el trabajo de Berges et al. [12] que asocia con SNOMED CT los elementos de los arquetipos pertenecientes al repositorio creado para nuestro trabajo [114]. Para ello han realizado un estudio de varias técnicas complejas basadas en subcadenas de caracteres sobre el campo *text* de la sección *ontology*. El mejor resultado lo han conseguido aplicando la técnica de Q-gramas. Con esto, se demuestra que, a pesar de que este tipo de tácticas no superan a otros tipos de métodos más complejos en resultados, suponen una alternativa eficiente y a tener en cuenta a la hora de proporcionar un conjunto de candidatos para la anotación considerablemente reducido.

4.2.2. Técnicas orientadas a la validación de las anotaciones automáticas

La validación de las anotaciones resultantes es la parte más crítica en este campo de investigación. La técnica ideal para hacerlo es frente a un *gold standard*, es decir, un conjunto de anotaciones de referencia elaborado por un grupo de expertos. Sin embargo, los repositorios de contenido abierto contienen una cantidad muy reducida de arquetipos con anotaciones a vocabularios estandarizados; y la construcción de un *gold standard* como tal es tedioso y costoso en tiempo y recursos humanos. En ausencia de un *gold standard*, las validaciones son difíciles y exclusivamente manuales. Sin embargo, a pesar de que una validación manual es imperfecta y no es ideal, es recomendable con el fin de sacar a la luz las fortalezas y limitaciones del método.

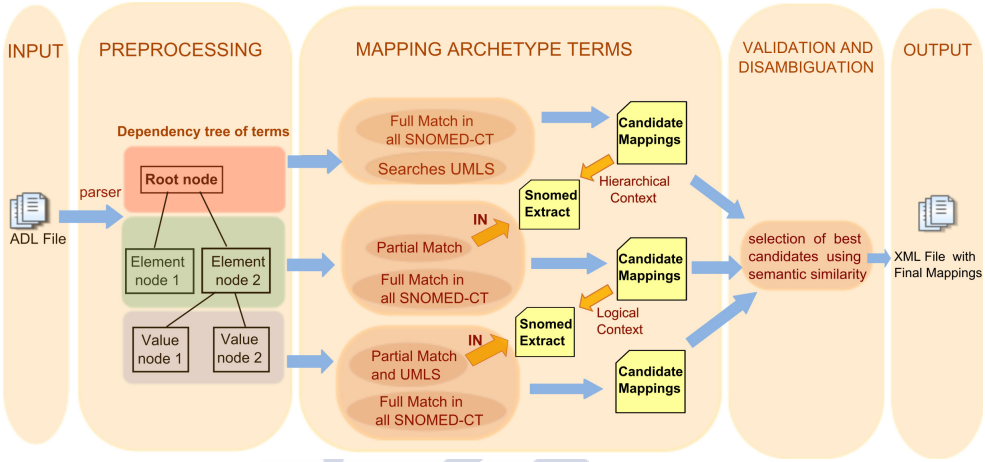


Figura 4.1: Fases generales del método propuesto para anotar los términos del arquetipo con conceptos de SNOMED CT.

4.3. El método de anotación propuesto

El método propuesto en esta tesis doctoral para anotar automáticamente fragmentos de arquetipos clínicos es el que se muestra en la figura 4.1. El algoritmo fundamentalmente es una combinación de técnicas léxicas y basadas en contexto. Las principales etapas involucradas en nuestra propuesta son:

- Pre-procesado del arquetipo como un árbol de términos.
- Anotación de términos del arquetipo con conceptos SNOMED CT.
- Validación y desambiguación del conjunto de anotaciones candidatas.

El proceso es completamente automático, garantizando la repetición de la anotación. A continuación, describimos el método con mayor profundidad de detalle, siguiendo las etapas mencionadas.

4.3.1. Pre-procesado del arquetipo

Como ya hemos comentado en el capítulo 2, los arquetipos clínicos se expresan mediante un lenguaje de definición denominado ADL (Archetype Definition Language). En la figura

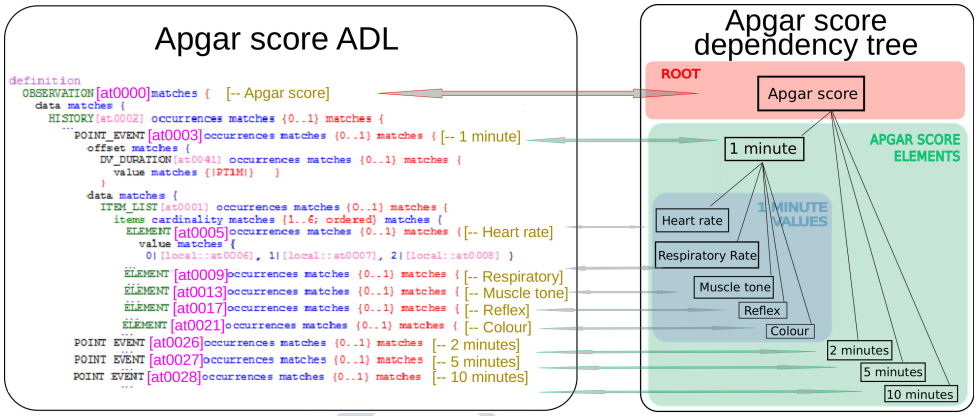


Figura 4.2: Extracción del árbol de dependencias a partir del fichero ADL que representa el arquetipo de *Apgar score*.

2.1 de dicho capítulo mostramos una representación del arquetipo de observación *Apgar Score*. Sobre él se comentan las tres secciones que lo constituyen: cabecera, cuerpo y ontología. La cabecera incluye la definición de los datos sobre el arquetipo; el cuerpo, organizado jerárquicamente, contiene la estructura y las restricciones de los conceptos clínicos necesarios para registrar los datos de los pacientes y que llamaremos términos; y la ontología incluye las definiciones de los términos empleados en el arquetipo y las correspondencias a conceptos de alguna terminología estándar, como SNOMED CT.

Para extraer la estructura jerárquica implícita en el cuerpo de los arquetipos, utilizamos un analizador sintáctico de ADL [21] que nos permitió generar automáticamente el árbol de dependencia de los términos contenidos en el archivo ADL del arquetipo. En la figura 2.2 del capítulo 2 puede verse esquemáticamente las entradas y salidas del analizador ADL. También la figura 4.1 muestra la etapa de pre-procesamiento de nuestro método y el árbol genérico de dependencia de términos que genera el analizador.

En términos generales, esta etapa posee el objetivo de simplificar la estructura del ADL y crear un archivo XML que se utilizará durante todo el desarrollo de la anotación. Sólo se procesan los nodos de las secciones *DATA* y *EVENTS* con significado clínico: el nodo *ROOT*, nodos *ELEMENT* y nodos *VALUE*. En la siguiente figura 4.2 se muestra cómo algunos de los términos del extracto son transformados en un árbol de dependencias, que mantiene la estructura definida en el arquetipo original.

4.3.2. Anotación de términos del arquetipo

El método propuesto se aplica una combinación de técnicas léxicas y basadas en contexto, siguiendo dos estrategias de forma secuencial. En primer lugar, describiremos las técnicas léxicas diseñadas e implementadas en esta parte de la tesis doctoral. En segundo lugar, detallaremos las técnicas basadas en contexto. Finalmente, nos centraremos en las estrategias seguidas para combinar las técnicas implementadas.

Técnicas Léxicas

Las técnicas léxicas se refieren a los métodos individuales de la sección 3.2 de cadenas de caracteres y, principalmente, a los basados en lenguaje.

Son técnicas que realizan la anotación buscando conceptos SNOMED CT que posean alguna descripción (es decir, un término preferido o un sinónimo) que sea similar léxicamente a los nombres de los términos del arquetipo. Primero pre-procesan tanto los términos del arquetipo como las descripciones de los conceptos de SNOMED CT. A continuación, realiza la equiparación total de los resultados del pre-procesamiento de ambos (Full Match en la figura 4.3).

En primer lugar, el pre-procesado de los nombres de los términos del arquetipo engloba las siguientes etapas:

- Filtrado de términos generales e imprecisos, como *comment*, *additional information* y *normal statement*.
- Tokenización de los nombres de los términos en sus palabras constituyentes (tokens), descartando las preposiciones.
- Combinación de los tokens con el fin de aumentar las posibilidades de encontrar un mapeo a algún concepto SNOMED CT. Los tokens se concatenan con otros que se definen a partir del contexto estructural (véase más adelante en la sección 4.3.2) del arquetipo y algunos tokens relevantes adicionales (por ejemplo, *observable* cuando el nombre del arquetipo coincide con una entidad observable de SNOMED CT). Por ejemplo, en la figura 4.3, el nombre *Apgar score*, de la raíz del arquetipo (ROOT), se combina con el nombre ELEMENT *1 minute*, generando el token *Apgar 1 minute*.

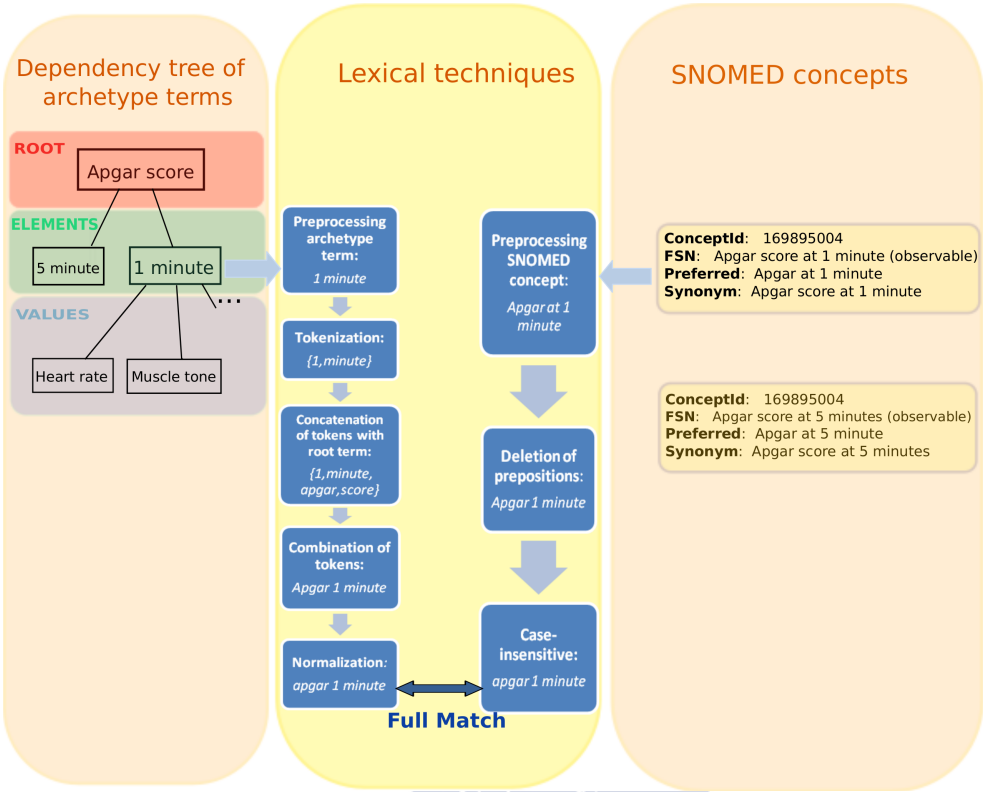


Figura 4.3: Etapas para anotar léxicamente los términos de un arquetipo con conceptos SNOMED CT. Un ejemplo de aplicación se muestra sobre el arquetipo de *Apgar Score*.

– Normalización de términos, que soluciona principalmente problemas de mayúsculas y minúsculas, plurales, singulares. También, la sustitución de abreviaturas de la *Word Equivalent Table*, la cual proporciona abreviaturas de uso común [61].

La coincidencia de un nombre de un término del arquetipo con el nombre de un concepto SNOMED CT comprende la correspondencia léxica tanto total y parcial. La correspondencia total se produce cuando un término del arquetipo, después de la normalización, es exactamente el mismo que alguna descripción SNOMED CT (término preferido o sinónimo). La coincidencia parcial tiene lugar cuando el nombre del término del arquetipo está contenido dentro de una cierta descripción SNOMED CT (véase figura 4.4, en la que se muestra un

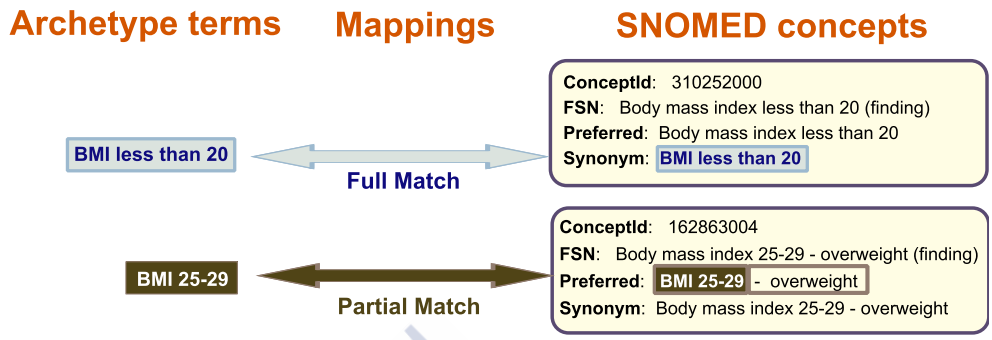


Figura 4.4: Un ejemplo de *correspondencia total* y *parcial*. El lado izquierdo muestra dos términos del arquetipo *body mass index*. El lado derecho muestra los conceptos de SNOMED CT que se anotan con los términos del arquetipo.

ejemplo de anotación por correspondencia léxica total y otro por correspondencia parcial para el arquetipo *body mass index*[2]).

Este método también utiliza recursos terminológicos externos cuando las técnicas léxicas propuestas no proporcionan anotaciones: el UMLS Metathesaurus [115] y MetaMap [110]. Ambos recursos ya han sido detallados en los capítulos 2 y 3 de esta tesis doctoral.

Técnicas basadas en el contexto

Estas se refieren a las técnicas relacionales de la sección 3.3 aplicadas de la siguiente forma: las taxonómicas se usan en los arquetipos y SNOMED CT, las mereológicas en el arquetipo y las lógicas en SNOMED CT.

Las técnicas basadas en contexto realizan la anotación descubriendo similitudes semánticas entre la estructura de agrupación de los términos del arquetipo, y las relaciones jerárquicas y lógicas definidas entre los conceptos SNOMED CT. Estas técnicas se basan en el **principio de vecindad**, que definimos a continuación.

```
SI
    Un término t de un arquetipo se anota correctamente con un
    concepto c de SNOMED CT,
Y
```

[2] http://www.usc.es/keam/TermArchetypes/Input_ADLs_Bindings/openEHR-EHR-OBSERVATION.body_mass_index.v3.adl

```
    El fragmento del arquetipo, al que pertenece el término t, se
    agrupa lógicamente en SNOMED CT,

ENTONCES
    Los términos relacionados estructuralmente en el arquetipo (es
    decir, anidados) con el término t son candidatos a ser anotados
    con alguno de los conceptos semánticamente relacionados con c
    en SNOMED CT.
```

Con el objetivo de extraer el contexto de SNOMED CT relevante para un término de un arquetipo, el método sigue el algoritmo de segmentación básica [145], previamente detallado en la sección 2.3.3. Este procede de la siguiente manera: se parte del conjunto de combinaciones de tokens del ADL obtenido con las técnicas léxicas, y se anotan estos tokens con conceptos SNOMED CT. A continuación, se generan extractos de SNOMED CT en torno a estos conceptos de SNOMED CT y a sus conceptos relacionados.

Distinguimos tres tipos de contextos, dos de ellos teniendo en cuenta la naturaleza de las dependencias, es decir, si las dependencias son entre términos del arquetipo, o si son entre conceptos SNOMED CT. El tercer tipo de contexto surge al diferenciar entre dependencias jerárquicas y lógicas.

En primer lugar, las dependencias entre los términos del arquetipo se pueden derivar de la estructura de este (es decir, de la jerarquía anidada, véase parte izquierda de la figura 4.3). El conjunto de términos ascendentes en la jerarquía del arquetipo con respecto a un término específico es lo que denominamos el **contexto estructural de ese término** específico. Por ejemplo, en la parte izquierda de la figura 4.5, podemos ver el árbol de dependencias de los términos definidos en el arquetipo *Tobacco use and exposure*[3]. Aquí, el contexto estructural del término *Never smoked (at0025)* es *Status (at0005)* y *Tobacco use (at0000)*.

En segundo lugar, las dependencias entre los conceptos SNOMED CT se pueden adquirir de forma recursiva al extraer el conjunto de todos los descendientes de un concepto específico en la jerarquía de SNOMED CT (es decir, lo que se denomina normalmente el cierre transitivo de relaciones *is a*). Este conjunto es el **contexto jerárquico del concepto** específico de SNOMED CT. De esta forma, la anotación se puede mejorar sustancialmente, haciendo equiparaciones léxicas parciales entre los términos del arquetipo y los conceptos SNOMED CT de ciertos segmentos extraídos (aquellos relevantes en el contexto de los términos del arquetipo),

[3] http://www.usc.es/keam/TermArchetypes/Input_ADLs_Bindings/openEHR-EHR-OBSERVATION.substance_use-tobacco.v7.adl

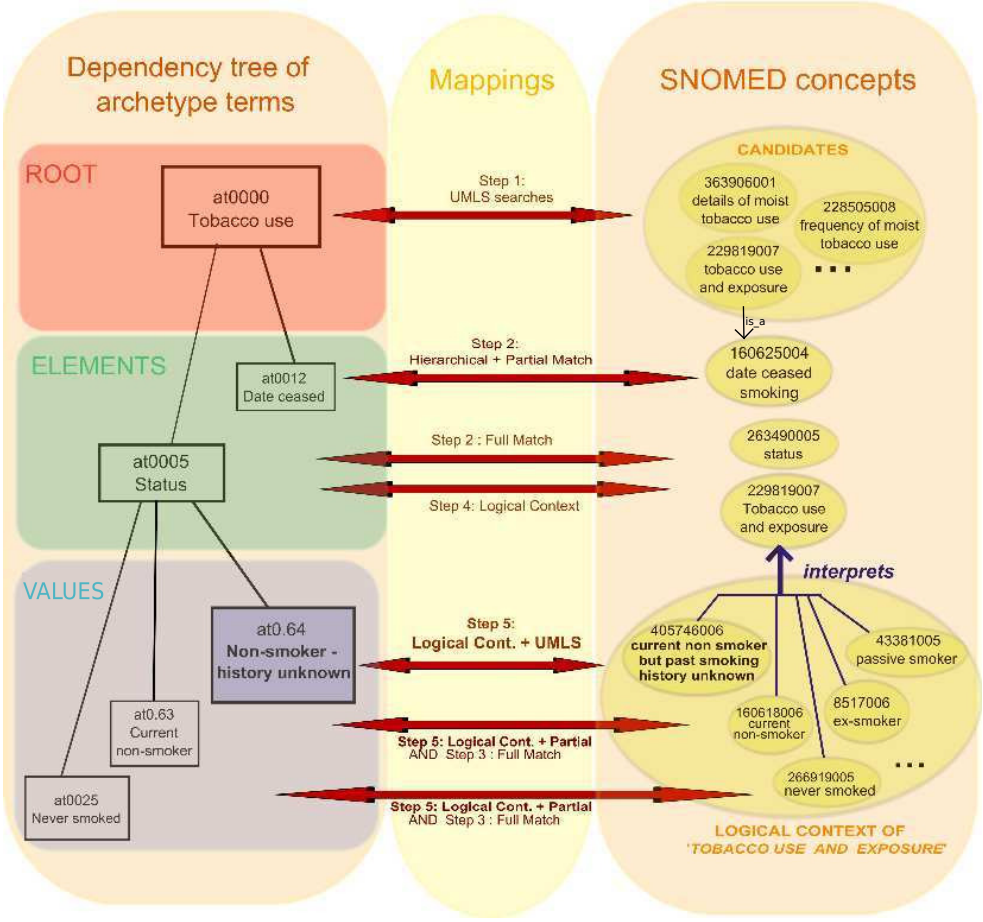


Figura 4.5: Un ejemplo de aplicación del método para parte del arquetipo *substance use tobacco*. El lado izquierdo muestra el árbol de dependencia del arquetipo. El lado derecho muestra el contexto jerárquico y lógico de *Tobacco use and exposure*.

en lugar de intentar equiparar de forma completa los términos ADL con todo SNOMED CT. Como un ejemplo, el lado derecho de la figura 4.5 muestra una parte del contexto jerárquico de *Tobacco use and exposure*.

En tercer lugar, las dependencias lógicas entre conceptos SNOMED CT se pueden obtener recorriendo el conjunto de las relaciones de SNOMED CT etiquetadas como *attribute* desde un concepto dado a otros conceptos. SNOMED CT contiene más de 50 relaciones de tipo atributo. Estas relaciones se incorporan en nuestro estudio ya que representan características cruciales de un concepto y, por ello, pueden complementar los otros contextos. El conjunto de conceptos relacionados lógicamente con un concepto específico SNOMED CT es lo que nosotros denominamos el **contexto lógico**.

La figura 4.5 muestra una pequeña parte del contexto lógico del *Tobacco use and exposure*; en concreto, el segmento de SNOMED CT relevante a dicho concepto, extraído a través de la relación de atributo *interprets*. Esta relación enlaza el hallazgo clínico *Tobacco use and exposure* con las entidades que se pueden observar para dicho hallazgo, como *ex-smoker* o *passive smoker*.

Combinación de técnicas léxicas y basadas en contexto

El método propuesto aplica una combinación de técnicas léxicas y basadas en contexto, siguiendo dos estrategias de forma secuencial. La primera estrategia sigue un enfoque de arriba hacia abajo (top-down). En concreto, recorre el árbol de dependencias del arquetipo desde el nodo *ROOT* hacia los nodos *VALUE*, pasando por los nodos *ELEMENT*. La estrategia aplica varias de las técnicas implementadas (correspondencia total, anotación basada en UMLS y MetaMap), hasta que consigue anotar el nodo *ROOT* con algún concepto candidato SNOMED CT. Por ejemplo, en la etapa 1 de la figura 4.5, la estrategia anota el término raíz *Tobacco use* con un conjunto de conceptos SNOMED CT, que llamamos conceptos candidatos, usando los recursos proporcionados por el UMLS. A continuación, cada candidato se expande, extrayendo automáticamente su contexto jerárquico de SNOMED CT. En la parte derecha de la figura se muestra el contexto jerárquico relevante al concepto *Tobacco use and exposure*.

Para cada nodo *ELEMENT*, se busca solamente una coincidencia parcial dentro del contexto jerárquico de SNOMED CT, extraído previamente. Por ejemplo, en la figura se muestra el término *Date ceased*, que ha sido anotado, por equipararlo parcialmente con el concepto de SNOMED CT *Date ceased smoking*, concepto incluido en el contexto jerárquico previamente extraído. A continuación, únicamente para aquellos elementos sin anotar, se intenta una equi-

paración completa dentro de todo SNOMED CT. Como se puede ver en el paso 2 en la figura 4.5, el término *Status* no se encuentra en el contexto jerárquico y por ello se busca en toda la base de datos de SNOMED CT.

Con el fin de encontrar una anotación para los nodos *VALUE*, se extrae el contexto lógico SNOMED CT para cada candidato obtenido previamente, y se lleva a cabo una equiparación parcial exclusivamente dentro de ese extracto lógico. Para aquellos nodos que todavía no hayan sido anotados con ningún concepto SNOMED CT durante la equiparación parcial, se realiza una búsqueda por todo SNOMED CT aplicando otras técnicas léxicas (correspondencia total, alineamiento basado en UMLS y MetaMap). En la figura 4.5 puede verse que los términos *current non-smoker* y *never smoked* han sido equiparados dentro del contexto lógico.

Por otra parte, la segunda estrategia sigue una aproximación de abajo hacia arriba (bottom-up), donde se recorre el árbol de dependencias desde los nodos *VALUE* hacia los nodos *ELEMENT*. Para los nodos *VALUE* que se han anotado con algún concepto SNOMED CT utilizando las técnicas léxicas de correspondencia total, se extrae el contexto lógico de SNOMED CT; y se generan nuevos candidatos SNOMED CT para los nodos *ELEMENT*. Por ejemplo, durante la aplicación de la primera estrategia, el término *VALUE Never smoked* había sido anotado con el concepto SNOMED CT *never smoked*. En esta fase del método, el contexto lógico de este concepto se extrae y se genera un nuevo candidato para el término *ELEMENT Status* (véase la etapa 4 en la 4.5). Con estos nuevos candidatos, se obtienen nuevos extractos SNOMED CT y se aplican de nuevo las técnicas léxicas parciales y basadas en UMLS para anotar los nodos *VALUE* (en el paso 5 en de la figura 4.5 se descubre la anotación para el término *Non-smoker history unknown*).

