Universidade de Santiago de Compostela
Departamento de Electrónica e Computación

TESIS DOCTORAL

ALINEAMIENTO Y VALIDACIÓN DE TERMINOLOGÍAS A GRAN ESCALA EN EL ÁMBITO MÉDICO

Presentada por:
María del Rosario Lalín Rodríguez

Codirigida por:
María Jesús Taboada Iglesias
Diego Martínez Hernández

Santiago de Compostela, Septiembre de 2011

Dr. **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

Dr. **Diego Martínez Hernández**, Catedrático de Escuela Universitaria del Área de Física Aplicada de la Universidad de Santiago de Compostela

HACEN CONSTAR:

Que la memoria titulada **Alineamiento y validación de terminologías a gran escala en el ámbito médico** ha sido realizada por Dña. **María del Rosario Lalín Rodríguez** bajo nuestra dirección en el Departamento de Electrónica e Computación de la Universidade de Santiago de Compostela, y constituye la Tesis que presenta para obtar al grado de Doctor.

Santiago de Compostela, Septiembre de 2011

Asdo: **María Jesús Taboada Iglesias**
Co-directora de la tesis

Asdo: **Diego Martínez Hernández**
Co-director de la tesis

Asdo: **Francisco Fernández Rivera**
Director del Departamento de Electrónica y Computación

Asdo: **María del Rosario Lalín Rodríguez**
Autora de la tesis

A mis padres y a Óscar

Agradecimientos

Agradecer a mis directores de tesis, Chus y Diego, la confianza depositada en mí y su dedicación durante el desarrollo de esta tesis y a Miguel Fernández el trabajo previo del que partí para realizarla. También, a los compañeros del Grupo de Ingeniería del conocimiento del Departamento de Electrónica y Computación.

Agradecer a la división de bases de datos de Elsevier BV Bibliographic, representado por Dr. Ian Crowlesmith, su ayuda al proporcionarnos acceso a Emtree para su evaluación. También a la National Library of Medicine por el acceso libre para acceder a los UMLS Knowledge Sources (UMLSKS).

Debo hacer mención también a los proyectos "Gestión de Terminologías Médicas para Arquetipos" (TIN2009-14159-C05-05) del Ministerio de Educación y Ciencia y "Adquisición semi-automática de conocimiento a partir de guías de práctica clínica" (TIN2006-15453-C04-02) de la Universidad de Santiago de Compostela que han permitido realizar esta investigación.

Por último, a mi familia y amigos, por su apoyo durante estos años y, en especial, a Óscar, por su empujón inicial y su paciencia y ánimo continuo desde entonces.

Septiembre de 2011

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Antecedentes

Durante las últimas décadas, la tecnología ha avanzado de forma tan espectacular que se ha abierto una puerta muy prometedora al intercambio, procesamiento y manipulación de grandes cantidades de información. Por ello, la comunidad científica se ha volcado en desarrollar nuevos métodos orientados a la automatización de la búsqueda, recuperación e intercambio de todo este volumen incesante de información. En especial, en el ámbito de la organización de la información y conocimiento, los proyectos de investigación se han centrado principalmente en desarrollar herramientas que facilitan el acceso a grandes colecciones de información.

Hace ya mucho tiempo que se viene realizando un esfuerzo considerable para desarrollar herramientas orientadas a la gestión de grandes volúmenes de información sobre libros, artículos, autores, ... Esto ha dado lugar a los llamados sistemas de organización de conocimiento (KOS), cuyo objetivo es estructurar el conocimiento sobre estas colecciones, recogiendo la terminología y las notaciones que usamos para definir y organizar los conceptos y objetos del mundo real. Para ello, se han diseñado vocabularios altamente estructurados, como tesauros, terminologías, redes semánticas u ontologías, donde, además de los términos en sí mismos, se incluye información de las relaciones entre ellos.

Uno de los usos principales de los KOS es la abstracción e indexación de la información, que busca el acceso eficiente a la información. En las últimas décadas, con la aparición de Internet, toda esa información se ha ido poniendo en línea y se ha hecho necesario pasar de sistemas que proporcionaban navegación por tabla de contenidos a sistemas de búsqueda por texto libre y términos. Un ejemplo es el portal web de Elsevier, el cual dispone de un

esquema de categorización para facilitar el acceso a sus títulos.

Otras organizaciones, en ámbitos muy diferentes, también han desarrollado sus propios sistemas de organización de conocimiento y clasificaciones que, en algunos casos, buscan convertirse en estándares a fin de unificar la referencia a esos conceptos por diferentes sistemas. Incluso, en un mismo ámbito, podemos encontrarnos con diferentes organizaciones que han ido desarrollado distintos sistemas enfocados a sus necesidades particulares, que tendrán similitudes y diferencias entre sí, pero que, en todo caso, representan conocimiento de gran interés. De ahí, surge la necesidad de interoperar entre ellos para incrementar la eficiencia de las búsquedas.

En resumen, el alineamiento de los modelos KOS es un proceso básico en diferentes ámbitos de las tecnologías de la información; por citar sólo algunos de ellos, en ingeniería ontológica, integración de la información, compartición de información (P2P), composición de servicios web, agentes autónomos y realización de consultas en la web.

1.2. Contexto

Todo lo dicho en el punto anterior es especialmente cierto en el ámbito biomédico, donde, por un lado, bibliotecas públicas (por ejemplo, National Library of Medicine) y editoriales (por ejemplo, Elsevier) y por otro, hospitales y asociaciones médicas han desarrollado a lo largo del tiempo sus sistemas de información. Consideremos, por ejemplo, los siguientes escenarios [ECH+94]:

- Un hospital puede requerir instalar una nueva herramienta de ayuda a la decisión, con el fin de revisar si hay contradicciones en los tratamientos indicados por los médicos (por ejemplo, con el fin de advertir al médico cuando éste prescriba un antiinflamatorio no esteroideo a un paciente con úlcera péptica). El sistema requiere acceder a diferentes fuentes de información existentes para extraer la información de medicamentos y las enfermedades que tratan e interoperar con el herramienta desarrollada.

- Un sistema de recuperación de artículos puede usar sinónimos para aumentar las posibilidades de obtener resultados en un proceso de búsqueda de dichos artículos. Una posible situación consistiría en acceder a una o varias fuentes independientes de información, que incluyan sinónimos de los términos, y relacionar esta información con la del sistema bibliográfico.

- Un sistema focalizado en el seguimiento de todos los datos del paciente,

incluyendo síntomas, requerirá usar fuentes genéricas de síntomas, medicamentos, enfermedades, pruebas médicas, ... a fin de mantener una uniformidad en su descripción.

- Un sistema para registrar ensayos controlados como parte del cuidado rutinario de paciente (como, por ejemplo, un centro de diálisis renal) requiere una fuente estándar de conceptos médicos, síntomas, medicamentos, etc., a fin de integrar y compartir los datos.

Estos escenarios reflejan diferentes áreas de aplicación del conocimiento médico donde el principal obstáculo es la ausencia de un lenguaje consistente y comprensivo para representar el dominio. La variabilidad en el uso del lenguaje, la ambigüedad, la vaguedad o las elipsis que pueden contener han impedido hasta ahora un uso profundo de la computación en medicina. De ahí surgió la necesidad de desarrollar los vocabularios y terminologías estándar, en algunos casos genéricos y en otros adaptados a un ámbito específico como pueden ser las enfermedades, los medicamentos o los resultados de laboratorio.

Las razones de este esfuerzo son, como es evidente, dar acceso a esa información de forma inequívoca tanto para la descripción de enfermedades y tratamientos como para la gestión estándar de las actividades de los hospitales o la generación de estadísticas; además, por supuesto, de proporcionar interoperabilidad semántica entre los diferentes recursos que pueden solaparse. El objetivo final a conseguir es el aumento de la calidad de los servicios médicos y la disminución de la duplicidad de esfuerzos.

Actualmente, muchos de estos recursos terminológicos son fácilmente accesibles a través de Internet, con el consiguiente problema del solapamiento: varios recursos pueden dar lugar a diferentes vistas del mismo dominio, por lo que se hace necesario implementar procesos que localicen las correspondencias entre ellos. Podemos indicar al menos 6 tareas en los que se hace evidente la necesidad de interoperabilidad entre las diferentes fuentes de información:

- Interacción persona-ordenador: para soportar la entrada de datos fácil e intuitiva y la formulación de consultas por usuarios clínicos.
- Archivo y recuperación: para almacenar y recuperar información clínica en historias clínicas.
- Mediación, distribución y reutilización: para permitir a la información ser compartida entre diferentes historias clínicas y sistemas de recuperación de información y soporte a la decisión.
- Indexado e inferencia: para hacer más fácil formular y compartir sistemas de ayuda a la decisión.

- Autorización y mantenimiento: para hacer posible construir, mantener y extender la propia terminología.

- Procesamiento de lenguaje natural: para facilitar la expresión y la comprensión en cualquier lenguaje de conceptos definidos usando el lenguaje natural.

1.3. Definición del problema

Proporcionar interoperabilidad entre diferentes recursos de conocimiento es también una tarea crítica para el intercambio eficiente de información en otras comunidades, como Ciencias de la Información ([ZC04], [W3C]), bases de datos ([DNH04], [SS06]) y ontologías ([KS03], [Noy04], [ES07]) y ha sido un tema extensamente tratado en la comunidad biomédica ([Doe01], [FBA+07], [VGHHT04], [Yu06], [ZMBB07]).

Independientemente del considerable volumen de investigación en diferentes campos, el alineamiento entre terminologías sigue siendo un problema complejo. Hay, entre otros, dos obstáculos que dificultan la interoperabilidad. Uno es el tratamiento informal de las relaciones en la terminología biomédica, que conduce a definiciones contradictorias y ambiguas [SCK+05]. El segundo es la falta de métodos automatizados que simplifiquen el proceso de alineamiento [DNH04]. En este trabajo, nos centramos en este segundo problema.

Una característica especial de las terminologías, frente a otros recursos, es que suele haber alguna referencia o estándar disponible, por lo que el uso de métodos léxicos en las primeras etapas del alineamiento produce alineamientos de alta calidad ([Doe01], [VGHHT04], [ZMBB07], [SS06]). Sin embargo, el enorme volumen de datos en el vocabulario dificulta la revisión manual de los mappings léxicos y es necesario un considerable esfuerzo humano para interpretarlos adecuadamente y garantizar la validez de las asignaciones léxicas resultantes [ZC04]. Por tanto, es imprescindible desarrollar métodos de validación automática que eviten al máximo la revisión manual, etapa que suele consumir un importante porcentaje de recursos en la mayoría de los proyectos. En resumen, el problema a resolver es proporcionar métodos para interpretar y evaluar automáticamente los alineamientos léxicos resultantes de forma más eficaz que la revisión manual.

En el presente trabajo, proponemos un método automático para alinear terminologías externas a la terminología de referencia en el ámbito biomédico, que es el Metathesaurus de UMLS [Bod04], [aHBM93]. Para ello, alinearemos una terminología llamada Emtree, un tesauro biomédico de gran tamaño,

desarrollado por Elsevier para indexar la base de datos bibliográfica Embase. El método diseñado también puede ser utilizado para alinear cualquier terminología a otra integrada en el Metathesaurus de UMLS.

Nuestro enfoque aplica una combinación secuencial de dos métodos básicos coincidentes con el fin de producir la alineación. En primer lugar, una técnica léxica identifica las cadenas similares entre la fuente y la terminología de destino. En segundo lugar, varias técnicas basadas en la semejanza estructural de las terminologías a equiparar validan la alineación léxica. Para ello, se explora la similitud entre categorías de nivel superior entre ambas terminologías y se aplica para descartar los alineamientos que, con mayor probabilidad, son semánticamente diferentes. A continuación, se explora la similitud entre las estructuras organizativas de los conceptos en ambas terminologías a fin de seleccionar los alineamientos semánticamente más similares. El método también aplica una combinación de técnicas de procesado de lenguaje natural, técnicas léxicas y técnicas semánticas para producir alineamientos complejos en aquellos casos donde no es posible encontrar alineamientos directos.

1.4. Estructura de la tesis

Esta memoria se organiza en los siguientes capítulos. En el capítulo 2, se presenta una descripción detallada del estado del arte sobre el alineamiento de terminologías. Se comienza presentando el problema de la interoperabilidad, para continuar definiendo las terminologías, dando ejemplos en el ámbito médico, entre ellas, Emtree y UMLS, que son las analizadas aquí. También se describe SKOS, el modelo definido por la W3C para la descripción de tesauros y taxonomías. A continuación, se define el proceso de equiparación entre terminologías y su principal problema, la heterogeneidad entre ellas. Seguidamente, se describen las diferentes técnicas existentes para la equiparación de terminologías y se explican varios ejemplos de su uso en la comunidad científica. Por último, se comentan las iniciativas actuales sobre la evaluación de los resultados del proceso de equiparación de terminologías y ontologías.

En el capítulo 3 se presentan los objetivos de este trabajo de tesis así como las limitaciones y restricciones que presenta. También se incluye el esquema general del método.

El capítulo 4 describe el método propuesto para el alineamiento de terminologías de gran tamaño. Para ello, se exploran con más detalle aquellos aspectos de las terminologías analizadas, Emtree y UMLS, que son utilizados en esta tesis. A continuación, se detalla cómo se lleva a cabo el alineamiento léxico de términos, usando un servicio del UMLS. Por último, se explica el proceso seguido para aquellos conceptos Emtree que no obtienen equipara-

ción léxica directa con ningún concepto del Metathesaurus de UMLS.

En el capítulo 5, se describe la segunda parte del método, la validación del alineamiento obtenido, descartando los alineamientos incompatibles por las categorías de alto nivel, y la posterior desambiguación, con el objetivo de escoger el mejor de ellos, o, cuando no es posible, aquellos que incluyen la mayor información sobre él. Por último, se construyen los alineamientos complejos, para aquellos casos donde no fue posible encontrar un alineamiento directo.

En el capítulo 6, se analizan con detalle los resultados obtenidos que avalan la propuesta metodológica. En primer lugar, se realiza un análisis cuantitativo a través de tablas que muestran los resultados obtenidos en cada fase. A continuación, se hace un análisis cualitativo con las conclusiones que se extraen en cada fase de proceso.

En el capítulo 7, se exponen las conclusiones de esta tesis haciendo especial énfasis en sus aportaciones y en las futuras líneas de investigación abiertas como continuación de esta tesis.

Finalmente, se incluyen 3 apéndices. En el apéndice A, Diccionario, se describen los términos que aparecen en la memoria de esta tesis. En el apéndice B, UMLS, se detallan los tipos semánticos y relaciones de la Semantic Network de UMLS. En el apéndice C, Implementación, se detalla cómo se ha desarrollado el proceso y su modelo de datos.

El último apartado de esta memoria destaca la bibliografía más importante utilizada en este trabajo.

Capítulo 2

Fundamentos

En este capítulo, revisaremos los fundamentos metodológicos que se han tenido en cuenta en esta tesis doctoral. Primero, veremos los elementos principales del problema de interoperatibilidad, empezando por su descripción y los problemas que supone. A continuación, expondremos qué es y de qué se compone una terminología, cómo se representa y las principales terminologías del ámbito médico. También analizaremos las principales técnicas de equiparación de terminologías y presentaremos varios ejemplos de su aplicación. Por último, expondremos las dificultades al realizar la evaluación de resultados.

2.1. Interoperabilidad

En general, podemos definir la interoperabilidad como la habilidad de un sistema para trabajar con otros sin un esfuerzo especial por parte del usuario y sin ninguna restricción de acceso. Por tanto, es una cualidad de gran importancia en el escenario actual, donde los ordenadores forman ya parte de nuestra vida. En general, hay dos formas de conseguirla:

- implementando los estándares publicados para la descripción de la interfaz de comunicación, o
- haciendo uso de un intermediario que convierta la interfaz de un producto en la de otro cada vez que sea necesario.

Un buen ejemplo de la primera aproximación es el conjunto de estándares que fueron desarrollados para la Web, como protocolos físicos (TCP/IP o HTTP) y lenguajes de marcado (HTML). El segundo tipo es ejemplificado en CORBA, que permite el desarrollo de sistemas distribuidos facilitando la invocación de métodos remotos bajo un paradigma orientado a objetos.

Si dos o más sistemas son capaces de comunicarse e intercambiar datos, se dice que tienen **interoperabilidad sintáctica**. Para conseguirla, son fundamentales las especificaciones de los formatos de datos y los protocolos de comunicación, como, por ejemplo, los estándares XML y SQL. La interoperabilidad sintáctica implica tanto un formato de datos común como un protocolo común para evitar las ambigüedades. Es la base para conseguir un mayor nivel de interoperabilidad.

Por otro lado, la **interoperabilidad semántica** es la capacidad de interpretar la información intercambiada de forma automática y con precisión a fin de producir resultados útiles según lo definido por los usuarios finales. Para lograrla, lo ideal es que ambas partes se sometan a un modelo de referencia común de intercambio de información. El contenido de las solicitudes está así inequívocamente definido, lo que se envía es lo mismo que lo que se entiende. Dentro de este concepto general, podemos distinguir dos tipos de aplicaciones principales:

- Las referidas a la comunicación entre personas y sistemas (interoperabilidad máquina-humano).
- Aquellas aplicaciones en las que la comunicación tiene lugar entre dos o más sistemas que intercambian información (interoperabilidad entre sistemas).

Para conseguir interoperabilidad semántica, es necesario que los datos y los metadatos estén estructurados. Cuando no es posible tener un marco común, se pueden establecer diferentes grados de interoperabilidad, al interpretar correctamente sólo una parte de la información transferida. En este caso, se establecen las correspondencias entre todas las entidades posibles de los entornos a interoperar.

¿Es posible el intercambio y la "comprensión"de la información entre distintos servicios públicos de salud, entre diferentes sistemas de información geográfica o entre gobiernos que no usan las mismas aplicaciones informáticas? Estas son algunas cuestiones a las que da respuesta la interoperabilidad semántica que, en un sentido más amplio, se refiere a la capacidad para compartir significado sin ambigüedades.

Alcanzar la interoperabilidad semántica es una tarea compleja, que afecta a múltiples niveles y funciones de los sistemas y que es objeto de estudio en ámbitos muy diferentes, como las bases de datos, equiparación de ontologías o tesauros, equiparación de catálogos (comercio electrónico), comunicación de agentes, integración de servicios web, ... En muchos casos, la implantación de un modelo común de alto nivel que defina los conceptos usados en modelos más específicos ha resultado imposible.

Debido a que la falta de interoperabilidad ocasiona costes muy altos, hay gran urgencia en mejorarla y son variados los esfuerzos que se están llevando a cabo [WPJ+].

2.1.1. Interoperabilidad en el ámbito médico

Las nuevas tecnologías se han introducido en los hospitales y laboratorios a un ritmo cada vez mayor y muchas de estas innovaciones tienen el potencial para interactuar sinérgicamente si se pueden integrar con eficacia. La información en el ámbito médico es enormemente compleja y cubre muy diferentes tipos de datos: administración de pacientes, información organizativa, datos clínicos y de laboratorio, enfermedades, medicamentos, etc ... De ahí que hayan surgido diferentes modelos para representar la información, en función de las necesidades concretas. Pero, para un acceso correcto a diferentes fuentes, es necesario buscar la interoperabilidad entre ellas.

La ausencia de interoperabilidad implica, por ejemplo, que los hospitales se ven obligados a recurrir a los grandes fabricantes que ofrecen suites de dispositivos compatibles, pero que no se especializan en un área. Además el intercambio de información entre diferentes servicios y diferentes hospitales puede hacerse difícil al ser sus sistemas diferentes. Resolver el problema de la interoperabilidad redundará en múltiples beneficios para los diferentes agentes implicados en el sistema sanitario:

- Los profesionales médicos de diferentes servicios podrán visualizar y acceder de forma integrada la información de los pacientes, lo que redundará en una mejor atención.
- Los profesionales médicos de diferentes hospitales podrán compartir información relevante sobre las enfermedades tratadas lo que potenciará la investigación.
- La innovación en el ámbito industrial se fomentará. Las pequeñas empresas podrán entrar en competencia y hacer productos especializados compatibles con los demás.
- Todo esto redundará en una mejora de la calidad del servicio para los pacientes.

Como ya comentamos en el capítulo de introducción, en la base de los sistemas biomédicos están los vocabularios controlados, terminologías u ontologías donde se reflejan todos los conceptos del dominio. En este caso, la interoperabilidad pasa por equiparar dichos vocabularios, para poder compartir la información ligada a ellos. Ese es el objetivo de nuestro método.

2.2. Terminologías

2.2.1. Definición

Una terminología es el vocabulario especial 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.

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 [Cim98], [RMJ+08]:

- **Contenido**: La terminología ha de incluir todos los términos necesarios para el desarrollo de la actividad (concept coverage), que han de ser exactos (term accuracy) y expresivos (term expressivity). 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 incambiable. 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 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 las decisiones y sistemas de recuperación de información. La terminología ha de modelar cuatro grandes tipos de funciones [Rec98]:

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

Cada función requiere diferentes formas de conocimiento, que se agrupan en 3 niveles [FCJ94]:

- 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, 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 el nombrado 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 como se enlazan las expresiones lingüísticas a los conceptos.

Se han desarrollado diferentes normas internacionales con el fin de hacer comprensibles y compatibles diferentes recursos tanto a nivel nacional como internacional. En este marco, la International Standard Organization (ISO) tiene como propósito fundamental promover a nivel mundial el desarrollo de la normalización con el objetivo de permitir el intercambio internacional de información en diferentes sectores. Entre ellos, se han publicado dos estándares relacionados con las terminologías:

- **ISO 2788** [Iso86] Guidelines for the establishment and development of monolingual thesauri. (Directrices para el establecimiento y el desarrollo de tesauros monolingües.) Esta norma abarca algunos aspectos de la selección de términos de indexación, los procedimientos para el control del vocabulario y, específicamente, el modo de establecer relaciones entre estos términos (particularmente aquellas relaciones que a priori se utilizan en los tesauros), así como la inclusión y supresión de términos, los métodos de compilación, la forma y el contenido de los tesauros, el uso de la automatización en el procesamiento de los datos, etc. Las indicaciones establecidas en esta norma aseguran una práctica uniforme en cada una de las áreas o entidades de indexación. Las técnicas descritas en su contenido se basan en principios generales que se aplican a cualquier materia.

- **ISO 5964** [Int85] Documentation for the establishment and development of multilingual thesauri. (Guía para el establecimiento y desarrollo de tesauros multilingües.) Las reglas ofrecidas en esta norma deben utilizarse en conjunto con la anterior, pues la mayoría de los métodos y recomendaciones consideradas en ésta son igualmente válidas para los tesauros multilingües. Esta norma se considera como un paso fundamental en el perfeccionamiento de la recuperación de la información y el logro de la compatibilidad entre los tesauros producidos por instituciones que indexan con términos seleccionados a partir de más de dos lenguajes naturales (idiomas). Su contenido abarca los problemas que

Figura 2.1: Esquema de relación entre término y concepto

pueden surgir durante la creación de un tesauro convencional. Incluye grados de equivalencia y no equivalencia de términos, equivalencia uno a muchos y ejemplos de visualización de tesauros.

Otro ejemplo es la ANSI/NISO Z39.19 [Nat05] Guidelines for the Construction, Format and Management of monolingual controlled vocabularies, cuyo objetivo principal es alcanzar consistencia en la descripción de los objetos para facilitar la recuperación. Incluye como formular descriptores, establecer relaciones entre términos y presentar información en pantalla.

2.2.2. Término, concepto y relaciones

Como decíamos, las terminologías están formadas por palabras o frases que representan el dominio. Estas palabras, también llamados términos o descriptores, se usan para dar nombre a los conceptos de ese dominio. 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. Esto está representado en la figura 2.1.

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 (IsA) establece el enlace entre una clase y sus miembros. Un ejemplo sencillo es [narrower term] IsA [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.

2.2.3. Terminologías en el ámbito biomédico

Desde hace más de 100 años, varias entidades en varios países han recopilado información médica, artículos, libros, ... y creado grandes bibliotecas. Posteriormente, surgió la necesidad de organizar todo ese conocimiento para facilitar el acceso y la búsqueda de información. Por ello, en los últimos 40 años, han sido varias las instituciones, sobre todo, americanas, que han desarrollado sus propias terminologías.

Algunas de las terminologías y proyectos más importantes han sido desarrollados por la Biblioteca Nacional de Medicina (National Library of Medicine, NLM) de los Estados Unidos. Esta es la mayor biblioteca médica del mundo, con más de 200 años de antigüedad, y donde se han realizado grandes esfuerzos para indexar y catalogar artículos de diversas publicaciones médicas. Es posible consultarla vía web en la dirección[1].

[1] http://www.ncbi.nlm.nih.gov/

A continuación, hacemos una breve reseña de algunas de las terminologías más importantes del ámbito biomédico.

MeSH

MEDLINE es posiblemente la base de datos de bibliografía médica más amplia que existe. Ha sido producida por el NLM y recoge más de 19 millones de referencias bibliográficas de los artículos publicados en unas 5.200 revistas médicas desde 1949 en Estados Unidos y más de 70 países, el 90 % en inglés. Cubre todos los campos de la medicina en sentido amplio: ciencias de la vida, ciencias conductuales (Sociología, Psicología, ...), ciencias químicas, bioingeniería, clínica, salud pública, políticas sanitarias y actividades educativas relacionadas con la salud. Cada registro de MEDLINE es la referencia bibliográfica de un artículo científico publicado en una revista médica, con los datos bibliográficos básicos de un artículo (título, autores, nombre de la revista, año de publicación) que permiten la recuperación de estas referencias posteriormente en una biblioteca o a través de software específico de recuperación. MEDLINE se actualiza diariamente a un ritmo de 2.000-4.000 nuevas referencias por día. PubMed es un motor de búsqueda de libre acceso a MEDLINE. Actualmente reúne más de 15.000.000 citas y está en marcha un proceso para la carga paulatina de citas anteriores a 1966.

MeSH es el vocabulario controlado creado desde 1960 por el NLM para indexar, catalogar y buscar información y documentos biomédicos y relacionados con la salud en MEDLINE/PubMed y otras bases de datos de NLM. Consiste en conjuntos de términos denominamos descriptores con una estructura jerárquica que permite buscar con varios niveles de especificidad. En la versión de 2009, había más de 25000 descriptores, que incluyen una breve descripción o definición, enlaces a los descriptores relacionados y una lista de sinónimos, por lo que, en total, incluye unos 160.000 términos. Hay, además, otros descriptores que aportan información sobre los descriptores anteriores, como las características de publicación (Publication Types), que dan información sobre qué es el término, u descriptores que se usan para indexado y catalogación como Geographics, Cualificadores y Supplementary Concept Records (SCR).

MeSH se almacena en ficheros XML que es posible solicitar en la página web de la NLM[2] o también es posible buscar directamente en ella desde esa página. La figura 2.2 muestra un ejemplo de la información de MeSH para el término *abdomen* y la jerarquía a la que pertenece se muestra en la figura 2.3.

[2]http://www.ncbi.nlm.nih.gov/mesh

MeSH Heading	Abdomen		
Tree Number	A01.047		
Annotation	GEN: prefer specifics; abdom muscles = ABDOMINAL MUSCLES but RECTUS ABDOMINIS is available; abdominal pain = ABDOMINAL PAIN; abrupt dis requiring emerg surg = ABDOMEN, ACUTE		
Concept 1 (Preferred)	**Abdomen**		
	Concept UI	M0000005	
	Scope Note	That portion of the body that lies between the THORAX and the PELVIS.	
	Semantic Type	T029 (Body Location or Region)	
	Term (Preferred)	Abdomen	
		Term UI	T000012
		Date	01-JAN-1999
		Lexical Tag	NON
		Thesaurus	NLM (1966)
Allowable Qualifiers	AB AH BS EM GD IR MI PA PH PP PS RE RI SU US VI		
Entry Combination	injuries:Abdominal Injuries		
Entry Combination	radiography:Radiography, Abdominal		
Date of Entry	19990101		
Unique ID	D000005		

Figura 2.2: Información sobre abdomen en MeSH

SNOMED CT

SNOMED es una nomenclatura médica desarrollada por el CAP (College of American Pathologists), desarrollada desde 2002, cuya extensa terminología basada en conceptos permite la captura de datos clínicos y la recuperación de información médica, la agregación de datos y aplicaciones de mensajería electrónica utilizando estándares internacionales. Diseñada para el desarrollo de aplicaciones de historias clínicas electrónicas, análisis de resultados clínicos y apoyo en la toma de decisiones médicas, SNOMED incluye información de enfermedades, síntomas, procedimientos, microorganismos, medicamentos, entre otros. Es considerada la terminología más completa, está disponible en varios idiomas y su uso es promovido por IHTSDO (International Health Terminology Standard Development Organization) [IHT] una organización sin fines de lucro, que la adquirió en 2007.

Actualmente, dispone de más de 300.000 conceptos médicos, con signi-

Figura 2.3: Jerarquía de abdomen en MeSH

ficados únicos y definiciones basadas en lógica formal organizados en jerarquías, con múltiples niveles de granularidad. Hay también más de 800.000 descripciones, incluyendo sinónimos que pueden ser usados para referirse a un concepto. Contiene aproximadamente 1.360.000 enlaces o relaciones semánticas entre sus conceptos, que proporcionan definiciones formales y otras características del concepto. Entre los tipos de relaciones, encontramos entre otras:

- **is-a**: permite definir la posición de un concepto en la jerarquía al indicar con A Is-A B que todas las instancias de A son también instancias de B. Por ejemplo, *Myocardial infarction* is-a *Myocardial disease*.
- **finding-site**: identifican las partes del cuerpo afectadas por una enfermedad y un procedimiento. Por ejemplo, *Injury of cornea* has finding-site *Corneal estructure*.
- **causative-agent**: permite indicar la relación de causa-efecto entre dos conceptos, por ejemplo, la causa de una enfermedad.

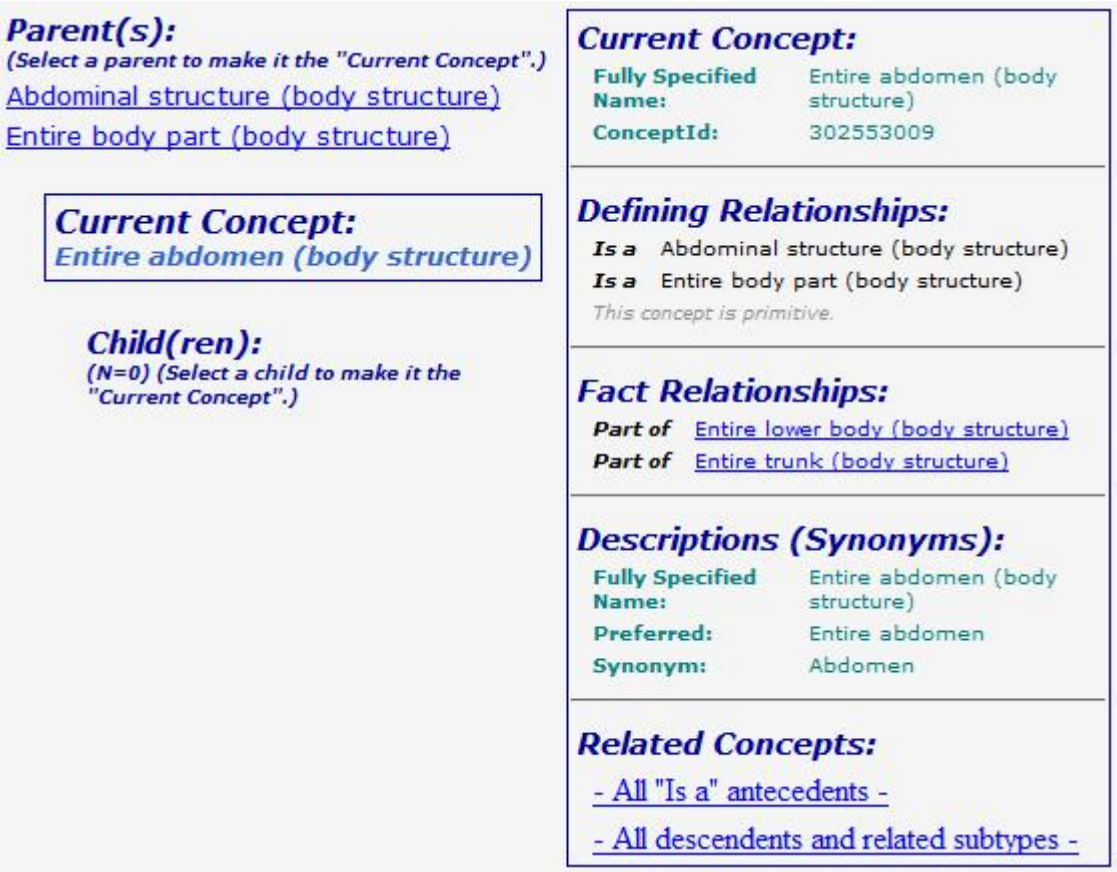

Figura 2.4: Entire abdomen en SNOMED CT

Por ejemplo, al buscar el termino *abdomen* se obtienen muchos términos como *Abdominal structure*, *Entire abdomen*, *Abdomen soft*, *Acute abdomen*, *Lower abdomen structure*, *Upper abdomen structure*, *Obese abdomen*, ... En la figura 2.4, mostramos el término *Entire abdomen*.

SNOMED CT trabaja para proporcionar enlaces explícitos (crossmaps) a otras clasificaciones relacionadas con la salud en uso, como por ejemplo, clasificaciones de diagnóstico (ICD-9-CM, ICD-10) o de intervenciones (OPCS-4). Es posible conseguirla tras suscribirse en la web del IHTSDO.

LOINC

El propósito de LOINC ® es facilitar el intercambio de los resultados clínicos proporcionando un conjunto de códigos universales y nombres para identificar principalmente las observaciones de laboratorio. El Instituto Regenstrief, Inc y la organización de la investigación informática mantienen la

base de datos LOINC, documentación de apoyo, y el programa de cartografía RELMA [3]. Desde su inicio en 1996, ha tenido buena acogida. Actualmente cuenta con más de 30.000 conceptos y está siendo utilizado por cada vez más entidades diferentes del ámbito de la salud (hospitales, asociaciones médicas, ...). También se está realizando una integración con SNOMED para asegurar una terminología de referencia clínica consistente y no ambigua construida sobre las fortalezas de cada uno. Está disponible en fichero Access 2007, fichero txt y pdf.

UMLS

El UMLS (Unified Medical Language System), también del NLM, [Bet05], [Bod04] es el resultado de desarrollar un método de integración de vocabularios médicos para solventar los dos inconvenientes principales de la recuperación de información: la variedad de nombres usados para expresar el mismo concepto y la ausencia de un formato estándar para distribuir terminologías. Incluye más de 130 terminologías de diferentes temáticas: resultados de laboratorio (LOINC), procedimientos y sistemas médicos (SNOMED, PT), enfermedades (MCA/MR, ICD), drogas, anatomía, genética (OMIM), enfermería, conceptos y otros muchos, hasta cubrir el dominio médico. En la figura 2.5, se muestra un esquema de los subdominios que contiene. El proceso de integración de nuevas terminologías continúa aún hoy, mientras el modelo se mantiene estable. Actualmente tiene 1.500.000 conceptos y más de 3 millones de términos, almacenados en ficheros planos pero con una estructura claramente definida.

NLM proporciona y distribuye los recursos UMLS (UMLS Knowledge Sources):

- **Metathesaurus**: es un gran vocabulario multipropósito y multilingüe que contiene conceptos biomédicos y del ámbito de la salud, sus nombres y las relaciones entre ellos. Incluye conceptos (que son significados específicos), términos (que son varios nombres sinónimos para conceptos de diferentes vocabularios, de los que se escoge uno como término preferido) y cadenas (que son las variantes léxicas para las cadenas). En la figura 2.6, se muestran para el concepto *Atrial Fibrilation* todos los elementos disponibles en UMLS: los conceptos (el preferido y los sinónimos), los términos, las cadenas y los átomos, cada uno con su correspondiente identificador. Los términos de diferentes vocabularios se relacionan entre sí de dos formas: como sinónimos, dentro del concepto, y a través de las relaciones entre conceptos. También incluye

[3] http://loinc.org/relma

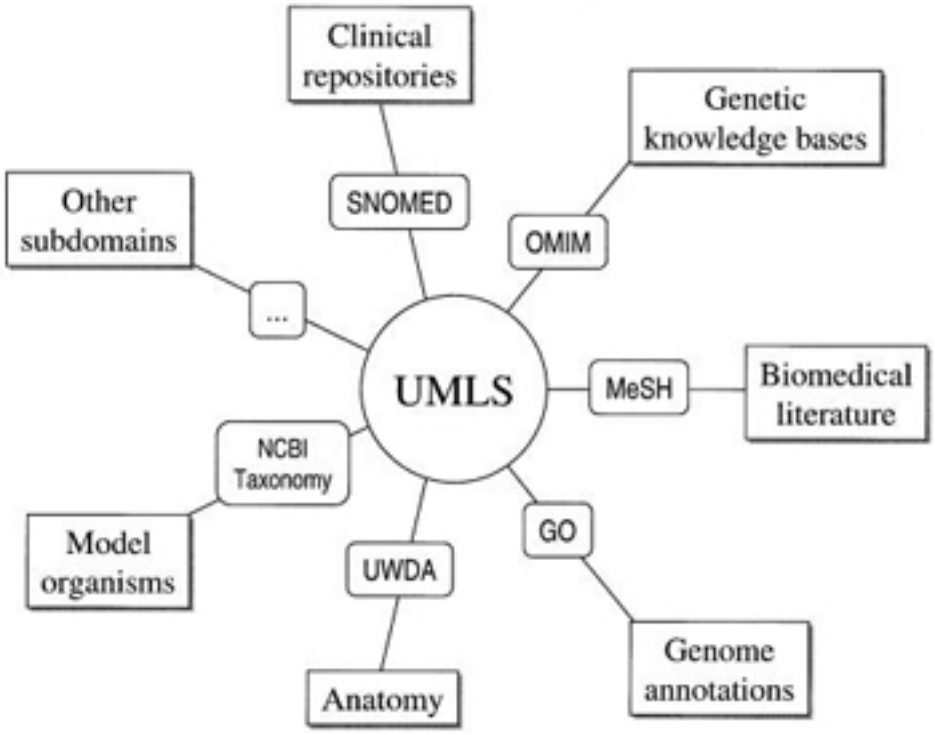

Figura 2.5: Subdominios integrados en UMLS [Bod04]

relaciones entre conceptos, que pueden ser heredadas de las terminologías originales o generadas específicamente por los editores. Pueden ser jerárquicas (como 'es-un', 'es-un-tipo-de', 'es-parte-de') o asociativas ('causado-por', 'localización-de'). También se han incluido las relaciones estadísticas de MeSH. Con todo ello, el significado de cada término viene dado por su origen, su definición o anotaciones, su contexto (su lugar en la jerarquía), sus sinónimos y sus relaciones con otros términos. Por último, podemos encontrar referencias a recursos externos que han sido heredades de algunas terminología externas. Por ejemplo, en MeSH se incluyen proteínas que tienen el identificador de GenBank.

- **Semantic Network (SN)**: proporciona una categorización consistente de todos los conceptos del Metathesaurus y un conjunto de relaciones útiles entre esos conceptos. Está formada por 135 tipos semánticos a los que se le asigna los conceptos, y 54 relaciones. Los tipos son los nodos de la red y las relaciones, los enlaces entre ellos. Esta red representa el dominio médico. Ejemplos de tipos y relaciones se pueden ver en las figuras 2.7 y 2.8. La lista completa de tipos semánticos y relaciones están en el apéndice B.

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 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 2.6: Concepto en UMLS [oMb]

- **SPECIALIST Lexicon**: es un lexicon (conjunto de palabras de un idioma y sus variaciones) en inglés de términos médicos. Cada entrada incluye información sintáctica, morfológica y ortográfica usada por técnicas de procesado de lenguaje natural. Contiene la forma básica del término, sus variantes morfológicas (género, número, ...), sus abreviaciones y acrónimos. También contiene una lista de morfemas del Latín y el Griego que son muy habituales en términos médicos.

El Metathesaurus refleja y preserva los significados, los conceptos y las relaciones de los vocabularios originales. Cuando dos vocabularios originales usan el mismo nombre para conceptos diferentes, Metathesaurus representa ambos significados e indica de qué vocabulario procede. Cuando el mismo concepto aparece en diferentes contextos jerárquicos en diferentes vocabularios, se incluyen todas las jerarquías. Cuando aparecen conflictos en la relaciones entre dos conceptos de diferentes vocabularios, se incluyen ambas vistas en el Metathesaurus. Por tanto, no representa una ontología global de biomedicina ni una sola vista consistente del mundo, sino que preserva los puntos de vista de los vocabularios origen, que pueden ser útiles para diferentes tareas.

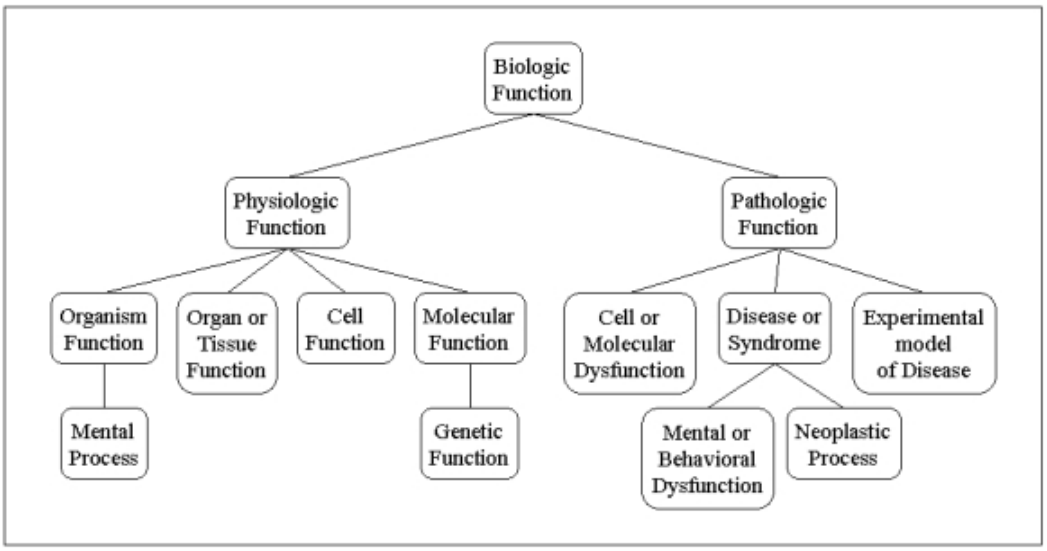

Figura 2.7: Ejemplo de tipos semánticos de la Semantic Network(UMLS) [oMc]

Todos estos recursos se hallan enlazados entre sí. Todos los conceptos del Metathesaurus tienen asignado, al menos, un tipo semántico de la Semantic Network. Muchas palabras o términos de varias palabras presentes en conceptos y cadenas en el Metathesaurus también aparecen en el Specialist Lexicon.

La Semantic Network reduce la complejidad del Metathesaurus al agrupar los conceptos en tipos semánticos. Sin embargo, para ciertos propósitos, 135 tipos son muchos y es mejor un conjunto más pequeño y general de agrupaciones semánticas [MBB01]. Para diseñarlas, se usaron los principios de validez semántica, parsimonia, integridad, exclusividad, naturalidad y utilidad. Esto dio lugar a sólo 15 grupos semánticos que incluye el 99.5 % de los conceptos. Algunos conceptos quedan en más de un grupo; en particular, 4913 están en más de 2 grupos de los cuales 16 están en 3 de ellos. Los grupos semánticos resultantes se muestran en la tabla 2.1 (versión 2003A). En la figura 2.9, se puede ver una representación parcial del resultado final.

Los recursos anteriores están accesibles en CD-ROM o vía web [4]. Se puede instalar en la máquina del usuario y acceder con consultas SQL o a través de una interfaz (Metamorphosys). Hay disponible, además, una API para que los desarrolladores creen sus propios programas. También se pueden acceder a través de internet [5]. El acceso es gratuito, los usuarios sólo han de firmar un

[4] http://www.nlm.nih.gov/research/umls/
[5] https://uts.nlm.nih.gov/home.html

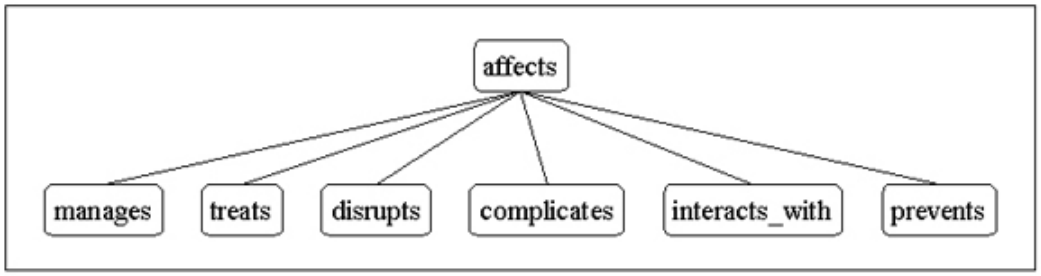

Figura 2.8: Ejemplo de relaciones de la Semantic Network(UMLS) [oMc]

acuerdo de licencia para acceder a ellos. A continuación se detallan algunas de estas herramientas.

- **UMLSKS**: es un conjunto de herramientas web interactivas y una interfaz que permite a usuarios y desarrolladores acceder a los recursos anteriores. Así, se puede realizar la consulta de cualquier término médico y devolverá todos los conceptos que haya en el Metathesaurus relacionados con él.
- **Metamorphosys**: programa que permite instalar y configurar UMLS en cualquier ordenador. Para ello, es necesario bajarse los ficheros del Metathesaurus, que ocupan actualmente más de 1G.
- **MetaMap**: accesible vía web o por API, permite comparar texto biomédico al Metathesaurus, descubriendo los conceptos referidos en el texto. Para ello, usa técnicas de proceso de lenguaje natural (NLP) y técnicas linguísticas computacionales. Se puede acceder vía web o descargar (MMTx).
- **SemRep**: accesible vía web por API, permite extraer relaciones definidas en la Semantic Network a partir de textos biomédicos, para lo cual usa MetaMap.
- **Herramientas léxicas**: diseñadas para ayudar en el procesamiento de lenguaje natural: un generador de variantes léxicas (LVG), un generador de cadenas normalizadas (Norm) y un generador de índices de palabras (Wordind).
- **Herramientas NLP SPECIALIST**: incluyen una serie de programas en java para ayudar a los desarrolladores en el procesado de textos médicos, manejando variaciones de las palabras que lo constituyen, así como las frases y sentencias en sí.

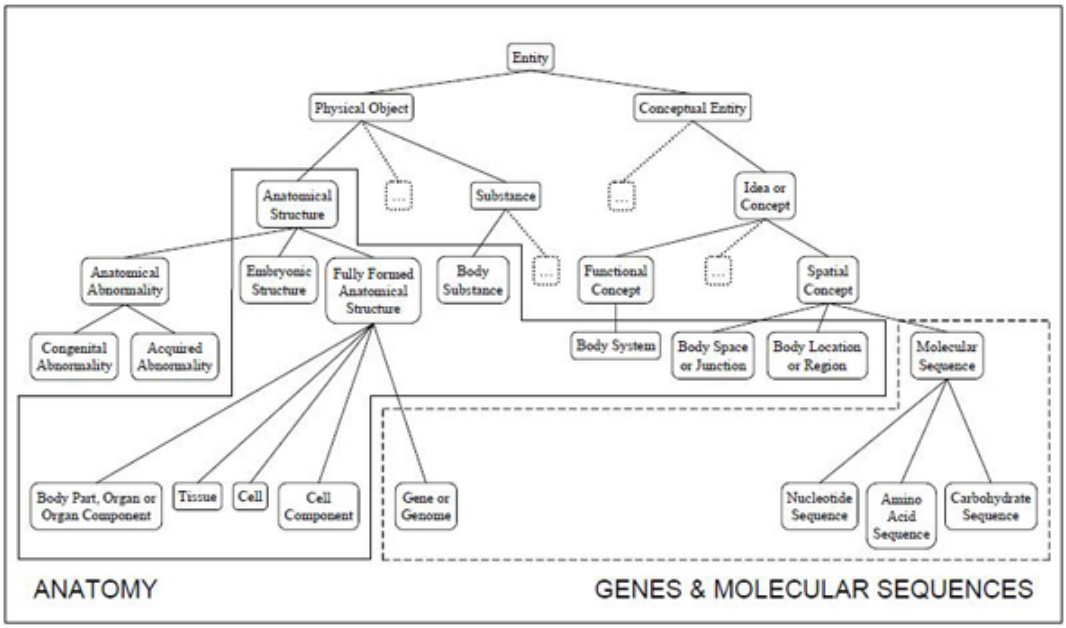

Figura 2.9: Semantic Network en grupos semánticos [MBB01]

- **BLAST (Basic Local Alignment Search Tool)**[6] [oMa]: encuentra regiones de similitud entre secuencias biológicas. Compara secuencias de proteínas o nucleótidos a bases de datos de secuencias y calcula la significación estadística de los resultados. Puede ser usado para inferir relaciones funcionales y evolutivas entre secuencias, por ejemplo, para identificar genes de la misma familia. Está disponible vía web o para descargar.

Estas herramientas son usadas por los investigadores en diferentes proyectos de investigación, algunos de los cuales veremos en apartados siguientes.

EMTREE

Esta es la terminología que empleamos en nuestro estudio. Emtree es el vocabulario controlado de ciencias de la vida y biomedicina de Elsevier, empresa líder en edición técnica en medicina y ciencias de la salud, con más de 6.000 libros y 2.000 revistas científicas y bases de datos.

Emtree [7] proporciona una descripción consistente y comprensiva de la información biomédica y facilita las búsquedas y la recuperación de información con alta precisión. En concreto, destaca su cobertura de drogas y

[6] http://blast.ncbi.nlm.nih.gov/

[7] http://www.elsevier.com/wps/find/bibliographicdatabasedescription.cws_home/707574/description

Grupos Semánticos	Nº Tipos	**Conceptos UMLS**	
		No	%
Activities and Behaviors	9	3224	0.4 %
Anatomy	11	34386	4.7 %
Chemical and Drugs	26	356211	48.8 %
Concepts and Ideas	12	17639	2.4 %
Devices	2	31092	4.3 %
Disorders	12	136389	18,7 %
Genes and Molecular Sequences	5	904	0.1 %
Geographic Areas	1	949	0.41 %
Living Beings	23	29699	4.1 %
Objects	5	6857	0.9 %
Occupations	2	890	0.1 %
Organizations	4	2124	0.3 %
Phenomena	6	4943	0.7 %
Physiology	9	27930	3.8 %
Procedures	7	81847	11.2 %
Total	134	735084	100.6 %

Tabla 2.1: UMLS:Agrupación de tipos semánticos

terminología médica. Contiene, al menos, 56.000 términos preferidos, 27.000 sobre drogas y medicamentos, (incluyendo todos los términos de MeSH), más de 230.000 sinónimos. Tiene estructura jerárquica, con broaders y narrowers, de hasta 12 niveles y consta de:

- **Índice alfabético:** Es la lista de todos los términos referidos a drogas y medicamentos, con códigos Emtree.
- **Estructura de árbol:** Es una lista jerárquica donde los términos son repartidos en 15 facets, como *Anatomical concepts*, *Organism names*, *Chemical and Drugs* o *Geographic names*.
- **Índice de términos permutados:** Proporciona un índice permutado de cada palabra contenida en Emtree, lo que permite facilitar las búsquedas y localizar términos de los que no se está seguro cómo se escriben.

Emtree ha sido usado desde 1974 por EMBASE, la base de datos biomédica de Elsevier, con 4.000 revistas activas. Contiene registros bibliográficos con citas, resúmenes e índices de artículos biomédicos, con un énfasis especial en drogas y medicamentos, que resulta de gran utilidad para los profesionales.

Tanto EMBASE como Emtree, se están actualizando continuamente y están disponibles vía web o también pueden ser obtenidos para uso por el usuario.

OMIM

OMIM (Online Mendelian Inheritance in Man) es un catálogo comprensivo de genes y de todas las enfermedades genéticas conocidas. Fue iniciada en 1966 por un médico y en 1985 se hizo la versión online con la colaboración de la NLM. En 2008, contaba con unos 18.900 entradas de texto y es actualizada continuamente, tras revisar las últimas publicaciones, a ritmo de 70 nuevas entradas y 700 modificadas cada mes. Se puede consultar en la web de la NLM [8].

OMIM tiene un entrada, con un número único, por cada gen o enfermedad sobre el que exista información, aunque no sea amplia. Cada entrada incluye el nombre del gen y su símbolo, junto con nombres y símbolos alternativos, descripción, localización cromosómica, métodos de clonación, herencia de patrones y variantes de alelos, entre otros campos. Las manifestaciones clínicas causadas por mutaciones y enfermedades son incluidas en la sección CS (Clinical Synopsis). Varios campos están en formato de texto libre, no estructurado. La información se añade incrementalmente, nunca se elimina lo ya escrito, si no que se añade en orden cronológico. También incluye referencias a otros recursos, como MEDLINE. Sus términos están incluidos en UMLS y SNOMED.

2.3. SKOS

Las terminologías vistas en el apartado anterior han sido creadas por equipos diferentes y, por tanto, su estructura e implementación será muy diferente. Sin embargo, para aumentar las posibilidades de interoperabilidad, lo ideal hubiera sido que estuvieran implementadas de la misma forma.

SKOS (Simple Knowlegde Organization System) [W3C], [PS] es una iniciativa del W3C que 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, folksonomías y otros vocabularios controlados similares. Por tanto, su objetivo es mejorar la funcionalidad e interoperabilidad de la Web. Al tratarse de una aplicación de RDF (Resource Description Framework) [9], SKOS permite

[8] http://www.ncbi.nlm.nih.gov/omim
[9] http://www.w3.org/RDF/

la creación y publicación de conceptos en la Web, así como su vinculación con datos en este mismo medio e incluso su integración en otros esquemas de conceptos.

Inicialmente, W3C desarrolló RDF, que proporciona una abstracción y sintaxis de datos para la web, OWL (Web Ontology Language), que da un lenguaje para modelado de datos, y SPARQL, un estándar para interacción de datos en la web. Estas tecnologías se están usando pero se vio que hacía falta organizar la gran cantidad de información no estructurada que hay en Internet y, de entrada, la creación de una ontología requiere la creación de un mapa detallado del dominio, lo cual no puede ser hecho automáticamente y es costoso. SKOS puede ser visto con una tecnología puente, entre el formalismo lógico riguroso de los lenguajes de ontologías y el caótico, informal y poco estructurado mundo de las herramientas de colaboración basadas en web. Su objetivo no es reemplazar los vocabularios existentes, sino permitir portarlos a un espacio común.

SKOS proporciona un camino estándar y de bajo coste para migrar los sistemas de conocimiento existentes a la Web Semántica, ya que aporta un lenguaje intuitivo y ligero. De esta forma, bibliotecas, museos, periódicos, portales de gobierno, empresas, aplicaciones de redes sociales y otras comunidades que manejan grandes colecciones de libros, noticias, glosarios, blogs y otros artículos pueden ahora utilizar SKOS para facilitar su búsqueda y aprovechar el poder de la vinculación de datos.

La primera versión se presentó en 2003, hubo varios borradores y la última es de agosto de 2009, que ya es un documento estable y puede usarse como referencia. Los documentos que ofrece la W3C [10] son una guía completa (SKOS Reference), una guía básica (SKOS Primer) y una lista de casos de uso y recopilación de peticiones (SKOS Use Cases and Requirements).

En SKOS básico, los conceptos están identificados con URIs, etiquetados con cadenas en uno o más lenguajes naturales, documentados con varios tipos de propiedades, semánticamente relacionados a otros en jerarquías informales y redes asociativas, y agregados en esquemas de conceptos.

En SKOS avanzado, los conceptos pueden ser mapeados a través de los esquemas de conceptos y agrupados en colecciones etiquetadas y ordenadas. Se pueden especificar relaciones entre etiquetas de concepto. Finalmente, el vocabulario SKOS puede ser ampliado en función de las necesidades de comunidades particulares o combinadas con otros vocabularios ya modelados.

[10] http://www.w3.org/2009/08/skos-reference/skos.html

2.3.1. Modelo de datos de SKOS

El modelo de datos SKOS está definido formalmente como una ontología OWL Full. Los datos son expresados con tripletas RDF y pueden ser codificadas usando sintaxis RDF. El sistema se ve como un esquema de concepto que comprende un conjunto de conceptos.

A continuación, se detallan los elementos principales del modelo SKOS.

Conceptos y esquemas de conceptos

Un concepto SKOS se representa con la clase **skos:Concept**, que es el elemento fundamental del vocabulario SKOS y puede ser visto como una idea o noción, una unidad de pensamiento, aunque como tal, puede ser subjetiva. Permite describir la estructura conceptual o intelectual del sistema y, por tanto, es independiente de los términos que se usen para definirlo.

También se permite definir un esquema de conceptos, **skos:ConceptScheme**, que es una agregación de uno o más conceptos SKOS, incluyendo las relaciones semánticas entre ellos, y generalmente se usan para representar o identificar las terminologías. Un concepto se asocia a un esquema con **skos:inScheme**. Un esquema puede tener uno o más conceptos cabecera **skos:hasTopConcept** que son aquellos conceptos que encabezan las estructuras jerárquicas dentro del esquema. Para los usuarios, éstos constituyen los puntos iniciales de búsqueda.

Etiquetas léxicas

Una etiqueta léxica es una cadena de caracteres UNICODE en un lenguaje natural dado, como inglés o japonés. SKOS permite la distinción entre las etiquetas preferente **skos:prefLabel**, alternativa **skos:altLabel** y oculta **skos:hiddenLabel** para cualquier recurso dado. Las etiquetas preferente y alternativa son útiles cuando se generan o crean representaciones legibles de un sistema de organización del conocimiento.

La etiqueta preferente se asocia a los conceptos para representar el término que será usado como descriptor. Sólo se admite una etiqueta preferente en cada idioma a un concepto.

Las etiquetas alternativas permiten asignar múltiples expresiones no preferentes a un concepto, para sinónimos, acrónimos, ... Estas expresiones enriquecen el vocabulario al permitir un mayor número de puntos de acceso al concepto. Otro ejemplo de etiquetas preferentes serían las variantes con errores ortográficos de otras etiquetas.

Las etiquetas ocultas son útiles cuando un usuario está interactuando con un sistema de organización del conocimiento a través de una función

de búsqueda basada en texto. El usuario puede, por ejemplo, introducir las palabras mal escritas cuando se trata de encontrar un concepto relevante. Así, la consulta mal escrita puede ser comparada con una etiqueta oculta y el usuario será capaz de encontrar el concepto relevante.

Notaciones

Los literales tipificados, esto es, una cadena UNICODE combinado con un tipo de dato URI, son usados comúnmente para denotar valores como interés, números decimales y fechas. En otras situaciones, puede ser necesario que el usuario defina sus propios tipos de datos. Una notación, **skos:notation**, se usa solamente con un literal tipificado donde el tipo de dato URI ha sido definido por el usuario. De esta forma, se asocia un concepto a su correspondiente entrada dentro de la terminología, clasificación o esquema en el que se identifican los elementos con determinados códigos o signaturas. Un mismo concepto puede tener varias notaciones aunque una notación solo debería ser asignada a un único concepto.

Documentación

Las notas permiten incluir información informal relativa a los conceptos SKOS. No hay restricción en la naturaleza de esta información, puede ser texto, hipertexto o una imagen.

Hay 7 propiedades en SKOS para asociar notas a conceptos, que, aunque pueden no cubrir todas las situaciones, sí son útiles en la mayor parte de ellas. Además, incluyen una serie de extensiones para definir tipos más específicos de notas. Son:

- **skos:note** se usa para propósitos generales. De él derivan todos los demás.
- **skos:changeNote** documenta cambios en el concepto, para administración y mantenimiento.
- **skos:definition** da una explicación completa del significado del concepto.
- **skos:editorialNote** suministra información para tareas administrativas de edición y publicación.
- **skos:example** permite incluir ejemplos de uso del concepto.
- **skos:historyNote** describe cambios significativos en el significado o la forma del concepto.