4.3.3. Validación y desambiguación

Con el fin de eliminar la ambigüedad y validar las anotaciones léxicas resultantes, se consideran que tienen preferencia aquellas anotaciones que incluyen conceptos SNOMED CT que son semánticamente consistentes con los términos anidados del arquetipo sobre el resto de las anotaciones obtenidas por otros métodos. Se aplican dos técnicas de forma secuencial en esta etapa. En primer lugar, el candidato SNOMED CT que posea un mayor número de descendientes compartidos con el árbol de dependencias de términos ADL, se supone que es semánticamente más similar y por lo tanto, será el candidato elegido. Por ejemplo, para el término *ROOT Tobacco use* se obtienen varios candidatos SNOMED CT: *tobacco use and exposure*, *details of moist tobacco use* y *frequency of moist tobacco use* (ver figura 4.5). En

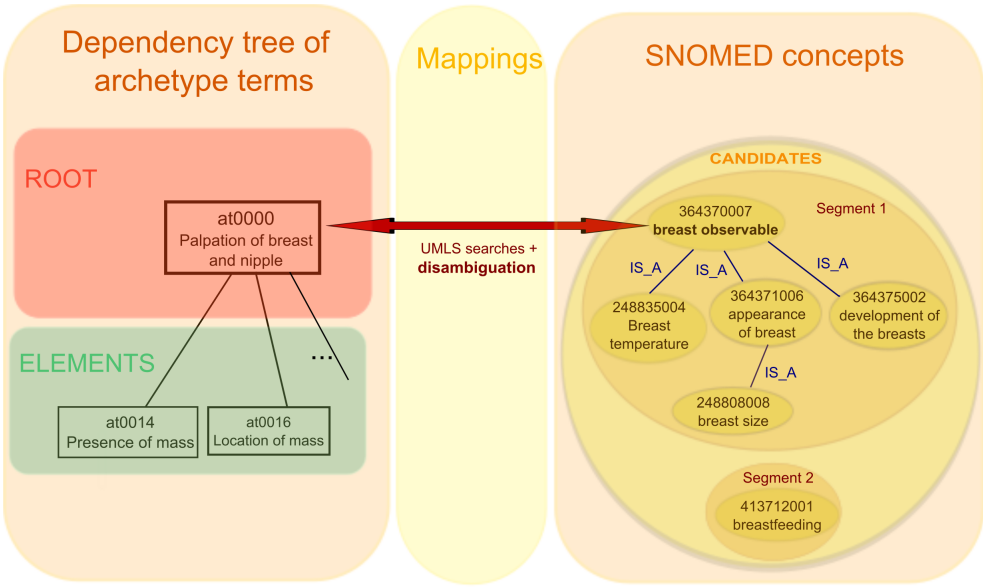


Figura 4.6: Ejemplo de desambiguación.

este caso, se selecciona el concepto *tobacco use and exposure*, ya que es el candidato con el mayor número de descendientes.

En segundo lugar, en ausencia de descendientes comunes, los segmentos de SNOMED CT extraídos para cada candidato se revisan con el objetivo de encontrar las relaciones jerárquicas entre los candidatos. En este caso, se elige como candidato SNOMED CT válido aquel concepto raíz del segmento que contiene el mayor número de candidatos. Por ejemplo, en el arquetipo openEHR-EHR *OBSERVATION.palpation_breast*[4], este método obtiene un conjunto de 6 conceptos candidatos para el término *Palpation of breast and nipple* (ver figura 4.6). Estos candidatos se agrupan en dos segmentos jerárquicos con cinco y un conceptos, respectivamente. Por último, se selecciona el concepto más general del segmento mayor: *breast observable*.

[4] `http://www.usc.es/keam/TermArchetypes/Input_ADLs_Bindings/openEHR-EHR-OBSERVATION.palpation_breast.v1.adl`

4.4. Procedimiento de evaluación del método propuesto

Para llevar a cabo la evaluación del método propuesto, usamos un conjunto de datos formado por 25 arquetipos del proyecto NHS Connecting for Health [112]. Este conjunto de datos ya no está disponible en el enlace indicado, pero se puede acceder a él desde una página web que hemos habilitado para proporcionar todo el material suplementario a este trabajo[114]. Todos los arquetipos de este conjunto son de tipo *observation* y contienen un nodo raíz que está asociado con una entidad SNOMED CT de tipo observable. De un total de 921 nodos en todos los arquetipos, sólo 37 (4 %) de ellos fueron enlazados a conceptos SNOMED CT por los diseñadores del arquetipo, lo que supone un número insuficiente de asociaciones para probar los métodos desarrollados. Por lo tanto, como tampoco se disponía ni de un *gold standard* ni de recursos suficientes para crearlos, hemos diseñado manualmente las anotaciones.

El procedimiento de evaluación utiliza estas anotaciones manuales para revisar de forma automatizada cada anotación generada por el método. En este contexto, la *precisión* se define como la fracción entre el número de nodos con alguna anotación correcta y el número total de nodos con anotaciones propuestas por el método; y el *recall* como la fracción entre el número de nodos con alguna anotación correcta y el número total de nodos con anotaciones relevantes (es decir, las anotaciones creadas manualmente).

4.5. Resultados

En esta sección detallaremos los resultados de la anotación manual, así como los referentes a la validación del método frente a la anotación manual de los 25 arquetipos. Una vez presentados los resultados, analizaremos estos en un apartado de discusión, y detallaremos las limitaciones encontradas en nuestro enfoque.

4.5.1. Anotación manual

Durante la anotación manual de los 25 arquetipos se crearon, en total, anotaciones para 477 nodos. Este número contrasta con el reducido conjunto de anotaciones iniciales incluidas en los arquetipos seleccionados (en total, 37). Todos los nodos anotados manualmente pertenecen a la sección *DATA* y *EVENTS*[5] de los 25 arquetipos. Estos son los nodos que se tienen

[5]Estas secciones de los arquetipos de tipo observación contiene el núcleo de la información del arquetipo; por ejemplo, la presión sistólica y diastólica en la medición de la presión arterial.

# de nodos totales	921
# de nodos con anotaciones	477
# de ROOT con anotaciones	25
# de ELEMENTs con anotaciones	188
# de VALUE con anotaciones	264
# de nodos no útiles	444
# nodos sin contenido clínico	170
# nodos comentario	57
# nodos no cubiertos por SNOMED CT	217
# nodos con anotaciones durante el diseño del arquetipo	37

Tabla 4.2: Características principales de los 25 arquetipos seleccionados.

en cuenta para evaluar el método automático: 25 *ROOTS*, 188 *ELEMENTS*, y 264 *VALUES* (ver tabla 4.2) [114]. Además, se han filtrado aquellos nodos que no contenían información clínica (en total, 170 nodos), los que eran comentarios generales del médico (57 nodos), o que no tenían ningún concepto SNOMED CT que representara su semántica (217 en total). Como resultado, 444 nodos de los arquetipos han sido considerados no útiles para anotar. Como no tenemos los recursos necesarios para realizar una anotación minuciosa, en casos de anotaciones ambiguas, los resultados alternativos se consideran también válidos.

4.5.2. Anotación automática

La tabla 4.3 muestra el resultado de aplicar el método desarrollado al conjunto de los veinticinco arquetipos, en términos de *precisión* y *recall*. Para cada tipo de término del arquetipo, la tabla detalla el número de anotaciones resultantes obtenidas por cada técnica utilizada, así como el resultado para la combinación de todas las técnicas. De los 477 nodos, la técnica léxica basada en la correspondencia total de los tokens anota todos los 25 nodos raíz (el 100 % de los términos *ROOT*), 115 nodos de tipo elemento (un 61 % de los términos *ELEMENT*), y el 196 nodos de tipo VALUE (74,2 %). La técnica basada en el contexto jerárquico combinada con la correspondencia parcial de tokens anota 18,6 % de nodos (35 de 188 nodos). Y, por último, la aplicación del contexto lógico, en combinación con la correspondencia parcial y el uso de técnicas de UMLS, consiguen anotar un 9,0 % de los nodos *ELEMENT* (17 de 188 nodos de elemento) y un 21,6 % de los nodos *VALUE* (57 de 264 valores). Estos resultados se

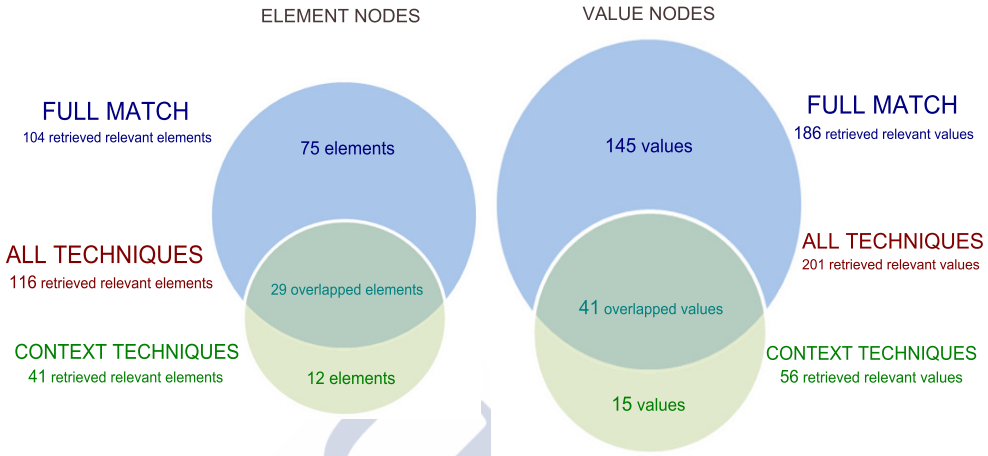


Figura 4.7: Número de nodos *ELEMENT* (lado izquierdo) y nodos *VALUE* (lado derecho) que han sido anotados usando técnicas de correspondencia total y basadas en contexto, así como la correspondencia entre ellos.

complementan con una *cobertura* de: 1,25 conceptos SNOMED CT por nodo *ELEMENT* y 1,21 por nodo *VALUE*.

Teniendo en cuenta todos los nodos, el enfoque sugerido logra una *precisión* de 96,1 % y un *recall* de 71,7 %. Conviene resaltar que el *recall* de las técnicas basadas en el contexto es menor que el de las técnicas léxicas, mientras que la *precisión* es mayor. Aún así, estos resultados son coherentes con la hipótesis inicial: hay algunas situaciones en las que los arquetipos agrupan información clínica lógicamente. Por lo tanto, en estos casos, la aplicación de métodos basados en similitud semántica en combinación con otras técnicas puede aumentar el *recall*. Por ejemplo, en la figura 4.5, para el nodo *ELEMENT status*, si aplicáramos una correspondencia total, se anotaría con el concepto SNOMED CT *status*. Sin embargo, la técnica basada en contexto recupera una anotación más precisa con un concepto que difiere en el nombre: *tobacco use and exposure*. La figura 4.7 muestra el número de nodos *ELEMENT* y *VALUE* que han sido asignadas a SNOMED CT, usando técnicas de correspondencia total y de contexto tanto por separado como en combinación.

Obsérvese que hay un solapamiento importante entre ambas técnicas. La correspondencia total recupera la mayoría de los nodos relevantes, mientras que las técnicas basadas en contexto son capaces de anotar conceptos con los nombres de los términos a pesar de no poseer ninguna semejanza léxica, que son imposibles de descubrir utilizando sólo técnicas de corres-

Tipo de término del arquetipo	# nodos	Técnicas	# nodos relevantes recuperados	# nodos recuperados	Precision %	Recall %
Nodos ROOT	25	Correspondencia total	12	12	100	48
		Búsquedas UMLS	13	13	100	52
Nodos ELEMENT	188	Correspondencia parcial en contexto jerárquico	32	35	91,4	17,0
		Correspondencia total	104	115	90,4	55,3
		Contexto lógico	15	17	88,2	8,0
		Todas las técnicas	116	125	92,8	61,7
Nodos VALUE	264	Correspondencia parcial y correspondencia UMLS en el contexto lógico	56	57	98,2	21,2
		Correspondencia total	186	196	94,9	70,5
		Todas las técnicas	201	206	97,6	76,1
Todos los nodos	477	Todas las técnicas	342	356	96,1	71,7

Tabla 4.3: Precisión y recall de las técnicas aplicadas.

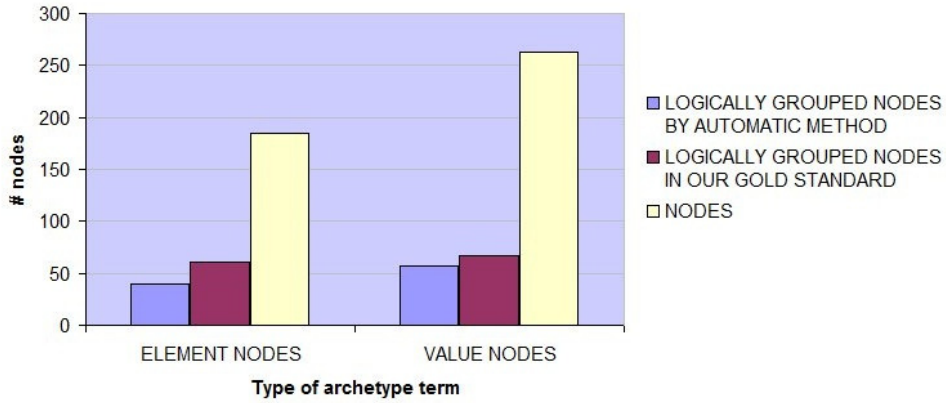


Figura 4.8: Nodos de los arquetipos agrupados lógicamente en SNOMED CT.

pondencia total. Pero el principal beneficio de la aplicación de técnicas basadas en contexto es la desambiguación y la validación de los resultados producidos por las técnicas léxicas. En total, 7 nodos raíz (28 %), 41 nodos de elemento (35 %) y 57 nodos de valor (una ratio del 28 %) son semánticamente desambiguadas y validados por el método.

Sin embargo, no todos los nodos arquetipo están lógicamente agrupados en SNOMED CT. En concreto, de un total de 185 nodos *ELEMENT*, solamente 61 están relacionados lógicamente en SNOMED CT; y de un total de 263 nodos *VALUE*, 67 nodos lo están. Por lo tanto, este método es capaz de validar automáticamente 41 de 61 nodos *ELEMENT* (67 %) y 57 de 67 nodos *VALUE* (85 %) (figura 4.8).

4.5.3. Discusión

A pesar de que tanto los arquetipos clínicos como SNOMED CT están enfocados a la estructuración de la información del paciente reflejando así las necesidades de la práctica clínica, sus puntos de vista pueden ser diferentes. Los arquetipos agrupan la información clínica para que sea registrada al mismo tiempo que tiene lugar el hecho clínico concreto. Por ejemplo, a fin de evaluar el bienestar de un neonato justo después de su nacimiento, el arquetipo de *Apgar score* (figura 2.1) registra las valoraciones acerca de la frecuencia cardiaca, el esfuerzo respiratorio, el color, el tono muscular y la sensibilidad de los reflejos en los minutos 1, 2, 5 y

10 después del nacimiento. Sin embargo, la estructura interna de SNOMED CT relaciona los conceptos entre sí lógicamente dentro de un dominio específico. Por ejemplo, los conceptos como *Apgar score at 1 minute* o *Apgar score at 5 minutes* se relacionan entre sí a través de una relación jerárquica con el concepto padre *Component of Apgar Score*. Sin embargo, conceptos como la frecuencia cardíaca y el esfuerzo respiratorio no están relacionados entre sí, ya que la primera es una función cardíaca mientras que la segunda es una característica del movimiento respiratorio. Por lo tanto, existen situaciones en las que la información clínica del arquetipo se agrupa lógicamente y otras en que no. Esto implica que, la aplicación de métodos basados en la similitud semántica en combinación con otras técnicas es beneficiosa para aumentar el *recall* de esas otras técnicas sólo en esas situaciones, así como para comprobar la validez de sus resultados.

A diferencia de otros trabajos en este campo, el método sugerido de correspondencia parcial encuentra anotaciones para los nombres del arquetipo sólo cuando hay un contexto jerárquico o lógico. Hay que remarcar que la aplicación de una correspondencia parcial con SNOMED CT completo podría proporcionar resultados ambiguos. Por ejemplo, el empleo de la correspondencia parcial del término *irregular* sobre todo SNOMED CT recuperará más de 72 conceptos. Por lo tanto, este método no sería ideal y se necesitarían técnicas adicionales para filtrar y eliminar la ambigüedad de los 72 resultados SNOMED CT. Sin embargo, esta técnica ofrece buenos resultados si se aplica en un contexto limitado. Por ejemplo, la correspondencia parcial del término *irregular* dentro del segmento de SNOMED CT que incluye sólo aquellos conceptos relacionados con el nodo padre (*Rhythm of respiration*) a través de la relación *interprets*, devuelve el mapeo válido (por ejemplo, *irregular breathing*).

Por otra parte, este estudio se basa en la suposición de que la mayoría de los elementos de arquetipos *OBSERVATION* se corresponden con conceptos *observable* en SNOMED CT. Lo mismo se espera para los nodos *VALUE* y los conceptos *finding*. Estos supuestos están fundamentados en la información que proporciona OpenEHR acerca de los arquetipos y terminologías[6]. Después de llevar a cabo los experimentos, se revisaron los términos raíz de la jerarquía SNOMED CT a las que pertenecían los conceptos que formaban parte del *gold standard* (figuras 4.9 y 4.10), y se confirmaron dichas suposiciones.

Una característica notable de los arquetipos analizados es que los nodos *ELEMENT* que toman valores booleanos se asignan con frecuencia a conceptos *finding*. Por ejemplo, en el arquetipo *Baby general observations*, el término *ELEMENT Blood in urine*, cuyos valores

[6] `http://www.openehr.org/wiki/display/healthmod/Archetypes+and+Terminology`

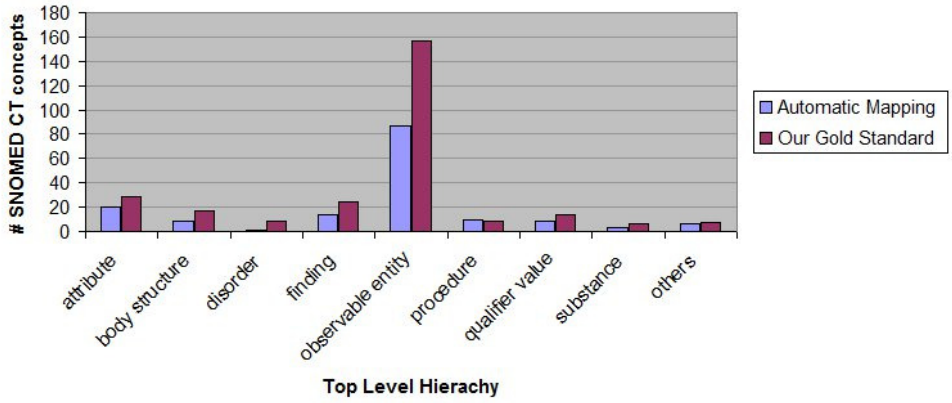


Figura 4.9: Conceptos raíz de la jerarquía de SNOMED CT a los que pertenecen los conceptos usados en el *gold standard* y a los que pertenecen los conceptos mapeados automáticamente a los términos *ELEMENT* del arquetipo.

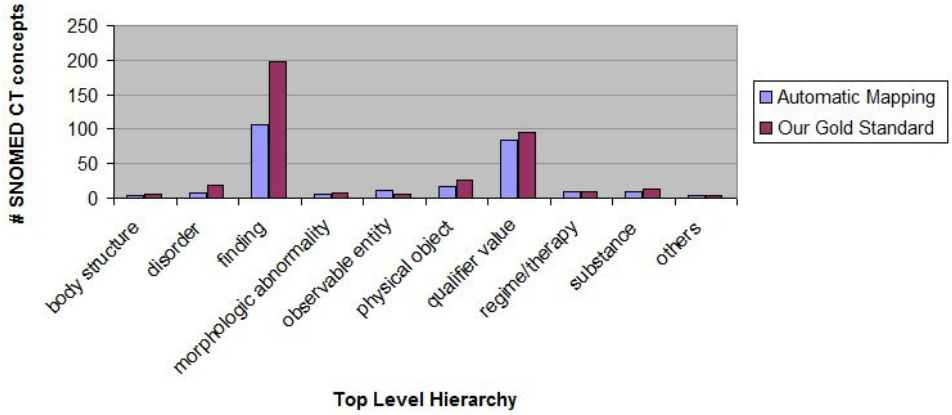


Figura 4.10: Conceptos raíz de la jerarquía de SNOMED CT a los que pertenecen los conceptos usados en el *gold standard* y a los que pertenecen los conceptos mapeados automáticamente a los términos *VALUE* del arquetipo.

asociados son *True* o *False*, se asigna al concepto SNOMED *Blood in urine (finding)*. Además, tanto en la figura 4.9 como en la 4.10, hay una cantidad importante de conceptos descendientes de los tipos semánticos generales *attribute* o *qualifier value*. Este tipo de conceptos se utilizan cuando el texto del término es demasiado general y el contexto del arquetipo no proporciona la suficiente información. Por ejemplo, en el arquetipo de *Apgar*, el término *ELEMENT Colour* que hace referencia al color de la piel, se anotó con el concepto *Colors (qualifier value)* porque no se ha podido encontrar un concepto más específico en SNOMED CT.

Limitaciones del Enfoque Propuesto

Los aspectos más críticos de la anotación de arquetipos con SNOMED CT se han discutido en enfoques anteriores. Se trata de cuestiones básicamente léxicas y terminológicas [128, 155]:

- deficiencias de normalización (por ejemplo, *waist:hip ratio* y *waist/hip ratio*),
- diferencias en la nomenclatura (por ejemplo, el término raíz del arquetipo *Cigarette* debería anotar la descripción SNOMED CT *Cigarette smoking tobacco*),
- la cobertura incompleta de sinonimia de SNOMED CT (por ejemplo, el término *Age commenced* del arquetipo *tobacco use* debería ser un sinónimo del concepto SNOMED CT *Age at starting smoking*),
- la ambigüedad en arquetipos y en SNOMED CT (por ejemplo, la diferencia entre los conceptos de SNOMED CT *feeding* y *feeding observable* no es precisa); y, por último,
- la necesidad de post-coordinación de conceptos para una anotación correcta del término del arquetipo (por ejemplo, el término *waist and hip circumference* necesita una post-coordinación de los conceptos pre-coordinados *waist circumference* y *hip circumference* conceptos obtenidos durante el proceso de anotación).

Estos problemas disminuyen notablemente los resultados de *recall*.

También se han detectado otros problemas en este estudio que provienen de deficiencias en los modelados de los arquetipos y de SNOMED CT [128, 155]. Por un lado, la falta de consenso en la estructuración y organización de los datos durante el diseño del arquetipo lleva a superposiciones e inconsistencias en su representación de información. Por otro lado, el

hecho de modelar arquetipos clínicos de forma separada de SNOMED CT conduce a diferencias entre ellos, tales como las ya comentadas, diferencias léxicas entre los nombres de los términos, la discrepancia intencionada en la semántica o la divergencia en el nivel de detalle de documentación.

Además, la cobertura SNOMED CT es incompleta, no sólo para los sinónimos sino también para las relaciones lógicas, lo que disminuye el *recall* del método propuesto. Por ejemplo, el rendimiento del método fue optimizado gracias a tener en cuenta la relación *interprets* entre un nodo *ELEMENT* (por ejemplo, el concepto *Heart rate*) y un nodo *VALUE* (*Fetal heart rate absent*). En las situaciones en las que la cobertura de las relaciones de SNOMED CT es incompleta, el método falla al intentar encontrar y validar las anotaciones léxicas.

El método propuesto no obtiene una *precision* ni *recall* óptimos en situaciones en las que:

– No es posible aplicar la información del contexto de forma adecuada. Por ejemplo, en el arquetipo *tobacco use*, el término *administration* se refiere a *route of administration*. Sin embargo, el arquetipo no incluye el término *route* requerido para construir la información de contexto apropiada.

– Cuando el contexto lo proporcionan las palabras clave (*key words*) de la sección *header* del arquetipo. Por ejemplo, el término *consumption* bajo el contexto de la palabra clave *smoking* y la palabra del término *ROOT tobacco* nos permitiría asignarla a una descripción del concepto *Tobacco smoking consumption (observable entity)* y validarlo dentro de la relación jerárquica.

– La información contextual principal es aportada por los términos que cuelgan del término *ROOT* del arquetipo, no por el término raíz en sí mismo. Por ejemplo, el arquetipo *postnatal assessment of mother* incluye la información modelada en subestructuras sobre el perineo, la micción o el pecho y los pezones, es decir, cada subestructura posee información de contexto independiente de las otras. Tal es el caso de arquetipos como *feeding*, *postnatal assessment of mother* o *tobacco use* para los que el *recall* de los nodos *ELEMENT* es muy bajo (38 %).

4.5.4. Estudio Comparativo con Otros Enfoques

Finalmente, los resultados este estudio se han comparado con los obtenidos con los métodos comentados previamente 4.2. Como puede verse en la tabla 4.4, cada enfoque utiliza

para sus experimentos diferentes tipos de arquetipos, distintos procedimientos de evaluación y diversas mediciones. Para minimizar este problema, sólo se han considerado dos medidas de evaluación: el *recall* (véase la definición en la sección 4.4) y la *cobertura* (es decir, el promedio de anotaciones por nodo, lo que nos da una idea acerca de la ambigüedad). Para algunos de estos enfoques, estas medidas han sido obtenidas a partir de otros valores extraídos de la documentación de referencia.

En los experimentos realizados por Yu et al. [175], los autores seleccionaron los arquetipos poseedores de un mayor número asociaciones con SNOMED creadas durante el proceso de desarrollo del arquetipo. En total, se utilizaron 7 arquetipos. Este enfoque tiene el *recall* más bajo (55 %) y la mayor *cobertura* (10 conceptos por nodo). La ambigüedad de los resultados podría reducirse aplicando técnicas de normalización. Por el contrario, la evaluación llevada a cabo por Lezcano et al. [85] emplea un mayor número de arquetipos; aunque los autores no especifican el procedimiento de evaluación. Este método logra 60 % de *recall* y una *cobertura* aceptable (2.4 conceptos por nodo). En el enfoque Qamar y Rector [128], la validez de los conceptos candidatos fue evaluada por expertos clínicos. Ellos sólo utilizaron 4 arquetipos, y el *recall* fue superior a los enfoques anteriores, ya que incluyen una amplia variedad de técnicas lingüísticas. La cobertura fue de 5,5 conceptos por nodo. En el trabajo de Berges et al. [12] demuestra que la técnica compleja basada en subcadenas que obtiene mejores resultados es la de Q-gramas. Para ello, han calculado el *recall* fijando distintos valores de *cobertura*, consiguiendo para una *cobertura* de valor 1, un *recall* de 25,26 % y para una *cobertura* de 10, un *recall* de 50,51 %.