- **skos:scopeNote** suministra alguna, posiblemente parcial, información sobre el significado del concepto, especialmente de cómo usarlo en un determinado campo de uso, por ejemplo, en procesos de indexación.

Las 6 últimas son subpropiedades de la primera, skos:note.

Relaciones semánticas

Las relaciones semánticas juegan un papel muy importante en la definición de conceptos, ya que el significado de un concepto no se define sólo por las palabras en sus etiquetas sino también por sus enlaces con otros conceptos. Son enlaces entre conceptos SKOS, donde el enlace es inherente al significado de los conceptos enlazados.

SKOS suministra 3 propiedades estándar, que reflejan las dos categorías básicas de relaciones (jerárquicas y asociativas):

- **skos:broader** y **skos:narrower** permite la representación de enlaces jerárquicos, como la relación entre un género y sus especies o, según la interpretación, un todo y sus partes. De esta forma, skos:broader se usa para enlazar a conceptos más genéricos y skos:narrower a conceptos más específicos.

- **skos:related** permite la representación de enlaces asociativos como la relación entre un tipo de evento y una categoría de entidades que participan en él.

Estas relaciones jerárquicas se definen sin la propiedad transitiva, que implica que si un concepto A es más genérico que B y B es más genérico que C, se podrá deducir que A es más genérico que C. SKOS también dispone de las relaciones transitivas, **skos:broaderTransitive** y **skos:narrowerTransitive**, con la finalidad de poder realizar inferencias e implementar algoritmos de consulta en aplicaciones de búsqueda. Por esta razón, es la relación transitiva la superior, siendo la no transitiva una subclase, tal como se ve en la figura 2.10.

La relación asociativa es una propiedad simétrica, es decir, si una concepto A tiene una relación asociativa con un concepto B, se puede deducir el concepto B la tiene con A. Sin embargo,no tiene la propiedad transitiva.

Colecciones

Las colecciones de conceptos SKOS, **skos:Collection**, son grupos ordenados y/o etiquetados de conceptos SKOS sin establecer relaciones semánticas explícitas que distorsionen la estructura jerárquica o asociativa del

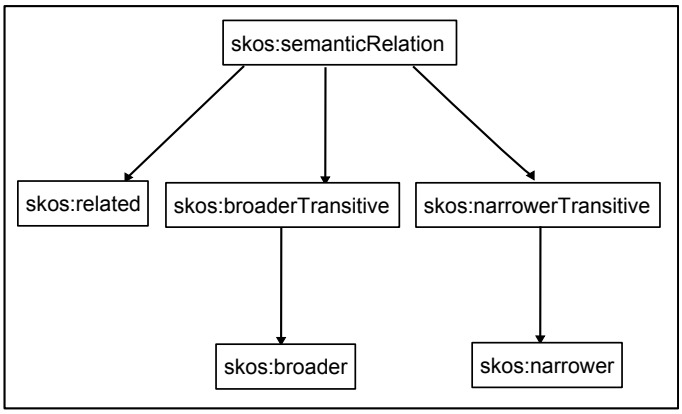

Figura 2.10: Relaciones semánticas en SKOS

modelo. Son útiles cuando un grupo de conceptos tiene alguna característica en común y es conveniente agruparlo bajo una única etiqueta. También cuando los conceptos pueden ser colocados en un orden determinado, **skos:OrderedCollection**. Para indicar que un concepto pertenece a una colección o a una lista, se usan **skos:member** y **skos:memberList**.

Los conceptos y las colecciones son disjuntos lo que no permite establecer relaciones semánticas entre ambos.

Es posible inferir los elementos de colección a partir de los elementos de una colección ordenada y también se permite incluir colecciones dentro de otras colecciones.

Propiedades de Equiparación (Mapping Properties)

Estas propiedades son usadas para establecer equiparaciones entre conceptos SKOS de diferentes esquemas de conceptos, de forma que el enlace es inherente al significado de los conceptos enlazados, igual que lo eran las relaciones semánticas.

SKOS ha definido las siguientes:

- **skos:closeMatch** enlaza 2 conceptos que son suficientemente similares como para ser intercambiados en las aplicaciones de recuperación de información. No es una propiedad transitiva.
- **skos:exactMatch** enlaza 2 conceptos que son intercambiables con una alto grado de similitud. Es transitiva y una subpropiedad de la anterior.
- **skos:broadMatch** y **skos:narrowMatch** permiten establecer relaciones jerárquicas entre 2 conceptos. Son subpropiedades de skos:broader y skos:narrower respectivamente.

- **skos:relatedMatch** se usa para establecer un alineamiento asociativo.

Estas propiedades no funcionan como las relaciones semánticas; por ejemplo, skos:broadMatch y skos:narrowMatch no implican la creación de una jerarquía, sino que sólo expresa cómo se relacionan dos conceptos de dos esquemas diferentes.

La relación entre estas propiedades y las relaciones semánticas puede verse en la figura 2.11.

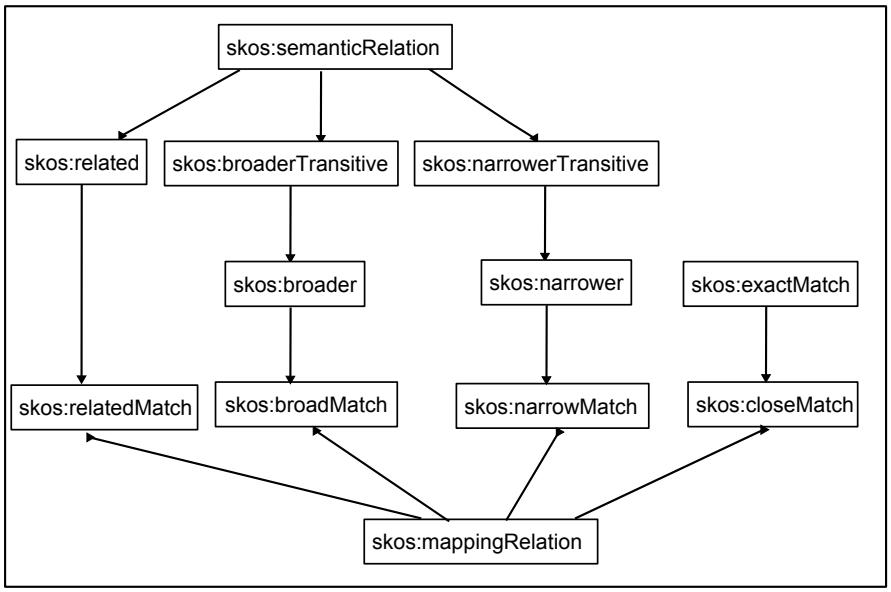

Figura 2.11: Relaciones semánticas y de correspondencia en SKOS

2.3.2. SKOS en terminologías médicas

Actualmente, existen proyectos que han usado SKOS en el ámbito biomédico aunque no se están empleando de forma extendida. Una de las aportaciones en este sentido es UMLS-SKOS[11] que propone una alternativa para representar el conjunto de conocimientos incorporados en el UMLSKS en el marco de las tecnologías de Web Semántica. También presenta una conceptualización de un algoritmo de transformación para producir una representación SKOS de UMLSKS que integra la Semantic Network y el Metathesaurus con todos los vocabularios de origen como un cuerpo unificado de conocimiento. Su propuesta se basa en la idea de que la representación formal y explícita de

[11] http://www.bioontology.org/umls-skos

cualquier cuerpo de conocimiento permite su interpretación inequívoca y precisa en los programas informáticos. Las consecuencias serían, por lo menos, tres:

- capacidad de comprobar automáticamente las inconsistencias y errores dentro de un conjunto grande y complejo del conocimiento
- la interpretación, integración y descubrimiento de información automatizada
- una mejor reutilización de la información y ampliar la base de conocimientos dentro de una comunidad distribuida y de colaboración de los investigadores.

2.4. Equiparación de terminologías

La equiparación (mapping) de terminologías consiste en identificar las correspondencias entre las entidades de esas terminologías; pero como éstas pueden diferir bastante de unas a otras, su localización es intrínsecamente problemática de automatizar.

Una primera opción para resolver el problema fue unificar las terminologías por fusión, dando lugar al UMLS, que hemos explicado en el apartado anterior. Esta aproximación se encontró con dos problemas principales [Doe01]:

- Las diferencias entre la semántica de términos y relaciones jerárquicas, y la superposición de términos dificultan su combinación. Además, esta solución es sólo semiautomática y muy costosa.
- Las terminologías suelen estar fuertemente implantadas en los sistemas que los usan y no es posible una migración a una nueva terminología.

Por tanto, el procedimiento que se impuso fue el de hallar las correspondencias entre las terminologías, su equiparación. El objetivo es reducir la heterogeneidad entre ellas, que viene dada, como ya se ha comentado, por su diferente grado de desarrollo, granularidad y objetivo. A continuación, indicamos los tipos más obvios de heterogeneidad:

- **Sintáctica**: ocurre entre dos terminologías que no están expresadas en el mismo lenguaje.
- **Terminológica**: ocurre debido a variaciones en los nombres que se refieren a las mismas entidades en ambas terminologías.

- **Conceptual**: hace referencia a diferencias en el modelado del dominio y pueden ser de diferentes tipos:
  - en **cobertura** ocurre cuando 2 terminologías describen diferentes regiones del mundo, posiblemente solapándose, en el mismo nivel de detalle o desde la misma perspectiva.
  - en **granularidad** ocurre cuando las terminologías hacen referencia a la misma región del mundo pero con diferente nivel de detalle.
  - en **perspectiva** ocurre cuando las terminologías hacen referencia a la misma región del mundo pero con diferente perspectiva. La información guardada puede ser diferente y las mismas palabras pueden usarse de diferente manera.
- **Semántica**: hace referencia a cómo las entidades son interpretadas por la gente. Este tipo es difícil de detectar por las computadoras.

La equiparación de terminologías busca básicamente encontrar las correspondencias entre ambas terminologías origen de forma que, por ejemplo, la consulta realizada en una pueda traducirse a la otra. Como vimos, podemos encontrar dos enfoques para conseguir la interoperabilidad entre las terminologías: la fusión y la alineación de las terminologías de origen. En el primer enfoque, se crea una versión única y coherente a partir de la fusión de las fuentes originales. Este enfoque fue seguido, como ya hemos comentado, para crear el UMLS y desarrollar una gran metaterminología que conciliase las diferencias de más de 130 fuentes de información biomédica. Sin embargo, esta solución es demasiado costosa con fuentes de gran tamaño, por lo que a menudo la solución más viable es mantener las fuentes originales por separado y añadir las correspondencias entre las fuentes. Esta segunda opción es, por ejemplo, la recomendada para el desarrollo de tesauros multilingües [Int85] y el seguido en muchas investigaciones en el ámbito biomédico [FBA+07], [VGHHT04], [ZMBB07], [SCG+03]. Además, gran parte del conocimiento médico se encuentra como texto sin estructura, por lo que también ha habido proyectos que han usado técnicas de procesado de lenguaje natural (NLP) para extraer el conocimiento estructurado de ellos.

A lo largo de este capítulo, se explican ejemplos de ambos métodos.

2.4.1. Definición del proceso de equiparación

El resultado del proceso de equiparación de terminologías es el conjunto de correspondencias (alineamientos) entre sus conceptos. Dado un concepto de la terminología origen, el alineamiento define qué conceptos de la terminología destino son equiparables con él. La forma más simple de definir un

alineamiento es como un par de conceptos (x,y) donde x es un concepto de la terminología origen e y es un concepto de la terminología destino.

El alineamiento óptimo, definido por Doer [Doe01], sería aquel que verifica las siguientes condiciones:

- Si un concepto c de la terminología A no tiene una equivalencia exacta con otro concepto de la terminología B, se busca, al menos, una equivalencia de un broader y una de un narrower a algún concepto en B.
- La equivalencia de broader ha de ser mínima, es decir, no debería haber ningún término en B que sea broader del término c y narrower del broader encontrado.
- La equivalencia de narrower ha de ser máxima, es decir, no debería haber ningún término en B que sea narrower del término c y broader del narrower encontrado.

Un proceso de equiparación reúne una o varias técnicas para obtener los conceptos de la terminología destino que se equiparan a los de la terminología fuente, usando para ellos los recursos computacionales disponibles en cada caso. Estas técnicas se detallan a continuación.

2.5. Técnicas para encontrar correspondencias

Independientemente del enfoque elegido para equiparar las terminologías, dos fuentes son compatibles cuando se conocen las correspondencias entre los conceptos que se solapan. El principal cuello de botella en la integración semántica [Noy04] es encontrar dichas correspondencias. El proceso manual es tedioso, lento e imperfecto (ya que la capacidad humana es limitada) y depende del operador (personas diferentes identifican correspondencias diferentes). Así, el uso de la tecnología adecuada con mayores niveles de automatización facilita, en gran medida, la tarea de encontrar las correspondencias.

En la literatura, existen varias técnicas para encontrar automáticamente correspondencias entre los conceptos que se solapan. De hecho, varias disciplinas, por ejemplo, las Ciencias de la Información ([ZC04], [W3C]), bases de datos ([DNH04], [SS06]) y ontologías ([KS03], [Noy04], [ES07]) han estado trabajando en esta tecnología, aunque la mayoría de ellos se han enfocado a fuentes de poco tamaño [DNH04].

Las técnicas de equiparación de ontologías se clasifican en función de la entrada de los algoritmos, las características del proceso (exacto o aproximado) y la salida del algoritmo. Puede usar la información que hay en la

propia terminología, los nombres y descripciones de los conceptos, y también la información estructural derivada de la relaciones entre ellos. También se pueden usar recursos externos. A continuación, explicamos las técnicas básicas.

2.5.1. Técnicas léxicas

Son las técnicas más básicas y usan las propiedades léxicas de los nombres de concepto para encontrar las correspondencias entre ellas. Algunas herramientas proporcionan técnicas de normalización, que reducen las cadenas que representan los nombres de los conceptos a un formato común, antes de buscar las equiparaciones. Los principales problemas que tienen estas técnicas son:

- **sinónimos**: palabras diferentes que representan el mismo concepto. En algunas terminologías, ya vienen incluidos los sinónimos dentro de la definición de un concepto.

- **homónimos**: la misma palabra representa diferentes conceptos. En este caso, la única manera de diferenciarlo es a través de otra información disponible, de tipo estructural (relaciones con otros conceptos) o contextual (atributos, descripciones).

A continuación, se indican las principales técnicas de este tipo.

Técnicas basadas en cadenas

Estas técnicas se basan en la estructura de las cadenas a comparar. Se pueden buscar combinaciones exactas de letras o palabras, o combinaciones similares. Estas técnicas son:

- **normalización**, que usa técnicas para 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 separadores. Con estas técnicas, se consiguen reducir las variaciones y aumentar los sinónimos. Los inconvenientes serían una posible pérdida de significado aunque se aumentan las posibilidades de encontrar resultados.

- **técnicas de subcadenas** con las que se mide la similitud de algunas combinaciones de letras comunes entre cadenas, como prefijos o sufijos.

- **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.
- **medidas estadísticas** que calculan la importancia de una palabra en una cadena.
- **comparaciones de camino** donde se compara también la secuencia de palabras del concepto.

Técnicas basadas en lenguaje

En las técnicas anteriores, se consideraban las cadenas como secuencias de caracteres. En las técnicas basadas en lenguaje, las cadenas se consideran textos que pueden ser separados en palabras y donde también se tiene en cuenta la forma en que se combinan, es decir, la estructura gramatical de la cadena.

Estos métodos usan técnicas de procesado de lenguaje natural (NLP) para facilitar la extracción de los términos. Comparando esos términos y sus relaciones, se obtiene la similitud entre los conceptos en ambas terminologías.

Existen varias técnicas. La primera de ellas es la normalización lingüística, similar a la normalización léxica que vimos en el apartado anterior, donde cada término es reducido a una forma estándar fácilmente reconocible. Para ello, se distinguen varios tipos de variabilidad de los términos, siempre sobre la base del mismo término:

- morfológico: variaciones en la forma y la función.
- sintáctico: variaciones en la estructura gramatical.
- semántico: hipernimia, hiponimia, ...

Para esta normalización, se usan analizadores que realizan los siguientes pasos:

- Tokenización: se corresponde con la normalización léxica, segmenta la cadena para reconocer signos de puntuación, números, mayúsculas, espacios en blanco.
- Lematización: se buscan derivaciones del término principal como marcas de género, plural, prefijos, sufijos, ...
- Extracción de términos: reconocen términos desde la repetición de frases similares morfológicamente en los textos y el uso de patrones.

- Eliminación de palabras sin significado: como preposiciones, conjunciones, artículos, ...

Recursos linguísticos

Las técnicas basadas en recursos lingüísticos usan recursos externos como diccionarios, lexicons o tesauros con el fin de mapear palabras sobre la base de las relaciones lingüísticas entre ellos (por ejemplo, sinónimos, hipónimos, ...). Esta es la base de las técnicas léxicas que proporciona la API UMLSKS, usando los sinónimos de los conceptos a fin de aumentar las posibilidades de encontrar los conceptos coincidentes sinónimos. En el ámbito biomédico, Fung et al. [FBA+07] han utilizado el UMLS para producir un mapeo entre terminologías.

Otro de los recursos más ampliamente usado en la bibliografía es WordNet [12]. Es una enorme base de datos léxica del idioma inglés, creada y mantenida desde 1985 por el Cognitive Science Laboratory de la Universidad de Princeton bajo la dirección del profesor de psicología George A. Miller. Agrupa las palabras en conjuntos de sinónimos llamados 'synsets', proporcionando definiciones cortas y generales, y almacenando las relaciones semánticas entre estos conjuntos de sinónimos. Dos términos se consideran similares cuando pertenecen a un synset común. Los synsets están unidos por relaciones léxicas y semántico-conceptuales como hiponimia (subclase), hipernimia (superclase), sinonimia, antonimia, causa, coordinación, vinculación, holonimia (conjunto de), meronimia (parte de) y atributo. Contiene 152.000 cadenas con 115.424 synset. Ejemplos de su uso se verán en el apartado 2.5.6.

2.5.2. Técnicas a nivel de estructura

Estas técnicas usan propiedades estructurales, tales como relaciones compartidas a través de ontologías, para encontrar las correspondencias entre los conceptos. En general, estas técnicas se utilizan en combinación con las técnicas léxicas, ya que aumenta el rendimiento global como se demuestra en [FBA+07], para validar las correspondencias léxicas.

Las técnicas basadas en estructura interna comparan las propiedades que definen los conceptos, su tipo y rango, su cardinalidad o multiplicidad y su transitividad o simetría. Son eficientes y fáciles de implementar. Las técnicas más habituales comparan las relaciones que el concepto tiene con otros conceptos en la terminología. Las relaciones jerárquicas constituyen la columna vertebral de las terminologías y ontologías y, por tanto, es posible

[12] http://wordnet.princeton.edu/

su comparación ya que su significado es siempre el mismo. Además, se han desarrollado medidas para comparar conceptos basadas en la estructura, la mayoría basadas en contar el número de conceptos entre ellos. Además de las relaciones jerárquicas, también se puede usar la relación part-of, aunque no está tan extendida como ellas ya que hay casos en que no es aplicable.

Además de comparar directamente las relaciones de los conceptos, se pueden usar otras aproximaciones:

- **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.

En algunas investigaciones [ZB07], se usan otras relaciones, tras un proceso previo de estudio en que se determina qué relaciones son equiparables.

2.5.3. Otras técnicas

En este apartado, veremos otra serie de técnicas que se pueden usar. En algunos casos, es posible usar técnicas llamadas estadísticas (extensionales, para otros autores [ES07]) donde se comparan los conjuntos de instancias de un concepto, por lo que necesitan un corpus de instancias grande tanto en la terminología origen como en la de destino. Estas son aplicables más para ontologías que para terminologías.

También hay una serie de métodos semánticos, como por ejemplo, usar una terminología intermedia que dará un contexto común a las terminologías a comparar. Será, en este caso, una terminología de carácter más general y con una buena cobertura del dominio de forma que servirá de ayuda para desambiguar entre posibles significados de términos.

2.5.4. Combinación de técnicas

Las técnicas anteriores rara vez se usan en solitario ya que, para mejorar la validez de los resultados obtenidos, las técnicas léxicas no suelen ser suficientes. Para ello, se recurre a las técnicas estructurales para adquirir más información. También se recurre a terminologías de referencia, bien en un determinado dominio como UMLS en el dominio médico, o Wordnet, para ámbito general.

El mismo método se puede aplicar varias veces ajustando los parámetros a fin de obtener los resultados mejores según el objetivo.

2.5.5. Ejemplos de equiparación de terminologías

Desde hace más de 20 años, la bibliografía sobre equiparación de terminologías y ontologías es muy amplia en el ámbito biomédico y fuera de él, aplicando una o varias de las técnicas vistas en los apartados anteriores. A continuación, hacemos una breve introducción de algunas de ellas.

Un método léxico automatizado

El objetivo del trabajo de Jennifer Sun [SS06] fue encontrar las equiparaciones entre 3 bases de datos de 3 hospitales estadounidenses usando dos recursos UMLS y LOINC, ya vistos en el apartado 2.2. El proceso realizaba primero un procesado de los términos para su normalización, a continuación el proceso de alineamiento y por último, una ordenación de las equiparaciones resultantes.

El proceso de normalización consistía en dejar el término en minúscula, separar sus partes (tokens) constituyentes, eliminar los repetidos y ordenarlos por frecuencia. A continuación, se usaron dos procedimientos para alineación de cadenas. Uno fue pasar al proceso de alineación los tokens separados, puntuarlos y luego recombinarlos dándole una puntuación. El otro fue concatenerlos primero y pasarlos al proceso de alineación como un todo y puntuarlo así.

El algoritmo utilizado se inspira en los algoritmos de alineación de secuencias de ADN, como BLAST. Dichos algoritmos usan una matriz para encontrar el mejor alineamiento entre 2 cadenas, el término de la terminología origen y el término del diccionario, donde las letras de uno se colocan en el eje Y y las del otro en el eje X y se marca con 1 las coincidencias. A continuación, el proceso representa como un nodo cada celda marcada con el 1 y se enlazan en una cadena empezando por la esquina superior izquierda. La puntuación se calcula como la suma de los nodos. El algoritmo busca el nodo con mayor puntuación y, siguiendo el camino desde el nodo a través de la matriz, se llega al mejor alineamiento.

Una vez obtenidas las correspondencias, se fijó el umbral que determinaba que la correspondencia era válida. Para ello, el algoritmo se ejecutó varias veces probando diferentes umbrales y el mejor balance entre sensibilidad y especificidad fue en 85 %. Por último, se reordenaron las correspondencias candidatas en función de varios parámetros: longitud del término del diccionario, posición, puntuación o por una combinación de ellos.

La evaluación se hizo con 200 términos obtenidos aleatoriamente de los 3 orígenes y, como ya se indicó antes, se aplicaron varias variantes del algoritmo. Se analizó cada correspondencia viendo los positivos verdaderos (corres-

pondencias correctas encontradas), los falsos positivos (una correspondencia era incorrecta pero estaba por encima del umbral), los falsos negativos (una correspondencia era correcta pero queda por debajo del umbral) y los negativos verdaderos. El estándar de referencia era una evaluación manual de las correspondencias por parte de los investigadores.

Según los autores, el sistema fue suficientemente flexible para permitir correspondencias de diferentes vocabularios, posiblemente de diferentes dominios. El sistema no dependía de la creación de lexicones especializados para realizar las correspondencias. Además, era completamente automático, no requiriendo preformateo, entrada manual o construcción de reglas o filtros. Por último, se evaluó con varios vocabularios del mundo real contra UMLS y LOINC.

Entre las limitaciones, está que ni UMLS ni LOINC tienen todo el vocabulario de resultados de laboratorio existente actualmente, por lo que algunos términos o abreviaturas no son reconocidos. Y que sólo se evaluó para términos de resultados de laboratorio, no para otros como genomas, enfermedades o medicamentos. Otra limitación es que sigue necesitando de un experto que confirme la correspondencia correcta de la lista de candidatos.

Combinación de técnicas léxicas y semánticas

La combinación de técnicas léxicas y semánticas es un procedimiento habitual ya que siempre hay disponible en las terminologías algún tipo de información semántica, como relaciones jerárquicas, que usa para afinar el alineamiento. Se ha usado en múltiples casos, de los que comentamos dos de los más interesantes.

En [FBA+07], los autores combinan técnicas léxicas y semánticas para generar los alineamientos entre 2 terminologías usando UMLS. Para ello, usan el MetaMap, ya mencionado en el apartado 2.2.3, pasándole como entrada los términos de la terminología de origen, que primero normaliza, y restringiendo las correspondencias obtenidas a los conceptos UMLS de la terminología de destino. Esto constituye la técnica léxica.

El alineamiento semántico lo hacen con el algoritmo IntraMap, que hace uso de las relaciones semánticas entre los conceptos de UMLS. Para cada término de la terminología origen, se buscan los conceptos de la terminología destino relacionados por sinonimia o relaciones explícitas. Si no encuentra, buscará en los antecesores del término origen. Si tampoco encuentra buscará en los antecesores de los hijos y si no, en los antecesores de los hermanos.

Ambos métodos tienen ventajas e inconvenientes. El método semántico es más preciso pero necesita de una base de conocimiento amplia, mientras que el léxico no.

Para evaluarlo, la terminología origen fue SNOMED CT y la de destino ICD9CM; y usó un estándar de referencia ya proporcionado por el College of American Pathologists. Se comparó la cobertura (porcentaje de términos SNOMED CT para los que se encuentra correspondencia), recall (porcentaje de correspondencias en el estándar de oro que fueron encontradas) y precision (porcentaje de correspondencias encontradas que también lo están en el estándar de oro).

Se evaluaron los resultados de ambas técnicas por separado y luego juntas. El método combinado superó a cada uno por separado, alcanzando una cobertura del 91 %, un recall del 43 % y una precisión del 27 %.

En [ZB07], se propone un método automático basado en reglas para alinear cuatro terminologías de Anatomía, FMA (Foundational Model of Anatomy) de la Universidad de Washington, GALEN (Generalized Architecture for Languages, Encyclopedias and Nomenclatures) desarrollada por la universidad de Manchester en el proyecto European Union AIM, MA (Adult Mouse Anatomical Dictionary) y NCI Thesaurus, usando una combinación de técnicas léxicas y estructurales. Hay dos formas de hacer la alineación, de forma directa alineando pares de ellas y de forma indirecta usando una terminología de referencia.

El primer método identifica los conceptos similares en las terminologías usando técnicas léxicas y estructurales y luego, basado en este alineamiento, se buscan correspondencias adicionales complejas entre grupos de conceptos solo en base a características estructurales. Por último, también se comparan las relaciones asociativas, también sólo en base a características estructurales.

En el primer paso, se busca la similitud léxica entre nombres de conceptos (y sinónimos) de las terminologías origen, que puede ser exacta o tras normalización. Para este último caso, se usa el programa de normalización incluido en el UMLS. En esta fase, se usa además el UMLS para incluir nuevas correspondencias entre conceptos que son sinónimos en él.

Para la validación estructural, el primer paso fue obtener las relaciones representadas en las terminologías, como 'is-a' y 'part-of' y sus inversas, 'inverse-is-a' y 'has-part'. A continuación, se normalizó su representación para poder compararlas, se generaron las relaciones inversas que no existían, se extrajeron las implícitas en los nombres de los conceptos y por inferencia desde otras relaciones. Luego, se identificaron las similitudes, caminos jerárquicos comunes entre los conceptos de las terminologías, y los conflictos, que pueden ser de dos tipos. El primer tipo es definido por la existencia de caminos en sentido contrario en una terminología y en otra ('part-of' en una y 'has-part' en otra). El otro tipo viene dado por las diferencias entre las categorías de alto nivel de las terminologías. Los resultados obtenidos fueron de un 88.8 % de similitud, un 1.3 % de conflictos y un 9.1 % de conceptos sin

similitud.

A continuación, se buscaron correspondencias entre conceptos que no se habían encontrado en el paso anterior, tanto con conceptos en la misma situación como con conceptos que sí tenían correspondencia. En el primer caso, dados dos conceptos X1 y X2 de una terminología y un concepto Y de la otra, si el conjunto de las correspondencias de los descendientes de X1 y X2 no son subconjuntos uno del otro y la unión de ambos es igual al conjunto de correspondencias de los descendientes de Y, entonces se establecía una correspondencia del concepto Y al grupo de conceptos (X1,X2), es decir, sería de tipo uno-a-muchos. En el segundo caso, dados dos conceptos X e Y con correspondencia, si sus hijos tenían el mismo número de conceptos sin correspondencia, se establecía una correspondencia muchos-a-muchos entre ambos grupos de conceptos. El número de conceptos en FMA es unas 3 veces el de GALEN de ahí que aparezcan muchas correspondencias muchos-a-uno o bien conceptos de FMA que no sean encontrados.

Por último, se alineaban relaciones asociativas. Para ello, había que tener en cuenta que estas relaciones podían diferir de unas terminologías a otras e, incluso, una relación en una terminología podía ser igual a una combinación de relaciones asociativas y jerárquicas en otra. Para cada relación asociativa entre conceptos de una terminología, se buscaban todos los caminos más cortos entre los conceptos correspondientes de la otra. Se ignoraban los caminos que contenían una relación y su inversa. De los pares obtenidos, se obtenían los patrones y se buscaban aquellos que aparezcan más veces.

Para el alineamiento indirecto, se buscaban las correspondencias de las terminologías con la de referencia y las correspondencias entre pares de terminologías podían ser derivados de ella. Para ello, una de ellas, FMA, fue usada como referencia y las otras dos, MA y NCI, se comparaban. Se realizaron los alineamientos directos entre FMA y MA y entre FMA y NCI de la cual se dedujo el alineamiento indirecto entre MA y NCI. Al ser FMA una ontología de gran tamaño, se consiguió la identificación de la mayor parte de correspondencias, 91.5 %. En las pruebas realizadas usando MA o NCI como referencia, los resultados fueron mucho menores al ser de mejor tamaño. También se vio que los sinónimos y relaciones adicionales que presenta FMA dieron lugar a correspondencias específicas. Así mismo, las diferencias en cobertura y representación del conocimiento entre FMA y las otras dos provocó que aparecieran también algunas correspondencias específicas. En estas, se vio que los alineamientos indirectos usando una terminología de referencia es factible y razonablemente eficiente.