En este trabajo, hemos creado nuestro propio *gold standard* para el proceso de evaluación (ver sección 4.4). En los experimentos, se utilizaron 25 arquetipos de tipo observación, llegando al *recall* más alto (71,7 %) y a una *precisión* muy significativo de 96,1 %. Alrededor del 30 % de los conceptos SNOMED anotadas a términos de los arquetipo se agruparon lógicamente. En tales casos, la aplicación de técnicas basadas en contexto en combinación con técnicas léxicas contribuyó a mejorar los resultados de los enfoques existentes [175, 85, 128]. Además, las técnicas basadas en contexto fueron de utilidad para validar los enlaces obtenidos a través de técnicas léxicas, resolver la ambigüedad de las anotaciones, y descubrir aquellas no detectadas por técnicas léxicas. La *cobertura* del proceso de alineación se redujo a 1,23 conceptos por nodo. En total, ciento cuatro términos de arquetipos (30,4 %) fueron validados semánticamente por el método.

Estudio	# de arquetipos	Arquetipo	Características del arquetipo	Metodología para la evaluación	# de nodos procesados	Recall %	Cobertura
Yu et al.[175]	7	NHS	OBSERVATIONS, EVALUATIONS, INSTRUCTIONS, y ACTIONS	Comparación entre las anotaciones candidatas y las existentes	147	55	10
Lezcano et al.[85]	40	OpenEHR	No se especifica	No se especifica	655[a]	60	2,4[b]
Qamar y Rector [128]	4	OpenEHR	OBSERVATIONS	Manualmente	122	68,9[c]	5,5[d]
Berges et al. [12]	25[e]	NHS	OBSERVATIONS	Validación contra nuestro *gold standard*	487	25,26 50,51 75,56	1 10 100
Investigación actual	25	NHS	OBSERVATIONS	Validación automática de las anotaciones contra nuestro *gold standard*	477	71,6	1,23

Tabla 4.4: Comparativa de los trabajos y resultados.

[a] Estos valores se derivan de la sección de Resultados de [85]

[b] 2,4 conceptos UMLS por nodo

[c] El valor 68.9 se obtuvo aplicando la siguiente fórmula: Los 84 nodos relevantes recuperados de la medida estadística llamada criteria2 (ver página 129 y 137 de [128]) dividida entre los 122 fragmentos relevantes del arquetipo (ver tabla 5.5, página 134).

[d] El valor 5.5 no aparece explícitamente en [128]. Este valor se obtuvo a través de la siguiente fórmula: Los 648 códigos de SNOMED CT remanentes después del filtrado divididos por los 118 fragmentos para los que se encontró al menos un vínculo con SNOMED CT (ver tabla 5.5, página 134 de [128])

[e] Utiliza el repositorio de arquetipos creado para nuestro trabajo [114]

En resumen, debido al alto número de nodos anotados, a la baja *cobertura*, a la alta *precisión* y a un cierto nivel de validación automática, se puede concluir que el rendimiento de las técnicas de anotación es satisfactorio y que el método reduce considerablemente el esfuerzo humano necesario para revisar y confirmar las anotaciones automáticas.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Capítulo 5

Una Propuesta para la Anotación Semántica de Guías de Práctica Clínica

Las guías de práctica clínica (GPC) constituyen una fuente sustancial e importante de conocimiento sobre las recomendaciones diagnósticas y terapéuticas basadas en la evidencia. El acceso a este conocimiento en el punto de atención al paciente mejora la calidad de la asistencia sanitaria y reduce costes innecesarios [46, 139]. Para poder implementar una GPC como una aplicación electrónica es esencial transformar el texto del documento de la GPC a un formato electrónico. Actualmente existen propuestas de lenguajes formales desarrollados específicamente para representar GPC [23, 62, 152], pero requieren que los ingenieros del conocimiento, con la ayuda de expertos clínicos traduzcan el texto médico a dicho lenguaje. Para solventar el problema, se han desarrollado herramientas informáticas, como AsbruView [80], Arezzo [55], Tallis [157] o TAT [153], a parte de las comentadas en la sección 3.5.2. Sin embargo, la adquisición manual de conocimiento de GPC sigue siendo compleja y laboriosa, a pesar de la ayuda proporcionada por algunas de estas herramientas. Por lo tanto, es evidente que hay una carencia de técnicas que automaticen, al menos en parte, la extracción del conocimiento contenido en los textos de las GPC.

Es frecuente el uso de técnicas de Procesamiento de Lenguaje Natural (PLN) para analizar de forma automática los textos referentes a registros e informes de pacientes [31]. Sin embargo, la mayoría de los métodos utilizados han sido desarrollados para sistemas específicos, por

lo que evaluar si tales métodos se pueden reutilizar fácilmente en nuevas aplicaciones es un trabajo muy relevante en el momento de realización de esta tesis, para valorar la transferencia real de dichos métodos al sector sanitario. En este capítulo vamos a exponer cómo hemos planteado la reutilización de varias herramientas de código abierto que funcionan como bloques constituyentes dentro de un nuevo sistema PLN. El objetivo es evaluar la aplicación de la tecnología de PLN actual a un nuevo dominio clínico: la adquisición automática de conocimiento sobre procedimientos diagnósticos y terapéuticos pertenecientes a la práctica clínica. Dichos procedimientos, como ya hemos comentado, se encuentran contenidos en documentos expresados en lenguaje natural.

5.1. Introducción

El PLN es una línea de investigación activa en el cuidado clínico [41, 40]. Además, juega un papel decisivo en la asistencia y mejora del proceso de atención médica, especialmente en relación con la información clínica almacenada en los registros de salud electrónicos, notas del médico o la literatura biomédica y expresada en lenguaje natural [18, 31]. La atención médica implica no solo la información del paciente, sino también el conocimiento clínico sobre las sugerencias del médico basadas en la evidencia sobre los procedimientos diagnósticos y terapéuticos. Es por ello que la tecnología de PLN puede ser crucial para aportar evidencias de las GPC en el punto de atención al paciente. Sin embargo, la ampliación del alcance de la tecnología actual de PLN biomédico no es trivial, ya que muchos de los sistemas desarrollados hasta el momento están centrados en una aplicación concreta [28]. Por lo tanto, se necesita una nueva investigación para determinar si tales métodos se pueden reorientar hacia nuevas aplicaciones y metas, obteniendo el mismo rendimiento.

El objetivo de este capítulo es evaluar la aplicación de la tecnología de PLN actual a un nuevo dominio: la adquisición automática de conocimiento de los procedimientos diagnósticos y terapéuticos de las GPC [161]. La idea principal es enriquecer estos documentos con una ontología, para hacerlos interpretables computacionalmente. Sin embargo, no está garantizada la fiabilidad de la tecnología de PLN y de las técnicas de ingeniería del conocimiento (IC) actuales. En primer lugar, las herramientas de pre-procesamiento de texto pueden generar errores cuando se utilizan en nuevos dominios y estos se pueden propagar hacia arriba creando más errores en los siguientes niveles [28]. En segundo lugar, el reconocimiento de entidades nominales (*Named Entity Recognition*, NER) implica encontrar entidades médicas en el texto

e interpretar su significado correctamente. Las terminologías son las principales fuentes de conocimiento que ayudan en el proceso de NER en el dominio clínico, debido a su extenso tamaño y a su fácil acceso. Aun así, el proceso de NER tiene que lidiar con el problema de la ambigüedad [39, 29, 48]. En tercer lugar, en contraste con los resultados favorables en NER, el avance en la investigación de extracción de relaciones ha sido menos productiva. Muchos trabajos realizados hasta la fecha incluyen esta tarea como parte de un sistema de extracción de información clínica completa [135, 29], y sólo existen unas pocas herramientas de código abierto disponibles a los usuarios, como SemRep [134, 51].

Para superar estos inconvenientes, en esta tesis se adaptaron varias herramientas para analizar documentos de GPC con el objetivo de identificar, en primer lugar, las entidades relacionadas con el diagnóstico y la terapia y, a continuación, las relaciones significativas entre estas entidades. El enfoque propuesto en esta tesis doctoral consiste en elaborar y aplicar una combinación secuencial de varios métodos utilizados tradicionalmente en PLN, para anotar gradualmente frases de la GPC con conceptos de una ontología. Como veremos más en detalle, hemos empleado 171 frases que describen procedimientos diagnósticos y terapéuticos. Estas oraciones fueron seleccionadas por un grupo de expertos con gran experiencia tras su participación en el proyecto de investigación HYGIA[1], proyecto que promovió la adquisición, la formalización y la adaptación del conocimientos de las GPC con el fin de describir las vías de atención al paciente.

5.2. Desarrollo de Herramientas de PLN

Las herramientas de PLN se pueden desarrollar utilizando técnicas de la Ingeniería de Conocimiento (IC), aprendizaje máquina (AM) o métodos híbridos [81]. Los sistemas de AM se han generalizado ya que los resultados se alcanzan en menor tiempo [70, 137]. Sin embargo, el aprendizaje automático es particularmente difícil si no hay disponibles los suficientes datos clínicos para el entrenamiento y, lo que sucede con frecuencia, los investigadores no poseen ningún tipo de recurso disponible. Hoy en día, los enfoques de la IC no tienen tanto éxito debido al esfuerzo manual que requieren y a la dificultad para ser reutilizados en otras aplicaciones. A pesar de ello, estos enfoques ofrecen buenos resultados principalmente en el reconocimiento de estructuras complejas. Este es el caso de muchos trabajos de investigación basados en el reconocimiento de entidades biomédicas y en la extracción de relaciones de

[1] http://banzai-deim.urv.net/~riano/TIN2006-15453/

informes de pacientes. Muchos de ellos aplican enfoques de la IC, utilizando analizadores sintácticos con reglas gramaticales específicas del dominio, lexicones computacionales y ontologías. El sistema RECIT [129] combina información sintáctica y semántica para extraer grafos conceptuales que expresan el significado de los componentes de las oraciones redactadas en lenguaje natural. MedLEE [39] es un sistema basado en reglas utilizado con éxito para procesar los informes de diferentes dominios (por ejemplo radiología, patología, electrocardiograma). SeReMeD [29] es un método utilizado para generar representaciones de conocimiento contenido en los informes de rayos X de tórax, apoyándose en UMLS. Hasta la fecha, también podemos encontrar y trabajos muy relevantes sobre la extracción de información a partir de textos de GPC. Kaiser et al. [73] han modelado procesos de tratamiento mediante la identificación automática de las partes relevantes de la guía a través de métodos de extracción de información. Servan et al. [146] simplifican la adquisición de conocimiento mediante la extracción automática del conocimiento de control (como la disgregación o secuenciación de acciones clínicas) utilizando patrones lingüísticos y una ontología que fue construida específicamente para este sistema. MapFace [48], un editor interactivo, está orientado a simplificar el enriquecimiento de documentos GPC con conceptos UMLS. cTAKES (clinical Text Analysis and Knowledge Extraction System) es un proyecto de código abierto para extracción de información de narrativas clínicas, que realiza análisis sintáctico, semántico y anotación de conceptos UMLS [142]. Por otra parte, YTEX se ha construido sobre cTAKES para anotar frases clínicas con conceptos de varias terminologías, incluyendo UMLS y sus terminologías integradas [43]. En los últimos años, diferentes estudios se han centrado en la extracción de información temporal en aplicaciones médicas, incluyendo temas desde la identificación de eventos, expresiones y relaciones temporales [164] hasta la normalización de expresiones temporales incompletas [159]. Un estudio reciente sobre el estado del arte en el reconocimiento de entidades sobre enfermedades concluyó que actualmente la mayoría de los sistemas que se desarrollan son aproximaciones híbridas que incluyen características generadas por reglas creadas a partir de datos de entrenamiento y de recursos externos, como el UMLS.

Hay cuatro pasos principales en PLN [28]: el pre-procesado del texto, NER, la extracción de contexto y extracción de relaciones. Actualmente, podemos encontrar herramientas disponibles para el pre-procesado de texto que nos proporcionan detectores de oraciones, tokenizadores, etiquetadores de categorías gramaticales, chunkers Treebank o analizadores Treebank. Muchas de ellas ya han sido comentadas en el capítulo 3, por lo que aquí sólo resumiremos las características más relevantes en relación con este capítulo. Ejemplos de estas herramientas

a destacar aquí son OpenNLP[2], que también forma parte del juego de herramientas integradas de software dedicado a PLN como GATE [26, 27], o Stanford[3], que se ha integrado en otros navegadores e interfaces de PLN. Con respecto a NER en ámbitos médicos, MetaMap [8, 103, 69] y el anotador del Bioportal [172, 106] son algunas de las herramientas más utilizadas hoy en día, a pesar de que también se aplican otros recursos de UMLS [52, 162]. UMLS [86], que proporciona la terminología para MetaMap, se compone de varias fuentes de conocimiento que proporcionan información terminológica como ya se ha comentado en el capítulo 2. La fuente de conocimiento más grande es Metathesaurus, que contiene información acerca de conceptos médicos, sinónimos y las relaciones entre ellos. Los tipos semánticos (TS) son un conjunto de categorías semánticas básicas utilizadas para clasificar los conceptos de Metathesaurus. Ejemplos de TS son *diagnostic procedure* o *therapeutic preventive procedure*. Los TS están relacionados entre sí a través de relaciones jerárquicas (*is-a*) y no jerárquicas (*diagnoses*, *treats*, ...) componiendo así la llamada red semántica (UMLS Semantic Network), tal y como se explica en la sección 2.3.2. Por otra parte, comprender el contexto del que se extrae una entidad es esencial para transformar texto en un formato electrónico. El reconocimiento del contexto puede implicar la identificación de una condición clínica, negaciones o antecedentes relacionados con la salud del paciente. Uno de los algoritmos más populares propuestos en la literatura para el reconocimiento contexto es NegEx [17, 19, 98], un algoritmo utilizado para identificar la negación en oraciones. NegEx se ha integrado en las últimas versiones de MetaMap. Por último, también está disponible para la investigación, SemRep [134, 16, 51], una herramienta orientada a extraer conocimiento en forma de predicados semánticos.

5.3. Extracción Automática de Procedimientos Diagnósticos y Terapéuticos

Con el fin de evaluar las tecnologías actuales de PLN aplicadas a la adquisición de conocimiento automatizado, se utilizaron una GPC pública sobre la insuficiencia cardiaca crónica (ICC)(*Chronic Heart Failure (CHF)*), varias herramientas de código abierto, y se aplicaron los criterios habituales de evaluación de rendimiento en investigación del PLN.

[2] `http://opennlp.apache.org/`
[3] `http://nlp.stanford.edu:8080/parser/`

5.3.1. El Texto

Para este caso de estudio, se utilizaron 171 frases de un documento redactado en lenguaje natural: la guía para el diagnóstico y tratamiento de la ICC publicada por la Sociedad Europea de Cardiología [105]. Un grupo de expertos seleccionó las oraciones que se procesarían en este trabajo con el objetivo de verificar la eficacia del nuevo enfoque. Inicialmente, los expertos subrayaron los contenidos clínicamente relevantes de la GPC y luego, las frases marcadas fueron anotadas manualmente con el contexto del diagnóstico que se refieren a: *CHF*, *CHF with Preserved Left Ventricular Ejection Fraction (PLVEF)* o *CHF in elderly patients*.

5.3.2. Las Herramientas de PLN de Código Abierto

Las herramientas de PLN de código abierto usadas en este estudio fueron las siguientes: los analizadores OpenNLP y Stanford, SemRep y el servicio UMLS para normalización de cadenas de caracteres. Todas estas herramientas han sido seleccionadas por los siguientes criterios: disponibilidad, alta calidad probada por otros investigadores, menciones en la literatura relacionada y la experiencia de uso de nuestro grupo de investigación.

5.3.3. Criterios de Evaluación de Resultados

La principal dificultad en la evaluación de la aplicación de técnicas de PLN sobre documentos clínicos es la ausencia de un *gold standard* sobre el dominio. Por lo tanto, hemos validado el enfoque automático con un *gold standard* realizado de forma manual, incluyendo frases anotadas por una persona anotadora capacitada clínicamente. Por otro lado, los indicadores de evaluación suelen interpretarse en términos de *precisión* y *recall*. La *precisión* es una medida de la exactitud de los elementos que se sugieren como entidades o relaciones, y se mide típicamente como la proporción de verdaderos positivos (elementos sugeridos correctamente) sobre todos los elementos sugeridos. *Recall* designa la proporción a la que se reconocen las entidades o relaciones y se mide generalmente como la relación de verdaderos positivos sobre todos los elementos que deben ser reconocidos. El rendimiento global se calcula generalmente a través de la *F-measure*, una media armónica entre *precisión* y *recall*.

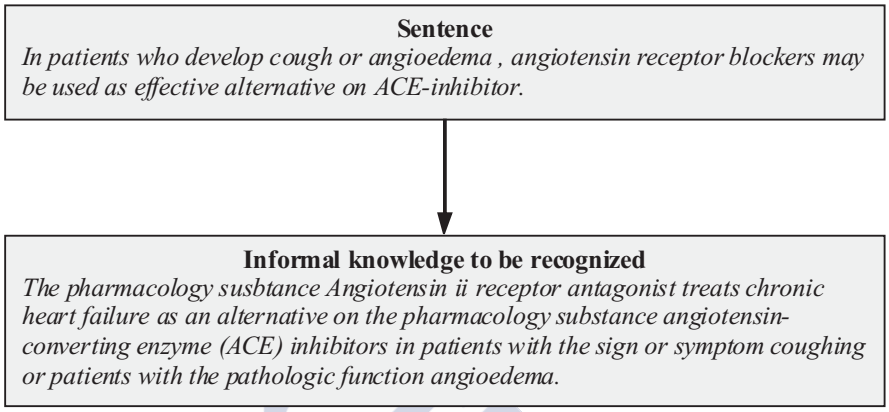


Figura 5.1: Oración de muestra y el conocimiento descriptivo a reconocer.

5.4. El Método Propuesto

Nuestro enfoque extrae entidades y relaciones de frases redactadas en lenguaje natural. Por **entidad**, nos referimos a un concepto clínico o evento: *disease*, *treatment*, *symptom and sign*, . . . , y por **relación**, a un predicado semántico sobre algún procedimiento diagnóstico o terapéutico. Estos predicados se componen de dos argumentos, es decir, dos entidades vinculadas en la frase. Además, para la generación de conocimiento computable se necesita comprender el contexto del que se extraen las entidades. Por ejemplo, el procesado de los procedimientos diagnósticos y terapéuticos requerirá no sólo reconocer el procedimiento que se va a aplicar al paciente, que viene dado por las entidades y sus relaciones, si no que también es necesaria la determinación de la condición clínica bajo la cual se establece el procedimiento. En la figura 5.1, podemos ver un ejemplo de oración extraída de la guía empleada en esta tesis doctoral, y el conocimiento expresado informalmente que debe ser reconocido. En esta situación particular, la frase implica un procedimiento terapéutico para la ICC, como una alternativa a otro procedimiento terapéutico, pero sólo en el contexto de los pacientes con un indicio particular o función patológica.

Con el fin de extraer las relaciones junto a su contexto de validez, nuestro enfoque produce, en primer lugar, un análisis sintáctico no excesivamente detallado; a continuación, un NER

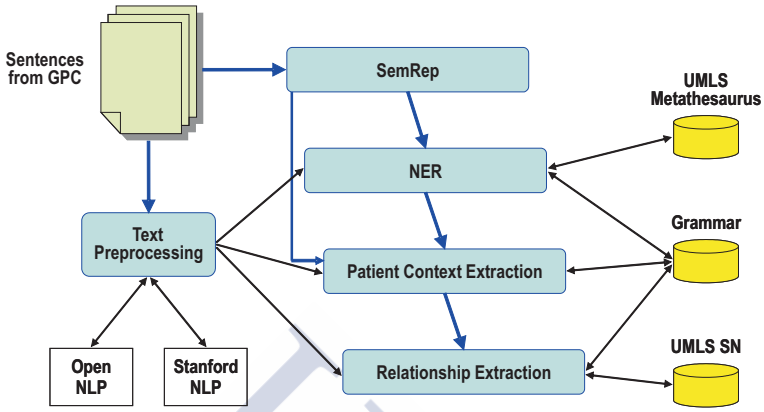


Figura 5.2: Bloques constituyentes de la propuesta de PLN en esta tesis doctoral para anotación de relaciones en una guía clínica.

basado en UMLS y, por último, una conversión de las estructuras sintácticas en relaciones semánticas. La Figura 5.2 muestra los principales componentes básicos de nuestro enfoque. En primer lugar, la herramienta SemRep[4] se utiliza para extraer información del conjunto de oraciones, en forma de predicados semánticos. SemRep produce dos tipos de resultados: un NER basado en UMLS utilizando la herramienta MetaMap ([8]) y un reconocimiento de predicados utilizando las relaciones de la red semántica de UMLS. En la figura 5.3, podemos ver los resultados de SemRep de la oración de la figura 5.1. A continuación, nuestro enfoque comprueba automáticamente los resultados y, en la ausencia de las entidades o predicados esperados, aplica los pasos utilizados clásicamente en PLN: pre-procesado de texto, NER, extracción de contexto y extracción de relaciones. La figura 5.4 muestra la aplicación de estos pasos básicos a la frase de muestra de la figura 5.1. A continuación, describimos en detalle nuestra propuesta, apoyándonos en el ejemplo propuesto.

5.4.1. Reconocimiento de Entidades y Extracción de Predicados Mediante SemRep

Utilizando SemRep, cada frase se asocia con un conjunto de conceptos UMLS y con un conjunto de predicados semánticos que relacionan los conceptos extraídos. Cada predi-

[4] http://skr.nlm.nih.gov/

SE|00000000||ti|1|text|**In patients who develop cough or angioedema, angiotensin receptor blockers may be used as an effective alternative on an ACE-inhibitor.**
SE|00000000||ti|1|entity|C0030705|**Patients**|podg|||patients||||1000|4|11
SE|00000000||ti|1|entity|C0010200|**Coughing**|sosy|||cough||||1000|25|29
SE|00000000||ti|1|entity|C0002994|**Angioedema**|patf|||angioedema||||1000|34| 43
SE|00000000||ti|1|entity|C0034787|**Angiotensin Receptor**|aapp,rcpt|||angiotensin receptor||||734|46|65
SE|00000000||ti|1|entity|C1280500|**Effect**|qlco|||effective||||888|94|102
SE|00000000||ti|1|entity|C1523987|**Alternative**|cnce|||alternative||||888|104|114
SE|00000000||ti|1|entity|C0003015|**Angiotensin-Converting Enzyme Inhibitors**|phsu|||ACE-inhibitor||||1000|122|134

SE|00000000||ti|1|relation|5|1|C0010200|Coughing|sosy|sosy|||cough||||1000|25|29|VERB|**PROCESS_OF**||17|23|1|1|C0030705|Patients|humn|humn|||patients||||1000|4|11
SE|00000000||ti|1|relation|5|1|C0002994|Angioneurotic Edema|patf|patf|||angioedema||||1000|34|43|VERB|**PROCESS_OF**||17|23|1|1|C0030705|Patients|humn|humn|||patients||||1000|4|11
SE|00000000||ti|1|relation|2|2|C0003015|Angiotensin-Converting Enzyme Inhibitors|phsu|phsu|||ACE-inhibitor||||1000|122|134|VERB|**ADMINISTERED_TO**||83|86|4|4|C0030705|Patients|humn|humn|||patients||||1000|4|11

Figura 5.3: Resultados de SemRep para la oración de la figura 5.1.

cado semántico se corresponde con algún tipo de relación perteneciente a la red semántica de UMLS. Ejemplos de tales son *diagnoses*, *treats*, *causes* o *location of*. Como solo estamos interesados en la extracción de los procedimientos diagnósticos y terapéuticos, nuestro enfoque se limita a algunos tipos semánticos de UMLS (*diagnostic*, *laboratory*, *therapeutic or preventive procedures*, *sign or symptoms*, *findings*, *diseases or syndromes and pathologic functions*, *pharmacologic substances*), dos tipos de predicados de la red semántica (*treats* y *diagnoses*) y un tipo de predicado no incluido en la red semántica (*has adverse effects*). Un ejemplo de la respuesta generada por SemRep para la oración de la figura 5.1 se puede ver en la figura 5.3. Siete entidades UMLS y tres relaciones son identificadas automáticamente por SemRep. Nuestro enfoque descarta resultados no deseados como, por ejemplo, las tres entidades y las tres relaciones resaltadas en negrita, ya que no están incluidos en los tipos semánticos o de predicados requeridos. En el caso de que no se recuperen las entidades y predicados esperados, nuestro enfoque aplica los pasos que se describen a continuación.