Para la evaluación, además de comparar el alineamiento directo con el indirecto, se comparó el resultado contra otro método y contra un estándar de oro establecido manualmente.

Entre las conclusiones del estudio, está la importancia del uso del conocimiento del dominio en el proceso de alineamiento. El alineamiento léxico depende de un modelo de reensamblaje léxico desarrollado para términos biomédicos por UMLS. La similitud estructural saca ventaja de las relaciones implícitas embebidas en los nombres de conceptos que se hacen explícitas como hemos visto. Esto también se usa en la evaluación.

2.5.6. Ejemplos de integración en UMLS

El proceso de integración de una nueva terminología en UMLS es definido por la NLM en 4 fases: análisis e inversión, inserción, edición humana y revisión de calidad. Es, por tanto, una labor intensa y sujeta a errores, por lo que varios proyectos han estudiado la forma de hacerlo más automatizado, entre ellos, los que explicamos a continuación.

Gene Ontology

En [SCG+03], Sarkar y sus colaboradores realizan un estudio preliminar la integración de GO (Gene Ontology) y examinan varias técnicas para integrar la ontología de genes (GO) con UMLS. Esta terminología está enfocada a desarrollar vocabularios estructurados para funciones moleculares, procesos biológicos y componentes celulares. Cada concepto GO está representado por una única cadena de texto y tiene un identificador único. Las técnicas usadas fueron:

- Equiparación de cadenas exactas. Es el método más simple y encuentra los conceptos UMLS que son exactamente iguales léxicamente a los términos GO.
- Norm. Utiliza uno de los métodos existentes en las herramientas léxicas de UMLS que convierte las cadenas de texto en una forma normalizada al quitar la parte de la palabra que son variaciones (género, número, acentos, ...). Se tratan con norm tanto los términos GO como los UMLS.
- MMTx. Como ya comentamos en el punto 2.2.2, esta es una herramienta del UMLS, para equiparar conceptos de textos biomédicos en UMLS. Al pasarle los términos GO, devuelve todas las cadenas candidatas contenidas en UMLS, las evalúa contra el término y produce una lista de resultados ordenada por longitud. En MMTx, se pueden indicar diferentes tipos de análisis. Los usados en este paper son Loose y Strict.
- BLAST. Usa la herramienta de UMLS, ya comentada en el punto 2.2.2, para comparar la similitud de secuencias de caracteres.

Para cada método, se evalúa su eficiencia en función de la precisión y el recall, para lo cual usan un estándar de oro facilitado por el NCI (National Cancer Institute). La precisión es medida como el número de correspondencias devueltas que están en el estándar de oro, dividido por el número de elementos devueltos en el alineamiento. El recall es el número de correspondencias devueltas que están en el estándar de oro, dividido por el número de correspondencias del estándar de oro.

El primer método tiene buena precisión ya que se ha buscado correspondencia léxica exacta, mientras que su recall es mucho menor, similar al MMTx-Strict. El segundo método tiene buenos resultados en ambos aspectos. Los métodos más complejos (BLAST y MMTx-Loose) tienen peores resultados.

Al usar como estandar de oro las correspondencias obtenidas con el NCI, que forma parte del Metathesaurus, sus resultados pueden ser sesgados ya que no se habrán obtenido correspondencias con conceptos que sí están en UMLS pero no en NCI.

Se hizo también una revisión de los falsos positivos, debido a la gran cantidad de sinónimos que presenta UMLS. De ellos, algunos ni siquiera estaban en el estándar de oro, otros era correctos pero estaban mal clasificados. Esto refleja la complejidad de la tarea. Su objetivo futuro era aplicar técnicas de procesado natural para crear las correspondencias.

OMIM

Como ya vimos en el apartado 2.2.2, OMIM es un vocabulario de genes y enfermedades genéticas, que tiene la particularidad de que la información está en formato texto. A continuación, vemos dos casos de su uso con UMLS.

En [CL03], extraen de OMIM los títulos y las variantes, obteniendo un total de 70.000 entradas, que se buscan en UMLS y SNOMED CT. El objetivo es añadir una estructura interna más formal a OMIM. Para ello, aplican la herramienta *norm* de UMLS para obtener las cadenas normalizadas. Una vez comparadas y obtenidos los conceptos UMLS y SNOMED, se analizan sus tipos semánticos, eliminando los que tengan tipos incompatibles. A continuación, se analizan las relaciones entre los conceptos obtenidos. Con ello, se determinan las relaciones semánticas a través de los conceptos OMIM iniciales.

En [HOTO04], el objetivo es indexar esa información en texto libre a través de UMLS. Para ello, se usa el subconjunto de tipos semánticos relacionados con anatomía y esos conceptos son buscados dentro de los textos de la Clinical Synopsis, guardando el punto del texto donde se inicia la concordancia y su longitud. Las concordancias más largas son seleccionadas. De

esta forma, los textos quedan indexados por los conceptos UMLS.

Uso de vocabularios genéricos (WordNet)

Recientemente, se ha realizado un estudio interesante, usar WordNet (que ya mencionamos en el apartado 2.3.1) para aumentar la integración de recursos en UMLS con la utilización de los sinónimos. Su objetivo era también ver cómo los parámetros "máximo número de sustituciones por término" y "máxima longitud del término" afectan al resultado.

Para ello, se usa MST (Minimal Standard Terminology), una terminología de términos gastro-intestinales ya integrada en UMLS. Contiene 1.944 términos, que representan 1.636 conceptos únicos y 289 de ellos tienen sinónimos. También tiene relaciones, entre ellas, 'part_of', 'has_location', 'manifestation_of' y 'treats'. Se usa también la herramienta norm del UMLS que ya comentamos en el apartado 2.2.3, que normaliza las cadenas. La versión de UMLS que usan no contiene la MST integrada inicialmente a fin de valorar la nueva integración.

En el proceso, primero buscan los términos de MST en UMLS tal cual (comparación exacta) y a continuación, los términos normalizados. Los que no obtengan equivalencia por ninguno de estos métodos pasan a la fase de sustitución de sinónimos. En ella, se construyen nuevos términos sustituyen palabras de los términos originales por sus sinónimos en Wordnet. Los sinónimos resultantes se buscan de nuevo en UMLS. Como antes, en caso de que no se obtenga nada por comparación exacta, se normalizarán.

Debido a la cantidad de sinónimos en WordNet, podría producirse una explosión combinatoria, sobre todo, para términos de muchas palabras. Esto puede llegar a consumir demasiados recursos computacionales, de ahí que se pretenda estudiar los parámetros de "máximo número de sustituciones por término" y "máxima longitud del término". Se realizaron pruebas combinando valores diferentes en ambos y se llegó a la conclusión de que bastaba con limitar el primero y que eso no afectaba de forma relevante al resultado obtenido.

Por último, se vio que usar WordNet permite un incremento del 11 % en el número de correspondencias encontradas. En trabajos futuros, el interés se centraba en usar las relaciones de WordNet (hiponimia, hipernimia, ...) y usar sustitución de varias palabras.

2.6. Evaluación de correspondencias entre terminologías

Evaluar es una tarea difícil, especialmente para terminologías de gran tamaño y dominios específicos. Es de gran ayuda contar con técnicos que validen los resultados, pero esto no siempre es posible. La creación de un estándar de referencia por parte de los especialistas que contenga todos los alineamientos correctos es una labor intensiva y cara.

En los ejemplos anteriores, ya vimos que, tras hacer el proceso de equiparación, se realizaba un proceso de evaluación definido ad-hoc por los desarrolladores del proyecto. En muchos casos, las correspondencias encontradas son evaluadas teniendo en cuenta varias medidas, que permiten determinar la fiabilidad del método:

- **recall** es el porcentaje de correspondencias válidas que son recuperadas.
- **precisión** es el porcentaje de correspondencias recuperadas que son válidas.

Un alto recall significa que se ha recuperado toda la información relevante, pero podría haber gran cantidad de resultados inútiles (lo que implicaría baja precisión). Una alta precisión significa que todo correspondencia encontrada es relevante, pero podrían faltar correspondencias relevantes (bajo recall).

Otra forma de evaluar un algoritmo es a través de una evaluación competitiva. Actualmente, hay al menos 2 competiciones de este tipo. Una de ellas, la KDDCup [13] forma parte de la International Conference on Knowledge Discovery and Data Mining y lleva organizándose desde 1997. En él, varios equipos de data mining del mundo compiten en resolver un problema práctico de alguna importancia.

Desde 2004, se viene celebrando una iniciativa similar, organizada por Ontology Aligment Evaluation Initiative (OAEI)[14], con el objetivo de establecer un consenso en los métodos para evaluación de alineamientos. Para ello, evalúan las ventajas y desventajas de los métodos que aparecen, comparan el rendimiento de las diferentes técnicas, incrementan la comunicación entre desarrolladores y ayudan a mejorar las técnicas existentes. Para ello, cada año organizan un evento y publican los resultados. Los organizadores proponen varios problemas de diferentes campos. Las ontologías, 2 o más por problema, están descritas en OWL-DL y vienen en formato RDF/XML. Los participantes sólo pueden aplicar un algoritmo y un conjunto de parámetros

[13] http://www.sigkdd.org/kddcup/index.php, http://www.kdd.org/kdd2010/kddcup.shtml
[14] http://oaei.ontologymatching.org/

a todos los problemas. Los organizadores evalúan los resultados enviados en precisión y recall contra los alineamientos de referencia.

Capítulo 3

Objetivos y esquema general

En este capítulo se describen los objetivos buscados en este tesis y se ofrece una visión general de la solución propuesta, exponiendo el alcance y las hipótesis de trabajo, así como las limitaciones que se han establecido.

3.1. Objetivos y requisitos

El objetivo general de esta tesis es el diseño, desarrollo e implementación de un método semiautomático para la equiparación y validación de terminologías que incluya la combinación de diferentes técnicas (léxicas, estructurales y de procesamiento de lenguaje natural). Este objetivo general puede desglosarse en varios objetivos específicos:

1. Diseño, desarrollo e implementación de la arquitectura del método para la equiparación de terminologías biomédicas de gran tamaño. Existen varias propuestas que ayudan al alineamiento de ontologías, principalmente léxicas. Sin embargo, siendo conocedores del inestimable valor de los avances obtenidos hasta el momento, la mayoría de las propuestas suponen puntos de vista parciales al problema. Por ello, el primer objetivo de la tesis ha sido desarrollar un método integrado de equiparación de terminologías biomédicas, que combinase diferentes aproximaciones, de forma semi-automática, con el fin último de producir un alineamiento con ciertas garantías de validez. Para llevar a cabo este objetivo, se planteó una arquitectura genérica para el alineamiento de las terminologías basada en los recursos disponibles en el momento, tales como las técnicas de equiparación léxicas proporcionadas por el servidor de conocimiento del UMLS o técnicas de procesamiento de lenguaje natural. La arquitectura ha integrado todas estas facilidades

para el alineamiento de las terminologías fuentes, con la mínima intervención del usuario final. Además, la arquitectura planteada se diseñó para permitir la adaptación de algunos de los elementos que la integran con el fin de obtener una mayor precisión y recall del alineamiento resultante.

2. Diseño, desarrollo e implementación de un procedimiento de validación y desambiguación automático del alineamiento resultante. La validación en el campo del alineamiento de terminologías y ontologías es un tema crucial y hoy en día se realiza, en muchos casos, de forma casi manual o frente a un estándar de referencia, en caso de existir. Por ello, el segundo objetivo de esta tesis doctoral ha sido el diseño, desarrollo e implementación de un procedimiento automático de validación de los resultados alcanzados durante la equiparación de las terminologías usando la combinación de técnicas propuestas. El objetivo de esta validación sería conocer el nivel de similitud semántica que presentan los alineamientos resultantes, permitiendo seleccionar los más adecuados.

A continuación, se describen un conjunto de requisitos que especifican más detalladamente las exigencias principales sobre los objetivos planteados.

1. El método propuesto deberá generar un alineamiento entre terminologías biomédicas de gran tamaño con mínima intervención del usuario. Para ello, combinará todas las técnicas y recursos necesarios para procesar las terminologías fuente y destino, y transformarlas en un formato manejable. El método deberá de ser lo más automático posible.

2. La solución propuesta deberá de ser escalable y proporcionar solución cuando se manejan terminologías de gran tamaño.

3. La solución propuesta deberá de ser genérica y reutilizable. Para ello, la arquitectura se planteará de forma modular, estableciendo una interfaz de intercambio de datos entre los módulos conectados. Esta visión favorece la combinación de nuevas técnicas para el alineamiento, con mínimos cambios y sin afectar al resto de los componentes.

4. La solución propuesta deberá introducir algún método automático de validación y desambiguación del alineamiento propuesto, que estime la similitud de cada alineamiento resultante. El método de validación debería basarse en principios bien fundamentados, que aseguren la calidad de la validación.

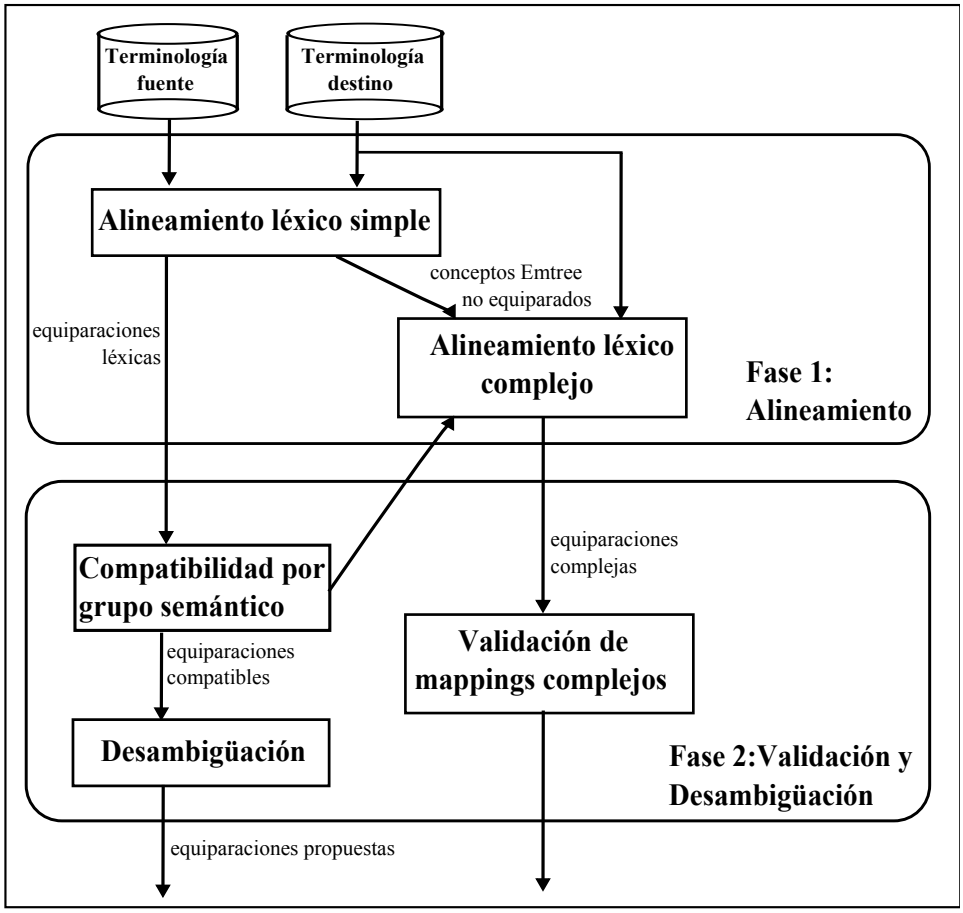

Figura 3.1: Esquema general del método propuesto

3.2. Esquema general del método propuesto

La metodología propuesta en esta tesis combina técnicas léxicas y estructurales, aprovechando determinadas características de las terminologías, como son la incorporación de sinónimos, la categorización de los conceptos y relaciones entre conceptos, sobre todo, la relación broader/narrower. El esquema general se muestra en la figura 3.1.

Como puede verse, el proceso consta de dos fases diferenciadas. En primer lugar, el alineamiento de conceptos de la terminología fuente a la terminología destino y, en segundo lugar, la validación y desambiguación de las equiparaciones obtenidas. La primera fase comienza realizando una equiparación léxica directa de todos los términos de la terminología fuente (términos preferidos y sinónimos) a la terminología destino. Tras ello, para aquellos términos que no se han equiparado léxicamente, se emplean técnicas de pro-

cesado de lenguaje natural para descomponer el término y realizar una nueva equiparación léxica de sus partes constituyentes a la terminología destino.

La segunda fase, Validación y Desambiguación, usa la similitud entre las categorías de las terminologías fuente y destino para descartar de forma automática las equiparaciones que son incompatibles con las facets de la terminología fuente. Posteriormente, aplicando técnicas basadas en proximidad estructural, que identifican similitudes entre conceptos, se validan y desambiguan los alineamientos léxicos. Por último, empleando técnicas de procesado de lenguaje natural y técnicas basadas en similitud estructural, se construyen equiparaciones complejas (BroadMatch y UnionMatch) para aquellos conceptos de la terminología fuente que no se han equiparado de forma léxica directa con ningún concepto de la terminología destino.

3.3. Hipótesis de trabajo y limitaciones de la solución propuesta

El conjunto de suposiciones en el que se basa esta tesis doctoral pone de relieve las distintas decisiones que se tomaron durante el diseño, desarrollo e implementación del método de alineamiento. A continuación detallaremos dichas suposiciones.

1. Las terminologías a equiparar. Hemos elegido Emtree como terminología fuente ya que se trata de una terminología biomédica de gran tamaño que se está utilizando para indexar bibliotecas digitales. El Metathesaurus UMLS es el mayor recurso del ámbito médico que, además, dispone de un conjunto de herramientas que facilitan mucho su consulta.

2. Las técnicas léxicas son adecuadas para abordar la fase inicial de alineamiento.

3. La validación automática es factible si las terminologías describen parcelas similares de conocimiento del dominio.

4. Las técnicas basadas en similitud estructural permiten validar el alineamiento léxico, así como discriminar entre los alineamientos alternativos.

5. Los factores de similitud asociados a cada alineamiento resultante permiten seleccionar alineamientos suficientemente relevantes entre diferentes alternativas.

6. Las técnicas de procesamiento de lenguaje natural son adecuadas para particionar términos que no han obtenido equiparación léxica y, a partir de ellos, construir alineamientos compuestos y/o complejos.

7. Las técnicas basadas en similitud estructural también permiten validar los alineamientos compuestos resultantes de la aplicación de técnicas basadas en lenguaje natural.

Una vez presentadas las suposiciones, a continuación describiremos el conjunto de hipótesis de trabajo en las que se basa la solución propuesta en esta tesis doctoral.

1. La combinación de diferentes técnicas de alineamiento y de procesado de lenguaje natural permite obtener un alineamiento de terminologías con mayor precisión y recall que si aplicásemos cualquiera de las técnicas individualmente.

2. Los alineamientos obtenidos por métodos léxicos pueden ser validados usando técnicas que aprovechan la similitud estructural para asociar un factor de similitud a cada alineamiento, factor que indica la semejanza entre los significados de los conceptos del alineamiento en cada terminología.

3. Los factores de similitud se pueden computar teniendo en cuenta la semejanza estructural entre las terminologías fuentes, dada por las relaciones entre los conceptos.

4. Las técnicas de procesamiento de lenguaje natural permiten describir alineamientos complejos en el caso de no existir alineamientos directos. Estos alineamientos complejos pueden ser útiles para facilitar la interoperabilidad de la información indexada por las terminologías fuentes.

Por último, expondremos algunas de las limitaciones del presente trabajo.

1. La terminología fuente debe representar la información a nivel de concepto y no de término.

2. El método propuesto requiere que la terminología destino esté integrada en el Metathesaurus del UMLS.

3. La terminología destino debe referirse a la misma parcela del dominio biomédico que la terminología fuente. En este caso, UMLS tiene la mayor extensión posible dentro de ese ámbito, al contener cientos de terminologías, de todas las temáticas posibles.

4. El método está limitado al idioma inglés, ya que todas las técnicas integradas en el mismo funcionan para dicho dominio. La aplicación al dominio del castellano debería incorporar nuevas técnicas mejoradas para nuestro idioma.

5. La terminología fuente está limitada a XML. Pero, debido a la independencia de los módulos integrados, permitiría ampliar el método a otro tipo de fuentes con pequeñas actualizaciones en la capa inicial de preprocesado.

Capítulo 4

Alineamiento de terminologías

4.1. Introducción

A pesar del considerable volumen de la investigación llevada a cabo en diferentes campos, el alineamiento de terminologías sigue siendo hoy en día una tarea compleja. Hay, entre otros, dos problemas fundamentales que dificultan la interoperabilidad. El primero de ellos es el tratamiento informal de las relaciones en las terminologías biomédicas, que conduce a definiciones contradictorias y ambiguas [SCK+05]. El segundo problema es la falta de métodos automatizados que simplifiquen el proceso de alineamiento [DNH04]. En esta tesis doctoral, nos centramos sólo en este segundo problema. Para llevarlo a cabo adecuadamente, hemos hecho una suposición importante, como ya lo hicieron otros trabajos anteriores como [ZMBB07] y [SCG+03]: "las terminologías a equiparar están correctamente diseñadas, por lo que las dificultades para descubrir sus equivalencias o similitudes provienen de la toma de decisiones llevada a cabo durante el proceso de diseño de ambas terminologías".

En este capítulo, comenzaremos definiendo formalmente lo que entendemos por alineamiento de terminologías. Seguidamente, profundizaremos en aquellos aspectos de las terminologías que se usan como fuentes experimentales en nuestro método, como son la agrupación de conceptos en categorías y las relaciones estructurales. A continuación, explicaremos el procedimiento general de nuestro método, describiendo en detalle los pasos necesarios para realizar el alineamiento. Finalmente, detallaremos los procesos de equiparación léxica de términos y conceptos e introduciremos la equiparación compleja de conceptos.

4.2. Definición de alineamiento

El alineamiento de terminologías se puede definir como el proceso consistente en determinar las correspondencias entre los términos o conceptos de dos terminologías, que llamaremos fuente y destino.

Un alineamiento de terminologías es una tupla A=(F,D,R), donde F denota los términos o conceptos de la terminología fuente, D denota los términos o conceptos de la terminología destino y R denota la relación del alineamiento.

A cada una de las relaciones o correspondencias se les suele llamar también alineamiento o "mapping". En el caso más sencillo, los alineamientos son correspondencias de equivalencias entre términos o conceptos con cardinalidad uno-a-uno. En este caso, un alineamiento se puede definir como una tupla (id, f_i, d_i, s), donde id es el identificador del alineamiento, f_i es el concepto o término de la terminología fuente, d_i es el concepto o término de la terminología destino y s es el grado de similitud de los conceptos o términos en el alineamiento.

Teniendo en cuenta el grado de similitud de las equivalencias en el alineamiento, se pueden distinguir diferentes tipos de equivalencias :

- **equivalencia exacta** establece una relación de equivalencia directa entre los conceptos o términos.
- **equivalencia de subsumisión**: establece una relación de inclusión entre conceptos o términos. Esta relación puede ser una equivalencia broader, cuando el concepto de la terminología destino es más general que el de la terminología fuente, o una equivalencia narrower.
- **equivalencia de composición**: relaciona un concepto o término a varios combinados de tipo booleano, como AND, OR o NOT.

Teniendo en cuenta esto, en esta tesis doctoral utilizaremos la siguientes definiciones de alineamientos:

- **alineamiento simple**: se define como un alineamiento con una relación de equivalencia exacta de cardinalidad uno-a-uno. Lo denotamos por ExactMatch (siguiendo la notación de SKOS).
- **alineamiento complejo**: se define como un alineamiento de cardinalidad uno-a-muchos, en el que existe una relación de equivalencia parcial (que denotaremos por BroadMatch) o de composición (UnionMatch). En los casos en los que no está definido el tipo de relación de equivalencia entre el concepto o término fuente y los conceptos o términos destino, hablaremos de alineamiento compuesto.

Term	**thorax**
Id	36153
Synonym	Chest rib cage thoracic index
Broader	body regions

Figura 4.1: Ejemplo de sinónimos en el concepto EMTREE *thorax*

4.3. Terminologías usadas

En nuestro estudio, hemos usado dos terminologías médicas introducidas previamente en el apartado 2.3.3 para probar la validez de nuestro método. Aquí detallaremos un poco más su composición y estructura, que se usarán como ejemplo para exponer tanto el procedimiento de alineamiento propuesto como el procedimiento de validación.

4.3.1. Emtree

Como ya se introdujo en el capítulo 2, la terminología EMTREE ha sido desarrollada por Elsevier para indexar EMBASE, una base de datos on-line biomédica y farmacológica. Tiene una estructura basada en concepto: todos los términos, que son estrictamente sinónimos entre sí, se agrupan en un concepto. Cada concepto tiene un término preferido (Preferred Term, PT), que es también el nombre del concepto, y un conjunto de sinónimos. En la figura 4.1, se muestra la información mínima sobre el concepto *thorax*, que tiene tres sinónimos *chest*, *rib cage* y *thoracic index*. Además, incluye información estructural a través de una relación broader entre conceptos, de forma que los términos quedan relacionados en una estructura de tipo árbol. Un ejemplo de esta relación puede verse en la figura 4.2 entre el concepto *thorax* y conceptos más específicos a *thorax*, como *diaphragm*, *male breast* y *mediastinum* que, a su vez, pueden estar relacionados con otros más específicos, como por ejemplo, *diaphragm muscle* es más específico que *diaphragm*. Finalmente, Emtree incluye otros tipos de información que no han sido explorados para este trabajo, como la fecha de creación o una relación genérica llamada "related".

Emtree es una terminología de gran tamaño y rica en sinónimos. Además los conceptos están organizados en una serie de categorías superiores, llamadas facets. Estas categorías fueron definidas por los diseñadores del proyecto y agrupan los conceptos según la parte del dominio a la que pertenecen. Ejemplos de facets son *Anatomical concepts* o *Chemicals and Drugs*. Ade-

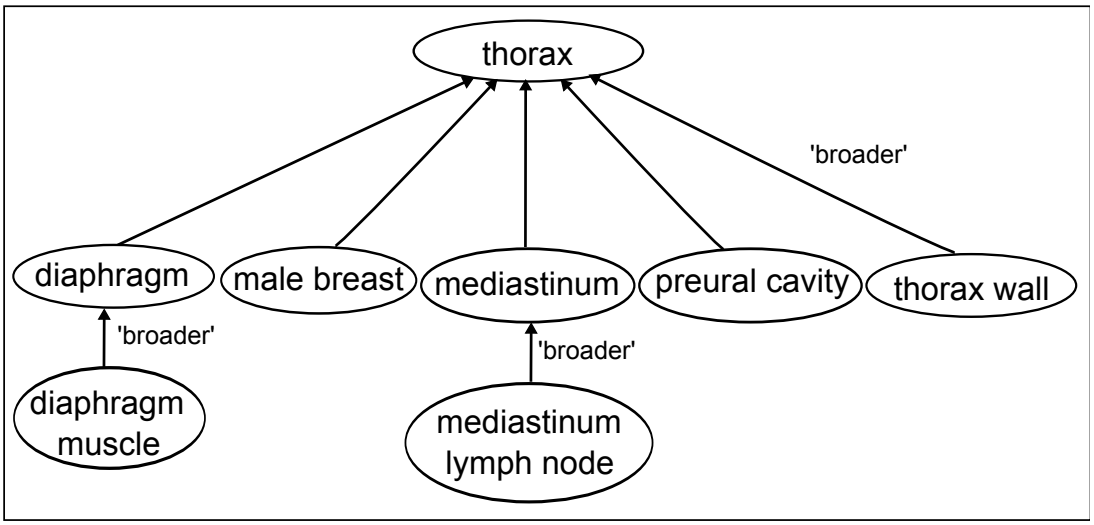

Figura 4.2: Ejemplo de relaciones 'broader' entre conceptos Emtree

más, Emtree tiene la particularidad de que un término puede estar en más de una de estas categorías, lo que supone el 9.8 % del número total de conceptos. Todos los conceptos que aparecen en más de una facet aparecen siempre en la facet *Anatomical concepts* y en alguna otra (el 61 % de las veces es *Biological phenomena and functions*). Algunos ejemplos pueden verse en la figura 4.3.

4.3.2. UMLS

Como ya se explicó en el apartado 2.3.3, el Metathesaurus UMLS es la unión de más de 130 terminologías, por lo que es de mucho mayor tamaño que Emtree, lo cual da lugar a una gran cantidad de sinónimos y a una gran granularidad, es decir, dispone en muchos casos de mayor número de conceptos para describir cada elemento del dominio.

Como en Emtree, el elemento principal es el concepto, que se identifica mediante un identificador único denominado CUI y que puede incluir gran cantidad de información, como una o varias definiciones (procedentes de las diferentes terminologías iniciales), sinónimos, narrowers, broader y otras relaciones. En la figura 4.4, se muestra la información disponible para el concepto *Entire ligament*, incluyendo sinónimos, broaders y narrowers. En la figura 4.5, se muestran las relaciones entre varios términos relacionados con *Ligaments*, tanto por relaciones de broader como de narrower. Debido a la gran granularidad, el número de broaders, pero sobre todo, de narrowers, que suele tener un concepto es alto.