Pre-procesado de Texto

Se utilizaron dos analizadores de código abierto (OpenNLP y Stanford) para proporcionar un análisis sintáctico sencillo (es decir, no muy detallado). Las etiquetas de categoría gramatical que utilizan los analizadores de Stanford y OpenNLP pertenecen al conjunto de etiquetas Penn Treebank[5]. Una vez obtenidos los árboles sintácticos utilizando los dos analizadores, nuestro método recorre los árboles resultantes, en busca de tuplas NP-VP y PP-NP-VP (véase el pre-procesado de texto en la Figura 5.4. En la parte superior se muestra un ejemplo con una tupla del segundo tipo). Estas son las tuplas que se procesan. Las tuplas NP-VP implican un sintagma nominal (NP), que generalmente representa el sujeto de la oración, y una frase verbal (VP), que normalmente representa el predicado de la oración, mientras que el tipo PP-NP-VP es una tupla NP-VP incluida en una frase pre-posicional (PP). Estas tuplas son interesantes para extraer relaciones semánticas binarias entre entidades médicas. El método compara los resultados de los dos analizadores y descarta aquellas tuplas que difieren de las esperadas, mejorando, de esta manera, el rendimiento de cada analizador individualmente. Para la oración de muestra, los dos analizadores proporcionan análisis correcto, pero para la siguiente oración *Breathlessness, ankle swelling, and fatigue are the characteristic symptoms and signs of chronic heart failure*, OpenNLP produce un análisis diferente de las tuplas esperados, mientras que el analizador de Stanford genera el análisis deseado. Por lo tanto, en esta última situación, el método elige el análisis de Stanford y se descarta el análisis OpenNLP.

NER

La mayoría de los patrones sintácticos de las entidades médicas pertenecientes a una guía de práctica clínica encajan en un sintagma nominal (NP). Dentro de una oración, puede haber dos tipos de NP ([52]): *NP máximas* o *compuestas*, es decir, los NP de longitud máxima y que contienen otras frases NP, PP o cláusulas relativas; los *NP básicos* o *simples*, es decir, aquellos NP mínimos que no incluyen otras frases nominales. Como no existe un nivel óptimo de NP para mapear a UMLS, nuestro enfoque intenta hacer coincidir las NP máximas (aquellas que no fueron mapeados con SemRep) utilizando el servicio UMLS NormalizeString. El objetivo de esta forma de procesado es la extracción del mayor nivel de semántica posible. En el caso de que no se produzca ninguna asignación para una NP máxima, ésta se divide en sus NP internas o constituyentes (es decir, nombres, adjetivos, ...) y se envía una solicitud a la base

[5] `http://web.mit.edu/6.863/www/PennTreebankTags.html`

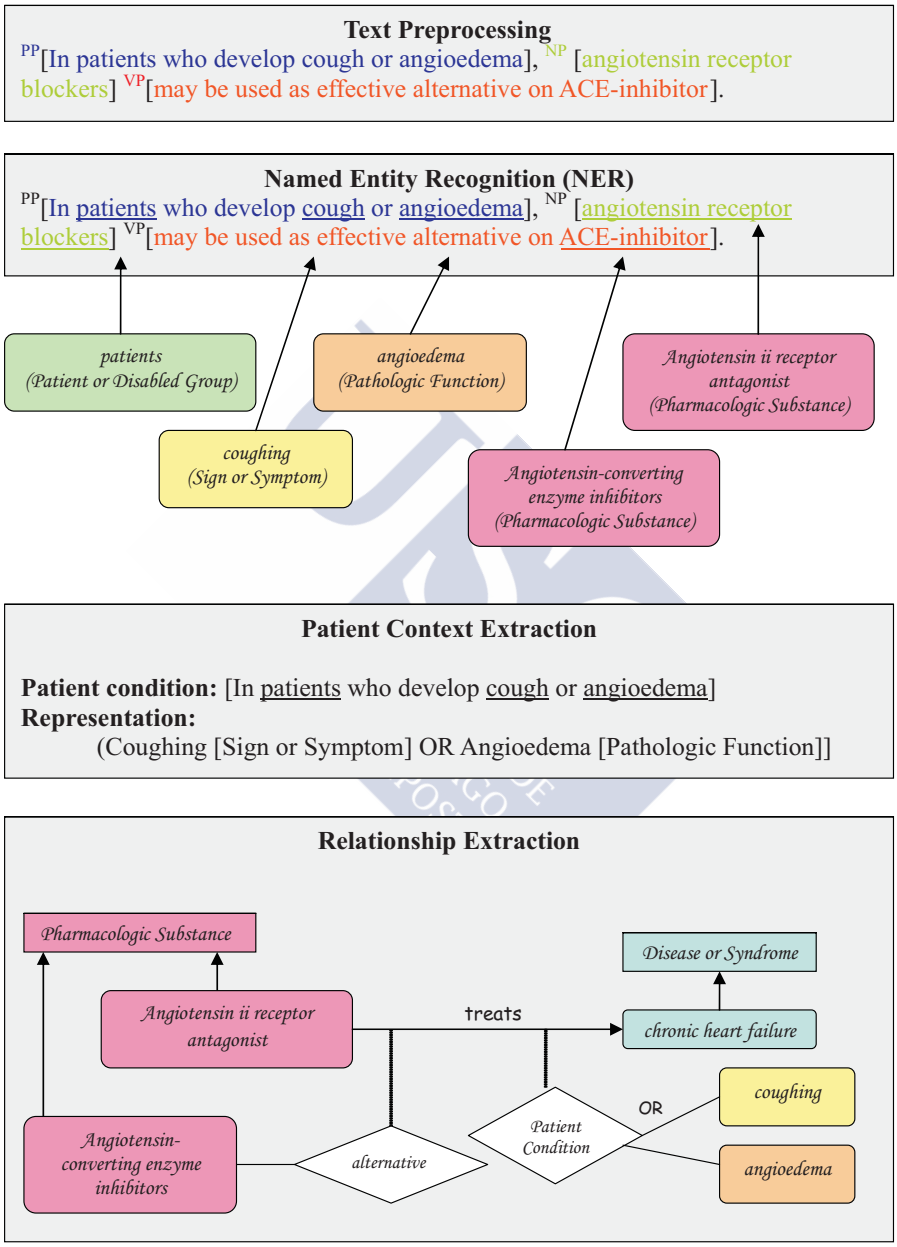


Figura 5.4: Un ejemplo con las etapas de PLN que sigue nuestra propuesta.

EXPRESIONES REGULARES TEXTUALES
in patients with *
* in patients who
in <Pathologic Function>[with | from | and | or <Pathologic Function>]
in * [pacientes tolerant | intolerant] to <Pharmacologic Substance>*

donde * puede ser un hallazgo, una función patológica o una sustancia farmacológica

Tabla 5.1: Expresiones regulares utilizadas para la extracción del contexto del paciente.

de datos UMLS con el fin de asignarles algún concepto UMLS. En el ejemplo de muestra, solo la frase nominal *angiotensin receptor blockers* es anotada con la sustancia farmacológica *angiotensin ii receptor antagonist* (Figura 5.4), ya que el resto de frases nominales fueron anotadas con éxito previamente por SemRep. Para el resto del texto, esta etapa no encontraría ninguna anotación más.

Extracción del Contexto del Paciente

El contexto del paciente, bajo el que se sugiere el diagnóstico o procedimiento terapéutico, se identifica automáticamente por medio de un conjunto de expresiones regulares que vienen marcadas por tipos NP o PP. Algunas de las expresiones regulares se pueden ver en la tabla 5.1.

Una vez que se extrae el contexto del paciente de la oración, el método identifica sus entidades centrales y los vínculos entre ellas, que pueden ser enlaces “Y” u “O”. Siguiendo con la frase de muestra (Figura 5.4), el contexto del paciente se representa por medio de dos conceptos: *Coughing* y *Angioedema*, unidos por la conectiva “O”.

Extracción de las Relaciones

La fase de extracción de relaciones se lleva a cabo en varias etapas. En primer lugar, el método identifica los conceptos relevantes en la oración a partir del conjunto de las entidades anotadas y las entidades del contexto. Las entidades que se corresponden con términos muy genéricos, como *symptom*, *sign*, *finding*, *procedure*, *therapy*, *patient*, . . . se consideran conceptos no relevantes, mientras que el resto de las entidades se consideran conceptos relevantes. En la oración del ejemplo, todas las entidades (junto con el contexto de diagnóstico de ICC) se clasifican como conceptos relevantes. A continuación, un conjunto de reglas predefinidas

identifica verbos, nombres y conceptos no relevantes como relaciones de la SN, así como las entidades de las oraciones como los argumentos en esas relaciones de *treats*, *diagnoses* o *has adverse effects*. Las relaciones sólo son posibles si el tipo semántico de las entidades candidatas coincide con los tipos semánticos de los argumentos de la relación. En la figura 5.4, la identificación de la relación *treats* se desencadena por el verbo *uses* y las entidades candidatas *angiotensin ii receptor antagonist (pharmacologic substance)* y *CHF (disease or syndrome)*. El siguiente paso, es identificar los atributos y añadirlos a la relación. Atributos posibles para la relación *treats* entre una sustancia farmacológica y una enfermedad o síndrome, son otras sustancias farmacológicas que se pueden administrar como alternativa o conjuntamente. En la oración de muestra, *angiotensin ii receptor antagonist* es una alternativa a *angiotensin-converting enzyme inhibitors* para la ICC. Finalmente, se añade el contexto del paciente a la relación.

5.5. Resultados

Con el objetivo de determinar la exactitud de este método propuesto, se llevó a cabo una evaluación en el dominio para el que fue desarrollado y que se describe brevemente a continuación. Para cada frase objeto de estudio, nuestro enfoque ha generado una o más entidades y relaciones que se han comparado con las entidades y las relaciones que fueron marcados manualmente. Además, se han considerado los siguientes criterios: una relación se interpreta como correcta si

– contiene las dos entidades descritas en la oración como argumentos, y un predicado, y

– las entidades y el predicado deben ser los mismos que los identificados manualmente

El contexto del paciente se considera correcto si las entidades y las expresiones resultantes son las mismas que las identificadas manualmente.

5.5.1. Pre-procesado de Texto

De 171 oraciones, los resultados de la identificación de las frases nominales utilizando cada analizador sintáctico (OpenNLP y el de Stanford) de forma independiente y después de una comprobación automatizada, han alcanzado el 81 % de *precisión*, 92 % de *recall* y 86 % de *F1-measure* para el analizador OpenNLP, y el 68 % de *precisión*, 96 % de *recall* y 80 %

F1-measure para el analizador de Stanford. Utilizando un enfoque combinado de los dos analizadores, nuestro método logra 86 % de *precisión*, 96 % de *recall* y 91 % de *F1-measure*.

Interpretación

Evaluando las discrepancias entre los analizadores sintácticos, hemos identificado diferentes situaciones. En primer lugar, la herramienta de Stanford analiza erróneamente las frases preposicionales que incluyen una coordinación, como *patients with PLVEF or diastolic dysfunction in CHF*. Este tipo de frases aparecen con frecuencia en las oraciones, siendo la razón principal de una *precisión* baja. En segundo lugar, algunos términos médicos que finalizan en *-y* se analizan incorrectamente como adverbios (como *ventriculectomy* o *aneurysmectomy*) provocando así, una disminución de la *precisión*. Sin embargo, en oraciones largas y complicadas que contienen frases subordinadas, el *recall* del analizador Stanford obtiene un valor mayor que el OpenNLP.

5.5.2. NER

De las 171 oraciones utilizadas en la evaluación, la extracción de entidades logra una *precisión* de alrededor de 83 % y un *recall* de 91 %, utilizando SemRep. En total, un 87 % de entidades se asignaron correctamente y completamente a los conceptos Metathesaurus correspondientes (*F-measure*), utilizando SemRep. Estos resultados se han mejorado sustancialmente cuando se aplicó el paso de NER para las frases nominales no mapeadas con SemRep: una *precisión* de 98 % y un *recall* de 95 %. En la Figura 5.5, se muestra la *precisión*, el *recall* y la *F-measure* del método SemRep, y la combinación de SemRep y nuestro NER para cada uno de los ocho tipos semánticos de nuestro dominio. La *precisión* de SemRep para la extracción de los *laboratory procedures*, *findings* y *pathologic functions* fue baja (alrededor del 35 %, 70 % y 75 %, respectivamente), mientras que el *recall* sólo fue bajo para *therapeutic procedures* (alrededor 60 %).

Interpretación

Durante la evaluación de las discrepancias entre el reconocimiento de entidades manual y automático, hemos detectado dos situaciones principales que disminuyen la *precisión* de SemRep. Teniendo en cuenta que SemRep utiliza MetaMap para reconocer conceptos del Metathesaurus y MetaMap ofrece las mejores entidades candidatas para cubrir el texto. En al-

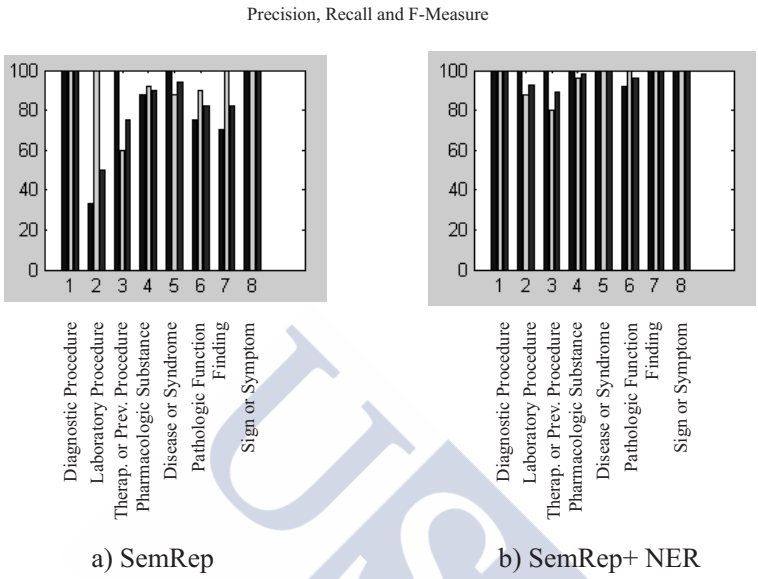


Figura 5.5: Resultados de aplicar los pasos generales de este estudio.

gunas ocasiones, MetaMap ofrece más de un candidato que cubra la misma parte del texto. Por ejemplo, para la frase nominal *s-creatinine*, MetaMap proporcionó dos mejores candidatos: *creatinine (biologically active substance, organic chemical)* y *creatinine (creatinine finding)*. En situaciones como esta, SemRep elige uno (quizás al azar). Hemos detectado que esta operación se repitió en varias ocasiones con las entidades correspondientes al tipo semántico de *laboratory procedures*. Esta fue la causa de una *precisión* baja. Una alternativa mejorada sería tener la posibilidad de configurar esta operación cuando MetaMap propone varios candidatos válidos. De esta manera, SemRep tendría en cuenta las heurísticas específicas de cada aplicación. En otras ocasiones, las entidades reconocidas por SemRep no coinciden con las proporcionadas por MetaMap. La causa puede ser que SemRep (vía MetaMap) está accediendo a una versión diferente UMLS a la que accede MetaMap cuando se ejecuta de forma independiente. En la frase de muestra escogida, SemRep reconoció la frase nominal *angiotensin receptor blockers* como *angiotensin receptor* (un receptor o una proteína), mientras que si se ejecuta MetaMap independientemente, esta frase es reconocida como *angiotensin ii recep-*

tor antagonist (substance). Una vez más, una alternativa mejorada sería tener la posibilidad de configurar esta operación con SemRep.

5.5.3. Extracción de Relaciones

De las 171 oraciones utilizadas en la evaluación, SemRep extrae 240 relaciones, logrando una *precisión* de alrededor del 80 %. En total, SemRep extrae 16 tipos diferentes de relaciones de la red semántica, recuperando principalmente relaciones de tipo *treats* (48 %), *process of* (24 %) y *coexist with* (8 %). Como solo nos interesaba la extracción de dos tipos de predicados de la red semántica (*treats* y *diagnoses*), en un principio, nuestro método sólo tuvo en cuenta los predicados *treats* proporcionados por SemRep (48 % de los totales). Por lo tanto, nuestro método descartó con acierto las relaciones cuyos argumentos no correspondían se con conceptos relevantes.

En total, se descartaron el 42 % de los predicados *treats* seleccionados dentro de los proporcionados por SemRep y, finalmente, sólo 86 relaciones *treats* fueron extraídas. Un ejemplo de predicado filtrado es: *angiotensin ii receptor antagonist TREATS patients*. Estos resultados se han mejorado sustancialmente al aplicar el segundo proceso de extracción de relaciones sobre las oraciones, en el que se han reconocido 137 relaciones adicionales. En total, en esta tesis hemos extraído 223 relaciones: 69 de tipo *diagnoses*, 151 de tipo *treats* y 3 de tipo *has adverse effects*, logrando una *precisión* del 82 % y un *recall* del 78 %.

Interpretación

A pesar de que la *precisión* de SemRep es alta, el número de predicados relevantes para nuestra aplicación es bajo: 39 % del total de predicados extraídos. Sin embargo, SemRep es una herramienta general orientada a reconocer una amplia gama de relaciones entre todos los conceptos UMLS identificados en la frase, mientras que nuestro método sólo genera tres tipos de predicados entre los conceptos principales. Más importante es el hecho de que nuestro método tiene en cuenta el contexto del paciente. Este último se lleva a cabo dividiendo la oración en dos partes: el contexto del paciente y la parte del predicado, siendo solo el último el utilizado para sugerir la relación. Pensamos que si SemRep se adaptase de la misma manera que nuestro método, el *recall* aumentaría sustancialmente. Por otro lado, el procesado de estructuras sintácticas *comparativas* dieron como resultado una identificación incorrecta de los conceptos relevantes, así como de predicados. Un ejemplo de oración *comparativa* es *anti-arrhythmic drugs other than beta-blockers*. Frases sintácticas complejas, incluyendo re-

ferencias y subordinadas relativas, provocaron una ausencia o una representación errónea de las relaciones. Otro motivo de errores en ambos métodos fue la representación inadecuada de oraciones negativas.

5.5.4. Extracción del Contexto del Paciente

De las 171 oraciones utilizadas, sólo 72 incluyen un contexto de pacientes (42 %). Nuestro método logra un 81 % de *precisión*, un 69 % de *recall* y un 75 % de *F1-measure* para la extracción de contexto.

Interpretación

Hemos detectado varias situaciones que disminuyen el rendimiento de nuestro método. Un *recall* bajo se debe principalmente a

- la falta del reconocimiento de entidades y
- la mala identificación del contexto por medio de expresiones regulares predefinidas.

Con el fin de resolver este inconveniente, en el futuro, tenemos la intención de explorar la posibilidad de utilizar también algunos predicados extraídos por SemRep, como *process of* o *coexist with*. En cuanto a la *precisión*, las expresiones sintácticas complejas llevaron a representaciones incorrectas o incompletas del contexto.

5.6. Ejemplo de Aplicación

Utilicemos las siguientes oraciones ([163]) para ejemplificar el método que acabamos de describir en el caso en el que SemRep no ofrezca respuesta:

Oración 1 *Routine diagnostic evaluation of patients with CHF includes complete blood count, S-electrolytes, S-creatinine, S-glucose, S-hepatic enzymes and thyroid function.*

Oración 2 *All patients with symptomatic chronic heart failure that is caused by systolic left ventricular dysfunction should receive an ACE-inhibitor.*

Oración 3 *Oxygen has no application in CHF.*

Como ya comentamos, esta propuesta aplica una combinación secuencial de métodos básicos típicos de la ingeniería del conocimiento, para anotar gradualmente el contenido relevante de un texto expresado en lenguaje natural con la terminología UMLS Metathesaurus. El proceso está dividido en una serie de pasos generales:

1. Usar el Metathesaurus para reconocer las entidades de la GPC.
2. Para cada oración, determinar el estado del paciente bajo las cuales el conocimiento descriptivo de la oración es válido.
3. Para finalizar, detecta las unidades de información central de la oración (conceptos destacados), para asociarlos con un conjunto de relaciones de la SN de UMLS.

NER

Esta parte hace uso del análisis sintáctico de la oración y del proceso de normalización lingüística. Veamos cómo implementamos este proceso:

Selección de la información relevante Después del marcado de la GPC por parte de los expertos, las oraciones se anotan con el contexto de diagnosis al que referencia. Por ejemplo, las oraciones de la guía, se anotan con *CHF*, *CHF with PLVEF*[6] o *CHF in elderly patients*.

Análisis sintáctico Después de extraer el árbol del análisis sintáctico, lo recorremos en busca de frases nominales (NP) simples, es decir, que no contienen otras frases nominales dentro, o compuestas, es decir, aquellas que incluyen otras NP en su subárbol, las llamadas preposicionales o coordinadas. La tabla 5.2 nos muestra estos casos para la oración 1 de ejemplo.

Anotación de las frases nominales con entidades médicas Cada NP se asocia con uno o más conceptos Metathesaurus. Cuando no se encuentra una asociación para la frase completa, esta se divide en sus frases nominales internas o en sus constituyentes morfológicos (nombres, adjetivos, . . .) y se vuelve a intentar la anotación. Veamos un ejemplo en la tabla 5.3 para la oración 2.

Desambiguación de las entidades médicas Que se lleva a cabo en base a la siguiente serie de reglas heurísticas de criterio sintáctico, semántico y terminológico:

[6] Preserved Left Ventricular Ejection Fraction

NP Preposicional [Routine diagnostic evaluation of patients with CHF] includes *NP Coordinado* [complete blood count, S-electrolytes, S-creatinine, S-glucose, S-hepatic enzymes and thyroid function]

Tabla 5.2: Resultados parciales del preprocesado de la oración 1 de ejemplo.

Término	Concepto Metathesaurus	Tipo Semántico
patients	C0030705	Patient or Disabled Group
symptomatic	C0231220	Functional Concept
chronic heart failure	C0264716	Disease or Syndrome
systolic left ventricular dysfunction	C1277187 C1963159	Pathologic Function Finding
ACE-inhibitor	C0003015	Pharmacologic Substance

Tabla 5.3: Resultados de anotación de las frases nominales con conceptos Metathesaurus.

- Regla sobre las frases nominales coordinadas: En el caso de que varias de las frases nominales que forman una coordinación estén asociadas con conceptos Metathesaurus que comparten el mismo TS, la frase de la coordinación ambigua se vinculará con un concepto que posea el mismo TS.
- Reglas sobre los tipos semánticos: Solo utilizamos un conjunto de los TS del total y además, establecemos prioridades entre ellos. Por ejemplo, *procedure* tiene preferencia sobre *substance*.
- Regla basada en comparación de cadenas de caracteres: Se establece una preferencia sobre aquellos términos que tienen una mayor similitud de caracteres.

Reconocimiento del Estado del Paciente

Para la identificación del estado del paciente, primero se normaliza la oración y luego se hace una representación semiformal.

Normalización Una vez que se identifican las expresiones regulares de la tabla 5.1 en las oraciones, se transforman en una forma normal consistente en el estado del paciente y en el conocimiento médico descriptivo, tal y como se puede observar en la tabla 5.4.

Estado del paciente	All patients with symptomatic chronic heart failure that is caused by systolic left ventricular dysfunction
Conocimiento médico	should receive an ACE-inhibitor

Tabla 5.4: Forma normal para la oración del segundo ejemplo.

Estado del paciente	**Representación**
All patients with symptomatic chronic heart failure that is caused by systolic left ventricular dysfunction	((Disease or Syndrome) chronic heart failure (Functional concept) Symptomatic) AND ((Pathologic Function) Left ventricular systolic dysfunction)

Tabla 5.5: Representación del estado del paciente para la segunda oración.

Representación del estado del paciente Después de identificar los conceptos principales del estado, se comprueba si están enlazados por coordinadas copulativas o disyuntivas. Continuando con la segunda oración de ejemplo, el estado del paciente se representa a través de dos conceptos (*Chronic heart failure* acompañado del modificador *symptomatic* y *Left ventricular systolic dysfunction*) enlazados por un *AND*. Por lo tanto, los pacientes bajo estas condiciones, *should receive an ACE-inhibitor*. Lo vemos en la tabla 5.5.

Extracción de las Relaciones

Este proceso se resume en las siguientes fases.

Identificación de los elementos relevantes de información Este paso es el encargado de identificar los conceptos que aportan información relevante dentro de la oración, para que puedan ser vinculados con un conjunto de relaciones predefinidas. Se descarta el texto que se corresponde con conceptos médicos muy generales ya que se considera que no aportan información destacada. En la oración ejemplo número 1, algunos conceptos se clasifican como conceptos relevantes (*CHF*, *complete blood count*, *Electrolytes measurement*, ...) y otros no relevantes como (*patient*, *evaluation procedure*, *diagno-*

sis). En la del ejemplo 2, solamente hay un concepto relevante (*ACE-inhibitor*), por lo que el contexto de diagnóstico (*CHF* en este caso) también se añade al conjunto de este tipo de conceptos.

Generación de pares de entidades Todas las entidades identificadas en el paso anterior se usan para generar los pares de entidades con la restricción de que deben estar en frases nominales independientes. En la oración 1 de ejemplo, se generan seis pares debido al elemento relevante *CHF* contenido en la frase preposicional y a los 6 elementos relevantes de la frase coordinada, segundo se ve en la tabla 5.2.

- Complete blood count ↔ Chronic heart failure
- Electrolytes measurement, serum ↔ Chronic heart failure
- Creatinine measurement ↔ Chronic heart failure
- Glucose measurement ↔ Chronic heart failure
- Measurement of liver enzyme ↔ Chronic heart failure
- Thyroid Function Tests ↔ Chronic heart failure

Identificación de las relaciones Cada par obtenido del paso anterior se asocia con un tipo de relación perteneciente a la SN de UMLS teniendo en cuenta los TS de los conceptos con los que se construye el par. En el ejemplo 1, se obtiene la relación *diagnoses* entre *Laboratory Procedure* y *Pathologic Function*. Sin embargo, en el ejemplo 2, hay dos posibles relaciones entre *Pharmacologic Substante* y *Pathologic Function* que son *diagnoses* y *treats*.

Desambiguación de relaciones Para desambiguar se utilizan los conceptos no relevantes. Por ejemplo, en el caso 2, la palabra *receive* se usa para descartar la relación *diagnoses*.

Detección de relaciones negadas Usamos una versión simplificada del algoritmo NegEx. Este está basado en expresiones regulares para identificar negaciones. Esto sería el caso del ejemplo 3.

Siguiendo los pasos que se acaban de comentar, cada oración se anota con una representación de su contenido clínico relevante. En la figura 5.6 podemos ver los resultados obtenidos sobre los ejemplos.

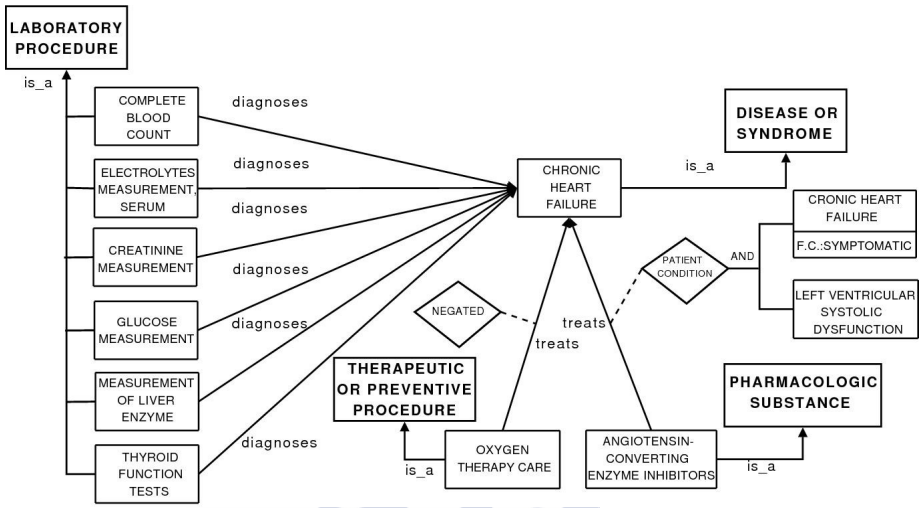


Figura 5.6: Resultados de aplicar el método a las oraciones de ejemplo.

5.7. Actualización del Contenido de las GPC

El uso de guías de práctica clínica electrónica requiere su actualización cada vez que la organización correspondiente publica una nueva versión de la GPC textual. Existen dos posibilidades para actualizar la GPC electrónica. La primera es repetir el proceso descrito sobre la GPC completa actualizada. Desde el punto de vista informático es la solución más sencilla en principio, aunque puede implicar cambios complejos cuando la GPC se integre con el sistema de información sanitario. La segunda opción consiste en revisar qué partes de la guía se han actualizado. Esta opción es más costosa, pero está más orientada al flujo de trabajo del personal sanitario, permitiendo disminuir su carga de trabajo. El personal sanitario requiere conocer los cambios y una opción como esta permitiría mostrarle los cambios de forma automatizada, sugiriendo las actualizaciones pertinentes a implantar en la GPC electrónica.

Por todo ello, en esta tesis hemos revisado cómo han ido evolucionando en el tiempo las GPC desarrolladas por la European Society of Cardiology (ESC) sobre el fallo cardíaco mediante el análisis de las GPC de 2005 [160], 2008 [77] y 2012 [105]. El objetivo prioritario ha sido estudiar el proceso de diseño de un herramienta orientada a facilitar al personal sanitario la revisión de las diferencias entre una guía y su actualización.

5.7.1. Análisis sobre la Evolución de GPCs sobre Fallo Cardíaco

En principio todas las guías desarrolladas por ESC sobre el fallo cardíaco reflejan ciertos temas básicos comunes: definición, síntomas, diagnóstico, pronóstico, tratamiento, monitorización ... Observando las tres guías, y a pesar de que representan el mismo conocimiento, es notable que hay una tendencia a ampliar el contenido, lo que implica una reestructuración de la información a lo largo de las versiones. Para todas las GPC, cabe destacar tres tablas que sirven para establecer una clasificación de la patología y que son referenciadas a lo largo de todo el ejemplar. Se trata de las tabla de las clases de recomendación que especifican la efectividad de un tratamiento, los niveles de evidencia que indican el motivo por el que se decide un tratamiento en base a estudios, consenso por parte de los expertos u otros motivos y, finalmente la clasificación de la patología que establece la New York Heart Association en función de la severidad de los síntomas.

Toda guía de ESC comienza con los apartados de definición de fallo cardíaco y de términos relacionados. En el caso de las guías de 2005 y 2008 poseen una estructura muy similar siguiendo los mismos apartados.

Las guías continúan con la especificación de los síntomas y los signos. En este punto comienza la evolución de la guía del 2005 hacia la del 2012, por ejemplo, en la inclusión de una tabla explicativa resumen de los síntomas y signos de la patología.

El siguiente tema tratado es la diagnosis. Todas coinciden en especificar el algoritmo de diagnosis en forma de diagrama. En el caso de la guía de 2005 incluye en este apartado todos los procedimientos y herramientas empleados para el diagnóstico, desde los análisis de laboratorio hasta las imágenes médicas. La guía de 2008 también sigue la misma pauta pero se amplia mucho más la información, por ejemplo, incorpora tablas que enumeran las anomalías posibles para los tipos de pruebas más generales. Al igual que la del 2008, la guía del 2012 sigue el mismo camino, pero además utiliza una sección diferenciada para los tipos de pruebas basados en imágenes médicas.

Con respecto al tratamiento, las guías de 2005 y 2008 diferencian entre farmacológico y no farmacológico, y entre dispositivos y cirugía. Con tratamiento no farmacológico se refiere al comportamiento y hábitos del paciente; sin embargo, la guía del 2012 la incluye en la sección de control holístico (*holistic management*) centrada en la educación, los hábitos y la monitorización del paciente. Dicha sección es independiente de los apartados de tratamiento. A cerca del farmacológico, por una parte, existen una serie de fármacos relevantes para el tratamiento de esta patología (*ACE-inhibitors*, *beta-blockers*, *ARB*, *diuréticos*...) a los que las

guías les dedican especial atención incluyendo tablas especiales que regulan la administración de estos fármacos. En el caso de la guía del 2012, los datos sobre estas drogas relevantes se incluyen en un anexo independiente al de la guía y la del 2008, para remarcar a mayores la importancia de estos fármacos, y utiliza un apartado independiente para el resto de medicamentos. Por otra parte, la guía del 2012 estructura el tratamiento farmacológico en función de la fracción de eyección (*ejection fraction*), y añade terapias perjudiciales o que no tienen su eficacia completamente demostrada. Las guías de 2005 y 2008 describen los métodos que involucran aparatos o cirugía en una única sección. Sin embargo, la guía de 2012 explica cada método en una sección separada.

La principal diferencia de la guía del 2005 con respecto al resto es la particularidad de la vejez del sujeto de observación a la hora del tratar del fallo cardíaco. Esta diferenciación incluye un tratamiento farmacológico específico, mientras que las otras guías tienen en cuenta estas condiciones especiales en el apartado de administración de la droga en cuestión. Y, lo más importante, esta condición incluye la información sobre las arritmias mientras que las otras guías las tratan de forma independiente a la edad.

Con respecto al contenido de las guías del 2008 y 2012 que amplían en mayor medida en comparación con la del 2005 estaría la etiología del fallo cardíaco, la contemplación de otros trastornos a mayores de la enfermedad primaria y el tratamiento de el fallo cardíaco agudo. Estas guías atienden también a posibles lagunas en la evidencia.

También existe cierta información incluida en las guías del 2008 y 2012 pero no en la del 2005 son las *key evidencie*. En estos apartados se hace referencia principalmente a estudios sobre los que se basa la recomendación del tratamiento. Para cada evidencia se especifica por ejemplo, el número de pacientes, los síntomas, el tratamiento y la evolución.

Una parte importante de las GPC son las tablas. Por una parte, ofrecen un resumen de cierto ámbito de información expresado normalmente con oraciones más sencillas y a las que se puede aplicar PLN. Por otra parte, son una forma de medir la similitud entre contenido de guías y de la información relevante de estas.

Podría decirse que el motivo principal por el que las guías de 2008 y 2012 son mucho más amplias que la del 2005 es que la del 2008 ya supone una continuación y revisión de la guía que se utiliza para esta tesis, a mayores de la guía *Executive summary of the guidelines on the diagnosis and treatment of acute heart failure: the Task Force on Acute Heart Failure of the European Society of Cardiology* y la del 2012 es una revisión de la del 2008. Otro motivo por el que pueden organizar la información de forma diferente es la referencia a otras GPC,

(por ejemplo, la guía del 2012 hace referencia la de *atrial fibrillation*) y no profundiza ni hace esfuerzo por una explicación clara de esta patología.

5.7.2. Análisis de la Similitud entre GPCs de Fallo Cardíaco

Para completar el estudio de evolución de las guías es importante observar el cambio que han sufrido las porciones de texto marcadas por los expertos clínicos del proyecto HYGIA.

Para ello, se ha analizado cada frase en la guía del 2005 y se ha comparado con su aparición en la del 2012. En función de la similitud entre contenidos se establecieron cinco niveles:

- Pieza de contenido exactamente igual: Si cada letra de la frase marcada en la GPC del 2005 coincide exactamente con otra de la del 2012.
- Pieza de contenido semánticamente igual pero ligeramente diferente en expresión: En el caso de que la frase del 2005 difiera en alguna palabra con respecto a la del 2012 debido al uso de algún sinónimo, alteración del orden de las palabras, ...
- Pieza de contenido semánticamente igual pero en diferente formato: En ocasiones, se citan frases que aparecen redactadas dentro del texto expositivo en 2005 (*ascertain presenting features: pulmonary oedema, exertional breathlessness, fatigue, peripheral oedema*) pero en 2012, la misma información aparece localizada en otra parte de la guía, como en forma de tabla (*Table 4: Signs and Symptoms*). Otro ejemplo, serían enumeraciones como *heart transplantation, ventricular assist devices, and artificial heart* que están repartidas a lo largo del contenido de una sección, *13. Coronary revascularization and surgery, including valve surgery, ventricular assist devices, and transplantation*, en este caso, y de sus subsecciones correspondientes.
- Pieza de contenido revisada: Cuando la mayor parte de la semántica del 2005 está contenida en la del 2012 pero con alguna modificación. Un ejemplo, lo podemos ver en 2005 para la dosis diaria *spironolactone 12.5-25 mg, eplerenone 25 mg* que en 2012 se especifica como *Spironolactone/eplerenone 12.5-25 mg*. Igual ocurre con los períodos de tiempo: bajo el mismo contexto, en 2005 aparece *check serum potassium/creatinine after 1 week*, mientras que en 2012 nos encontramos con *Check blood chemistry at 1 and 4 weeks after starting/increasing dose*. Un último caso sería la diferencia de criterios entre guías, en 2005 podríamos hallar *avoid NSAIDs and coxibs* y en 2012 *Avoid NSAIDs unless essential*.

# piezas de contenido totales GPC 2005	274
# de piezas de contenido exactamente igual	46
# de piezas de contenido semánticamente igual pero ligeramente diferente en expresión	32
# de piezas de contenido semánticamente igual pero en diferente formato	132
# de piezas de contenido revisada	43
# de piezas de contenido omitida	21

Tabla 5.6: Niveles de Similitud entre el texto utilizado en la GPC del 2005 y 2012.

– Pieza de contenido omitida: Cuando no es posible encontrar el contenido de 2005 en 2012.

En la siguiente tabla 5.6 podemos ver estadísticas sobre los niveles de similitud.

Hay que destacar que parte importante del texto de la GPC del 2005 que coincide exactamente igual con la del 2012, es debido a coincidencias con títulos de secciones o contenidos de tablas en las que no es frecuente que existan oraciones completas redactadas. Las frases de contenido semánticamente igual pero ligeramente diferente en expresión son cadenas de texto más largas. Es importante tener en cuenta la clasificación de contenido semánticamente igual pero en diferente formato debido al nivel de expresividad del lenguaje natural puesto que la misma información se puede representar de muchas formas posibles. Además, tanto los períodos de tiempo como las dosis de administración de medicación han cambiado sutilmente a lo largo de los años, por lo que, no sería posible una equiparación exacta a pesar de que la semántica es la misma. Con respecto a los contenidos no reflejados, debido a que este análisis se ha realizado por miembros del grupo de investigación y no personal clínico experimentado, es posible que la ESC haya incluido esta información en la guía del 2012. Esta suposición se basa en la evidencia de que como fruto del análisis se han encontrado nuevos hipónimos, hiperónimos, sinónimos y siglas.

5.7.3. Alineamiento entre Diferentes Versiones de una Misma GPC

De la misma forma que es importante realizar una revisión y actualización de las GPC a lo largo del tiempo, también es relevante la identificación automática del contenido relevante en las nuevas guías. Para ello, sería interesante localizar en las guías actualizadas la información marcada por los expertos en las guías iniciales.

Un trabajo que estudia el modelado de GPC es el de Serban et at. [147]. Este modelado se realiza a partir de patrones lingüísticos que realizan una asociación entre un fragmento del texto y una representación formal del conocimiento que contiene. Puntualizan que solo una parte de la GPC necesita actualización y, además, basándose en estudios previos sugieren procesar por separado los diferentes tipos de conocimiento presentes en la guía, entre ellos los procedimientos, sobre los que trabaja también nuestro estudio. Comentemos a continuación el algoritmo que crearon de forma abreviada: partiendo del lenguaje natural, identifican los términos contenidos en la ontología clínica y en un lexicón no médico para etiquetarlo y a continuación, sustituyen esas partes con la categoría a la que pertenecen en la ontología. Así, van consiguiendo patrones a distintos niveles de abstracción (formados por términos médicos a nivel de palabra y por conceptos a nivel de oración) hasta conseguir una expresión compacta y que supone una secuencia de conceptos médicos enlazados a través de las relaciones de control definidas en la ontología. Consideran que una representación ejecutable de una GPC consiste en las acciones y en las relaciones de control referenciadas en la guía.

Una ventaja de este método es que reduce la ambigüedad ya que utiliza una ontología propia con el inconveniente de que no se puede aplicar en todos los ámbitos. Otro problema es que la creación de patrones lingüísticos con significado completo no puede ser totalmente automatizado. Al igual que nuestros trabajos, crean su propio *gold standard*. Sin embargo, la principal diferencia es que este trabajo se centra en generar una ontología para crear, además de validar, los patrones candidatos para una categoría de textos con reglas de formato bastante estrictas, lo que, de nuevo, reduce la ambigüedad a costa de reducir flexibilidad en formato de oraciones y contenido del texto.

Por nuestra parte, para lograr una identificación automática del contenido relevante en las GPC y basándonos en el estudio realizado sobre comparación de guías, se podría trabajar en un futuro en el procesado que se describe a continuación.

El primer paso a ejecutar, por su facilidad de implementación y eficiencia, sería la identificación de piezas de texto que son exactamente iguales en ambas guías. En el capítulo 3 se han definido técnicas para el alineamiento de cadenas de caracteres, teniendo en cuenta que, debido a la longitud del texto contenido en la guía destino, las técnicas de recuperación de información podrían tener más peso en esta fase.

A continuación, se introducirían en el proceso los recursos lingüísticos para poder emplear características como la sinonimia o la flexión de las palabras en conjunción con otras técnicas de alineamiento a nivel individual.

Llegados a este punto deberíamos tener un conjunto bastante fiable de alineaciones en el caso de que se aplique un criterio que tenga como condición una gran similitud entre cadenas de caracteres. Aquí sería el momento adecuado para plantearse la aplicación de técnicas estructurales.

Observando la información marcada por los expertos clínicos, vemos que gran parte está incluida dentro de tablas o diagramas. Esto tiene sentido puesto que son estructuras que remarcan datos relevantes y facilitan su procesado. Por otra parte, examinando desde fuera el proceso manual de comparación de guías, nos hemos dado cuenta de que es muy importante el contexto en el que se encuentra la pieza de texto que queríamos localizar. Por lo tanto, nos interesa conocer e intentar relacionar aquellas secciones, tablas o diagramas de las guías en las que esté incluida la información relevante inicial.

Una forma de intentar vincular los contextos en los que están incluidos el texto a localizar es utilizar las alineaciones ya realizadas: situar a qué sección pertenecen en el origen y a cuál en el destino, comprobar si esos contextos son equivalentes e intentar asociar el texto que está bajo el mismo contexto en origen con el contexto ligado en destino.

El principal inconveniente con el que nos encontraremos será que, a partir de este momento del proceso, las alineaciones que no sean exactas no tienen por qué ser erróneas. Para ello habrá que contemplar situaciones como cambios en las dosis o en los períodos de aplicación del tratamiento o control. También podría haber modificaciones en los criterios clínicos, por ejemplo, en la enumeración de medicamentos, condiciones . . . Como ayuda a esta parte, entraría en juego el análisis sintáctico de las frases y la relación entre sus elementos constituyentes utilizando alineaciones a una ontología, incluyendo las relaciones lógicas, y patrones como los que ya se aplican en este estudio.

En conclusión, las técnicas enunciadas y aplicadas en esta tesis servirán de base para una continuidad del trabajo centrado en un ámbito tan complejo como es el lenguaje natural.

Capítulo 6

Conclusiones y Trabajo Futuro

El problema principal con el que tienen que enfrentarse los sistemas informáticos clínicos es el de alcanzar la interoperabilidad. La gran heterogeneidad en formato y contenido que presenta la documentación médica es el principal impedimento. Por eso es necesario el encontrar un formato común a través del cual se puedan comunicar los sistemas ofreciendo un servicio que sea lo más rápido y eficiente posible. Durante la consulta de un paciente por parte de un especialista sanitario se debe permitir un fácil acceso tanto al historial del sujeto de observación como a documentación de apoyo al diagnóstico. Por este motivo creamos una aproximación con la intención de avanzar en la solución de este obstáculo.

En esta tesis se trata de alcanzar este objetivo teniendo en cuenta todos los tipos de documentos clínicos en cuanto a estructuración se refiere. Cada uno de ellos tiene sus propias desventajas para las que hay que aplicar diferentes soluciones en función de las características específicas de cada uno.

Como recurso de información semiestructurada trabajamos con los arquetipos OpenEHR. El objetivo de estos es conseguir una representación íntegra de la historia clínica electrónica (HCE) de una manera significativa, inequívoca y precisa. Por lo tanto, es esencial una estandarización de la representación de información relacionada con el paciente para que sea detallada y completa. El uso de la terminología estructurada SNOMED CT proporciona un método más apropiado para expresar términos de datos clínicos no ambiguos, computables e interoperables. El uso de esta terminología en los arquetipos, les proveería de la interoperabilidad que buscamos. Actualmente, con la excepción de un pequeño número de casos en los repositorios de arquetipos de contenido abierto, los enlaces con vocabularios estándar son

poco frecuentes. Por otra parte, hay terminologías médicas estándar, como SNOMED CT, que incluyen alrededor de 300.000 conceptos médicos, por lo que la asignación manual de vuelve muy lenta y requieren una gran cantidad de recursos humanos. Además, arquetipos clínicos y SNOMED CT, poseen puntos de vista diferentes en cuanto a estructuración de la información se refiere. Por eso es de gran utilidad el identificar las coincidencias entre el diseño de ambos.

El otro tipo de recurso con el que trabajamos es el de las guías de práctica clínica (GPC), cuya representación del conocimiento no posee ni modelado ni estructura. El objetivo de los documentos de GPC es asistir a los médicos en el cuidado apropiado de los pacientes y mediante una redacción en lenguaje natural (LN). La principal ventaja del LN es su expresividad que, a su vez, supone un inconveniente considerable de cara al modelado de la GPC. Dentro del procesado de GPC, es complicado identificar cuál es la información relevante que los clínicos usarán en la atención al paciente, por lo que hemos necesitado de expertos que nos dirigieran en el proceso de extracción de información.

A día de hoy las herramientas existentes para extraer la información están orientadas a ofrecer buenos resultados cuando son aplicadas en los dominios bajo los que fueron creadas pero no necesariamente bajo nuestro ámbito de aplicación. Tal es el caso de SemRep, cuya intención es extraer una amplia gama de predicados entre los conceptos UMLS que fueron identificados por MetaMap en el texto. SemRep logra una alta precisión, pero cuando se aplica a un dominio específico, como es el caso de estudio este trabajo, la cobertura (*coverage*) de los predicados extraídos considerados como relevantes disminuye sustancialmente. Una adaptación adecuada de SemRep que implique un pequeño esfuerzo de implementación es importante pero aún no ha sido completamente resuelto.

Para finalizar, un problema común a la mayoría de los trabajos relacionados con la extracción de conocimiento, es falta de un *gold estándar* contra el cual poder evaluar nuestros resultados.

6.1. Conclusiones

A continuación vamos a enumerar las siguientes conclusiones extraídas durante el desarrollo de esta tesis doctoral:

– El uso de métodos basados en similitud semántica, por ejemplo el uso del contexto del arquetipo, en combinación con otras técnicas, principalmente aquellas asentadas en el uso de métodos basados en caracteres, recursos lingüísticos o de información

estructural, pueden ser beneficiosos tanto para aumentar el *recall* de las otras técnicas, como para comprobar la validez de sus resultados y poder desambiguarlos.

– A pesar de lo evolucionadas que están las técnicas individuales basadas en la similitud léxica y de caracteres, hemos comprobado que la *precisión* mejora cuando se aplican en conjunto con otros métodos que tienen en cuenta más información de las entidades y del contexto en el que están encuadradas dentro de la fuente de información clínica.

– Verificación de que el proceso manual de anotación, a pesar de lo tedioso que puede llegar a ser, facilita la deducción del procedimiento automático que se implementa posteriormente y permite profundizar en el diseño de las fuentes de información. Partiendo de un conocimiento profundo de las técnicas de reconocimiento de entidades y relaciones, la creación de nuestro propio *gold estándar*, nos ha ayudado a descubrir particularidades y, sobre todo relaciones no documentadas entre las fuentes de información, que son clave a la hora de aplicar las técnicas relacionales.

– La información de estado del paciente en la GPC es imprescindible para extraer conocimiento consistente de la GPC. Puesto que en este trabajo se procesan los procedimientos diagnósticos y terapéuticos, estos no tendrían sentido si no se enmarcaran dentro del contexto del estado del paciente.

– Este trabajo confirma los beneficios de adaptar herramientas de código abierto para la extracción de predicados terapéuticos y de diagnóstico relevantes a partir de las GPC tal como es SemRep.

– El análisis sintáctico facilita en gran medida el procesado del LN contenido en las guías. Entre otras ventajas, nos permite identificar el contenido que hace referencia a entidades clínicas mediante las frases nominales, características específicas de estas entidades gracias a los adjetivos o adverbios y las relaciones entre entidades mediante los verbos o ciertas expresiones. Además, si combinamos el análisis con patrones lingüísticos, se puede localizar mejor el texto objetivo a identificar.

– La aplicación de técnicas apropiadas, la elección de herramientas y recursos adecuados al objetivo del trabajo, así como, el establecimiento de premisas fundamentadas nos han permitido obtener unos buenos resultados. Por ejemplo, en función del recurso de información que se va a anotar, se utiliza una terminología u otra teniendo en cuenta

si se adapta mejor o peor. Además, se garantiza la comunicación entre terminologías debido a la integración de UMLS. En estos trabajos, SNOMED CT se adapta muy bien a la historia clínica, puesto que es uno de sus objetivos, pero para las GPC que incluyen información sobre diagnóstico y tratamiento, en principio, su cobertura es más limitada, y usar UMLS es lo más práctico y sencillo por cobertura y herramientas de acceso.

6.2. Aportaciones

El trabajo sobre anotación de modelos de datos clínicos supone las siguientes contribuciones:

- Método completamente automatizado para anotar arquetipos de tipo OBSERVATION con conceptos SNOMED CT.
- Validación automática para las anotaciones realizadas a través de otras técnicas específicas y reduce la ambigüedad de los términos de los arquetipos generando asociaciones más específicas y adecuadas utilizando el principio de vecindad. Creemos que esta estrategia es un enfoque prometedor para relacionar arquetipos con SNOMED CT en combinación con otras técnicas.
- Hallazgo de una relación existente entre el contexto de un arquetipo y el contexto de SNOMED CT. Un arquetipo está constituido por agrupaciones de términos unos dentro de otros. Pues bien, en el caso en el que el contenido del arquetipo sea una observación específica, no un compendio de varias con diferente origen, los términos *ELEMENT* estarán relacionados entre sí mediante una relación jerárquica *is a* de la misma forma con la que se relacionan entre sí los conceptos de SNOMED CT con los que se anotan. Por otro lado, la agrupación de términos *VALUE* por parte de un *ELEMENT* en la mayoría de las ocasiones coincide en la relación lógica de *interprets*, siendo los conceptos de SNOMED CT participantes en la relación con los que se anota el arquetipo.
- Aumento del *recall*, utilizando la información de contexto de las dependencias entre términos del arquetipo para aumentar el de las técnicas léxicas. En concreto, nuestro método utiliza el contexto estructural de los términos dentro de la jerarquía arquetipo.

Con respecto a la anotación semántica de guías de práctica clínica podemos decir que aporta lo siguiente:

– Una combinación de diferentes herramientas de PLN que sirve como ejemplo y método prometedor para codificar automáticamente las recomendaciones de diagnóstico y tratamiento mediante asignaciones a una terminología estandarizada.

– Propuesta de un método basado en la tecnología PLN actual para la adquisición automática de conocimiento de los procedimientos diagnósticos y terapéuticos de las GPC.

– Propuesta de un método para la extracción del contexto sobre el estado del paciente reflejado en las GPC y que complementa a la información diagnóstica y terapéutica.