Los conceptos del Metathesaurus se clasifican también utilizando un con-

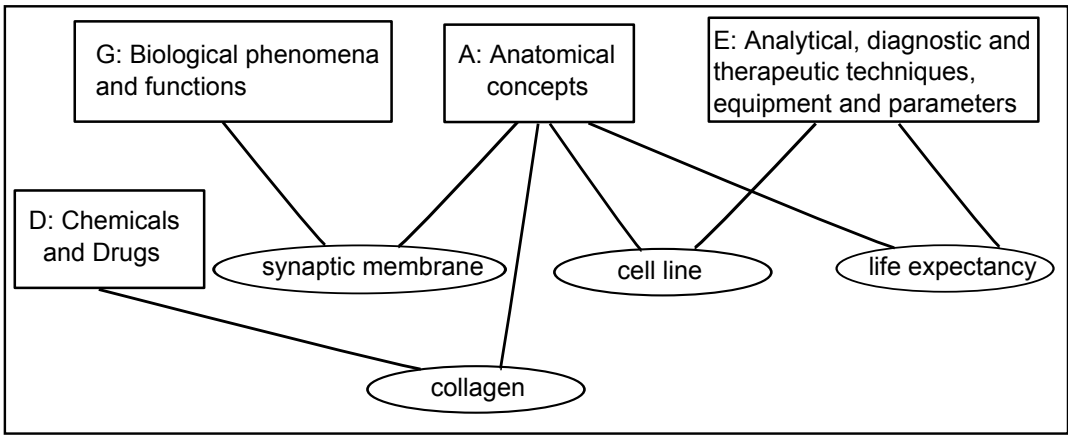

Figura 4.3: Ejemplo de conceptos Emtree en más de una facet

junto de categorías semánticas básicas llamadas Tipos Semánticos (TS), como *Anatomical Structure* o *Chemical*. En total, incluye 134 TSs, lo cual para algunos propósitos, resultan excesivos. Por ello, se realizó un estudio [MBB01] para estudiar las similitudes entre ellos y, como resultado, los TS se agruparon, a su vez, en 15 categorías de alto nivel, denominadas Grupos Semánticos (GS), tales como *Anatomy* o *Disorders*.

4.4. Procedimiento general de alineamiento léxico

En este apartado, nos centraremos en presentar el procedimiento general de alineamiento léxico, que detallaremos en las siguientes secciones. Aunque, en el momento de realización del trabajo de esta tesis doctoral, teníamos alguna herramienta orientada al alineamiento de ontologías (como Protégé-Prompt[1]) para alinear terminologías, aquí no pudieron usarse dado el gran tamaño de las terminologías que nos interesaban; es decir, las herramientas disponibles estaban pensadas para ontologías pequeñas.

El alineamiento léxico permite obtener las correspondencias entre los conceptos de la terminología fuente y de la terminología destino que vienen descritos por las mismas (o similares) etiquetas. En la figura 4.6, se muestra el esquema general del proceso.

El procedimiento parte de la terminología fuente, una terminología de gran tamaño (p.ej., el fichero Emtree que contiene en torno a 46.000 términos

[1]http://protege.stanford.edu/plugins/prompt/prompt.html

ConceptName	**Entire ligament**
CUI	C1269080
Semantic Type	Body Part, Organ, or Organ Component
Synonym	[SO]Ligaments of joints Entire ligament (body structure) Entire ligament
Broader	Ligaments Entire body as a whole
Narrower	Structure of sternopericardial ligament Structure of superior sternopericardial ligament Structure of inferior sternopericardial ligament Entire superior sternopericardial ligament Entire inferior sternopericardial ligament Entire sternopericardial ligament

Figura 4.4: Ejemplo de concepto en UMLS

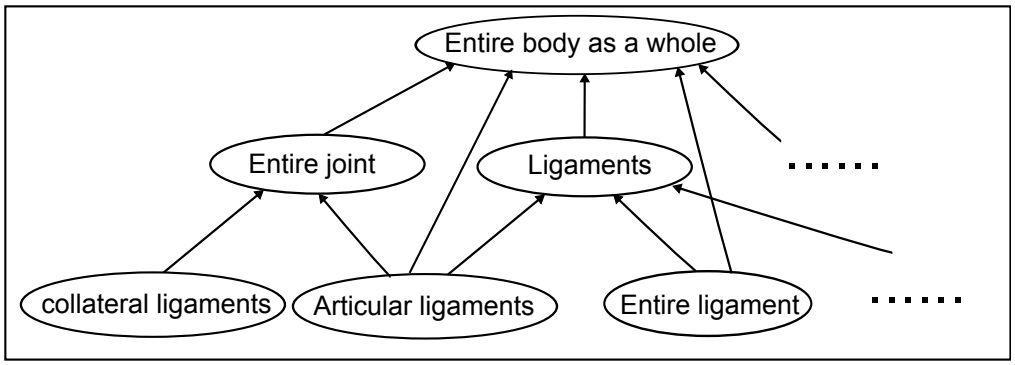

Figura 4.5: Ejemplo de relaciones 'broader' entre conceptos UMLS

(del orden de 20MB)). Dada la dificultad que lleva consigo el procesamiento de terminologías de gran tamaño, la primera fase divide la terminología fuente en varios ficheros, uno por cada categoría de alto nivel (llamadas facets en Emtree). Tras esta fase, hay términos que pueden aparecer en más de una categoría de alto nivel por lo que aparecerán repetidos en más de un fichero. Tras el particionamiento, cada fichero contendrá todos los términos que componen una parte del dominio organizados jerárquicamente.

La siguiente fase consiste en equiparar léxicamente todos los términos de la terminología fuente (preferido y sinónimos) con los términos de la terminología destino en UMLS. En la figura 4.6, se ve el ejemplo del concepto Emtree *thorax* descrito por un PT (que se equipara léxicamente a 3 conceptos UMLS) y 3 sinónimos, *Chest* que se equipara con 2 conceptos UMLS, *rib cage* con 1 concepto y *thoracic index* que no se equipara a ninguno. A con-

tinuación, se realiza la agrupación de todas las equiparaciones del concepto fuente (los correspondientes a su término preferido y a sus sinónimos). Así, por ejemplo, el concepto Emtree *thorax* queda equiparado con 5 conceptos UMLS.

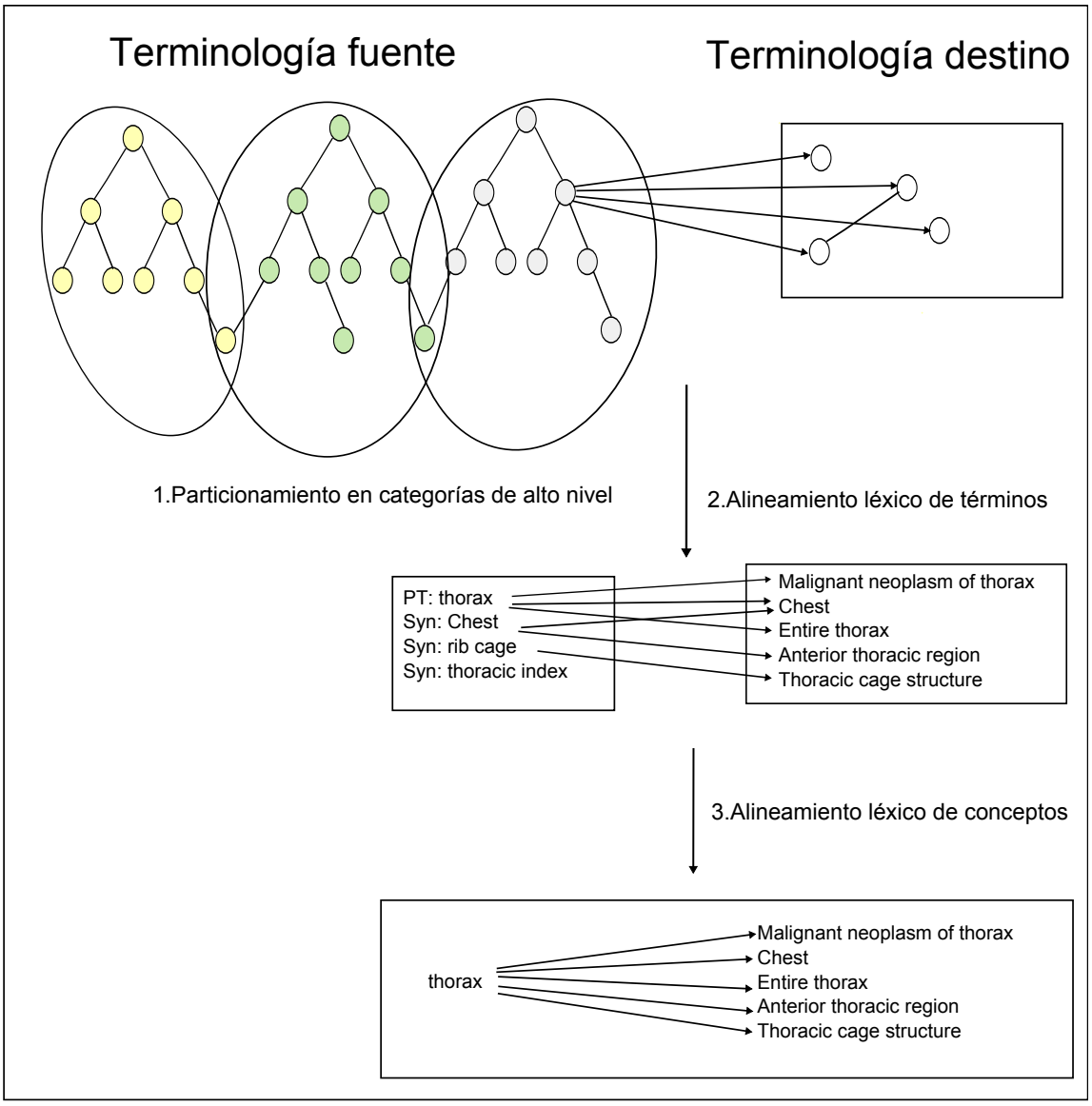

Figura 4.6: Esquema del proceso de Alineamiento léxico

4.5. Criterios de selección del servicio de alineamiento léxico

Para llevar a cabo el alineamiento léxico de términos a través del Metathesaurus, usamos uno de los servicios web del UMLSKS, a través de la API que el NLM tenía disponible, llamada KssApi. Esta API proporcionaba varios métodos para realizar consultas directas al Metathesaurus sobre un término y obtener así toda la información de los conceptos similares léxicamente a él. Estos métodos se diferencian en cómo se realiza la extracción de información, si mediante una comparación exacta o con algún tipo de normalización. Los métodos disponibles son:

- Word: separa las palabras constituyentes del término de entrada y devuelve los conceptos UMLS que las contienen.
- ApproximateMatch: obtiene aquellos conceptos UMLS cuyos sinónimos difieren poco de los términos de entrada. La diferencia puede estar en la ausencia o variación de letras.
- ExactMatch: devuelve sólo los conceptos UMLS cuyas etiquetas coinciden exactamente con el término buscado.
- NormalizedString: busca el término tras ser normalizado (se explica con más detalle a continuación).
- NormalizedWord: obtiene los conceptos UMLS que contienen cada palabra del término normalizada por separado.
- RightTruncation: trunca el término por la derecha antes de realizar la búsqueda.
- LeftTruncation: trunca el término por la izquierda antes de realizar la búsqueda.

Con el fin de seleccionar el método más apropiado para la equiparación léxica de términos, se observó que, para el mismo conjunto de términos de entrada, Right y Left Truncation daban resultados arbitrarios, Word y NormalizedWord necesitaban demasiado tiempo y recuperaban un número alto de conceptos incorrectos y ApproximateMatch devolvía conceptos muy dispares. Así, las únicas dos opciones eran ExactMacth y NormalizedString, y elegimos la segunda por ser más flexible a variaciones léxicas, ya que ExactMatch sólo recupera coincidencias exactas.

En particular, NormalizeString, antes de la equiparación de términos, normaliza las cadenas de entrada, eliminando las diferencias léxicas (mayúsculas, inflexión, variantes ortográficas o puntuación). El método devuelve todos los conceptos UMLS que contengan la cadena de entrada normalizada como sinónimo. Para ello, divide cada palabra en sus constituyentes, las pone en minúsculas, las convierte a su forma sin inflexiones y ordena las palabras por orden alfabético.

4.6. Alineamiento léxico de términos

En este apartado, nos centraremos en detallar el método de alineamiento léxico que hemos implementado para obtener los conceptos del Metathesaurus que son léxicamente equivalentes a cada término de la terminología de entrada. Recordemos que dos términos se consideran equivalentes léxicamente cuando sus nombres son idénticos tras las normalización.

En primer lugar, para cada concepto de la terminología fuente, el método envía una solicitud a la base de datos UMLS para todos los términos que definen el concepto (es decir, PT y sinónimos). El servicio NormalizeString permite identificar los conceptos UMLS que contienen un sinónimo con la misma forma normalizada que la cadena de entrada y puede devolver cero, uno o varios conceptos UMLS equivalentes léxicamente. En el primer caso, decimos que no hay ninguna equiparación; en el segundo caso, el proceso de equiparación es simple (se obtiene un sólo concepto UMLS que se equipara léxicamente con el término de entrada); en el tercer caso, el proceso de equiparación es ambiguo (varios conceptos UMLS son léxicamente equivalentes al concepto de la terminología fuente). A continuación, veremos ejemplos de todos estos casos.

Consideraremos que cada alineamiento es un resultado del proceso de equiparación léxica entre un término de la terminología fuente y un concepto UMLS recuperado. En nuestra aproximación, un alineamiento se puede expresar como una quíntupla de la siguiente forma:

$$(\text{id}, C_{Fuente}, t_{Fuente}, C_{umls}, \text{m})$$

donde

- id permite identificar cada alineamiento de forma unívoca
- C_{Fuente} es el identificador del concepto de la terminología fuente
- t_{Fuente} es el término preferido o sinónimo del concepto de la terminología fuente que se ha equiparado léxicamente a algún término del concepto UMLS

- C_{umls} es el CUI, el identificador del concepto UMLS
- m es el tipo de alineamiento, en este caso, alineamiento léxico simple de término (representado por 'lt').

Ejemplos de ausencia de alineamientos se pueden dar tanto en términos de una sola palabra como *subendothelium* como en términos de múltiples palabras como *face, nose and sinuses*, o *eyelid muscle*. Ejemplos de equiparaciones simples se pueden ver en la figura 4.7:

- El término preferido *animal anatomy* no tiene asignado ningún alineamiento directo a UMLS, pero el sinónimo *animal structures* se equipara léxicamente al concepto UMLS *Animal Structures*.

$$(3, 675550, animal\ structures, C0003058,' lt')$$

- El término Emtree *beak* es equivalente léxicamente con *Beak*.

$$(361, 62163, beak, C0004895,' lt')$$

- El término *cornea vascularization* se equipara con *Vascularization of cornea*, que pertenece al grupo semántico *Disorders*.

$$(17534, 462737, cornea\ vascularization, C0474344,' lt')$$

En los dos primeros ejemplos, puede observarse que los conceptos equiparados son conceptos del grupo semántico *Anatomy*, mientras que, en el tercer ejemplo, las categorías de nivel superior a las que pertenecen los conceptos del alineamiento son diferentes en Emtree y en UMLS: *cornea vascularization* pertenece a *Anatomical concepts* mientras que *Vascularization of cornea* es una enfermedad. Por tanto, este alineamiento léxico se considerará incorrecto durante la fase de validación.

Los alineamientos ambiguos son lo más habitual debido a que UMLS es la fusión de varios vocabularios diferentes y contiene también una gran cantidad de sinónimos. Esto mejora los resultados de las consultas aunque también puede añadir imprecisión. Podemos hablar entonces de dos tipos de alineamientos ambiguos:

- Alineamientos semánticamente muy diferentes, que son fáciles de identificar ya que, aunque léxicamente son similares, pertenecen a categorías semánticas muy diferentes.

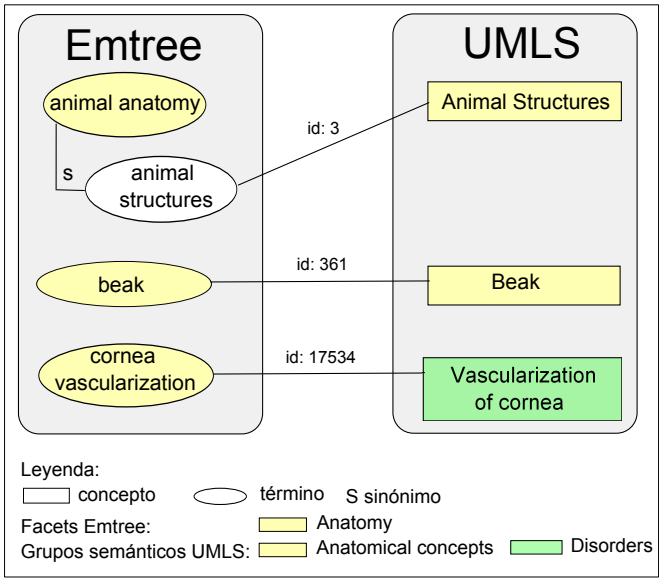

Figura 4.7: Ejemplos de equiparaciones simples

- Alineamientos semánticamente similares por proximidad de los conceptos en las terminologías fuente.

A continuación, mostramos algunos ejemplos de ambos.

Caso Body Regions

Como se puede ver en la figura 4.8, el término Emtree *body regions*, de la facet *Anatomical concepts*, se equipara léxicamente con 3 conceptos UMLS que pertenecen al mismo grupo semántico *Anatomy*. Si consideramos que el alineamiento representa conceptos equivalentes, entonces teniendo en cuenta el significado de los conceptos de Emtree, *Body part*, *Body Regions* y *Entire body region* también serían equivalentes en UMLS y esto no es cierto. Es decir, si nosotros proporcionamos estas equiparaciones como correctas, estamos modificando implícitamente UMLS, lo que puede conllevar a errores de interpretación de dicha fuente. Consideramos, por tanto, que dichos alineamientos son ambiguos. Además, dichos conceptos son semánticamente similares en UMLS:

$$(3184, 668756, \textit{body regions}, C0005898,' lt')$$
$$(3185, 668756, \textit{body regions}, C0229962,' lt')$$
$$(3186, 668756, \textit{body regions}, C1280064,' lt')$$

Caso bacterium

En la figura 4.8, hay alineamientos ambiguos para el término preferido *bacterium* y para el sinónimo *bacteria* ya que ambos se equiparan léxicamente

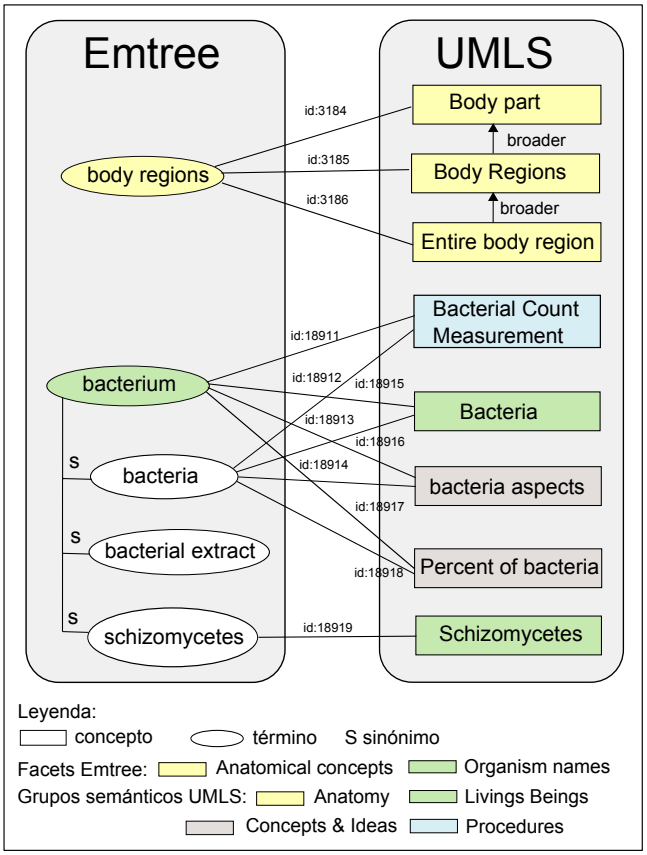

Figura 4.8: Ejemplo de alineamientos ambiguos: *body regions* y *bacterium*

con los conceptos UMLS *Bacterial Count Measurement*, *Bacteria*, *bacteria aspects* y *Percent of bacteria*, de los cuales el primero pertenece al grupo semántico *Procedures*, *Bacteria* al grupo *Livings Beings* y los dos últimos, a *Concepts and Ideas*. Hay un alineamiento simple del sinónimo *schizomycetes* al concepto UMLS *Schizomycetes*, que pertenece al tipo semántico *Bacterium* igual que *Bacteria*. No hay alineamiento para el término EMTREE *bacterial extract*.

Los alineamientos son:

$$(18911, 4567, bacterium, C0004611,' lt')$$
$$(18912, 4567, bacterium, C1510439,' lt')$$
$$(18913, 4567, bacterium, C0004618,' lt')$$
$$(18914, 4567, bacterium, C2347473,' lt')$$
$$(18915, 4567, bacteria, C0004611,' lt')$$
$$(18916, 4567, bacteria, C1510439,' lt')$$
$$(18917, 4567, bacteria, C0004618,' lt')$$

$$(18918, 4567, \textit{bacteria}, C2347473,' lt')$$
$$(18919, 4567, \textit{schizomycetes}, C0036340,' lt')$$

Caso back

En la figura 4.9, el término preferido *back* se equipara léxicamente con 4 conceptos de 3 grupos semánticos diferentes. De ellos, puede verse similitud con los conceptos *Back* y *Lumbosacral Region*, que pertenecen al grupo semántico *Anatomy*, mientras que *Back problem* hace referencia a una enfermedad, grupo semántico *Disorders*, y *Dorsal* pertenece a *Concepts and Ideas*. Además, 2 de sus sinónimos tienen cada uno un alineamiento simple a un nuevo concepto y el último sinónimo *sacrococcyx* no se equipara léxicamente con ninguno.

Los alineamientos son:

$$(7524, 4517, \textit{back}, C0460009,' lt')$$
$$(7525, 4517, \textit{back}, C1995000,' lt')$$
$$(7526, 4517, \textit{back}, C0004600,' lt')$$
$$(7527, 4517, \textit{back}, C0205095,' lt')$$
$$(7528, 4517, \textit{back}, C0579085,' lt')$$
$$(7529, 4517, \textit{back}, C1281593,' lt')$$
$$(7530, 4517, \textit{lumbosacral region}, C0024094,' lt')$$
$$(7531, 4517, \textit{sacrococcygeal region}, C0036035,' lt')$$

Caso spine

En la figura 4.9, para el término *spine* se muestran equiparaciones léxicas tanto para el término preferido como para los sinónimos a conceptos UMLS de 2 grupos semánticos diferentes. Algunos de los conceptos UMLS son equivalentes léxicamente a varios de los términos Emtree. Por ejemplo, el concepto *Vertebral column* (C0037949) es equivalente al término preferido *spine* y a los sinónimos *dorsal spine*, *spinal column* y *vertebral column*.

Los alineamientos son:

$$(7987, 45401, \textit{spine}, C0037949,' lt')$$
$$(7988, 45401, \textit{spine}, C0150920,' lt')$$
$$(7989, 45401, \textit{spine}, C1267072,' lt')$$
$$(7990, 45401, \textit{spine}, C1280065,' lt')$$
$$(7991, 45401, \textit{dorsal column}, C0458459,' lt')$$
$$(7992, 45401, \textit{dorsal spine}, C0037949,' lt')$$
$$(7993, 45401, \textit{spinal column}, C0037949,' lt')$$
$$(7994, 45401, \textit{spinal column}, C1267072,' lt')$$
$$(7995, 45401, \textit{vertebral column}, C0037949,' lt')$$
$$(7996, 45401, \textit{vertebral column}, C1267072,' lt')$$

(7997, 45401, *vertebral column*, $C0346667,'lt'$)
(7998, 45401, *vertebral column*, $C0347311,'lt'$)

En esta etapa del método, además de extraer las equivalencias léxicas de los términos Emtree a conceptos UMLS, para cada concepto UMLS, se extrae información extra necesaria para la siguiente etapa de validación-desambiguación. En particular, se extrae el CUI (identificador del concepto en UMLS), su tipo semántico, su lista de sinónimos y los conceptos UMLS vinculados con el concepto por relaciones de tipo broader y narrower. En la figura 4.10, para el término *Lumbosacral region*, se extraen sus 2 sinónimos, 2 broaders y 4 narrowers. Debido al gran tamaño de UMLS, en muchos de los términos, el número de sinónimos, broaders y narrowers recuperados es muy alto, sobre todo, el de estos últimos.

A la información anterior, el método añade el grupo semántico, a partir del tipo semántico, usando una tabla de pertenencia de tipos a grupos semánticos. Así, en el caso de la figura, el tipo semántico *Body Location or Region* pertenece al grupo semántico *Anatomy*.

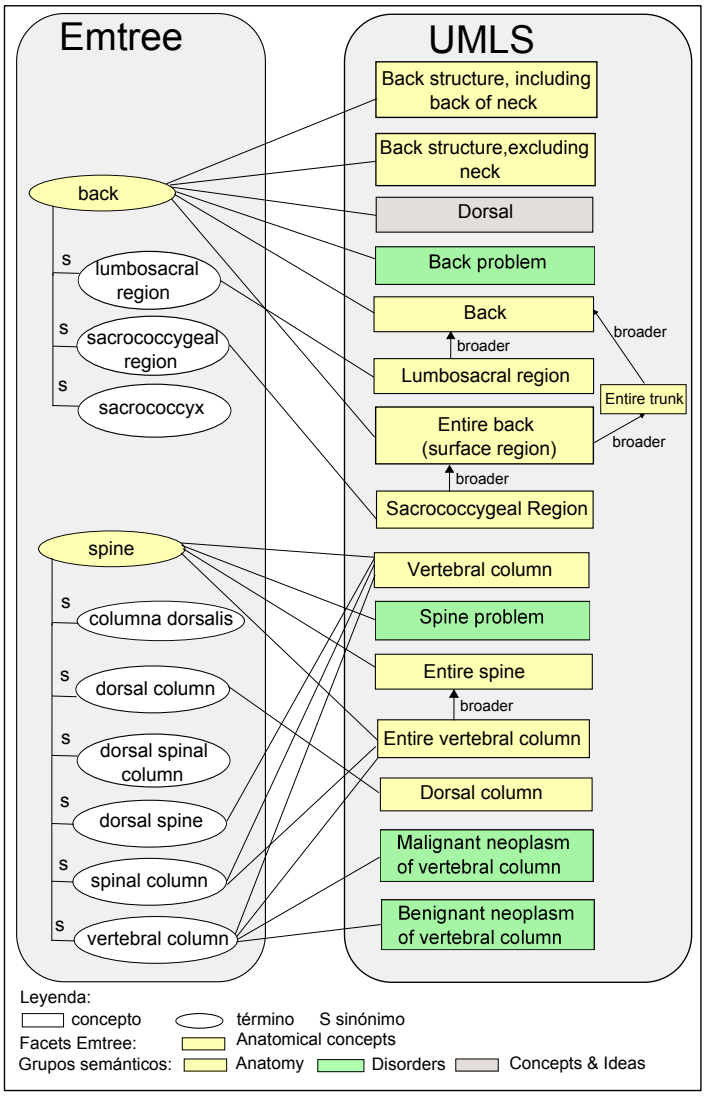

Figura 4.9: Ejemplos de alineamiento ambiguos: *back* y *spine*

ConceptName	**Lumbosacral Region**
CUI	C0024094
Semantic Group	Anatomy
Semantic Type	Body Location or Region
Synonym	Lumbosacral region structure (body structure) Lumbosacral region structure
Broader	Back Entire lower back
Narrower	Lumbar Region Skin structure of lumbosacral region Subcutaneous tissue structure of lumbosacral region Entire lumbosacral region

Figura 4.10: Información recuperada de UMLS

4.7. Alineamiento léxico de conceptos

Una vez obtenidas las correspondencias entre los términos de las terminologías, la etapa siguiente consiste en alinear sus conceptos, ya que la unidad con significado en el tipo de terminologías consideradas es el concepto. Para ello, el algoritmo agrupa todas las correspondencias léxicas de los términos de la terminología de entrada (incluyendo tanto los términos preferidos como los sinónimos) a cada concepto UMLS en un único alineamiento.

De esta forma, un alineamiento de concepto contiene la información del concepto de la terminología de entrada y del concepto UMLS que se equiparan léxicamente. Se puede definir como una tupla de cuatro elementos:

$$(\text{id}, C_{Fuente}, C_{umls}, \text{m})$$

donde

- id permite identificar cada alineamiento de concepto de forma unívoca.
- C_{Fuente} es el identificador del concepto de la terminología fuente que es el homólogo al concepto de la terminología destino en el alineamiento.
- C_{umls} es el CUI, el identificador del concepto de la terminología destino en UMLS homólogo al concepto de la terminología fuente en el alineamiento.
- m es el tipo de alineamiento, en este caso, alineamiento léxico simple de concepto (representado por 'lc').

El concepto *spine* visto en el apartado anterior quedaría equiparado, tal y como se muestra en la figura 4.11, a 7 conceptos UMLS. Como puede verse, en vez de tener 4 alineamientos al concepto UMLS *Vertebral column*, pasamos a tener sólo uno. Los alineamientos del concepto *spine* quedan así:

$$(7987, 45401, C0037949, 'lc')$$
$$(7988, 45401, C0150920, 'lc')$$
$$(7989, 45401, C1267072, 'lc')$$
$$(7990, 45401, C1280065, 'lc')$$
$$(7991, 45401, C0458459, 'lc')$$
$$(7997, 45401, C0346667, 'lc')$$
$$(7998, 45401, C0347311, 'lc')$$

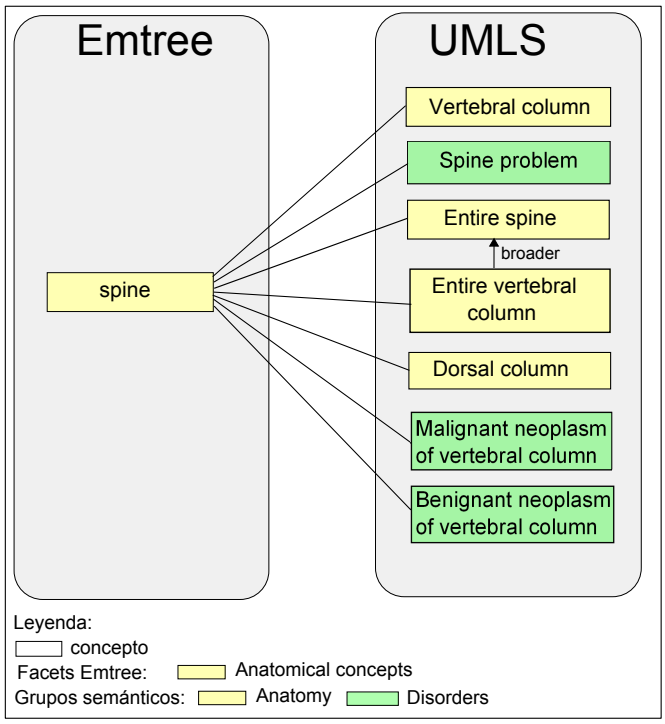

Figura 4.11: Ejemplo de alineamiento de conceptos: *spine*

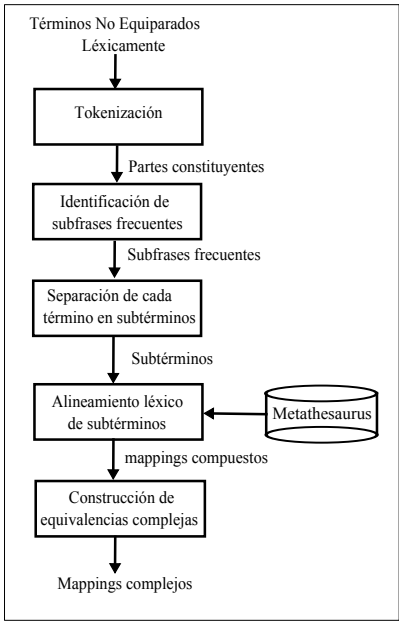

Figura 4.12: Esquema del proceso de Alineamiento Complejo basado en técnicas NLP

4.8. Alineamiento léxico compuesto

Aquellos conceptos de la terminología fuente para los que no se obtiene ninguna correspondencia léxica con algún concepto de la terminología destino en UMLS, y que están descritos por algún término con más de una palabra, se analizan sintácticamente usando técnicas de procesado de lenguaje natural. El objetivo es dividirlos en subfrases que se puedan equiparar con UMLS a fin de proporcionar, al menos, un alineamiento compuesto, es decir, una correspondencia entre un concepto de la terminología fuente y varios UMLS. Este tipo de alineamiento resulta interesante en aplicaciones como las tratadas en esta tesis doctoral, aplicaciones en las que las terminologías (o tesauros) indexan grandes colecciones de información.