6.3. Limitaciones de Nuestro Trabajo

Los aspectos más críticos de nuestro trabajo se derivan de la aproximaciones léxicas y terminológicas, que ya habían sido detectadas por otros trabajos previos al nuestro, tal y como hemos comentado en el capítulo 4. Además, las técnicas estructurales diseñadas limitan los resultados en el caso de realizar anotaciones sobre arquetipos que están compuestos por un agrupamiento de observaciones de diferente ámbito. Estas condiciones no nos permiten sacar el partido esperado a las técnicas estructurales. Además, muchos arquetipos incluyen un apartado de protocolo en el que se especifica las condiciones bajo las que se realizan las observaciones al paciente. El protocolo incluye información relacionada con instrumentación, posición del paciente, ... En este caso, las técnicas estructurales tampoco responderían según lo esperado cuando se aplican al árbol jerárquico del arquetipo.

Un factor limitante al abanico que abarca el enfoque centrado en la anotación de GPC es que sólo se aplica la extracción de conocimiento descriptivo en los procedimientos de diagnóstico y terapéuticos. Esta es una parte importante del conocimiento de las GPC, pero otro contenido importante es el conjunto de acciones a llevar a cabo en una situación particular. Por ejemplo, las acciones para el tratamiento de ICC son evaluar la gravedad de los síntomas, determinar la etiología de la ICC, ... La extracción y representación de este tipo de acciones es particularmente difícil con los recursos disponibles, ya que muchas acciones médicas no están incluidas en estos recursos o se clasifican en categorías semánticas que no son adecuadas para nuestros propósitos. Esto lleva a una adquisición automatizada con muchos errores. Además, este tipo de conocimiento a menudo incluye estructuras de control, tales como la separación o secuenciación de acciones, cuya extracción es necesaria que se realice al mismo tiempo.

6.4. Trabajo Futuro

Enriquecer el contexto sobre el que se realizan las asociaciones entre los arquetipos y SNOMED CT mejoraría el *recall* y proporcionaría conceptos más específicos de SNOMED CT que favorecerían las técnicas lógicas y la validación. En el futuro, se planea probar nuestra metodología con la información de contexto de arquetipo utilizada por otros enfoques [85, 128] (es decir, términos de un nivel superior) y las palabras clave de la sección *header* del arquetipo.

Por otra parte, los resultados de este método han demostrado que existe una importante dependencia jerárquica entre los términos *ELEMENT* y los términos *VALUE*. En particular, se ha detectado que el número de *ELEMENTS* que hacen referencia a partes del cuerpo asociadas con *VALUEs* a través de la relación *finding site* en SNOMED CT es notable. Por otra parte, la relación *finding site* tiene una cobertura más completa en SNOMED CT que *interprets*. Por lo tanto, este método podría mejorar al utilizar otras relaciones de definición de SNOMED CT y otros tipos de arquetipos, tales como *ACTIONs*, *EVALUATIONs* e *INSTRUCTIONs*.

Con respecto a la anotación de GPC, trabajos futuros explorarán otras alternativas para extraer el conocimiento relacionado con el conjunto de acciones a llevar a cabo en una situación particular así como la secuenciación de estas acciones. Estas alternativas consistirán, por ejemplo, en el uso de estructuras de control o patrones lingüísticos ([146, 125]).

Otro objetivo de cara al futuro es tener en cuenta otras formas de adaptación de SemRep que incluyan las ventajas del método descrito aquí: la detección de los conceptos relevantes y la extracción de contexto paciente antes de la extracción de los predicados.

Por otra parte, el éxito de las GPC electrónicas no sólo depende de la calidad del conocimiento utilizado para representar y ejecutar las recomendaciones terapéuticas y de diagnóstico, sino también del grado de interoperabilidad con la historia clínica del paciente. La intención de algunas terminologías, como SNOMED CT, y de los arquetipos es facilitar el intercambio de las historias clínicas. Una investigación futura también analizará terminologías y arquetipos como mediadores para integrar GPC electrónicas y registros de pacientes, quizás mediante la construcción de una ontología intermedia que represente el conocimiento completo de la GPC.

Por último, la evaluación de nuestro método en diferentes textos clínicos puede ayudarnos a determinar el ámbito de aplicación del método. También se probará en el futuro la capacidad de adaptación del método, con el fin de extraer información fenotípica de notas clínicas.

APÉNDICE A

FUENTES DE INFORMACIÓN

Este apéndice contiene comparativas en las formas de almacenamiento de información clínica, así como, ejemplos de cómo almacenan el conocimiento clínico diferentes fuentes de información

A.1. Fuentes no estructuradas

La figura A.1 muestra una comparación entre un historial clínico electrónico (HCE) y un historial clínico (HC) tradicional.

A.2. Fuentes semiestructuradas

Ejemplo de un arquetipo CEN 13606 referente la observación de presión sanguínea[1].

A.2.1. Arquetipos OpenEHR

Puesto que nuestro trabajo se basa en OpenEHR referenciamos aquí los arquetipos utilizados ubicados en el sitio habilitado por nuestro grupo de investigación [114]. Los cuales, además de proporcionarnos un ejemplo de modelo de arquetipos OpenEHR, permiten acceder libremente a la anotación manual realizada y servir de *gold estándar* para otros trabajos, tal como el de Berges et al. [12].

[1] `https://github.com/openEHR/adl-archetypes/blob/master/Reference/ISO_13606/Spanish_MOH/ADL_14/CEN-EN13606-ENTRY.PresionSanguinea.v1.adl`

HISTORIA CLÍNICA ELECTRÓNICA	HISTORIA CLÍNICA TRADICIONAL
Inviolabilidad	
Por medio de firma digital, inserción de hora y fecha automática y técnicas de Back up adecuadas	Puede llegar a rehacerse total o parcialmente sin poder comprobarlo
Secuencialidad de la información	
Garantizada por mecanismos de campos auto-numéricos e inserción de hora y fecha automática	Es más difícil preservarla si no está previamente foliada
Reserva de la información privada del paciente	
Garantizada por mecanismos de seguridad informáticos	Garantizada por mecanismos de control del archivo
Accesibilidad	
En todo momento vía *internet, wireless y wap*	Utilizable en un solo lugar
Disponibilidad	
Personal habilitado debe poder acceder a toda o parte de la información que se requiera	Dependiendo de la accesibilidad a los Archivos físicos
Riesgo de pérdida de información	
Seguridad garantizada con una correcta política de resguardo de la información (back-up)	Frecuentemente extraviada, posibilidad de microfilmarse
Integridad de la información clínica	
La informatización racional garantiza que la información de un paciente no esté atomizada	La pérdida o extravío origina que se abra otra historia clínica para un mismo paciente.
Durabilidad	
Inalterable en el tiempo para su consulta	Sufre deterioro con el tiempo, por su propio uso
Legibilidad	
Siempre legible	Algunas veces ilegible
Legalidad y valor probatorio	
Garantizado por la firma digital y el inserción de hora y fecha automática	Garantizado sí está bien confeccionada, clara, foliada y completa
Identificación del profesional	
Por la firma digital	Por la firma holográfica y el sello con la matrícula
Temporalidad precisa	
Garantizada con la inserción de hora y fecha automática del servidor local	A veces con fecha y hora
Garantía de la autoría	
Identifica en forma inequívoca a quien generó la información mediante la firma digital	Por medio de la firma manual y sello que a veces suele faltar
Redundancia	
Posibilidad reducida: información duplicada	Posibilidad alta: información duplicada
Errores de consignación	
Menor número de errores	A veces inexacta
Estandarización de datos	
Ingreso estandarizado de datos	Organizada según necesidad de cada servicio
Costos de personal administrativo	
Operada por los mismos profesionales	Requiere personal para el mantenimiento
Costos de papel	
Bajo, sólo cuando se requiera imprimirla	Alto
Tiempo de Consulta y de búsqueda de estudios complementarios	
Más corto	Más largo
Orientaciones en la terapéutica, recordatorios y alertas	
Puede incorporar alertas y reglas informatizadas	
Disponibilidad de los datos para estadísticas	
Inmediata	Mediante tediosos procesos
Búsqueda de información y separación de datos por distintos ítem	
Fácil y accesible	Dificultosa, poco confiable y costosa
Robo de la historia clínica	
Imposible si hay una política de seguridad informática. Si se perdiese se recupera del backup	Si se roba o se pierde es imposible de recuperarla

Figura A.1: Comparación detallada de una HCE y de una HC tradicional

A.2.2. Arquetipo CEN 13606

```
archetype (adl_version=1.4)
    CEN-EN13606-ENTRY.PresionSanguinea.v1

concept
    [at0000]

language
    original_language = <[ISO_639-1::es]>

description
    original_author = <
        ["email"] = <"jamaldo@upv.es">
        ["name"] = <"Grupo de Informática Médica (IBIME)">
        ["organisation"] = <"Universitat Politècnica de Valencia">
        ["date"] = <"20131108">
    >
    lifecycle_state = <"Draft">
    other_contributors = <"Arturo Romero, Ministerio de Sanidad, Servicios Sociales
       e Igualdad","Pablo Serrano, Hospital de Fuenlabrada">
    details = <
        ["es"] = <
            language = <[ISO_639-1::es]>
            keywords = <"CMDIC">
        >
    >

definition
    ENTRY[at0000] occurrences matches {1..1} matches {  -- Presión sanguínea
        items existence matches {0..1} cardinality matches {0..*; ordered; unique}
                                                                        matches{
            CLUSTER[at0008] occurrences matches {0..1} matches {
                    -- Medida de presión sanguínea
                parts existence matches {0..1} cardinality
                        matches {2..4; ordered; unique} matches {
                    ELEMENT[at0001] occurrences matches {1..1} matches {  --Sistólica
                        value existence matches {1..1} matches {
                            PQ[at0005] occurrences matches {1..1} matches {
                                    -- Medida sistólica
                                units existence matches {1..1} matches {
                                    CS[at0009] occurrences matches {0..1} matches {--CS
                                        codeValue existence matches {0..1}
                                            matches {"mm[Hg]"}
                                        codingSchemeName existence matches {0..1}
                                            matches {"UCUM"}
                                    }
                                }
                                value existence matches {1..1} matches {|0.0..<1000.0|}
                            }
```

```
            }
        }
        ELEMENT[at0002] occurrences matches {1..1} matches {  --Diastólica
            value existence matches {1..1} matches {
                PQ[at0006] occurrences matches {1..1} matches {
                       -- Medida diastólica
                    units existence matches {1..1} matches {
                        CS[at0010] occurrences matches {0..1} matches {--CS
                            codeValue existence matches {0..1}
                               matches {"mm[Hg]"}
                            codingSchemeName existence matches {0..1}
                               matches {"UCUM"}
                        }
                    }
                    value existence matches {1..1} matches {|0.0..<1000.0|}
                }
            }
        }
        ELEMENT[at0003] occurrences matches {0..1} matches {
               -- Presión arterial media
            value existence matches {0..1} matches {
                PQ[at0007] occurrences matches {1..1} matches {
                       -- Media de medidas
                    units existence matches {1..1} matches {
                        CS[at0011] occurrences matches {0..1}
                                                          matches {--CS
                            codeValue existence matches {0..1}
                               matches {"mm[Hg]"}
                            codingSchemeName existence matches {0..1}
                               matches {"UCUM"}
                        }
                    }
                    value existence matches {1..1} matches {|0.0..750.0|}
                }
            }
        }
        ELEMENT[at0004] occurrences matches {0..1} matches {  --Posición
            value existence matches {0..1} matches {
                SIMPLE_TEXT[at0012] occurrences matches {0..1} matches {
                       -- Posición
                    originalText existence matches {0..1}
                       matches {"Standing","Sitting","Reclining","Lying"}
                }
            }
        }
    }
    structure_type existence matches {1..1} matches {
        CS[at0014] occurrences matches {1..1} matches {  --
            codeValue existence matches {0..1} matches {"STRC01"}
            codingSchemeName existence matches {0..1}
```

```
                        matches {"CEN/TC251/EN13606-3:STRUCTURE_TYPE"}
                    }
                }
            }
        }
    }

ontology
    term_definitions = <
        ["es"] = <
            items = <
                ["at0000"] = <
                    text = <"Presión sanguínea">
                    description = <"La medida por cualquier método (invasivo o no
                                    invasivo) de la presión arterial sistémica que
                                    se considera que representa la presion sanguínea
                                    sistémica real">
                >
                ["at0001"] = <
                    text = <"Sistólica">
                    description = <"Objeto de tipo ELEMENT">
                >
                ["at0002"] = <
                    text = <"Diastólica">
                    description = <"Objeto de tipo ELEMENT">
                >
                ["at0003"] = <
                    text = <"Presión arterial media">
                    description = <"Objeto de tipo ELEMENT">
                >
                ["at0004"] = <
                    text = <"Posición">
                    description = <"Objeto de tipo ELEMENT">
                >
                ["at0005"] = <
                    text = <"Medida sistólica">
                    description = <"Objeto de tipo PQ">
                >
                ["at0006"] = <
                    text = <"Medida diastólica">
                    description = <"Objeto de tipo PQ">
                >
                ["at0007"] = <
                    text = <"Media de medidas">
                    description = <"Objeto de tipo PQ">
                >
                ["at0009"] = <
                    text = <"CS">
                    description = <"Objeto de tipo CS">
                >
```

```
            ["at0010"] = <
                text = <"CS">
                description = <"Objeto de tipo CS">
            >
            ["at0011"] = <
                text = <"CS">
                description = <"Objeto de tipo CS">
            >
            ["at0012"] = <
                text = <"Posición">
                description = <"Objeto de tipo SIMPLE_TEXT">
            >
            ["at0008"] = <
                text = <"Medida de presión sanguínea">
                description = <"Objeto de tipo CLUSTER">
            >
        >
    >
>
constraint_definitions = <
    ["es"] = <
        items = <
        >
    >
>
term_binding = <
>
constraint_binding = <
>
```

A.3. Fuentes estructuradas

A.3.1. Estándar ISO-25964

Veamos un ejemplo[2] del estándar en el que se define un tesauro **iso25964:Thesaurus**. Cada entrada del tesauro comienza con la etiqueta **iso25964:ThesaurusConcept** que posee un identificador, sus traducciones a diferentes idiomas mediante etiquetas léxicas e información extra. Una vez definidos los conceptos, es posible la definición de relaciones entre los conceptos haciendo uso del identificador. Estas relaciones pueden ser jerárquica (de generalización en este caso a través de la etiqueta **iso25964:HierarchicalRelationship**), permite componer términos más complejos a partir de conceptos sencillos definidos en la ontolo-

[2] http://www.niso.org/schemas/iso25964/example_multi_lingual_10-10T06-29.xml

gía (**iso25964:CompoundEquivalence**) y construir conceptos *virtuales* mediante la etiqueta **iso25964:SplitNonPreferredTerm**

```
<iso25964:ISO25964Interchange
     xsi:schemaLocation="http://iso25964.org/iso25964-1_v1.4.xsd">
   <iso25964:Thesaurus>
      <iso25964:identifier>iso25964Serial1</iso25964:identifier>
      <dc:coverage>worldwide</dc:coverage>
      <dc:creator>ISOTC46/SC9/WG8</dc:creator>
      <iso25964:date>2011-08-08</iso25964:date>
      <iso25964:created>2011-07-31</iso25964:created>
      <iso25964:modified>2011-08-08</iso259 64:modified>
      <dc:description xml:lang="en">
           a thesaurus illustrating how to serialize diverse features
           of the ISO 25964 data model
      </dc:description>
      <dc:format>text/xml</dc:format>
      <dc:language>en</dc:language>
      <dc:language>fr</dc:language>
      <dc:language>de</dc:language>
      <dc:language>es</dc:language>
      <dc:publisher>NISO</dc:publisher>
      <dc:rights>tbd</dc:rights>
      <dc:source>ISO 25964-1 :Thesauri</dc:source>
      <dc:title>
         Serialization Example 1 illustrating diverse features
      </dc:title>
      ...
      <iso25964:ThesaurusConcept>
         <iso25964:identifier>C9</iso25964:identifier>
         <iso25964:PreferredTerm>
            <iso25964:lexicalValue xml:lang="en">
               coal
            </iso25964:lexicalValue>
            <iso25964:identifier>L9</iso25964:identifier>
         </iso25964:PreferredTerm>
         <iso25964:PreferredTerm>
            <iso25964:lexicalValue xml:lang="fr">
               charbon
            </iso25964:lexicalValue>
            <iso25964:identifier>L9.fr</iso25964:identifier>
         </iso25964:PreferredTerm>
         <iso25964:CustomNote>
            <iso25964:lexicalValue xml:lang="fr">
               section 8.5
            </iso25964:lexicalValue>
         </iso25964:CustomNote>
         <iso25964:CustomNote>
            <iso25964:lexicalValue xml:lang="en">
               clause8.5
```

```
        </iso25964:lexicalValue>
    </iso25964:CustomNote>
</iso25964:ThesaurusConcept>
<iso25964:ThesaurusConcept>
    <iso25964:identifier>C10</iso25964:identifier>
    <iso25964:PreferredTerm>
        <iso25964:lexicalValuexml:lang="en">
            mining
        </iso25964:lexicalValue>
        <iso25964:identifier>L10</iso25964:identifier>
    </iso25964:PreferredTerm>
    <iso25964:PreferredTerm>
        <iso25964:lexicalValue xml:lang="fr">
            exploitation mini\‘{\e}re
        </iso25964:lexicalValue>
        <iso25964:identifier>L10.fr</iso25964:identifier>
    </iso25964:PreferredTerm>
    <iso25964:CustomNote>
        <iso25964:lexicalValuexml:lang="fr">
            section 8.5
        </iso25964:lexicalValue>
    </iso25964:CustomNote>
    <iso25964:CustomNote>
        <iso25964:lexicalValue xml:lang="en">
            clause 8.5
        </iso25964:lexicalValue>
    </iso25964:CustomNote>
</iso25964:ThesaurusConcept>
...
<iso25964:ThesaurusConcept>
    <iso25964:identifier>C16</iso25964:identifier>
    <iso25964:PreferredTerm>
        <iso25964:lexicalValue xml:lang="en">
            houses
        </iso25964:lexicalValue>
        <iso25964:identifier>L19.en</iso25964:identifier>
    </iso25964:PreferredTerm>
    <iso25964:PreferredTerm>
        <iso25964:lexicalValue xml:lang="fr">
            maison
        </iso25964:lexicalValue>
        <iso25964:identifier>L19.fr</iso25964:identifier>
    </iso25964:PreferredTerm>
    <iso25964:PreferredTerm>
        <iso25964:lexicalValue xml:lang="de">
            Haus
        </iso25964:lexicalValue>
        <iso25964:identifier>L19.de</iso25964:identifier>
    </iso25964:PreferredTerm>
    <iso25964:PreferredTerm>
```

```
            <iso25964:lexicalValue xml:lang="es">
               casas
            </iso25964:lexicalValue>
            <iso25964:identifier>L19.es</iso25964:identifier>
         </iso25964:PreferredTerm>
         <iso25964:CustomNote>
            <iso25964:lexicalValue xml:lang="en">
               clause 6.5.1
            </iso25964:lexicalValue>
         </iso25964:CustomNote>
      </iso25964:ThesaurusConcept>
      ...
      <iso25964:HierarchicalRelationship>
         <iso25964:role>BT</iso25964:role>
         <iso25964:hasHierRelConcept>C13</iso25964:hasHierRelConcept>
         <iso25964:isHierRel Concept>C12</iso25964:isHierRelConcept>
      </iso25964:HierarchicalRelationship>
      <iso25964:CompoundEquivalence>
         <iso25964:UFPlus>SL1</iso25964:UFPlus>
         <iso25964:USEPlus>L9</iso25964:USEPlus>
         <iso25964:USEPlus>L10</iso25964:USEPlus>
      </iso25964:CompoundEquivalence>
      <iso25964:SplitNonPreferredTerm>
         <iso25964:lexicalValue xml:lang="en">
            coal mining
         </iso25964:lexicalValue>
         <iso25964:identifier>SL1</iso25964:identifier>
      </iso25964:SplitNonPreferredTerm>
</iso25964:ISO25964Interchange>
```

A.3.2. SKOS

Extracto de información de MeSH[3] relacionado con la presión arterial y codificado mediante SKOS. Cada etiqueta **rdf:Description** especifica un concepto. Y para cada concepto especifica el término preferido **skos:prefLabel**, otras etiquetas **skos:altLabel**, **skos:hiddenLabel**. Las etiquetas **skos:annotation** y **skos:scopenote** permiten añadir información extra del concepto. También refleja cambios importantes en el concepto mediante la etiqueta **skos:-historyNote**. La etiqueta **skos:related** especifica relaciones jerárquicas entre conceptos, por ejemplo el concepto *Pressure, Blood* sería padre de los conceptos *Hypertension* con identificador D006973 y el concepto *Hypotension* con identificador D007022. Podemos ver que el ejemplo contiene no solo el espacio de nombres de SKOS, si no también de RDF del que está basado y de MESH.

[3] http://thesauri.cs.vu.nl/eswc06/mesh/rdf/meshdata.rdf

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE rdf:RDF [
    <!ENTITY mesh 'http://www.nlm.nih.gov/mesh/2006#'>
    <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
    <!ENTITY skos 'http://www.w3.org/2004/02/skos/core#'>
]>

<rdf:RDF
    xmlns:mesh="&mesh;"
    xmlns:rdf="&rdf;"
    xmlns:skos="&skos;"
    xml:lang="en">

<rdf:Description rdf:about="&mesh;D001794"
    mesh:dateCreated="1999-01-01"
    mesh:dateRevised="2005-08-01"
    mesh:recordAuthorizer="ags"
    mesh:recordMaintainer="ags"
    mesh:recordOriginator="NLM"
    skos:prefLabel="Blood␣Pressure">
  <mesh:activeMeSHYear>2004</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2005</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2006</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2006A</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2006B</mesh:activeMeSHYear>
  <skos:hiddenLabel>Pressure, Blood</skos:hiddenLabel>
  <skos:hiddenLabel>Pressure, Diastolic</skos:hiddenLabel>
  <skos:hiddenLabel>Pressure, Pulse</skos:hiddenLabel>
  <skos:hiddenLabel>Pressure, Systolic</skos:hiddenLabel>
  <skos:hiddenLabel>Pressures, Systolic</skos:hiddenLabel>
  <skos:related rdf:resource="&mesh;D006973"/>
  <skos:related rdf:resource="&mesh;D007022"/>
  <skos:annotation>GEN; note specifics; "arterial" pressure = BLOOD
      PRESSURE and not also ARTERIES unless a specific artery;
      pressure within a specific vessel: coord vessel /physiol(IM)
      + BLOOD PRESSURE (NIM); do not add SYSTOLE; DIASTOLE; or
      PULSE unless particularly discussed; with diseases coord IM
      with disease /physiopathol (IM), not /blood (IM):
      Manual 23.28; blood pressure vs HYPERTENSION &amp;
      HYPOTENSION: Manual 23.27+
  </skos:annotation>
  <skos:scopeNote>PRESSURE of the BLOOD on the ARTERIES and other
      BLOOD VESSELS.
  </skos:scopeNote>
</rdf:Description>

<rdf:Description rdf:about="&mesh;D006973"
    mesh:dateCreated="1999-01-01"
    mesh:dateRevised="2004-07-07"
    mesh:recordAuthorizer="sjn"
```

```
    mesh:recordMaintainer="lkt"
    mesh:recordOriginator="NLM"
    skos:altLabel="Blood␣Pressure,␣High"
    skos:prefLabel="Hypertension">
  <mesh:activeMeSHYear>2005</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2006</mesh:activeMeSHYear>
  <skos:hiddenLabel>Blood Pressures, High</skos:hiddenLabel>
  <skos:hiddenLabel>High Blood Pressure</skos:hiddenLabel>
  <skos:hiddenLabel>High Blood Pressures</skos:hiddenLabel>
  <skos:related rdf:resource="&mesh;D000959"/>
  <skos:related rdf:resource="&mesh;D014655"/>
  <skos:annotation>not for intracranial or intraocular pressure;
      relation to BLOOD PRESSURE: Manuel23.27; Goldblatt kidney
      or Goldblatt hypertension is HYPERTENSION, GOLDBLATT see
      HYPERTENSION, RENOVASCULAR; hypertension with kidney
      disease is probably HYPERTENSION, RENAL, not HYPERTENSION;
      venous hypertension: index under VENOUS PRESSURE (IM) &amp;
      do not coordinate with HYPERTENSION
  </skos:annotation>
  <skos:scopeNote>Persistently high systemic arterial BLOOD
      PRESSURE. Based on multiple readings (BLOOD PRESSURE
      DETERMINATION), hypertension is currently defined as when
      SYSTOLIC PRESSURE is consistently greater than 140 mm Hg or
      when DIASTOLIC PRESSURE is consistently 90 mm Hg or more.
    </skos:scopeNote>
</rdf:Description>

<rdf:Description rdf:about="&mesh;D007022"
    mesh:dateCreated="1999-01-01"
    mesh:dateRevised="1997-06-20"
    mesh:recordAuthorizer="SJN"
    mesh:recordMaintainer="TGC"
    mesh:recordOriginator="NLM"
    skos:altLabel="Blood␣Pressure,␣Low"
    skos:prefLabel="Hypotension">
  <mesh:activeMeSHYear>1998</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>1999</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2000</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2001</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2002</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2003</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2004</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2005</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2006</mesh:activeMeSHYear>
  <skos:hiddenLabel>Blood Pressures, Low</skos:hiddenLabel>
  <skos:hiddenLabel>Hypotensions</skos:hiddenLabel>
  <skos:hiddenLabel>Low Blood Pressure</skos:hiddenLabel>
  <skos:hiddenLabel>Low Blood Pressures</skos:hiddenLabel>
  <skos:annotation>only blood pressure; not for intracranial pressure;
      relation to BLOOD PRESSURE: Manual 23.27+
```

```
  </skos:annotation>
  <skos:related rdf:resource="&mesh;D019462"/>
  <skos:scopeNote>Abnormally low blood pressure seen in shock but
       not necessarily indicative of it. (Dorland, 28th ed)
  </skos:scopeNote>
</rdf:Description>

<rdf:Description rdf:about="&mesh;D014690"
     mesh:dateCreated="1999-01-01"
     mesh:dateEstablished="1970-01-01"
     mesh:dateRevised="2003-07-09"
     mesh:historyNote="70"
     mesh:publicMeSHNote="70"
     mesh:recordAuthorizer="sjn"
     mesh:recordMaintainer="nns"
     mesh:recordOriginator="NLM"
     skos:altLabel="Blood␣Pressure,␣Venous"
     skos:prefLabel="Venous␣Pressure">
  <mesh:activeMeSHYear>2004</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2005</mesh:activeMeSHYear>
  <mesh:activeMeSHYear>2006</mesh:activeMeSHYear>
  <skos:hiddenLabel>Blood Pressures, Venous</skos:hiddenLabel>
  <skos:hiddenLabel>Pressure, Venous</skos:hiddenLabel>
  <skos:hiddenLabel>Pressure, Venous Blood</skos:hiddenLabel>
  <skos:hiddenLabel>Pressures, Venous</skos:hiddenLabel>
  <skos:hiddenLabel>Pressures, Venous Blood</skos:hiddenLabel>
  <skos:hiddenLabel>Venous Blood Pressure</skos:hiddenLabel>
  <skos:hiddenLabel>Venous Blood Pressures</skos:hiddenLabel>
  <skos:hiddenLabel>Venous Pressures</skos:hiddenLabel>
  <skos:historyNote>Blood Pressure (1966-1969)</skos:historyNote>
  <skos:historyNote>Veins (1966-1969)</skos:historyNote>
  <skos:annotation>IM GEN only; NIM for pressure within a specific
       vessel; CENTRAL VENOUS PRESSURE &amp;amp; PORTAL PRESSURE are also
       available; venous hypertension: index under VENOUS PRESSURE (IM)
       &amp;amp; do not coord with HYPERTENSION
  </skos:annotation>
  <skos:scopeNote>The blood pressure in the VEINS. It is usually
       measured to assess the filling PRESSURE to the HEART VENTRICLE.
  </skos:scopeNote>
</rdf:Description>

</rdf:RDF>
```

A.3.3. OWL

Extracto de la terminología de SNOMED CT en formato OWL para conceptos relacionados con la presión sanguínea. En este ejemplo, la etiqueta **owl:ObjectProperty** define las

relaciones de SNOMED CT. Con **owl:Class** especificamos los conceptos y de quién descienden con **rdfs:subClassOf**. Para reflejar que el concepto que se define tiene una relación descriptiva con otro concepto, se utilizan las etiquetas: **owl:Restriction** para comenzar la descripción de la relación del concepto con otro, **owl:onProperty** para establecer la relación que ya fue previamente definida y **owl:someValuesFrom** con quién está relacionado.

```
<?xml version="1.0"?>

<!DOCTYPE rdf:RDF [
    <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
    <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
    <!ENTITY owl2xml "http://www.w3.org/2006/12/owl2-xml#" >
    <!ENTITY snomedct "http://www.ihtsdo.org/snomedct.owl#" >
    <!ENTITY Finding_of_blood
        "http://www.ihtsdo.org/snomedct.owl#Finding_of_blood," >
       ....

<rdf:RDF
       xmlns="http://comlab.ox.ac.uk/modules/SNOMED_module_mappings_nci.owl#"
   xml:base="http://comlab.ox.ac.uk/modules/SNOMED_module_mappings_nci.owl"
   xmlns:Host_defense="&snomedct;Host_defense,"
   xmlns:Localized_bone_cyst="&snomedct;Localized_bone_cyst,"
   xmlns:Closed_fracture_of_humerus="&snomedct;Closed_fracture_of_humerus,"
   xmlns:Finding_of_blood="&snomedct;Finding_of_blood,"
      .....

    <!-- http://www.ihtsdo.org/snomedct.owl#Finding_site -->
    <owl:ObjectProperty rdf:about="&snomedct;Finding_site">
      <rdfs:label xml:lang="en"
        >Finding site (attribute)</rdfs:label>
    </owl:ObjectProperty>

    <!-- http://www.ihtsdo.org/snomedct.owl#Has_definitional_manifestation -->
    <owl:ObjectProperty rdf:about="&snomedct;Has_definitional_manifestation">
      <rdfs:label xml:lang="en"
        >Has definitional manifestation (attribute)</rdfs:label>
    </owl:ObjectProperty>

    <!-- http://www.ihtsdo.org/snomedct.owl#Has_interpretation -->
    <owl:ObjectProperty rdf:about="&snomedct;Has_interpretation">
      <rdfs:label xml:lang="en"
        >Has interpretation (attribute)</rdfs:label>
    </owl:ObjectProperty>

    <!-- http://www.ihtsdo.org/snomedct.owl#Interprets -->
    <owl:ObjectProperty rdf:about="&snomedct;Interprets">
      <rdfs:label xml:lang="en"
        >Interprets (attribute)</rdfs:label>
```

```
</owl:ObjectProperty>

<!-- http://www.ihtsdo.org/snomedct.owl#Occurrence -->
<owl:ObjectProperty rdf:about="&snomedct;Occurrence">
  <rdfs:label xml:lang="en"
    >Occurrence (attribute)</rdfs:label>
</owl:ObjectProperty>

<!-- http://www.ihtsdo.org/snomedct.owl#RoleGroup -->
<owl:ObjectProperty rdf:about="&snomedct;RoleGroup">
  <rdfs:label xml:lang="en">RoleGroup</rdfs:label>
</owl:ObjectProperty>

<!-- http://www.ihtsdo.org/snomedct.owl#Blood_pressure -->
<owl:Class rdf:about="&snomedct;Blood_pressure">
  <rdfs:label xml:lang="en"
    >Blood pressure (observable entity)</rdfs:label>
  <rdfs:subClassOf rdf:resource="&snomedct;Fluid_pressure"/>
  <rdfs:subClassOf rdf:resource="&snomedct;Vascular_measure"/>
  <rdfs:subClassOf rdf:resource="&snomedct;Vital_sign"/>
</owl:Class>

<!-- http://www.ihtsdo.org/snomedct.owl#Diastolic_blood_pressure
 -->
<owl:Class rdf:about="&snomedct;Diastolic_blood_pressure">
  <rdfs:label xml:lang="en"
    >Diastolic blood pressure (observable entity)</rdfs:label>
  <rdfs:subClassOf rdf:resource="&snomedct;Blood_pressure"/>
</owl:Class>

<!-- http://www.ihtsdo.org/snomedct.owl#Systolic_blood_pressure
 -->
<owl:Class rdf:about="&snomedct;Systolic_blood_pressure">
  <rdfs:label xml:lang="en"
    >Systolic blood pressure (observable entity)</rdfs:label>
  <rdfs:subClassOf rdf:resource="&snomedct;Blood_pressure"/>
</owl:Class>

<!-- http://www.ihtsdo.org/snomedct.owl#Blood_pressure_finding
 -->
<owl:Class rdf:about="&snomedct;Blood_pressure_finding">
  <rdfs:label xml:lang="en"
    >Blood pressure finding (finding)</rdfs:label>
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <rdf:Description
            rdf:about="&snomedct;Cardiovascular_pressure_AND/OR_pulse_finding"/>
        <rdf:Description
            rdf:about="&snomedct;Finding_of_cardiovascular_measurement"/>
```

```
        <rdf:Description rdf:about="&snomedct;Vital_signs_finding"/>
        <owl:Restriction>
            <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
            <owl:someValuesFrom>
              <owl:Restriction>
                <owl:onProperty rdf:resource="&snomedct;Finding_site"/>
                <owl:someValuesFrom
                    rdf:resource="&snomedct;Structure_of_cardiovascular_system"/>
              </owl:Restriction>
            </owl:someValuesFrom>
        </owl:Restriction>
        <owl:Restriction>
            <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
            <owl:someValuesFrom>
              <owl:Restriction>
                <owl:onProperty rdf:resource="&snomedct;Interprets"/>
                <owl:someValuesFrom
                    rdf:resource="&snomedct;Blood_pressure"/>
              </owl:Restriction>
            </owl:someValuesFrom>
        </owl:Restriction>
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</owl:Class>

<!-- http://www.ihtsdo.org/snomedct.owl#Endocrine_hypertension -->
<owl:Class rdf:about="&snomedct;Endocrine_hypertension">
  <rdfs:label xml:lang="en"
    >Endocrine hypertension (disorder)</rdfs:label>
  <rdfs:subClassOf rdf:resource="&snomedct;Secondary_hypertension"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
      <owl:someValuesFrom>
        <owl:Restriction>
            <owl:onProperty rdf:resource="&snomedct;Finding_site"/>
            <owl:someValuesFrom
                rdf:resource="&snomedct;Systemic_arterial_structure"/>
        </owl:Restriction>
      </owl:someValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
      <owl:someValuesFrom>
        <owl:Restriction>
            <owl:onProperty
                rdf:resource="&snomedct;Has_definitional_manifestation"/>
```

```
            <owl:someValuesFrom
                 rdf:resource="&snomedct;Finding_of_increased_blood_pressure"/>
          </owl:Restriction>
        </owl:someValuesFrom>
      </owl:Restriction>
    </rdfs:subClassOf>
  </owl:Class>

  <!-- http://www.ihtsdo.org/snomedct.owl#Finding_of_increased_blood_pressure -->
  <owl:Class
      rdf:about="&snomedct;Finding_of_increased_blood_pressure">
    <rdfs:label xml:lang="en"
      >Finding of increased blood pressure (finding)</rdfs:label>
    <owl:equivalentClass>
      <owl:Class>
        <owl:intersectionOf rdf:parseType="Collection">
          <rdf:Description
              rdf:about="&snomedct;Abnormal_blood_pressure"/>
          <owl:Restriction>
              <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
              <owl:someValuesFrom>
                <owl:Class>
                  <owl:intersectionOf rdf:parseType="Collection">
                    <owl:Restriction>
                      <owl:onProperty
                          rdf:resource="&snomedct;Has_interpretation"/>
                      <owl:someValuesFrom
                          rdf:resource="&snomedct;Abnormal"/>
                    </owl:Restriction>
                    <owl:Restriction>
                      <owl:onProperty
                          rdf:resource="&snomedct;Interprets"/>
                      <owl:someValuesFrom
                          rdf:resource="&snomedct;Blood_pressure"/>
                    </owl:Restriction>
                  </owl:intersectionOf>
                </owl:Class>
              </owl:someValuesFrom>
          </owl:Restriction>
          <owl:Restriction>
              <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
              <owl:someValuesFrom>
                <owl:Class>
                  <owl:intersectionOf rdf:parseType="Collection">
                    <owl:Restriction>
                      <owl:onProperty
                          rdf:resource="&snomedct;Has_interpretation"/>
                      <owl:someValuesFrom
                          rdf:resource="&snomedct;Increased"/>
                    </owl:Restriction>
```

```
                    <owl:Restriction>
                      <owl:onProperty
                          rdf:resource="&snomedct;Interprets"/>
                      <owl:someValuesFrom
                          rdf:resource="&snomedct;Blood_pressure"/>
                    </owl:Restriction>
                  </owl:intersectionOf>
                </owl:Class>
              </owl:someValuesFrom>
            </owl:Restriction>
            <owl:Restriction>
              <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
              <owl:someValuesFrom>
                <owl:Restriction>
                  <owl:onProperty
                      rdf:resource="&snomedct;Finding_site"/>
                  <owl:someValuesFrom
                      rdf:resource="&snomedct;Structure_of_cardiovascular_system"/>
                </owl:Restriction>
              </owl:someValuesFrom>
            </owl:Restriction>
          </owl:intersectionOf>
        </owl:Class>
      </owl:equivalentClass>
    </owl:Class>

  <!-- http://www.ihtsdo.org/snomedct.owl#Neonatal_hypertension -->
   <owl:Class rdf:about="&snomedct;Neonatal_hypertension">
     <rdfs:label xml:lang="en"
       >Neonatal hypertension (disorder)</rdfs:label>
     <owl:equivalentClass>
       <owl:Class>
         <owl:intersectionOf rdf:parseType="Collection">
           <rdf:Description
               rdf:about="&snomedct;Hypertensive_disorder,_systemic_arterial"/>
           <rdf:Description
               rdf:about="&snomedct;Neonatal_cardiovascular_disorder"/>
           <owl:Restriction>
               <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
               <owl:someValuesFrom>
                 <owl:Restriction>
                   <owl:onProperty rdf:resource="&snomedct;Finding_site"/>
                   <owl:someValuesFrom
                       rdf:resource="&snomedct;Systemic_arterial_structure"/>
                 </owl:Restriction>
               </owl:someValuesFrom>
           </owl:Restriction>
           <owl:Restriction>
               <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
               <owl:someValuesFrom>
```

```
                <owl:Restriction>
                  <owl:onProperty
                      rdf:resource="&snomedct;Has_definitional_manifestation"/>
                  <owl:someValuesFrom
                      rdf:resource="&snomedct;Finding_of_increased_blood_pressure"/>
                </owl:Restriction>
              </owl:someValuesFrom>
            </owl:Restriction>
            <owl:Restriction>
              <owl:onProperty rdf:resource="&snomedct;RoleGroup"/>
              <owl:someValuesFrom>
                <owl:Restriction>
                  <owl:onProperty rdf:resource="&snomedct;Occurrence"/>
                  <owl:someValuesFrom rdf:resource="&snomedct;Neonatal"/>
                </owl:Restriction>
              </owl:someValuesFrom>
            </owl:Restriction>
          </owl:intersectionOf>
        </owl:Class>
      </owl:equivalentClass>
    </owl:Class>