Los conceptos de la terminología fuente pueden ser enumeraciones como *face, nose and sinuses*, frases como *bones of the leg and foot* o bien sintagmas nominales como *arm blood vessel*. El proceso mostrado en la figura 4.12 comienza separando las frases en sus partes constituyentes (tokens), usando técnicas de procesado de lenguaje natural. A continuación, haciendo uso de la información estructural de las terminologías, se identifican las unidades léxicas de los términos que son las unidades con semántica propia en el dominio. El siguiente paso es equiparar léxicamente estas subfrases a conceptos UMLS y obtener así alineamientos complejos. En la figura 4.13, puede verse

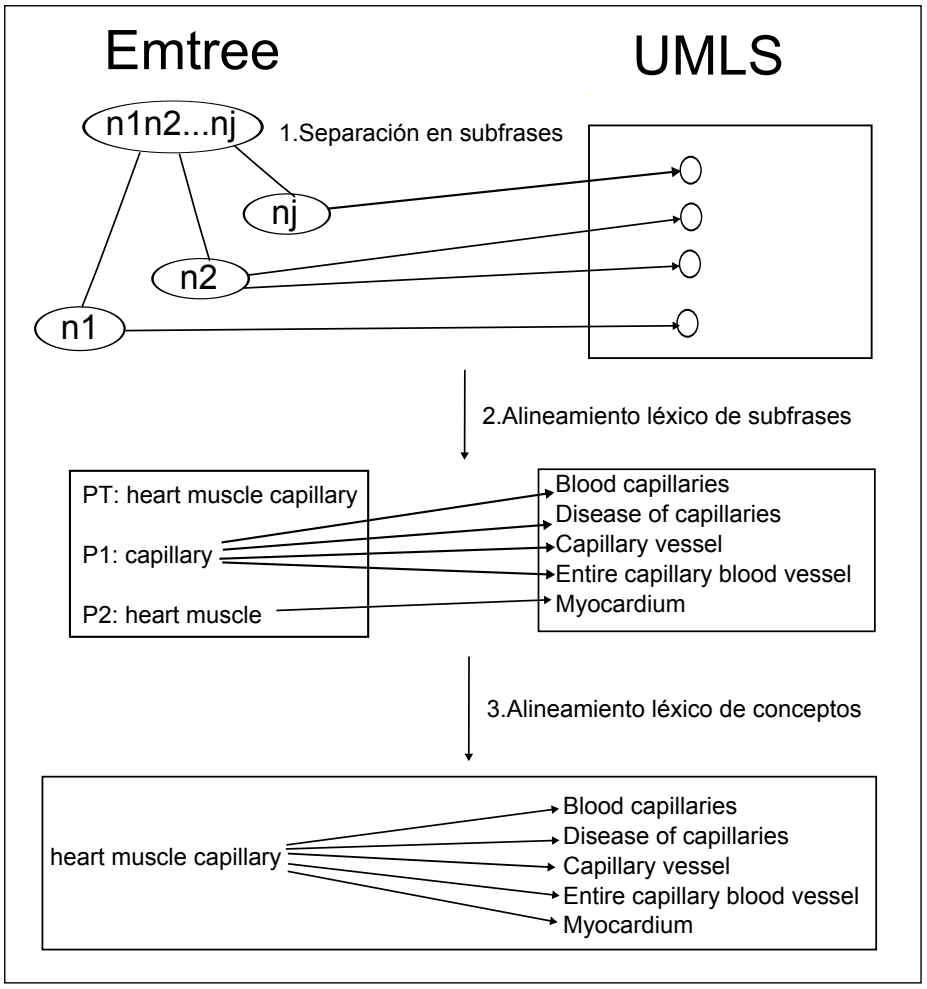

Figura 4.13: Figura del Alineamiento léxico complejo

una figura explicativa del proceso, donde un concepto es particionado en sus subfrases, que se buscan en UMLS para obtener las equiparaciones léxicas a ellas. Finalmente, los conceptos UMLS obtenidos quedan equiparados al concepto Emtree.

A continuación, se explican con detalle las etapas del proceso.

Etapa 1: Análisis sintáctico de los términos (Tokenization)

El proceso de separar una frase en sus partes constituyentes (tokens) se llama tokenización (tokenization). Una opción sería separar todas las palabras que forman el término e intentar equipararlas por separado. Esta forma tiene varios inconvenientes, por ejemplo, se buscarían preposiciones, deter-

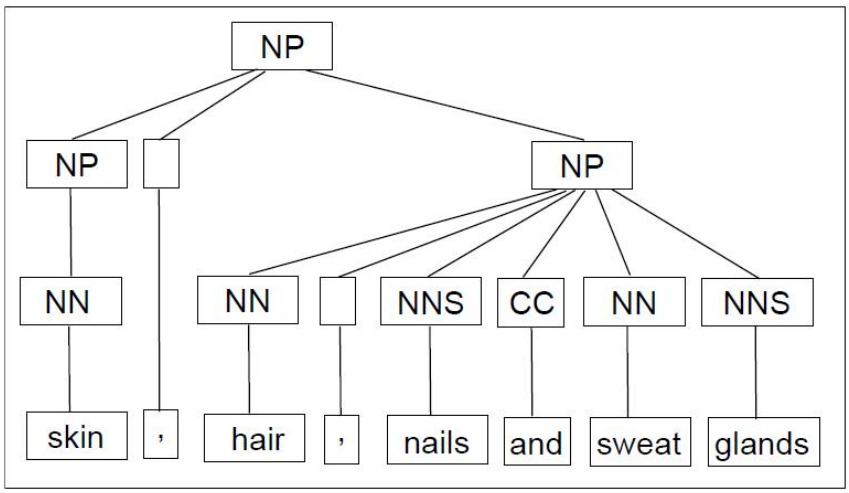

Figura 4.14: Análisis sintáctico del término *skin, hair, nails and sweat glands*

minantes o adjetivos que no tienen significado propio; también se buscarían palabras sueltas cuando realmente el significado puede estar en la agrupación de palabras. Para intentar hacer una tokenización lo más fiel posible a la semántica del término original, utilizamos técnicas de procesado de lenguaje natural para intentar extraer subtérminos con significado.

El proyecto OpenNLP[2] proporciona una estructura para la coordinación de varios proyectos necesarios para el Procesado de Lenguaje Natural. Como parte de ella, han definido una librería Java que implementa los elementos básicos de ese procesado. Esta librería Java es la que hemos usado para analizar los términos Emtree y obtener las frases nominales (NP) que las componen. Cada término preferido se analiza utilizando la librería Java, que realiza el análisis sintáctico y nos devuelve sus partes constituyentes. En la mayoría de los casos, será una frase válida formada por un nombre (NN), que determina el significado básico del término, y por un determinante (DT) u otras palabras adyacentes como adjetivos (o nombres que funcionan como adjetivos) que complementan el significado. En la figura 4.14, puede verse el resultado para el término Emtree *skin, hair, nails and sweat glands*. Si el término sólo contiene nombres, obtendrá cada uno por separado. Por ejemplo, *arm blood vessel* se descompone en los 3 nombres como se muestra en la figura 4.15. En la figura 4.16, se muestra el caso de *digestive tract blood vessel*, donde *digestive* es identificado correctamente como adjetivo.

El resultado de esta etapa es la separación de cada término en sus partes constituyentes según OpenNLP, incluida la identificación del tipo de palabra

[2]http://incubator.apache.org/opennlp/

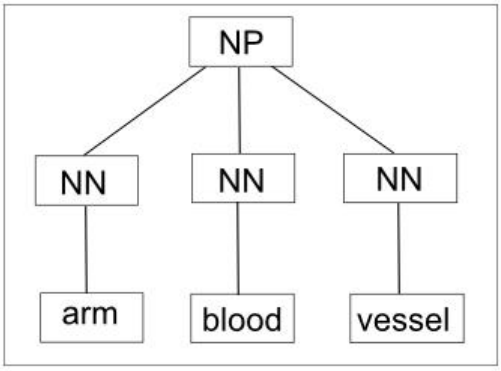

Figura 4.15: Análisis sintáctico del término *arm blood vessel*

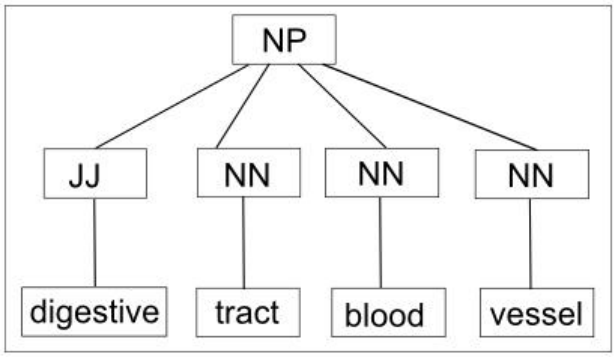

Figura 4.16: Análisis sintáctico del término *digestive tract blood vessel*

(adjetivo, nombre, ...) y descartando palabras sin significado como artículos, preposiciones, conjunciones o adverbios.

Etapa 2: Identificación de unidades léxicas

La mayoría de los términos de la terminología fuente estarán formados por 2 o más nombres. En ocasiones, algunas partes de los términos, formadas por dos o más palabras, aparecerán frecuentemente en otros términos de la terminología, indicando que dicha agrupación constituye una unidad con semántica propia en el dominio. Por ello, consideramos este grupo de palabras como unidades léxicas. Esto nos permitirá descomponer los términos en unidades léxicas, con el fin de equipararlos con UMLS. Un ejemplo de unidad léxica es *blood vessel*, que aparece frecuentemente formando parte de otros términos Emtree, como *arm blood vessel* o *optic nerve blood vessel*. Por ello, consideramos *blood vessel* como una unidad léxica y por tanto, indivisible.

El proceso procesa todos los términos y las subfrases obtenidas aplicando

las técnicas de procesado de lenguaje natural e identifica las unidades léxicas analizando la frecuencia de colocación de las palabras que componen sus términos dentro de EMTREE.

Etapa 3: Separación de cada término en sus partes

Seguidamente, se realiza la división definitiva del término, teniendo en cuenta tanto los constituyentes obtenidos por el procesado inicial como las unidades léxicas identificadas en la etapa anterior y las relaciones jerárquicas entre los conceptos Emtree. Se omiten todos los determinantes, conjunciones y preposiciones. A continuación, se revisan los casos más representativos.

Las frases nominales que incluyen sólo un nombre son indivisibles, por tanto ya no es posible obtener equiparación léxica. Casos como *antenna*, *scolex* o *diabetogenesis*.

Las frases nominales que incluyen 2 palabras (un nombre y otra palabra) se parten en dos. Por ejemplo, el término original *pia artery* se parte en dos tokens *pia* y *artery*.

Las frases nominales que incluyen más de 2 palabras se dividen aplicando los siguientes criterios en este orden. Primero, se parten teniendo en cuenta los broaders del término en la terminología fuente que se pretende equiparar. Para cada término, el algoritmo analiza su frase nominal con el objetivo de identificar algún broader en él. Si el término incluye los nombres de uno o más broaders, el resultado serán los nombres de los broaders más las palabras restantes. Por ejemplo, los broaders del término *capillary basement membrane* son *basement membrane* y *microvasculature* y el término se parte en *capillary* y *basement membrane*. En la figura 4.17, se muestra el análisis sintáctico del concepto *heart left ventricle muscle*. Aquí, aplicamos directamente la comprobación de sus broaders que son *heart left ventricle* y *heart muscle* y estos son las subfrases que consideramos.

Una vez que todos los términos han sido particionados, se identifican las unidades léxicas (por ejemplo, *blood vessel*) y se asume que es más probable encontrarlas en UMLS. Los términos que no incluyen ningún broader se descomponen teniendo en cuenta estas unidades léxicas. Por ejemplo, *digestive track blood vessel* se parte en *digestive track* y *blood vessel*.

Cuando no hay broaders útiles, el término queda descompuesto en las frases nominales identificadas por OpenNLP. Por ejemplo, el término *skin, hair, nails and sweat glands*, que tiene como broader *anatomical concepts*, queda dividido en *skin*, *hair*, *nails* y *sweat glands*.

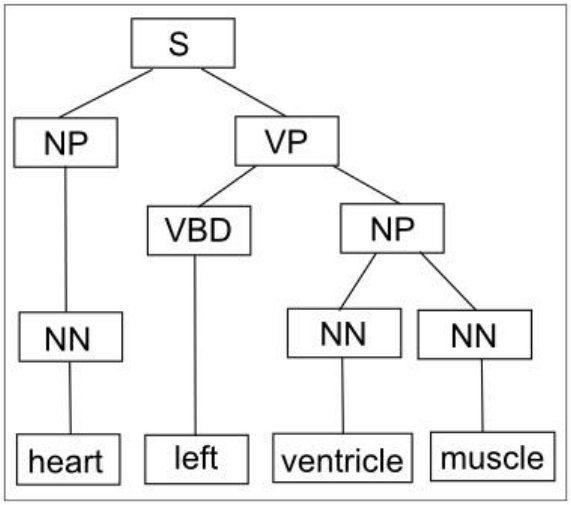

Figura 4.17: Análisis sintáctico del término *heart left ventricle muscle*

Etapa 4: Alineamiento de las partes constituyentes

A continuación, se realiza el alineamiento léxico de los tokens resultantes en UMLS usando, como en el apartado 4.3.1, la opción NormalizeString. Igual que entonces, por cada token se pueden obtener 0, 1 o más conceptos UMLS. Si el alineamiento es completo, todos se equiparan léxicamente con algún concepto UMLS. Si es parcial, alguno de ellos no se equipara a ningún concepto.

Los alineamientos se definen como una tupla de cinco elementos:

$$(\text{id}, c_{Fuente}, tk^1, c^1_{UMLS}, tk^2, c^2_{UMLS}, ..., tk^n, c^n_{UMLS}, \text{m})$$

donde

- id permite identificar cada alineamiento de forma unívoca
- c_{Fuente} es el identificador del concepto de la terminología fuente que contiene el token que es homólogo al concepto UMLS
- tk^1, tk^2,..., tk^n son las partes constituyentes del término que describe el concepto fuente
- c^1_{UMLS},c^2_{UMLS},..., c^n_{UMLS} son los conceptos UMLS que se equiparan a cada una de las partes constituyentes tk^1, tk^2,..., tk^n respectivamente.
- m es el tipo de alineamiento, en este caso, c indica que se trata de un alineamiento compuesto.

De esta forma, el concepto Emtree queda equiparado a varios conceptos UMLS, a través de varios alineamientos, que provienen de subtérminos diferentes y por tanto se considera un alineamiento compuesto, en contraposición, del alineamiento léxico simple obtenido en las fases anteriores.

En la figuras 4.18, 4.19 y 4.20, se muestran algunos ejemplos que reflejan los diferentes casos vistos en este método. En la figura 4.18, el concepto *pia artery* está formado por dos nombres por lo que esas serán sus partes constituyentes *pia* y *artery*. El primero de ellos se equipara léxicamente al concepto UMLS *Cisplatin/Doxorubicin/Ifosfamide protocol*, del grupo semántico *Procedures* y al *RICTOR gene* de *Genes and Molecular Sequences*. Como puede observarse, estos conceptos no parecen tener el mismo significado que *pia*; serán validados en el capítulo siguiente. El concepto Emtree *artery* se equipara a dos conceptos UMLS, *Arteries* y *Procedures on Arteries*, el primero de *Anatomy* y el segundo de *Procedures*.

El concepto *heart muscle capillary* ilustra el caso en que usamos los broaders para realizar el particionado del término. Está formado por 3 nombres, pero se revisan sus broaders y uno de ellos es la unidad léxica *heart muscle*, por lo que ésta será una de sus partes constituyentes, la otra será el nombre que queda, *capillary*. Tanto *pia artery* como *heart muscle capillary* tienen un alineamiento completo ya que todas sus partes se equiparan léxicamente a 1 o más conceptos UMLS. Estos son los alineamientos resultantes:

(361, 37420, 'artery', C0003842, 'artery', C0397581, 'pia', C0063264, 'pia', C1836945, 'c')
(5276, 58417, 'heart muscle', C0027061, 'capillary', C0006901, 'capillary', C0155765, 'capillary', C0935624, 'capillary', C1280521, 'c')

En la figura 4.19, se muestra el ejemplo *mesentery blood vessel*, que se descompone teniendo en cuenta la unidad léxica, *blood vessel*. Por tanto, los subtérminos son *blood vessel* y *mesentery*. El primero se equipara a *Blood vessels* y a *Entire blood vessel*, donde además, éste es broader de aquel. La segunda parte constituyente se equipara únicamente a *Mesentery*.

Por último, en la figura 4.20, se muestra el ejemplo *muscle fiber membrane*, que está formado por 3 nombres y donde ninguna de las técnicas puede ser aplicada ya que ninguna palabra o combinación de palabras coincide con un broader. Tampoco contiene ninguna unidad léxica. Por tanto, queda particionada en *muscle*, *fiber* y *membrane*. La parte *muscle* se equipara léxicamente a 2 conceptos UMLS, *Muscle* y *Set of muscles*, *membrane* se equipara a *Membrane* y a *Tissue membrane* que es narrower del anterior, ambos del grupo semántico *Anatomy*, y a *Membrane Device Component* del grupo *Procedures*. La palabra *fiber* consigue equiparaciones a 8 conceptos

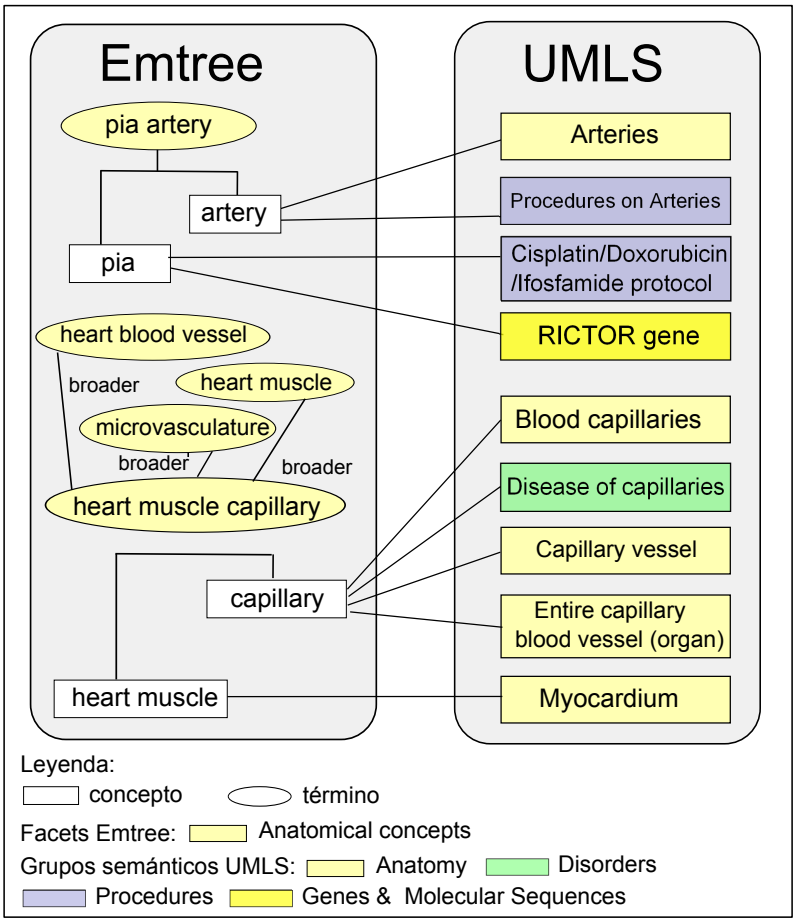

Figura 4.18: Ejemplos de alineamiento léxico de tokens(1)

UMLS de diferentes grupos semánticos: *Tissue fiber* de *Anatomy*, *Dietary fiber*, de *Objets*, *Fiber* de *Chemical and Drugs*, *Plant Fiber* y *Hemp fiber* de *Living Beings*, *Fiber - RoleCole* y *Fiber brand of calcium polycarbophil* de *Concepts and Ideas* y, por último, *Fiber Device Component* de *Procedures*.

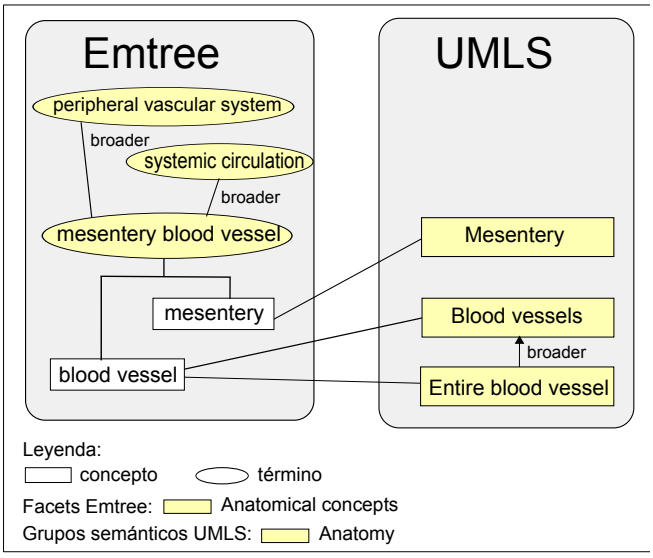

Figura 4.19: Ejemplos de alineamiento léxico de tokens(2)

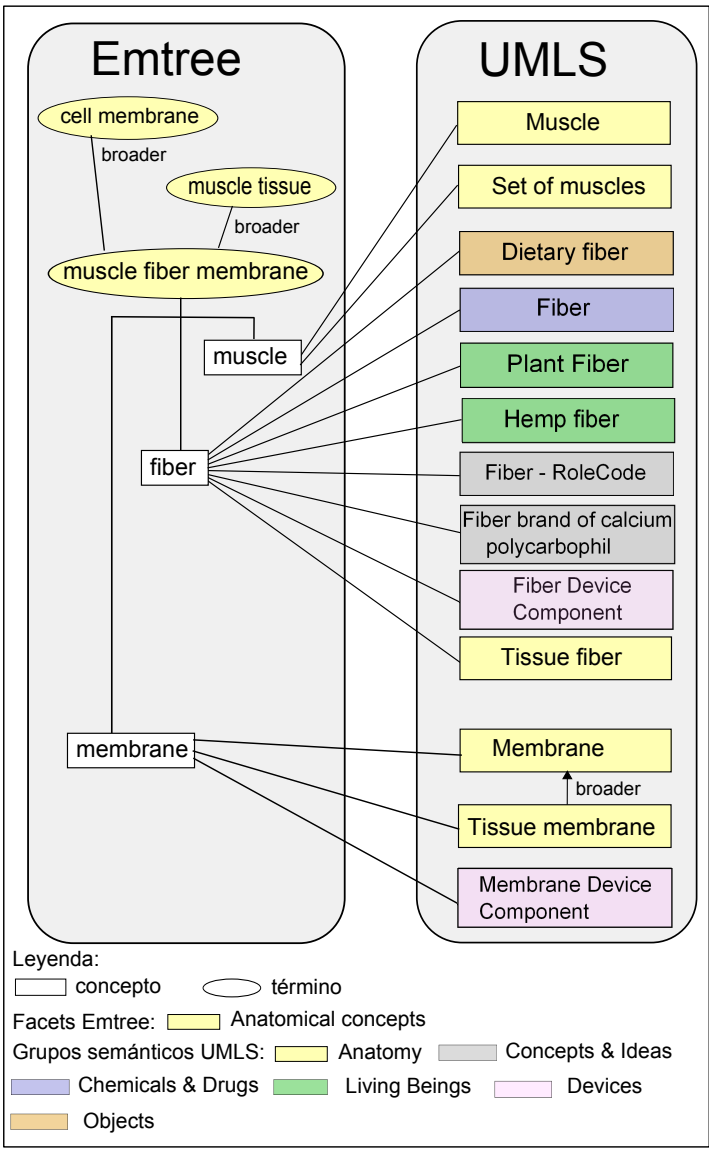

Figura 4.20: Ejemplos de alineamiento léxico de tokens(3)

Etapa 5: Construcción de equivalencias complejas

En esta etapa, se analizan las equiparaciones léxicas obtenidas para las partes constituyentes de cada concepto fuente. Se observa que la estructura del concepto original determina la adecuación de las equiparaciones obtenidas. Así, los conceptos de tipo enumeración formados por una serie de sintagmas nominales relacionados por una preposición "y", es decir, términos con patrones sintácticos de la forma *NP ,NP* y NP*, se equiparan al conjunto de los conceptos UMLS obtenidos para cada uno de los subtérminos, lo que denominaremos un alineamiento complejo UnionMatch, ya que el concepto fuente de existir en UMLS sería la unión de todos ellos, tal como fue descrito en [Doe01]. En la figura 4.21, se ven dos ejemplos de este tipo: *embryonic, fetal and placental structures* y *hamsters and gerbils*.

El concepto Emtree *embryonic, fetal and placental structures* se descompone en los subtérminos *embryonic structures*, *fetal structures* y *placental structures* ya que, al analizar sintácticamente se obtienen 3 adjetivos y un nombre unidos por una preposición y por tanto, se entiende como una enumeración donde los 3 adjetivos afectan al nombre. Cada uno de estos subtérminos se equipara léxicamente a un concepto UMLS, y son respectivamente *Embryonic Structures*, *Fetal Structures* y *Placenta*, todos del grupo semántico *Anatomy*. El concepto Emtree *hamsters and gerbils*, de la facet *Organism names*, está formado por 2 nombres que, por tanto, serán los subtérminos. El subtérmino *hamsters* se equipara léxicamente a *Hamsters* del grupo semántico *Living Beings* y a *allergy testing hamster* de *Procedures*. El subtérmino *gerbils* se equipara léxicamente a *Gerbils* de *Living Beings*, *Gerbil antigen* de *Chemicals and Drugs* y *allergy testing gerbil* de *Procedures*.

En el resto de casos, el concepto inicial se ha particionado en subtérminos que agrupan uno o más palabras del término original, sea usando información de broaders, unidades léxicas o directamente y, por tanto, se puede hablar de un alineamiento complejo BroadMatch, de forma que el concepto fuente de existir en UMLS sería un concepto más específico que el concepto UMLS al que se equipara el núcleo del sintagma nominal, donde la especificidad viene determinada por los atributos que acompañan al sintagma nominal. De este tipo, son todas las equiparaciones explicadas en la etapa anterior y que aparecen en las figuras 4.18, 4.19 y 4.20.

Se trata, por tanto, de alineamientos complejos. Según el tipo de información que estos alineamientos aporten con respecto al significado global del concepto de la terminología fuente, se pueden clasificar en:

- **BroadMatch**: cuando el concepto de la terminología fuente queda equiparado a conceptos de significado más general en la terminología destino.

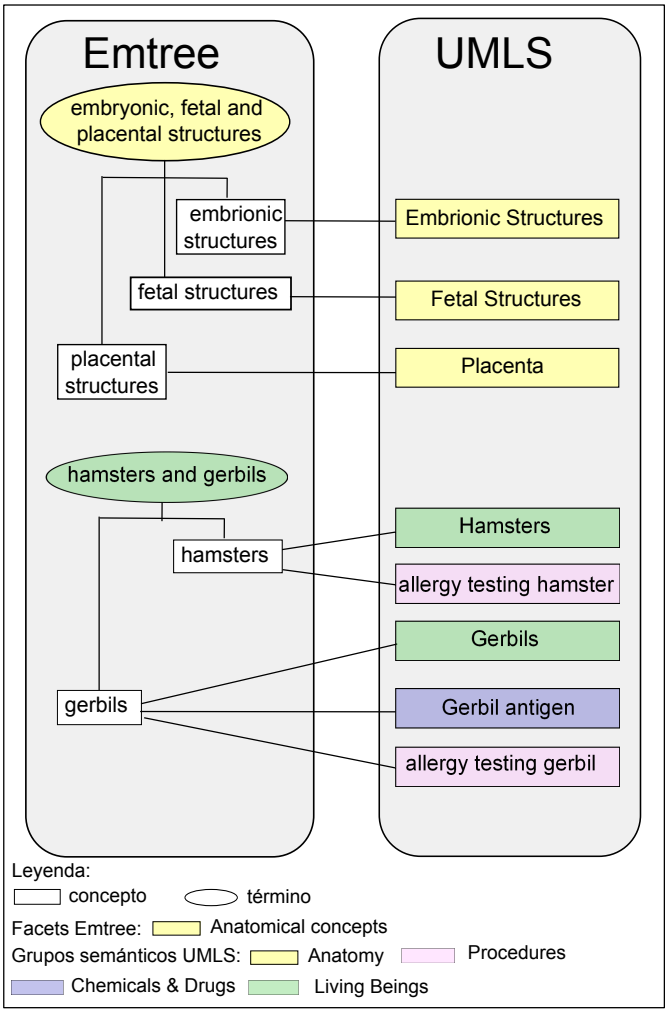

Figura 4.21: Ejemplos de alineamiento complejo UnionMatch

- **UnionMatch**: cuando el concepto de la terminología fuente queda equiparado a varios conceptos de la terminología destino cada uno de los cuales tiene parte de su significado, incluso pudiendo solaparse.

Los alineamientos se definen de la siguiente manera:

$$(\text{id}, C_{Fuente}, c^1_{UMLS}, c^2_{UMLS}, \ldots, c^n_{UMLS}, \text{m})$$

donde

- id permite identificar cada alineamiento de forma unívoca
- C_{Fuente} es el identificador del concepto de la terminología fuente

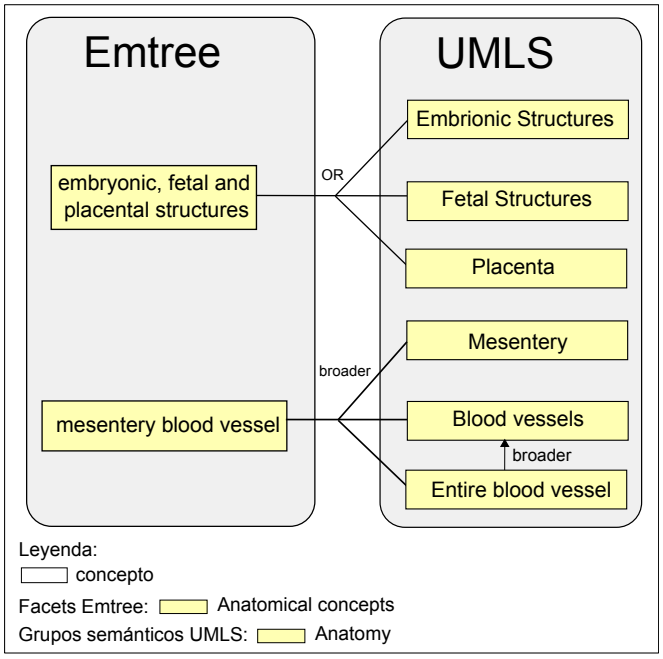

Figura 4.22: Alineamientos complejos

- c^1_{UMLS},c^2_{UMLS},..., c^n_{UMLS} son los conceptos UMLS que se equiparan a alguna de las subfrases del concepto
- m es el tipo de alineamiento, b si es un alineamiento complejo BroadMatch y u para un alineamiento complejo UnionMatch.

Para algunos de los ejemplos vistos, las tuplas quedan así:

$$(361, 37420, C0003842, C0397581, C0063264, C1836945,' b')$$
$$(5276, 58417, C0027061, C0006901, C0155765, C0935624, C1280521,' b')$$
$$(17238, 668831, C0446378, C0032043, C0013948,' u')$$
$$(20823, 668894, C0018557, C2097388, C0017462, C1446514, C2097387,' u')$$

En la figura 4.22, se muestra cómo queda finalmente la equiparación para un mapping complejo UnionMatch como *embryonic, fetal and placental structures* y para uno BroadMatch como *mesentery blood vessel*, que quedan equiparados a todos los conceptos obtenidos.

En este punto, cada concepto de la terminología Emtree está equiparado léxicamente a uno o varios conceptos UMLS, bien al concepto directamente o a alguna de sus partes constituyentes. Al ser un alineamiento léxico, los conceptos UMLS pueden no ser realmente equivalentes semánticamente al concepto Emtree, que es lo que buscamos. Por tanto, el siguiente paso de

nuestro proceso es realizar la validación de estos alineamientos usando la información semántica existente en las terminologías.

Capítulo 5

Validación del alineamiento

5.1. Introducción

La mayor parte de las técnicas utilizadas actualmente para equiparar ontologías o terminologías son fundamentalmente léxicas [ES07] o hacen uso de recursos externos, como Wordnet o UMLS, en busca de sinónimos y variaciones léxicas de los conceptos [HGH+09]. Algunos también explotan la estructura y/o semántica de las terminologías fuentes [FBA+07]. Hoy en día, pues, la equiparación automática del alineamiento entre terminologías usadas en aplicaciones reales es posible gracias al reciente número de técnicas disponibles. Sin embargo, estas técnicas no son fiables al cien por cien, pudiendo dar lugar a errores que introducen inconsistencias entre las fuentes y ambigüedades en los alineamientos obtenidos.

Una aportación significante de esta tesis es proporcionar técnicas de validación para detectar de forma automática estos errores, apoyándonos en que éstos salen a la luz cuando se tiene en cuenta la estructura de las fuentes. Así, en esta tesis hemos desarrollado técnicas que usan las propiedades estructurales de las fuentes para caracterizar y validar las equiparaciones léxicas obtenidas previamente.

5.2. Fundamentos de las técnicas propuestas

Las técnicas propuestas en esta tesis doctoral para validación automática de alineamientos entre terminologías están fundamentadas en los siguientes principios:

1. **Principio de compatibilidad**: las terminologías fuentes y destino deberán ser compatibles con el alineamiento resultante. Por ejemplo, en la

figura 4.8, el alineamiento léxico obtenido equipara el concepto Emtree *bacterium* al concepto UMLS *Bacteria aspects*. En Emtree, *bacterium* es un concepto de tipo organismo, mientras que en UMLS *bacteria aspects* está clasificado como un concepto o idea. El significado de ambos conceptos es diferente y, por tanto, son disjuntos; en el caso hipotético de considerar que ambos fuesen equivalentes, las fuentes (Emtree y UMLS) y el alineamiento entre las fuentes serían incompatibles, ya que estaríamos considerando como homólogos conceptos con significados diferentes en las dos terminologías.

2. **Principio de similitud estructural**: Si un concepto c_1 de la terminología fuente se equipara correctamente a un concepto c_2 de la terminología destino, entonces los conceptos que están próximos a c_1 (por sus relaciones estructurales) probablemente serán homólogos a los conceptos estructuralmente próximos a c_2. En la figura 5.1, se muestra el concepto en la terminología Emtree *back*, que es un concepto más específico que *body regions* y más general que *back muscle* y *spine* en la terminología Emtree. Dicho concepto se equipara léxicamente, entre otros, con el concepto UMLS *Back*. El concepto Emtree *body regions* se equipara léxicamente, entre otros, al concepto UMLS *Body Regions*, que, a su vez, es el broader de Back. En los narrowers, también se observan casos similares como el concepto Emtree *back muscle* que se equipara léxicamente con el concepto UMLS *Entire skeletal muscle of back* que, a su vez, está próximo a *Back*. Por tanto, los conceptos *body regions* y *back muscle* que están próximos al concepto Emtree *back* son homólogos a los conceptos que están próximos a su homólogo en UMLS *Back*.

3. **Principio de precisión**: El alineamiento resultante no debería proporcionar más de una equiparación para el mismo concepto fuente o el mismo concepto destino; es decir, un concepto en la terminología fuente (o destino) no puede ser homólogo a varios en la terminología destino (o fuente) en equiparaciones simples. Se distinguen dos casos:

   a) El mismo concepto fuente (o destino) se equipara a dos o más conceptos con significados claramente diferentes. En este caso, al menos una o más equiparaciones violan el principio de compatibilidad. Ejemplo: el concepto Emtree *gill*, de la facet *Anatomical concepts*, se equipara léxicamente con los conceptos UMLS *Gill structure* del grupo semántico *Anatomy* y el concepto *gill unit of measure* del grupo *Concepts and Ideas*. Por tanto, este último concepto no cumple el principio de compatibilidad. La resolución

de las ambigüedades la resolveremos identificando las categorías de alto nivel que son similares semánticamente, y considerando incorrectas las equiparaciones entre conceptos pertenecientes a categorías diferentes semánticamente.

b) El mismo concepto fuente (o destino) se equipara a dos o más conceptos con significados muy similares.

1) Si la similitud entre los conceptos viene dada por relaciones estructurales (narrower, broader), entonces los conceptos más similares se consideran redundantes. En la figura 5.1, esto puede verse para el concepto Emtree *body regions*, que se equipara léxicamente a los conceptos *Body Regions*, *Body part* y *Entire body region* que, a su vez, están relacionados directamente entre sí, por lo consideraremos que dos de ellos son redundantes. En este caso, aplicaremos métodos basados en similitud estructural para identificar estas redundancias.
2) Si la similitud no viene determinada por las relaciones estructurales, consideraremos sólo dos casos: que las equiparaciones estén en conflicto (y una de ellos sea incorrecta); o que ambas sean correctas y la equiparación resultante debería ser la unión lógica de los conceptos equiparados. En la figura 5.1, se muestra el término Emtree *spine* que se equipara léxicamente con *Entire spine*, *Entire vertebral column*, *Dorsal column* y *Vertebral column*. *Entire vertebral column* es narrower de *Entire spine*, por lo que alguna de ellas puede ser redundante pero no existe una relación directa en UMLS con los otros 2 conceptos. Dado que un concepto Emtree es un conjunto de términos equivalentes que se agrupan en un concepto, se puede suponer que la equiparación resultante será la unión lógica de 3 conceptos UMLS: *Vertebral column*, *Dorsal column* y el que resulte más similar del par (*Entire spine*, *Entire vertebral column*) aplicando el principio de similitud estructural.

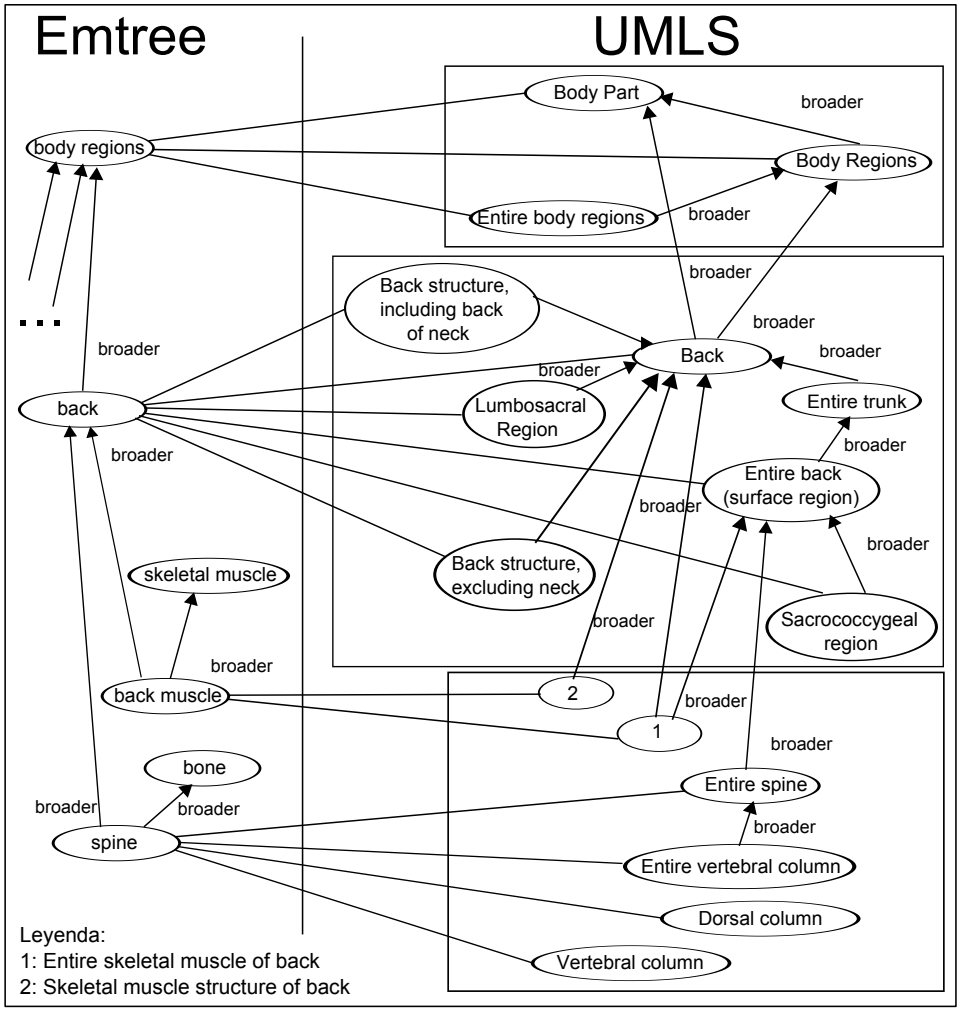

Figura 5.1: Equiparaciones léxicas y proximidad estructural para el concepto Emtree *back*

5.2.1. Similitud estructural entre conceptos

El principio de similitud estructural tiene en cuenta que el significado de los conceptos viene representado, en gran medida, por el lugar que éstos ocupan en la red de relaciones de las terminologías y, por ello, explota la semejanza de las relaciones estructurales entre los conceptos de las terminologías fuente y destino. Para poder medir la similitud entre conceptos, hacemos uso de las técnicas básicas de segmentación de ontologías [SR06]. Un segmento de una ontología es un fragmento coherente de ésta, incluyendo un conjunto de conceptos y todas las relaciones de la ontología que los definen. Los algoritmos básicos para extracción de segmentos parten de uno o varios conceptos y crean un extracto que contiene a estos conceptos y todos los relacionados a través de la estructura de uniones de la ontología. En nuestro particular caso, la estructura de uniones considerada serán las relaciones estructurales de tipo broader y narrower. Así, un extracto de un concepto estará formado por el propio concepto y los conceptos más generales y más específicos directamente relacionados con el propio concepto. Como ejemplo, en la figura 5.1, mostramos el extracto para el concepto *back*, que incluye el propio concepto y los conceptos broader *body regions* y narrowers *back muscle* y *spine*. La figura muestra también los conceptos UMLS que se equiparan a cada uno de estos conceptos del extracto.

Agrupación de los conceptos UMLS equiparados a un concepto fuente

En muchos casos, el conjunto de conceptos UMLS a los que se equipara un concepto fuente están relacionados directamente entre sí por relaciones de tipo broader/narrower. Esto puede apreciarse en la figura 5.2, donde conceptos UMLS a los que se equipara el concepto Emtree *body regions* están relacionados entre sí: *Body Part* es el broader de *Body Regions*, que a su vez, es el broader de *Entire body region*.

Teniendo en cuenta lo anterior, nuestro método agrupa en una estructura que llamamos clúster a todos los conceptos UMLS que se equiparan léxicamente con un concepto fuente y que están relacionados entre sí mediante relaciones broader/narrower. El clúster se enriquece con todas la relaciones tipo narrower/broader, que existen en el Metathesaurus de UMLS, entre los conceptos asociados en el clúster. Tras esta agrupación, un concepto fuente estará equiparado a conceptos que forman ninguno, uno o incluso dos clusters. La no existencia de clusters indica que los conceptos UMLS no tienen relación entre sí y están en partes diferentes de la jerarquía.

Los conceptos que forman un clúster, a su vez, tienen una fuerte relación

entre sí; en algunos casos, incluso pueden ser muy similares. Será importante medir el nivel de similitud de esos conceptos, con el fin de seleccionar el más adecuado para el alineamiento. En la figura 5.2, se muestran en colores distintos ejemplos de clusters, entre ellos, los formados por los conceptos *Back* y *Lumbosacral Region*, *Entire back (surface region)* y *Sacrococcygeal region* y, por último, *Entire spine* y *Entire vertebral column*.

Como decíamos, los conceptos dentro de un clúster serán muy similares y, por tanto, en el proceso de desambiguación posterior, serán usados para elegir el mejor alineamiento de los obtenidos.

Cálculo de factores de similitud

Teniendo en cuenta el principio de similitud estructural, podemos calcular un factor de semejanza o similitud para cada uno de los alineamientos léxicos obtenidos previamente. Definimos dicho factor como una media de los factores de similitud calculados teniendo en cuenta los sub-extractos y super-extractos del concepto. En concreto, se obtienen como la proporción entre el número de conceptos equiparados en el sub-segmento/super-segmento del concepto fuente y el número de conceptos fuente totales en dicho sub-segmento/super-segmento. Así, para cada alineamiento, el factor de similitud para los sub-extractos de los conceptos se define como:

$$similitudsub(m(c_1, c_2, l)) = \frac{|conceptos\ equiparados\ en\ el subextracto\ para\ c_1|}{|subextracto\ para\ c_1|} \tag{5.1}$$

El factor de similitud para los super-extractos de los conceptos se define como:

$$similitudsuper(m(c_1, c_2, l)) = \frac{|conceptos\ equiparados\ en\ el\ superextracto\ para\ c_1|}{|superextracto\ para\ c_1|} \tag{5.2}$$

Por último, el factor de similitud es la media de los dos factores anteriores:

$$similitud(m(c_1, c_2, l), extracto) = \frac{similitudsub + similitudsuper}{2} \tag{5.3}$$

En la figura 5.1, ya mostramos las relaciones para el concepto *back*. A partir de ellas, obtenemos los factores de sub-extracto y super-extracto. En cuanto a los broaders, en Emtree, *back* tiene un único broader *body regions*, que en UMLS se equipara con 3 conceptos UMLS, que son *Body Regions*,

Body Part y *Entire body region*. El concepto *back* se equipara léxicamente con 6 conceptos UMLS, de los cuales sólo *Back* tiene como broaders los que equiparan a body regions (*Body Regions* y *Body Part*). Esto da un factor de similitud de super-extractos de 100 % para el alineamiento (*back*, *Back*). Esto se muestra en la figura 5.3.

Los otros 5 conceptos a los que se equipara léxicamente *back* tienen como broaders otros términos que no son ninguno de esos tres, *Body Regions*, *Body Part* y *Entire body region*, por tanto, su factor de similitud de super-extracto es 0. En concreto, *Lumbosacral Region* y *Back structure, excluding neck* por ser narrower del principal *Back*, *Entire back (surface region)* es narrower no directo de *Back*, *Sacrococcygeal region* a su vez es narrower de *Entire back (surface region)*. El concepto *Back structure, including back of neck* no tiene relación directa con ninguno de los conceptos UMLS.

En cuanto a los narrowers, el concepto Emtree *back* tiene dos alineamientos, *back muscle* y *spine*. El concepto *Back* tiene factor de similitud de sub-extractos de 50 % ya que de los dos narrowers del Emtree, *spine* y *back muscle*, sólo el segundo se equipara léxicamente a dos conceptos *Entire skeletal muscle of back* y *Skeletal muscle structure of back* que son narrowers de *Back*. En cambio, para *Entire back (surface region)*, el factor de similitud de sub-extractos es 100 % ya que tiene como narrowers *Entire spine*, equiparado léxicamente a *spine*, y *Entire skeletal muscle of back*, a *back muscle*. Estos dos casos pueden ver en la figura 5.4. Por último, *Lumbosacral Region*, *Sacrococcygeal region*, *Back structure, including back of neck* y *Back structure, excluding neck* no tienen narrowers vinculados a los narrowers de *back* por lo que su factor de similitud de sub-extractos es 0.

En la figura 5.5, se reúne toda la información explicada en los apartados anteriores. Además, puede apreciarse algún caso especial. Por ejemplo, los conceptos UMLS que se equiparan al concepto Emtree *back muscle* no tienen factor de similitud de sub-extractos, lo cual indica que no pudo ser calculado al ser conceptos hoja (sin narrowers). Por otra parte, los conceptos UMLS *Vertebral column* y *Dorsal column* no muestran relación con ninguno de los conceptos del gráfico, ya que pertenecen a zonas estructurales alejadas.

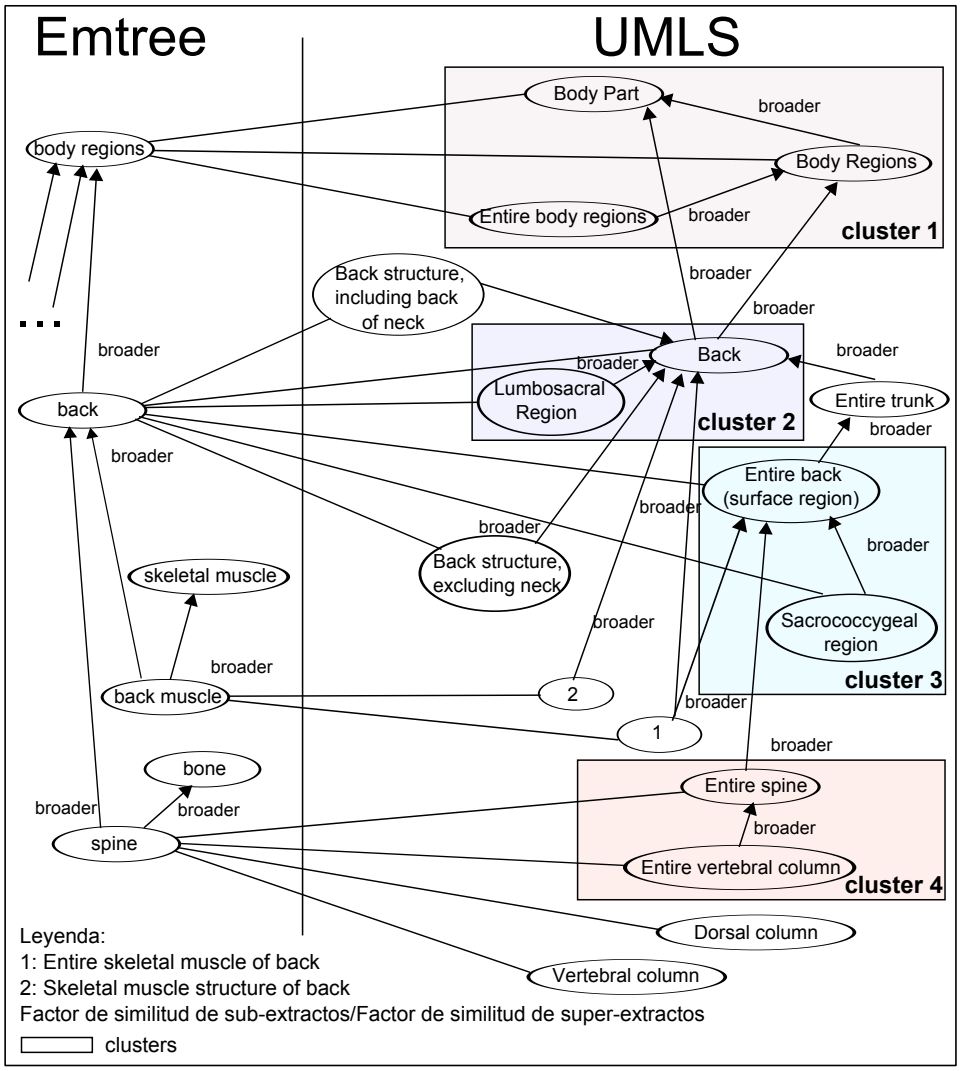

Figura 5.2: Clusters para el concepto Emtre *back*

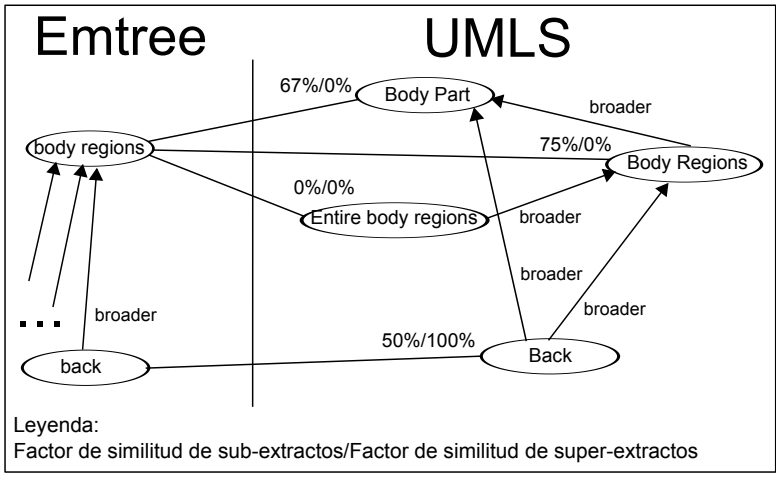

Figura 5.3: Factores de similitud de los sub-extratos y super-extractos para los conceptos *body regions* y *back*

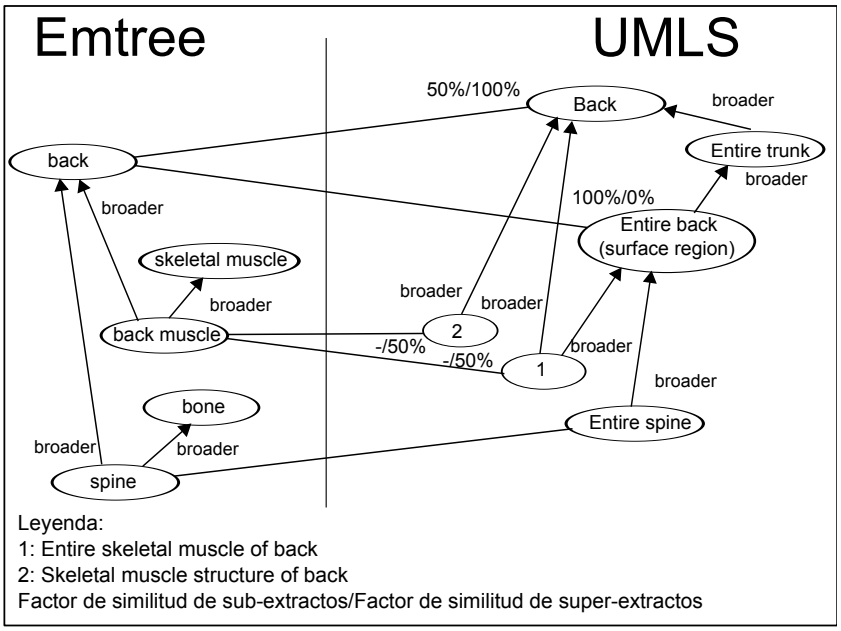

Figura 5.4: Factor de similitud teniendo en cuenta el sub-extracto para el concepto Emtree *back*

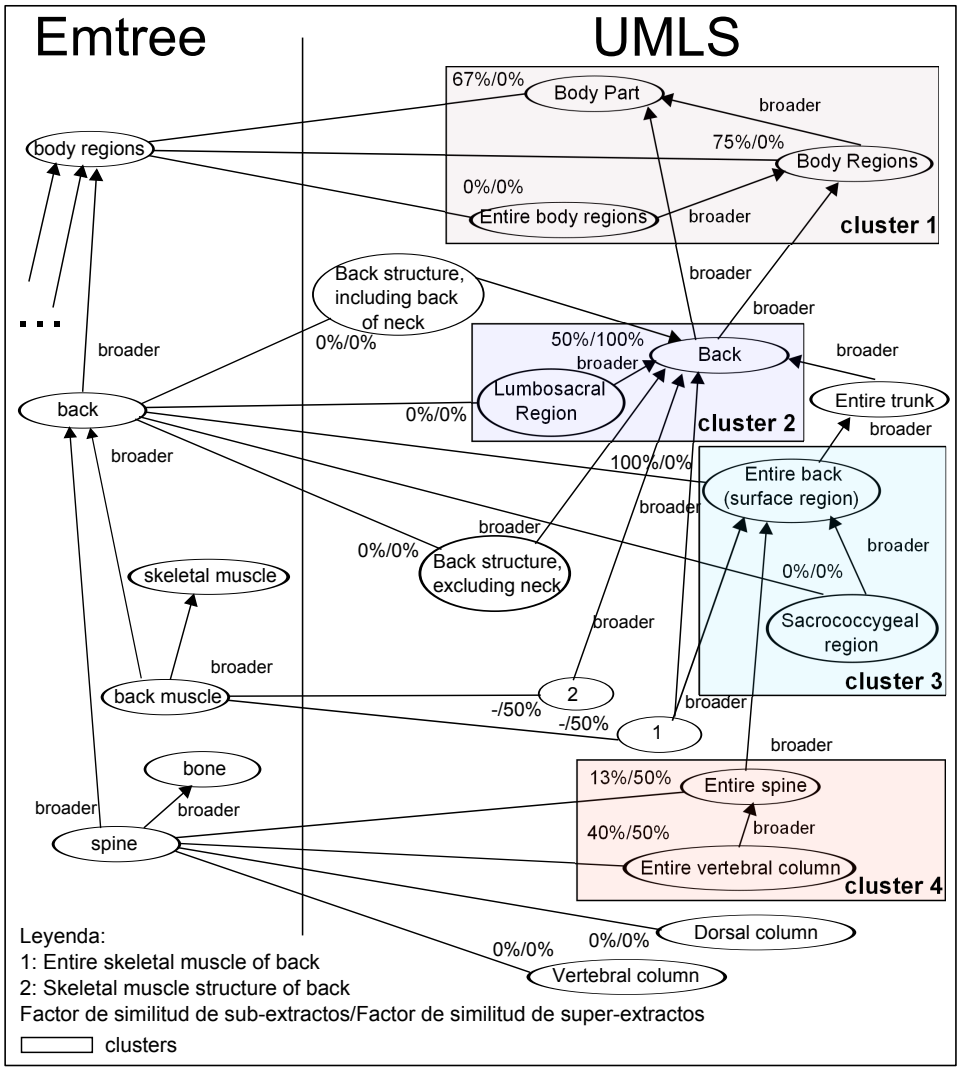

Figura 5.5: Clusters y factores de similitud para el concepto Emtree *back*

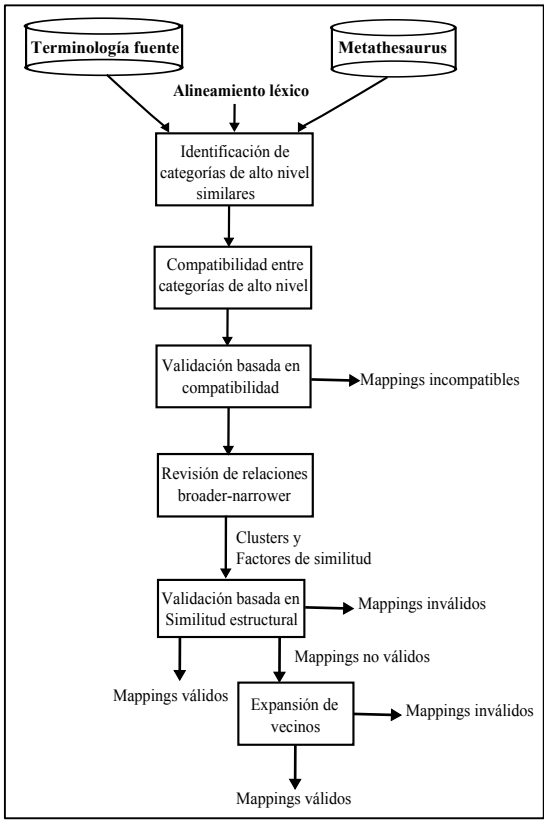

Figura 5.6: Esquema del proceso de Validación y Desambigüación

5.3. Procedimiento general de validación y desambiguación

Nuestro método aplica los principios explicados en el apartado anterior usando la información semántica presente en las terminologías para poder validar el alineamiento léxico obtenido en el capítulo anterior. Las etapas de este proceso se muestran esquematizadas en la figura 5.6.

Para aplicar el principio de compatibilidad, tenemos en cuenta que los conceptos en ambas terminologías están clasificados en categorías, que han sido fijadas por los diseñadores de las terminologías a fin de facilitar su tratamiento. Cada categoría engloba todos los conceptos de una parte determinada del dominio. Categorías habituales en el ámbito biomédico pueden ser Anatomía o Conceptos anatómicos, Enfermedades, Medicinas, Síntomas, etc. La definición de unas u otras categorías varía en función de los objetivos para las que fue definida la terminología, pero, al tratarse del mismo dominio, hacemos la suposición inicial de que las categorías en ambas terminologías

pueden ser similares, probablemente no todas, pero sí las principales.

Por tanto, inicialmente se realiza la identificación de las categorías de alto nivel en ambas terminologías que son compatibles (es decir, que agrupan conceptos similares). Para ello, se analizan cuantitativamente los alineamientos obtenidos estableciendo, para cada categoría fuente, el porcentaje de alineamientos a cada una de las categorías destino. De esta forma, cuando ese porcentaje supera un umbral, se considera que las categorías de ambas terminologías son compatibles.

A continuación, se consideran compatibles los alineamientos donde los conceptos las terminologías fuente y destino pertenecen a categorías similares. El resto de alineamientos se consideran incompatibles y, por tanto, inválidos.

Tras esta primera validación de los alineamientos basada en la compatibilidad de las terminologías fuente y destino, muchos conceptos de la terminología fuente aún estarán equiparados a varios conceptos destino; por tanto, será necesario realizar la identificación de las ambigüedades y su ponderación para elegir el mejor de los alineamientos posibles. Para ello, se aplica el principio de similitud estructural, usando la información contenida en las relaciones broader-narrower entre los conceptos de ambas terminologías. Se analizan los alineamientos, obteniendo los sub-extractos y super-extractos y calculando los factores de similitud y marcando los clusters (agrupaciones de conceptos UMLS relacionados directamente). Por último, se realiza la validación basada en similitud estructural eligiendo el mejor de los alineamientos para cada concepto de la terminología fuente. Como resultado, normalmente habremos seleccionado un sólo alineamiento como el mejor, pero aún pueden darse casos en que aparezcan más de uno.

Por último, viendo la mayor granularidad de UMLS, que hace que 2 conceptos directamente relacionados en la terminología fuente queden equiparados a conceptos relacionados indirectamente (a través de otros) en UMLS, realizamos una revisión de los alineamientos descartados en la etapa anterior para recuperar aquellos donde se encuentren estas relaciones indirectas. De esta forma, podemos afirmar que, para los alineamientos considerados válidos por este proceso, el concepto de la terminología fuente y el concepto UMLS serán similares semánticamente ya que se ha comprobado que están en un punto similar de la jerarquía.

5.4. Compatibilidad entre las fuentes y el alineamiento

5.4.1. Identificación de similitud entre categorías de alto nivel

El principio de compatibilidad establece que en cada alineamiento los conceptos han de ser homólogos, es decir, han de tener un significado similar, lo cual puede comprobarse en primer lugar por la categoría de alto nivel a la que pertenecen ambos. Así, por ejemplo, si un concepto Emtree es una parte anatómica y el concepto UMLS en el alineamiento es una enfermedad, el alineamiento entre ambos conceptos no se considerará válido.

En esta etapa del proceso, en primer lugar, identificamos la similitud entre las categorías de alto nivel de ambas terminologías lo cual permitirá determinar automáticamente si un alineamiento es incompatible. Para obtener la similitud entre categorías de alto nivel, partimos del alineamiento léxico de conceptos y calculamos, para cada categoría fuente, el porcentaje de alineamientos obtenidos con cada categoría destino. Se establece un umbral o porcentaje mínimo de coincidencia para considerar similares las categorías fuente y destino.

Así, por ejemplo, se confirma esta similitud entre la facet *Anatomical concepts* de Emtree y el grupo semántico *Anatomy* de UMLS con un porcentaje de alineamientos léxicos del 83 % y, también, entre la facet *Organism names* y *Living Beings* con un porcentaje del 96 %. En otros casos, sin embargo, no se ha encontrado correspondencia clara entre algunas de las facets de Emtree y algún GS UMLS, ya que para ningún par de categorías se obtiene un porcentaje superior al umbral. Esto sucede con categorías que no han sido diseñadas con los mismos criterios en Emtree y en UMLS por diferentes objetivos o diferente alcance de la terminología. Por ejemplo, Emtree es una terminología usada para indexar una base de datos on-line biomédica y tiene una facet *Types of article or study* para organizar conceptos sobre los tipos de artículos y de bibliografía. En UMLS, este tipo de conceptos están repartidos entre diferentes grupos semánticos. En casos como éste, no será posible localizar el grupo semántico que es más similar a una facet dada, y los alineamientos de esta facet no podrán ser validados. Aún así, el porcentaje de conceptos Emtree con alineamientos validables sigue siendo alto ya que, para las facets con mayor número de conceptos, sí se encuentra el grupo semántico compatible.

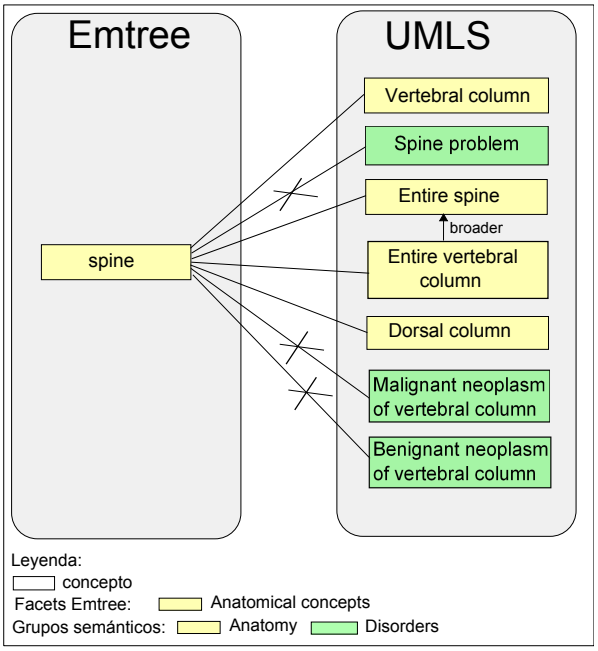

Figura 5.7: Ejemplo de agrupación de conceptos: *spine*

5.4.2. Validación basada en compatibilidad

Una vez determinada la similitud entre categorías de alto nivel, se validan automáticamente los alineamientos léxicos descartando todos aquellos cuyos conceptos no pertenezcan a categorías de alto nivel compatibles.

En la figura 5.7, se observa el caso *spine*, perteneciente a la facet *Anatomical concepts*. El grupo semántico compatible es *Anatomy* por lo que se descartan los tres alineamientos léxicos a los conceptos del grupo semántico *Disorders* (se muestran en la figura marcados).

5.5. Desambiguación usando información estructural

Como resultado de la etapa anterior, cada concepto fuente queda equiparado a uno o varios conceptos destino del grupo semántico compatible. Por tanto, puede afirmarse que los conceptos de ambas terminologías pueden ser bastante homólogos al ocupar posiciones similares dentro de su estructura y será necesario desambiguar entre varios alineamientos correspondientes al mismo concepto fuente (o destino).

Para ello, aplicamos el principio de similitud estructural, por el cual si un

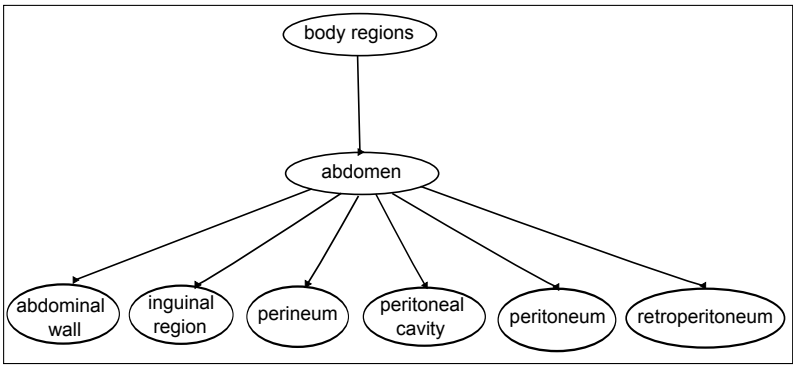

Figura 5.8: Ejemplo de relaciones "broader-narrower"en EMTREE

concepto fuente se equipara correctamente a un concepto destino, los conceptos próximos al concepto fuente estarán equiparados a conceptos próximos al concepto destino. Para ello, usamos las relaciones broader y narrower presentes en ambas terminologías ya que establecen qué conceptos son más generales/específicos que otros. Así, por ejemplo, en la figura 5.8, se muestran las relaciones para *abdomen*, que tiene como broader *body regions* y como narrowers *abdominal wall*, *inguinal region*, *perineum*, *peritoneal cavity*, *peritoneum* y *retroperitoneum*.

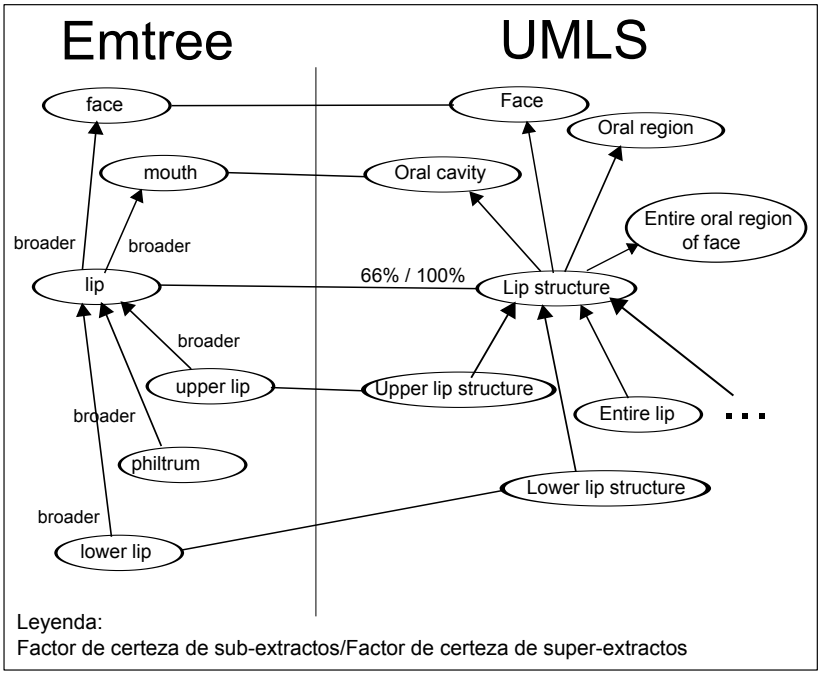

Figura 5.9: Factores de similitud para *lip*

5.5.1. Identificación de ambigüedades

En las etapas siguientes, la información estructural se usa para elegir el mejor alineamiento de todos los obtenidos para cada concepto. En primer lugar, se calculan los factores de similitud de cada alineamiento, en función de la localización de los conceptos en las jerarquías de sus terminologías. En segundo lugar, se exploran las relaciones existentes entre los conceptos destino de todos los alineamientos que comparten el mismo concepto fuente para la detección de alineamientos redundantes.

Cálculo de factores de similitud estructural

El primer paso para determinar la similitud semántica de un concepto fuente y un concepto destino es tener en cuenta sus extractos y buscar la coincidencia entre ellos. Esto nos puede confirmar si realmente estamos hablando del mismo concepto al analizar si la estructura de la vecindad que los rodea es similar. Para ello, se comparan las estructuras que surgen al tener en cuenta los extractos de los conceptos fuente y destino de los diferentes alineamientos léxicos resultantes. En la figura 5.9, se muestran los extractos Emtree y UMLS, junto a las equiparaciones léxicas para el concepto *lip*.

El proceso se lleva a cabo de la siguiente manera. Para cada alineamiento

léxico, se calculan los factores de similitud definidos en el apartado 5.2.1. Para ello, se crean los extractos, tanto super-extractos y sub-extractos, para cada concepto fuente y destino y se computa el porcentaje de similitud entre ellos. Además de los casos vistos en la figura 5.1, analizaremos aquí el caso del alineamiento entre el concepto Emtree *lip* y el concepto UMLS *Lip structure*, mostrado en la figura 5.9. El super-extracto del concepto Emtree *lip* contiene los conceptos *mouth* y *face*, sus broaders, y el sub-extracto contiene sus narrowers *upper lip*, *philtrum* y *lower lip*. En el super-extracto del concepto UMLS *Lip structure* están los conceptos *Face*, *Oral region*, *Oral cavity* y *Entire oral region of face*. En su sub-extracto, están los conceptos *Malignant neoplasm of mouth*, *Mouth Neoplasms*, *Operation on lip*, *Skin of lip*, *Subcutaneous tissue structure of lip*, *Entire lip*, *Upper lip structure*, *Lower lip structure* y *Orbicularis oris muscle structure*. Todos los broaders de *lip* (*face* y *mouth*) se equiparan a algún broader de *Lip*, por tanto el factor de similitud del super-extracto de *lip* es del 100 %. También encontramos equiparaciones léxicas entre algunos conceptos más específicos de *lip* (*upper lip* y *Upper lip structure* y entre *lower lip* y *Lower lip structure*). El concepto *philtrum* no tiene equiparación léxica, por lo que el factor de similitud del sub-extracto es del 66.7 %.

Agrupación de conceptos destino equiparados a un concepto fuente

El siguiente paso consiste en la agrupación de todos los conceptos destino que comparten el concepto fuente a través de los alineamientos léxicos y entre los cuales hay relaciones directas, creando lo que llamamos clusters.

El clúster muestra una relación muy estrecha entre sus conceptos indicando que son muy similares, y por tanto, pueden dar lugar a alineamientos redundantes. Un ejemplo claro son los 3 conceptos obtenidos para *body regions*, donde *Body Part* es broader de *Body Regions* y éste es broader de *Entire body regions*. Este caso se muestra en la figura 5.1. Los tres son muy similares semánticamente, por lo que los tres alineamientos resultantes son redundantes.

En ocasiones, en cambio, en términos que pueden parecer tan similares como éstos, no aparecen esas relaciones. La razón puede tener que ver con el hecho de que UMLS es la unión de muchos vocabularios de los cuales también se importan las relaciones. Los términos que forman los clusters normalmente estaban relacionados en el vocabulario original. Un ejemplo es *cardiovascular system* se equipara léxicamente a *Cardiovascular system*, *Entire cardiovascular system* y *Vascular System*, que no forman clúster, como se ve en la figura 5.11.

Dos conceptos destino serán tanto más similares cuanto más similares

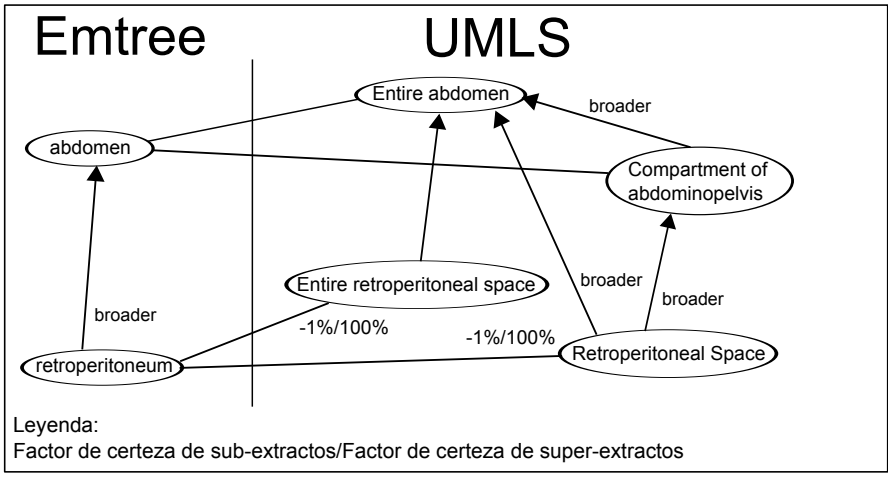

Figura 5.10: Índice Comparativo para retroperitoneum

sean sus extractos. Para aquellos alineamientos correspondientes al mismo concepto fuente, con factor de similitud mayor que 0, se establece el índice comparativo como el porcentaje de conceptos comunes en sus extractos. Cuanto más diferentes sean los conceptos destino, menor número de elementos tendrán en común, y por tanto, el índice será más bajo. En este caso, puede ser interesante considerar ambos alineamientos válidos, ya que aportan más información. Se calculan dos índices, uno para los super-extractos y otro para los sub-extractos, en caso de que los términos dispongan de ellos. Para calcularlos, se comparan los broaders/narrowers de los conceptos destino y el índice será el número de elementos comunes en ambos entre el número de elementos menor. En la figura 5.10, puede verse el caso del concepto Emtree *retroperitoneum*, que se equipara léxicamente a los conceptos UMLS *Retroperitoneal Space* y *Entire retroperitoneal space*, el primer concepto tiene como broader *Entire abdomen* y *Compartment of abdominopelvis*, y el segundo concepto solamente el primero de ellos. Por tanto, el número de broaders comunes es 1 y el número mínimo de broaders es 1, por tanto el índice es 100 %. El concepto *retroperitoneum* no tiene narrowers por lo que no se calcula índice comparativo de narrowers. Este índice se usará en el proceso de desambiguación, que veremos en el apartado siguiente.

En el proceso de creación de clusters, pueden darse varios casos, que se detallan a continuación con ejemplos. En ellos, incluimos también los factores de similitud obtenidos en el paso anterior.

Todos los conceptos obtenidos están relacionados entre sí

En la figura 5.1, se muestran el ejemplo *body regions* que se equipara léxicamente a 3 conceptos UMLS, a su vez, están relacionados directamente

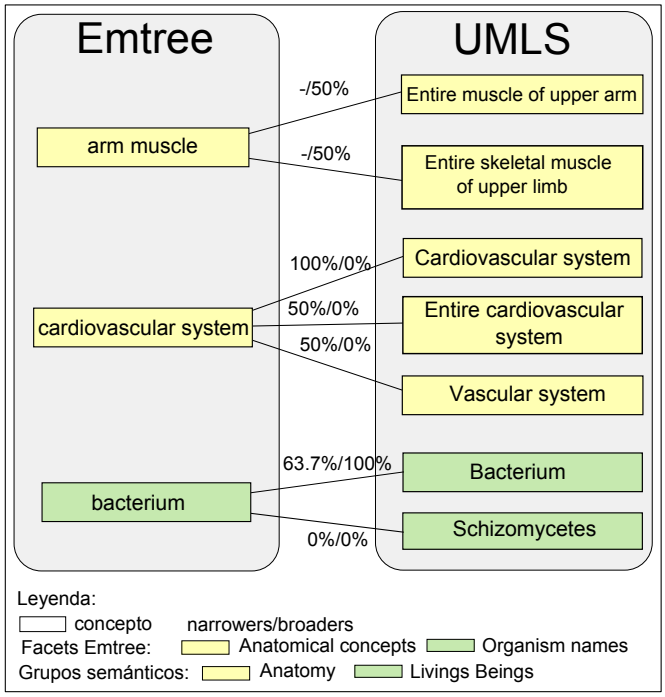

Figura 5.11: Ejemplo de conceptos sin clúster

entre sí, ya que *Body Part* es broader de *Body Regions* que, a su vez, es broader de *Entire body region*. Por tanto, forman un clúster.

Alguno de los conceptos no está relacionado

En la figura 5.1 se muestran 2 casos, *spine* y *back*. Para *spine*, de los 4 conceptos que quedaron de la fase anterior (figura 5.7), *Entire spine* es broader de *Entire vertebral column*, los otros dos conceptos, *Vertebral column* y *Dorsal column*, no tienen relaciones entre ellos ni con los otros.

El concepto *back* tiene 2 clusters de 2 conceptos cada uno. El primero de ellos está formado por los conceptos *Back* y *Lumbosacral Region*, que son broader y narrower respectivamente y el segundo por *Entire back (surface region)* y *Sacrococcygeal Region* que tienen la misma relación. A su vez, los conceptos de ambos clusters no están relacionados directamente entre sí por lo que se definen 2 clusters y no 1. Además, quedan 2 conceptos que no forman parte de ninguno de ellos, *Back structure, excluding neck* y *Back structure, including back of neck*. Esta distribución se usará en el paso siguiente en la elección del mejor alineamiento.

No hay relaciones estructurales entre conceptos

La figura 5.11 muestra los casos *arm muscle*, *cardiovascular system* y *bacteria*, cuyos conceptos UMLS no tienen relaciones broader/narrower entre

sí. En concreto, *arm muscle* tampoco tiene factor de similitud de sub-extracto ya que no tiene narrowers.

5.5.2. Validación basada en similitud

En la etapa anterior, aplicamos la compatibilidad en las categorías de alto nivel para descartar aquellos alineamientos que no cumplían el principio de compatibilidad. Por ello, en esta fase del proceso, cada concepto fuente se encuentra equiparado a uno o más conceptos destino que pertenecen a un grupo semántico compatible, lo cual ya indica que hay cierta similitud entre ellos. A continuación, se necesita desambiguar entre esos alineamientos para elegir el más similar, para lo cual se usa la información obtenida en los pasos anteriores: factores de similitud y clusters.

Para ello, se aplica el siguiente algoritmo. Para cada conjunto de alineamientos al mismo concepto fuente, se analizan sus factores de similitud y la existencia o no de clusters y se selecciona el alineamiento que cumple las siguientes condiciones:

- Para cada clúster, se selecciona el alineamiento correspondiente al concepto destino con el índice de similitud máximo del clúster.

- Para cada alineamiento, si el concepto destino forma parte de un clúster, si se ha encontrado alguna similitud (es decir, factor de similitud mayor que 0) y es la similitud máxima, el alineamiento se considera válido. En caso contrario, se considera no válido. De los demás miembros del clúster, con Índice de Similitud inferior al máximo, se consideran válidos aquellos en los que alguno de los Índices Comparativos -por Broaders y por Narrowers- con el concepto o conceptos con mayor Índice de Similitud sea menor que 75 %. El Índice Comparativo menor a 75 % nos indica que ese concepto y el ya elegido como válido realmente no son tan similares y, por tanto, ambos aportan información relevante.

- Para cada alineamiento donde el concepto destino no forme parte de ningún clúster, se consideran válidos todos los que tienen factor de similitud mayor que 0, es decir, aquellos en los cuales se ha encontrado alguna similitud con la estructura de la terminología destino. Un factor de similitud 0 indica que no se ha encontrado similitud estructural entre el concepto Emtree y el UMLS y, por tanto, el alineamiento ha de ser descartado.

En los ejemplos del apartado anterior, figura 5.5, los resultados quedan como se muestra en la figura 5.12:

- *body regions*: el concepto con más puntuación es *Body Regions* con un 37.5 %. La correspondencia narrower con *Body Part* es del 100 %, ya que los narrowers de ambos coinciden, por lo que este concepto no se incluye. *Entire body region* tiene un indice de 0 por lo que queda descartado.

- *back*: se elige el concepto de más puntuación *Back* y también *Lumbosacral Region* ya que el Índice Comparativo es del 50 %.

- *spine*: se elige el concepto de más puntuación *Entire vertebral column* y a *Entire spine* ya que el Índice Comparativo es del 0 %. Del otro clúster, no se considera válido ninguno al tener factor de similitud 0.

- *arm muscle* no tiene clúster por lo se eligen los dos conceptos al tener un Índice de Similitud mayor que 0 en ambos conceptos.

- *cardiovascular system* no tiene clúster por lo que se consideran válidos los tres conceptos al tener un Índice de Similitud mayor que 0.

- *bacterium* selecciona *Bacterium* al tener índice mayor que 0.

De esta forma, se descartan todos los alineamientos de los conceptos destino que no tienen ninguna similitud estructural directa con el concepto Emtree y también aquellos que, formando parte del clúster, no tienen la mayor puntuación. Aún así, habrá casos donde un concepto fuente aún quede equiparado a más de un concepto destino, cuando los índices de similitud no sean excluyentes.

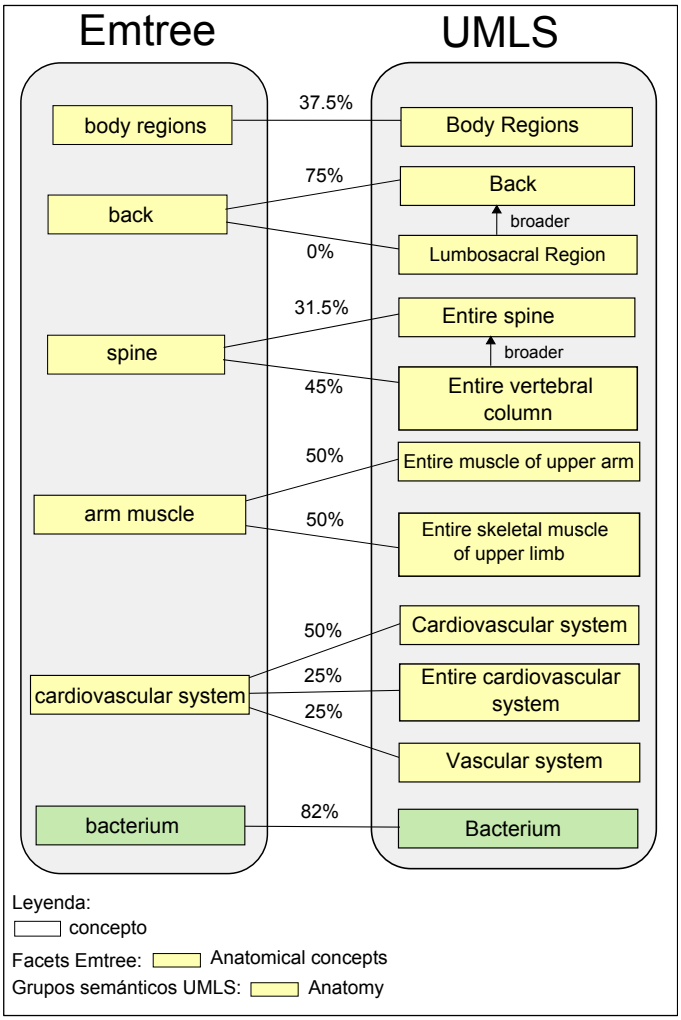

Figura 5.12: Resultados de similitud con Índice de Similitud

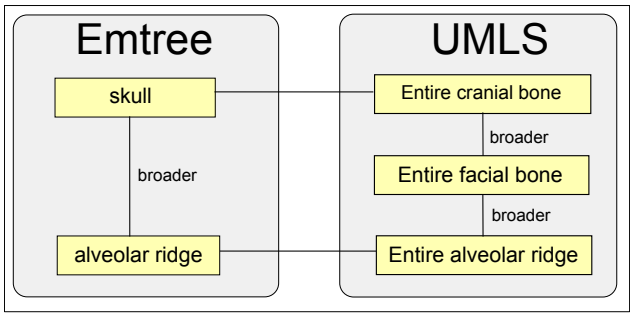

Figura 5.13: Similitud en nivel 2: *Abdomen - Abdominal cavity*

5.5.3. Expansión de vecinos

Las terminologías presentan diferentes granularidades, es decir, diferentes niveles de detalle, según las necesidades para las que fueron creadas. Esto hace que algunas relaciones inmediatas, que se valoraban con los factores de sub-extracto y super-extracto de primer nivel en el apartado anterior, no coincidan pero sí se puedan encontrar coincidencias explorando las relaciones broader y narrower del concepto destino. Esto puede apreciarse gráficamente en el ejemplo de la figura 5.13. Uno de los conceptos que se equiparan léxicamente con *alveolar ridge* es *Entire alveolar ridge*. Este término fue considerado no válido por tener factor de similitud 0. En cambio, uno de sus broaders es *Entire facial bone* que, a su vez tiene como broader *Entire cranial bone* que es equiparado léxicamente a *skull*, broader del concepto *alveolar ridge*. Por tanto, hay una relación indirecta en UMLS entre *Entire alveolar ridge* y *Entire cranial bone* que se corresponde con una relación en Emtree, debido al mayor nivel de detalle de UMLS.

Para aquellos alineamientos descartados en el paso anterior, se analizan automáticamente los broaders del concepto destino hasta un número de niveles en la estructura fijado comparándolos con los conceptos destino que se equiparan directamente a los broaders de los conceptos Emtree. Si se encuentra alguna coincidencia, el alineamiento pasa a considerarse válido ya que se habrá comprobado una relación semántica entre ellos.

Para el caso concreto de alineamiento entre Emtree y UMLS, se aplicó a este proceso una restricción operativa. Debido a la gran cantidad de relaciones existentes en UMLS, al explorar los broaders o narrowers, aparecen ciclos. Por ello, se estableció la restricción de que sólo se puede pasar una vez por cada nodo de la jerarquía. De esta manera, se recuperan un conjunto de equiparaciones entre conceptos Emtree y UMLS para las que la similitud semántica es tan buena como los resultados del algoritmo del apartado anterior ya que, aunque las relaciones sean indirectas, son debidas a la granularidad