    <rdf:Description>
      <rdfs:subPropertyOf rdf:resource="&snomedct;Direct_substance"/>
      <owl:propertyChain rdf:parseType="Collection">
        <rdf:Description rdf:about="&snomedct;Direct_substance"/>
        <rdf:Description rdf:about="&snomedct;Has_active_ingredient"/>
      </owl:propertyChain>
    </rdf:Description>
</rdf:RDF>
```

Bibliografía

[1] *CIMI Wiki*. Disponible en: `informatics.mayo.edu/CIMI/index.php/Main_Page` [Último acceso: septiembre 2015], 2013.

[2] *HIMSS Dictionary of Healthcare Information Technology Terms, Acronyms and Organizations, Third Edition*. Healthcare Information and Management Systems Society. 2013.

[3] Allones, J.L.: *Métodos semánticos automatizados de apoyo a la gestión y a la interoperabilidad de la información clínica*. Tesis de Doctorado, Departamento De Electrónica y Computación, Universidad de Santiago de Compostela, 2014.

[4] Apkon, M. y P. Singhaviranon: *Impact of an electronic information system on physician workflow and data collection in the intensive care unit*. Intensive Care Medicine, 27(1):122–130, 2001.

[5] Arano, S.: *La ontología: una zona de interacción entre la Lingüística y la Documentación*. Hipertext.net, 2, 2003.

[6] Aronson, A. R.: *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*. Proceedings / AMIA ... Annual Symposium. AMIA Symposium, páginas 17–21, 2001.

[7] Aronson, A. R.: *MetaMap: Mapping Text to the UMLS Metathesaurus*, 2006.

[8] Aronson, Alan R. y François-Michel Lang: *An overview of MetaMap: historical perspective and recent advances*. JAMIA, 17(3):229–236, 2010.

[9] Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin y Gavin Sherlock: *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. Nature Genetics, 25(1):25–29, Mayo 2000.

[10] Ayatollahi, Haleh, Peter A. Bath y Steve Goodacre: *Paper-based versus computer-based records in the emergency department: Staff preferences, expectations, and concerns*. Health Informatics Journal, 15(3):199–211, 2009.

[11] Beale, Thomas y Sam Heard: *GeHR in Australia - The Good electronic Health Record - openEHR :: future proof and flexible EHR specifications*. Informe técnico, Commonwealth Department of Health and Ageing (DoHA), 2009.

[12] Berges, Idoia, Jesus Bermudez y Arantza Illarramendi: *Binding SNOMED CT Terms to Archetype Elements: Establishing a Baseline of Results*. Methods of Information in Medicine, 54(1):45–49, 2015.

[13] Bhatia, N., N. H. Shah, D. L. Rubin, A. P. Chiang y M. A. Musen: *Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap*. En *Proceedings of the medical informatics association (AMIA) annual symposium*, 2008.

[14] Bodenreider, O.: *The Unified Medical Language System UMLS: integrating biomedical terminology*. Nucleic Acids Research, 32:267–270, Enero 2004.

[15] Bodenreider, O y AT McCray: *Exploring semantic groups through visual approaches*. Journal of Biomedical Informatics, 36(3):414–432, 2003.

[16] Cameron, Delroy, Ramakanth Kavuluru, Thomas C. Rindflesch, Amit P. Sheth, Krishnaprasad Thirunarayan y Olivier Bodenreider: *Context-driven automatic subgraph creation for literature-based discovery*. Journal of Biomedical Informatics, 54:141–157, 2015.

[17] Chapman, W., W. Bridewell, P. Hanbury, G. F. Cooper y B. G. Buchanan: *A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries*. Journal of Biomedical Informatics, 35(5):301–310, 2001.

[18] Chapman, W. y K. Cohen: *Current issues in biomedical text mining and natural language processing (Guest Editorial)*. Journal of Biomedical Informatics, 42(5):757–759, 2009.

[19] Chapman, Wendy Webber, Dieter Hillert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E. Chapman, Mike Conway, Melissa Tharp, Danielle L. Mowery y Louise Deléger: *Extending the NegEx Lexicon for Multiple Languages*. En *MEDINFO 2013 - Proceedings of the 14th World Congress on Medical and Health Informatics, 20-13 August 2013, Copenhagen, Denmark*, páginas 677–681, 2013.

[20] Chen, Danqi y Christopher D Manning: *A fast and accurate dependency parser using neural networks*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1:740–750, 2014.

[21] Chen, Rong y Gunnar O. Klein: *The openEHR Java Reference Implementation Project*. En Kuhn, Klaus A., James R. Warren y Tze Yun Leong (editores): *MedInfo*, volumen 129 de *Studies in Health Technology and Informatics*, páginas 58–62. IOS Press, 2007.

[22] Cimino, J. J.: *Desiderata for controlled medical vocabularies in the twenty-first century*. Methods of information in medicine, 37(4-5):394–403, Noviembre 1998.

[23] Clercq, P. A. de, J. A. Blom, H. H. M. Korsten y A. Hasman: *Approaches for creating computer-interpretable guidelines that facilitate decision support*. Artificial Intelligence in Medicine, 31(1):1–27, Mayo 2004. Review article.

[24] Clinical Practice Guidelines, Institute of Medicine (U.S.). Committee on, M.J. Field y K.N. Lohr: *Guidelines for Clinical Practice: From Development to Use : Summary*. National Academy Press, 1992.

[25] Cunningham, Hamish, Diana Maynard, Kalina Bontcheva y Valentin Tablan: *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. En *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[26] Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica

Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li y Wim Peters: *Text Processing with GATE (Version 6)*. GATE, 2011.

[27] Cunningham, Hamish, Valentin Tablan, Angus Roberts y Kalina Bontcheva: *Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics*. PLOS Computational Biology, 9(2):e1002854, Febrero 2013.

[28] Demner-Fushman, D., W. Chapman y C. McDonald: *What can natural language processing do for clinical decision support?* Journal of Biomedical Informatics, 42(5):760–772, 2009.

[29] Denecke, K.: *Semantic structuring of and information extraction from medical documents using the UMLS*. Methods of Information in Medicine, 47(5):425–534, 2008.

[30] Doan, AnHai, Natalya F. Noy y Alon Y. Halevy: *Introduction to the Special Issue on Semantic Integration*. SIGMOD Rec., 33(4):11–13, Diciembre 2004.

[31] Doan, Son, Mike Conway, TuMinh Phuong y Lucila Ohno-Machado: *Natural Language Processing in Biomedicine: A Unified System Architecture Overview*. En Trent, Ronald (editor): *Clinical Bioinformatics*, volumen 1168 de *Methods in Molecular Biology*, páginas 275–294. Springer New York, 2014.

[32] Dolin, R.H., L. Alschuler, S. Boyer, C. Beebe, F.M. Behlen, P.V. Biron y A. Shabo: *HL7 Clinical Document Architecture, Release 2*. J Am Med Inform Assoc, 13(1):30–39, 2006.

[33] Ehrler, Frédéric, Antoine Geissbühler, Antonio Jimeno y Patrick Ruch: *Data-poor categorization and passage retrieval for gene ontology annotation in Swiss-Prot*. BMC bioinformatics, 6(Suppl 1):S23, 2005.

[34] EN 13606 Association: *CEN/ISO EN13606 EHR*, 2008. Disponible en: `http://www.en13606.org/` [Último acceso: septiembre 2015].

[35] England National Health Service: *NHS Logical Record Architecture*. Disponible en: `www.kith.no/upload/6407/HelsIT-2011_T2-2_Laura_Sato.pdf` [Último acceso: septiembre 2015].

[36] Euzenat, Jérôme y Pavel Shvaiko: *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edición, 2013.

[37] Fellbaum, Christiane: *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[38] Friedman, C., P. O. Alderson, J. H. Austin, J. J. Cimino y S. B. Johnson: *A general natural-language text processor for clinical radiology*. Journal of the American Medical Informatics Association : JAMIA, 1(2):161–174, Marzo 1994.

[39] Friedman, C., L. Shagina, Y. Lussier y G. Hripcsak: *Automated encoding of clinical documents based on natural language processing*. Journal of the American Medical Informatics Association, 11(5):392–402, 2004.

[40] Friedman, Carol y Noémie Elhadad: *Natural Language Processing in Health Care and Biomedicine*. En Shortliffe, Edward H. y James J. Cimino (editores): *Biomedical Informatics*, páginas 255–284. Springer London, 2014.

[41] Friedman, Carol, Thomas C. Rindflesch y Milton Corn: *Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine*. Journal of Biomedical Informatics, 46(5):765–773, 2013.

[42] Fung, Kin Wah, Olivier Bodenreider, Alan R. Aronson, William T. Hole y Suresh Srinivasan: *Combining Lexical and Semantic Methods of Inter-terminology Mapping Using the UMLS*. En Kuhn, Klaus A., James R. Warren y Tze Yun Leong (editores): *MedInfo*, volumen 129 de *Studies in Health Technology and Informatics*, páginas 605–609. IOS Press, 2007.

[43] Garla, Vijay N y Cynthia Brandt: *Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification*. Journal of the American Medical Informatics Association, 20(5):882–886, 2013.

[44] Gennari, J.H., M.A. Musen, R.W. Fergerson, W.E. Grosso, M. Crubezy, H. Eriksson, N.F. Noy y S.W. Tu: *The evolution of Protégé: an environment for knowledge-based systems development*. International Journal of Human-Computer Studies, 58(1):89–123, 2003.

[45] Graham, Ian D., Susan Beardall, Anne O. Carter, Judith Glennie, Paul C. Hébert, Jacqueline M. Tetroe, Finlay A. McAlister, Silvia Visentin y Geoffrey M. Anderson: *What is the quality of drug therapy clinical practice guidelines in Canada?* Canadian Medical Association Journal, 165(2):157–163, 2001.

[46] Grimshaw, J. M. y I. T. Russel: *Implementing clinical practice guidelines: can guidelines be used to improve clinical practice?* Effective Health Care, 8:1–12, 1994.

[47] Grol, Richard, Johannes Dalhuijsen, Siep Thomas, Cees Veld, Guy Rutten y Henk Mokkink: *Attributes of clinical guidelines that influence use of guidelines in general practice: observational study*. BMJ, 317(7162):858–861, Septiembre 1998.

[48] Gschwandtner, T., K. Kaiser, P. Martini y S. Miksch: *Easing semantically enriched information retrieval: An interactive semi-automatic annotation system for medical documents*. International Journal of Human Computer Studies, 68(6):370–385, 2010.

[49] Hearst, Marti A.: *Multi-paragraph Segmentation of Expository Text*. En *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, páginas 9–16, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[50] Hersh, W. R. y L. C. Donohoe: *SAPHIRE International: a tool for cross-language information retrieval*. En *AMIA Annual Symposium proceedings*, páginas 673–677, USA, 1998. AMIA.

[51] Hristovski, Dimitar, Dejan Dinevski, Andrej Kastrin y Thomas C. Rindflesch: *Biomedical question answering using semantic relations*. BMC Bioinformatics, 16:6, 2015.

[52] Huang, Y., H.J. Lowe, D. Klein y R.J. Cucina: *Improved Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon*. JAMIA, 12(3):275–285, 2006.

[53] IHTSDO: *SNOMED CT Starter Guide*. Disponible en: `http://www.snomed.org/starterguide.pdf` [Último acceso: septiembre 2015], 2014.

[54] IHTSDO and openEHR Begin Collaborative Work Programme: *OpenEHR (Internet)*. Disponible en: `http://www.openehr.org/292-OE.html` [Último acceso: enero 2013].

[55] InferMed, Arezzo Technical White Paper, Technical report. Disponible en: `http://www.infermed.com/` [Último acceso: septiembre 2015], 2007.

[56] Information Society European Union., European Commission. Directorate General for the y Media.: *Semantic interoperability for better health and safer healthcare: research and deployment roadmap for Europe. Semantic health report January 2009*. EUR-OP, 2009.

[57] International, HL7: *Health Level 7 International*. Disponible en: `http://www.hl7.org/` [Último acceso: septiembre 2015], 1987.

[58] International Health Terminology Standards Development Organisation: *SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms, (Internet)*. Disponible en: `http://www.ihtsdo.org/snomed-ct/` [Último acceso: septiembre 2015].

[59] International Health Terminology Standards Development Organisation: *SNOMED Clinical Terms User Guide*. Disponible en: `http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_UserGuide_Current-en-US_INT_20130731.pdf` [Último acceso: septiembre 2015], 2013.

[60] International Health Terminology Standards Development Organisation: *SNOMED Clinical Technical Implementation Guide*. Disponible en: `http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_TechnicalImplementationGuide_Current-en-US_INT_20140813.pdf?ok` [Último acceso: septiembre 2015], 2014.

[61] (Internet) Unified Medical Language System: *SNOMED CT Release Files*. Disponible en: `http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html` [Último acceso: septiembre 2015].

[62] Isern, D. y A. Moreno: *Computer-based execution of clinical guidelines: A review*. International Journal of Medical Informatics, 77(12):787–808, 2008.

[63] ISO: *Health Informatics, Electronic Health Record Definition, Scope, and Context*. ISO ISO/TR 20514, International Organization for Standardization, Geneva, Switzerland, 2005.

[64] ISO: *Health informatics – Electronic health record communication – Part 1: Reference model*. ISO 13606-1:2008, International Organization for Standardization, Geneva, Switzerland, 2008.

[65] ISO: *Health informatics – Electronic health record communication – Part 2: Archetype interchange specification*. ISO 13606 - 2:2008, International Organization for Standardization, Geneva, Switzerland, 2008.

[66] ISO: *Health informatics – Harmonized data types for information interchange*. ISO ISO 21090:2011, International Organization for Standardization, 2011.

[67] ISO: *Health informatics – Electronic health record communication – Part 2: Archetype interchange specification*. ISO ISO/WD 13606 - 2, International Organization for Standardization, Geneva, Switzerland, 2012.

[68] J. Field, Marilyn y Kathleen N. Lohr: *Clinical Practice Guidelines: Directions for a New Program*. The National Academies Press, 1990.

[69] Jimeno-Yepes, Antonio y Alan R. Aronson: *Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation*. En Luo, Gang, Jiming Liu y Christopher C. Yang (editores): *IHI*, páginas 733–736. ACM, 2012.

[70] Jin, Y., R.T. McDonald, K. Lerman, M.A. Mandel y et al.: *Automated recognition of malignancy mentions in biomedical literature*. BMC Bioinformatics, 7:492, 2006.

[71] Jonquet, Clement, Nigam H. Shah y Mark A. Musen: *The open biomedical annotator*. Summit Trans Bioinformatics, páginas 56–60, 2009.

[72] Kaiser, Frieda, James Angus y Helen Stevens: *e-MS Clinical Document Architecture Implementation Guide*, Diciembre 2004.

[73] Kaiser, K., C. Akkaya y S. Miksch: *How can information extraction ease formalizing treatment processes in clinical practice guidelines? A method and its evaluation*. Artificial Intelligence in Medicine, 39(2):151–163, 2007.

[74] Kaiser, Katharina y Silvia Miksch: *Versioning Computer-interpretable Guidelines: Semi-automatic Modeling of 'Living Guidelines' Using an Information Extraction Method*. Artificial Intelligence In Medicine, 46(1):55–66, Mayo 2009.

[75] Kalfoglou, Yannis y Marco Schorlemmer: *Ontology Mapping: The State of the Art*. The Knowledge Engineering Review, 18(1):1–31, Enero 2003.

[76] Kashyap, Vipul y Amit Sheth: *Semantic and Schematic Similarities Between Database Objects: A Context-based Approach*. The VLDB Journal, 5(4):276–304, Diciembre 1996.

[77] Kenneth, D.: *ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008*. European Heart Journal, 29:2388–2442, 2008.

[78] Kim, J. D., T. Ohta, Y. Tateisi y J. Tsujii: *GENIA corpus - a semantically annotated corpus for bio-textmining*. Bioinformatics, 19(suppl 1):i180–i182, 2003.

[79] Knight, Kevin y Steve K Luk: *Building a large-scale knowledge base for machine translation*. En *AAAI*, volumen 94, páginas 773–778, 1994.

[80] Kosara, Robert y Silvia Miksch: *AsbruView: Capturing Complex, Time-Oriented Plans Beyond Flow-Charts*. En Olivier, Paul et al. (editor): *Diagrammatic Representation and Reasoning*. Springer, Berlin, 2002.

[81] Krauthammer, M. y G. Nenadic: *Term identification in the biomedical literature*. Journal of Biomedical Informatics, 37(6):512–526, 2004.

[82] Kumar, Anand, Paolo Ciccarese, Barry Smith y Matteo Piazza: *Context-based task ontologies for clinical guidelines*. En Pisanelli, Domenico M. (editor): *Ontologies in Medicine. Proceedings of the Workshop on Medical Ontologies*, volumen 102 de *Studies in Health Technology and Informatics*, páginas 81–94, LMI, Department of Computer Science, University of Pavia, Italy. anand.kumar@ifomis.uni-leipzig.de, 2004.

[83] Köhler, Sebastian, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C. M. Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. FitzPatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane A. Hurst, Johanna Jähn,

Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem Ouwehand, Soo Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O. M. Wilkie, Caroline F. Wright, Anneke T. Vulto van Silfhout, Nicole de Leeuw, Bert B. A. de Vries, Nicole L. Washington, Cynthia L. Smith, Monte Westerfield, Paul N. Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis y Peter N. Robinson: *The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data*. Nucleic Acids Research (NAR), 42(Database-Issue):966–974, 2014.

[84] LePendu, Paea, Srinivasan Iyer, Cédrick Fairon y Nigam H. Shah: *Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes*. J. Biomedical Semantics, 3(S-1):S5, 2012.

[85] Lezcano, Leonardo, Salvador Sánchez-Alonso y Miguel Angel Sicilia: *Associating clinical archetypes through UMLS metathesaurus term clusters*. Journal of medical systems, 36(3):1249–1258, 2012.

[86] Lindberg, D., B. Humphreys y A. Mc Cray: *The Unified Medical Language System*. Methods of Information in Medicine, 32:281–91, 1993.

[87] London University College: *UCL Chronic Disease Management*. Disponible en: `http://www.ucl.ac.uk/` [Último acceso: septiembre 2015].

[88] Macedo, N.AM., N.AG. Hilares, J.P.P. Quispe y R.A Matutti: *Electronic Health Record: Comparative analysis of HL7 and open EHR approaches*. En *Health Care Exchange (PAHCE), 2010 Pan American*, páginas 105–110, Marzo 2010.

[89] Manning, Christopher D., Prabhakar Raghavan y Hinrich Schütze: *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[90] Martínez-Costa, C, Menárguez Tortosa M. y Fernández Breis J.T: *An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes*. Journal of Biomedical Informatics, 43(5):736–746, 2010.

[91] McBride, B: *The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS*. En *In Handbook on Ontologies*, páginas 51–66, 2004.

[92] McCray, A. T.: *An upper-level ontology for the biomedical domain*. Comparative and Functional Genomics, páginas 80–84, 2003.

[93] McCray, A. T., S. Srinivasan y A. C. Browne: *Lexical methods for managing variation in biomedical terminologies*. Proceedings of the Annual Symposium on Computer Application in Medical Care, páginas 235–239, 1994.

[94] McEntyre, J. y D. Lipman: *PubMed: bridging the information gap*. CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne, 164(9):1317–1319, Mayo 2001.

[95] McGuinness, D. L. y F. van Harmelen: *OWL Web Ontology Language*. Disponible en: `http://www.w3.org/TR/2004/REC-owl-features-20040210/` [Último acceso: septiembre 2015], 2004.

[96] Medicine, National Library of: *Medical Subject Headings {MeSH}*. Disponible en: `http://www.nlm.nih.gov/mesh/meshhome.html` [Último acceso: septiembre 2015], 2003.

[97] Medicine (US), Bethesda (MD). National Library of: *6. SPECIALIST Lexicon and Lexical Tools*. UMLS Reference Manual, Septiembre 2009. `http://www.ncbi.nlm.nih.gov/books/NBK9680/`.

[98] Mehrabi, Saeed, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C. Max Schmidt, Hongfang Liu y Mathew Palakal: *DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx*. Journal of Biomedical Informatics, 54:213–219, Abril 2015.

[99] Meizoso García, María, José Luis Iglesias Allones, Diego Martínez Hernández y María Jesús Taboada Iglesias: *Semantic similarity-based alignment between clinical archetypes and SNOMED CT: an application to observations*. International journal of medical informatics, 81(8):566–578, 2012.

[100] Miles, A. y S. Bechhofer: *SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009*. Disponible en: `http://www.w3.org/TR/2009/REC-skos-reference-20090818/` [Último acceso: septiembre 2015], 2009.

[101] Miles, Alistair y Sean Bechhofer: *SKOS Simple Knowledge Organization System Reference*. Informe técnico, W3C, 2009.

[102] Miller, Dan Brickley Eric: *RDF: Resource Description Framework*. Disponible en: `http://www.w3.org/RDF/` [Último acceso: septiembre 2015], 2004.

[103] Mork, James G., Dina Demner-Fushman, Susan Schmidt y Alan R. Aronson: *Recent Enhancements to the NLM Medical Text Indexer*. En *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, páginas 1328–1336, 2014.

[104] Murata, Masaki, Toshiyuki Kanamaru y Hitoshi Isahara: *Automatic synonym acquisition based on matching of definition sentences in multiple dictionaries*. En *Computational Linguistics and Intelligent Text Processing*, páginas 293–304. Springer, 2005.

[105] Murray, J.: *ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012*. European Heart Journal, 33:1787–1847, 2012.

[106] Musen, Mark A., Natalya Fridman Noy, Nigam H. Shah, Patricia L. Whetzel, Christopher G. Chute, Margaret Anne D. Storey y Barry Smith: *The National Center for Biomedical Ontology*. JAMIA, páginas 190–195, 2012.

[107] Na, Jin Cheon y Hock Leng Neoh: *Effectiveness of UMLS semantic network as a seed ontology for building a medical domain ontology*. Aslib Proceedings, 60(1):32–46, 2008.

[108] National E-Health Transition Authority: *Clinical Knowledge Manager*. Disponible en: `http://dcm.nehta.org.au/ckm/` [Último acceso: septiembre 2015].

[109] National E-Health Transition Authority: *National E-Health Transition Authority (Internet)*. Disponible en: `http://www.nehta.gov.au/` [Último acceso: septiembre 2015].

[110] National Library of Medicine: *MetaMap*. Disponible en: `http://metamap.nlm.nih.gov/` [Último acceso: septiembre 2015].

[111] Nguyen, Viet Cuong: *A Study on Statistical Generation of a Hierarchical Structure of Topic-information for Multi-documents*. Tesis de Doctorado, Japan Advanced Institute of Science and Technology, 2011.

[112] NHS Connecting for Health: *Archetypes from: NHS Connecting For Health*. Disponible en: `https://svn.connectingforhealth.nhs.uk/svn/public/nhscontentmodels/TRUNK/cm/archetypes/gen/html/index.html` [Último acceso: enero 2013].

[113] NHS Connecting for Health: *NHS Connecting for Health (Internet)*. Disponible en: `http://www.connectingforhealth.nhs.uk/` [Último acceso: marzo 2013].

[114] NHS Connecting for Health: *Archetype dataset*. Disponible en: `http://www.usc.es/keam/TermArchetypes/input.html` [Último acceso: septiembre 2015], 2011.

[115] NLM: *UMLS Reference Manual*. Bethesda (MD): National Library of Medicine (US), Septiembre 2009. `http://www.ncbi.nlm.nih.gov/books/NBK9676/`.

[116] Noy, Natalya F: *Semantic integration: a survey of ontology-based approaches*. ACM Sigmod Record, 33(4):65–70, 2004.

[117] Noy, Natalya F.: *Semantic Integration: A Survey of Ontology-based Approaches*. SIGMOD Rec., 33(4):65–70, Diciembre 2004.

[118] Ohno-machado, Lucila, John H. Gennari, Shawn Murphy, Nilesh L. Jain, D. Sc, Samson W. Tu, Diane E. Oliver, Edward Pattison-gordon, Robert A. Greenes, Edward H. Shortliffe y G. Octo Barnett: *The GuideLine Interchange Format: A Model for Representing Guidelines*. Journal of the American Medical Informatics Association, 5:357–372, 1998.

[119] OpenEHR Foundation: *OpenEHR: An open domain-driven platform for developing flexible e-health systems*. Disponible en: `http://www.openehr.org` [Último acceso: septiembre 2015], 2013.

[120] OpenEHR Specification Program: *The openEHR Archetype Model: Archetype Definition Language - ADL 2*. Informe técnico, The OpenEHR Foundation, 2014.

[121] OpenEHR Specification Program: *The openEHR Archetype Model: Archetype Object Model*. Informe técnico, The OpenEHR Foundation, 2014.

[122] Ortiz, A.M. and Universitat Autónoma de Barcelona, Laboratori de Lingüística Informática: *Diseño e implementación de un lexicón computacional para lexicografía y traducción automática*. Universitat Autónoma de Barcelona, Laboratori de Lingüística Informática, 2000.

[123] Parcero, Estíbaliz, Monserrat Robles y José A. Maldonado: *Implementación de técnicas de string matching y selección semántica aproximada en un motor de normalización terminológica*. Disponible en: `https://riunet.upv.es/bitstream/handle/10251/17464/Memoria.pdf?sequence=1` [Último acceso: septiembre 2015], 2012.

[124] Patel, Vimla L., Vanessa G. Allen, José F. Arocha y Edward H. Shortliffe: *Research Paper: Representing Clinical Guidelines in GLIF: Individual and Collaborative Expertise*. JAMIA, 5(5):467–483, 1998.

[125] Peleg, M. y S. Tu: *Design patterns for clinical guidelines*. Artificial Intelligence in Medicine, 47(1):1–24, 2009.

[126] Physicians The Royal College of: *The Royal College of Physicians*. Disponible en: `https://www.rcplondon.ac.uk/` [Último acceso: septiembre 2015].

[127] Prud'hommeaux, Eric y Andy Seaborne: *SPARQL Query Language for RDF*. W3C Recommendation, 2008.

[128] Qamar, R: *Semantic mapping of clinical model data to biomedical terminologies to facilitate data interoperability*. Tesis de Doctorado, Faculty of Engineering and Physical Sciences: University of Manchester, 2008.

[129] Rassinoux, A. M., R.H. Baud y J. R. Scherrer: *A multilingual analyser form medical texts*. Disponible en: `http://mbi.dkfz-heidelberg.de/helios/doc/nlp/Rassinoux94b.html` [Último acceso: septiembre 2015], 1994.

[130] Raychaudhuri, Soumya, Jeffrey T. Chang, Patrick D. Sutphin y Russ B. Altman: *Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature*. Genome Research, 12(1):203–214, 2002.

[131] Rector, A. L.: *Thesauri and formal classifications: terminologies for people and machines*. Methods of information in medicine, 37(4-5):501–509, Noviembre 1998.

[132] Reeve, Lawrence H.: *Semantic Annotation and Summarization of Biomedical Text*. Tesis de Doctorado, Drexel University, Julio 2007.

[133] Reeve, Lawrence H. y Hyoil Han: *CONANN: An Online Biomedical Concept Annotator*. En Boulakia, Sarah Cohen y Val Tannen (editores): *DILS*, volumen 4544 de *Lecture Notes in Computer Science*, páginas 264–279. Springer, 2007.

[134] Rindflesch, T.C. y O. Bodenreider: *Advanced library services: developing a biomedical knowledge repository to support advanced information management applications*. Informe técnico, The Lister Hill National Center for Biomedical Communications. LHNCBC-TR-2006-001, 2006.

[135] Rindflesch, T.C. y M. Fiszman: *The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text*. Journal of Biomedical Informatics, 36(6):462–477, 2003.

[136] RN, Shiffman, Shekelle P, Overhage JM, Slutsky J, Grimshaw J y Deshpande AM: *Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization*. Annals of Internal Medicine, 139(6):493–498, 2003.

[137] Roberts, A., R. Gaizauskas, M. Hepple, G. Demetriou y et al.: *Building a semantically annotated corpus of clinical texts*. Journal of Biomedical Informatics, 42(5):950–66, 2009.

[138] Rosenbloom, Trent S., Randolph A. Miller, Kevin B. Johnson, Peter L. Elkin y Steven H. Brown: *A model for evaluating interface terminologies*. J Am Med Inform Assoc, 15(1):65–76, 2008 Jan-Feb.

[139] Rosser, W.W., D. Davis y E. Gilbart: *Promoting effective guideline use in Ontario.* JAMC, 165(2):181–182, Julio 2001.

[140] Roukema, J., R. K. Los, S. E. Bleeker, A. M. van Ginneken, J. van der Lei y H. A. Moll: *Paper versus computer: feasibility of an electronic medical record in general pediatrics*. Pediatrics, 117(1):15–21, Enero 2006.

[141] Ruch, Patrick, Julien Gobeill, Christian Lovis y Antoine Geissbühler: *Automatic medical encoding with SNOMED categories*. BMC Medical Informatics and Decision Making, 8(S-1):S6, 2008.

[142] Savova, G.K., J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler y C.G. Chute: *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, 2010.

[143] Schleyer, Titus, Heiko Spallek y Pedro Hernández: *A qualitative investigation of the content of dental paper-based and computer-based patient record formats*. Journal of the American Medical Informatics Association : JAMIA, 14(4):515–526, 2007.

[144] Schloeffel, P., T. Beale, G. Hayworth, S. Heard y H. Leslie: *The relationship between CEN 13606, HL7, and OpenEHR*. Proc HIC, 2006.

[145] Seidenberg, Julian y Alan Rector: *Web ontology segmentation: analysis, classification and use*. En *WWW '06: Proceedings of the 15th international conference on World Wide Web*, páginas 13–22, New York, NY, USA, 2006. ACM Press.

[146] Serban, R., A. ten Teije, F. van Harmelen, M. Marcos y C. Polo-Conde: *Extraction and use of linguistic patterns for modelling medical guidelines*. Artificial Intelligence in Medicine, 39(2):137–149, 2007.

[147] Serban, Radu, Annette ten Teije, Frank van Harmelen, Mar Marcos y Cristina Polo-Conde: *Ontology-Driven Extraction of Linguistic Patterns for Modelling Clinical Guidelines*. En Miksch, Silvia, Jim Hunter y ElpidaT. Keravnou (editores): *Artificial Intelligence in Medicine*, volumen 3581 de *Lecture Notes in Computer Science*, páginas 191–200. Springer Berlin Heidelberg, 2005.

[148] Shah, Nigam H., Nipun Bhatia, Clement Jonquet, Daniel Rubin, Annie P. Chiang y Mark A. Musen: *Comparison of concept recognizers for building the Open Biomedical Annotator*. BMC bioinformatics, 10 Suppl 9, 2009.

[149] Shiffman, Richard N, Bryant T Karras, Abha Agrawal, Roland Chen, Luis Marenco y Sujai Nath: *GEM: a proposal for a more comprehensive guideline document model using XML*. Journal of the American Medical Informatics Association, 7(5):488–498, 2000.

[150] Shiffman, Richard N., George Michel, Michael Krauthammer, Norbert E. Fuchs, Kaarel Kaljurand y Tobias Kuhn: *Writing Clinical Practice Guidelines in Controlled Natural Language*. En *Proceedings of the 2009 Conference on Controlled Natural Language*, CNL'09, páginas 265–280, Berlin, Heidelberg, 2010. Springer-Verlag.

[151] Siering, Ulrich, Michaela Eikermann, Elke Hausner, Wiebke Hoffmann-Esser y Edmund A. Neugebauer: *Appraisal Tools for Clinical Practice Guidelines: A Systematic Review*. PLoS ONE, 8(12):e82915, Diciembre 2013.

[152] Simalatsar, Alena y Giovanni De Micheli: *TAT-based Formal Representation of Medical Guidelines : Imatinib Case-study*. En *Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2012)*, IEEE Engineering in Medicine and Biology Society Conference Proceedings, New York, 2012. IEEE.

[153] Simalatsar, Alena, Wenqi You, Verena Gotta, Nicolas Widmer y Giovanni De Micheli: *Representation of Medical Guidelines with a Computer Interpretable Model*. International Journal on Artificial Intelligence Tools (IJAIT), 23(3), 2014.

[154] Simera, I., D. Moher, J. Hoey, K. F. Schulz y D. G. Altman: *A catalogue of reporting guidelines for health research*. European Journal of Clinical Investigation, 40(1):35–53, 2010.

[155] Späth, Melanie Bettina y Jane Grimson: *Applying the archetype approach to the database of a biobank information management system*. I. J. Medical Informatics, 80(3):205–226, 2011.

[156] Stearns, M. Q., C. Price, K. A. Spackman y A. Y. Wang: *SNOMED clinical terms: overview of the development process and project status*. Proceedings / AMIA ... Annual Symposium. AMIA Symposium, páginas 662–666, 2001.

[157] Steele, R. y J. Fox: *Tallis PROforma Primer - Introduction to PROforma Language and Software with Worked Examples*. Informe técnico, Advanced Computation Laboratory, Cancer Research, London, UK, 2002.

[158] Sun, Jennifer Y. y Yao Sun: *A System for Automated Lexical Mapping*. Journal of the American Medical Informatics Association, 13:334–343, Febrero 2006.

[159] Sun, Weiyi, Anna Rumshisky y Ozlem Uzuner: *Normalization of relative and incomplete temporal expressions in clinical narratives*. Journal of the American Medical Informatics Association, 2015.

[160] Swedberg, K.: *ESC Guidelines for the diagnosis and treatment of chronic heart failure:executive summary (update 2005)*. European Heart Journal, 26:1115–1140, 2005.

[161] Taboada, M., M. Meizoso, D. Martínez, D. Riaño y A. Alonso: *Combining open-source natural language processing tools to parse clinical practice guidelines*. Expert Systems, 30(1):3–11, 2013.

[162] Taboada, Maria, Rosario Lalín y Diego Martínez Hernández: *An Automated Approach to Mapping External Terminologies to the UMLS*. IEEE Transactions on Biomedical Engineering, 56(6):1598–1605, 2009.

[163] Taboada, María, María Meizoso, David Riaño, Albert Alonso y Diego Martínez: *From Natural Language Descriptions in Clinical Guidelines to Relationships in an Ontology*. En Riaño, David, Annette ten Teije, Silvia Miksch y Mor Peleg (editores): *Knowledge Representation for Health-Care. Data, Processes and Guidelines*, volumen 5943 de *Lecture Notes in Computer Science*, páginas 26–37. Springer Berlin Heidelberg, 2010.

[164] Tang, Buzhou, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C. Denny y Hua Xu: *A hybrid system for temporal information extraction from clinical text*. JAMIA, páginas 828–835, 2013.

[165] The AGREE Collaboration: *Instrument voor beoordeling van richtlijnen. (AGREE)*, 2001. `http://www.cbo.nl/product/richtlijnen/folder20021023121843/agreeinvulform.pdf`.

[166] The openEHR Foundation: *OpenEHR Clinical Knowledge Manager*. Disponible en: `http://www.openehr.org/knowledge/` [Último acceso: septiembre 2015].

[167] Turner, Tari, Marie Misso, Claire Harris y Sally Green: *Development of evidence-based clinical practice guidelines (CPGs): comparing approaches*. Implementation Science, 3(1):45, 2008.

[168] Utiyama, Masao y Hitoshi Isahara: *A Statistical Model for Domain-independent Text Segmentation*. En *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, páginas 499–506, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[169] Vlayen, Joan, Bert Aertgeerts, Karin Hannes, Walter Sermeus y Dirk Ramaekers: *A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit*. International Journal for Quality in Health Care, 17(3):235–242, 2005.

[170] W3 Consortium: *OWL 2 Web Ontology Language Guide*, 2004. Disponible en: `http://www.w3.org/TR/owl-guide/` [Último acceso: septiembre 2015].

[171] W3 Consortium: *OWL 2 Web Ontology Language Document Overview (Second Edition)*, 2012. Disponible en: `http://www.w3.org/TR/2012/REC-owl2-overview-20121211/` [Último acceso: septiembre 2015].

[172] Whetzel, P. L., N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache y M. A. Musen: *BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications*. Nucleic Acids Research, 39(suppl):W541–W545, Junio 2011.

[173] Woolf, S. H., R. Grol, A. Hutchinson, M. Eccles, J. Grimshaw, S. H. Woolf, R. Grol, A. Hutchinson, M. Eccles y J. Grimshaw: *Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines*. BMJ, 318(7182):527–30+, 1999.

[174] Wu, Zhibiao y Martha Palmer: *Verbs Semantics and Lexical Selection*. En *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, páginas 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[175] Yu, Sheng, Damon Berry y Jesús Bisbal: *An Investigation of Semantic Links to Archetypes in an External Clinical Terminology through the Construction of Terminological "Shadows"*. En *Proceedings of the IADIS International Conference on e-Health*, 2010.

[176] Zeng, Marcia Lei y Lois Mai Chan: *Trends and issues in establishing interoperability among knowledge organization systems*. Journal of the American Society for Information Science and Technology, 55:377–395, Marzo 2004.

[177] Zou, Qinghua, Wesley W Chu, Craig Morioka, Gregory H Leazer y Hooshang Kangarloo: *IndexFinder: a method of extracting key concepts from clinical texts for indexing*. En *AMIA Annual Symposium Proceedings*, volumen 2003, página 763. American Medical Informatics Association, 2003.

Índice de figuras

Índice de tablas

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA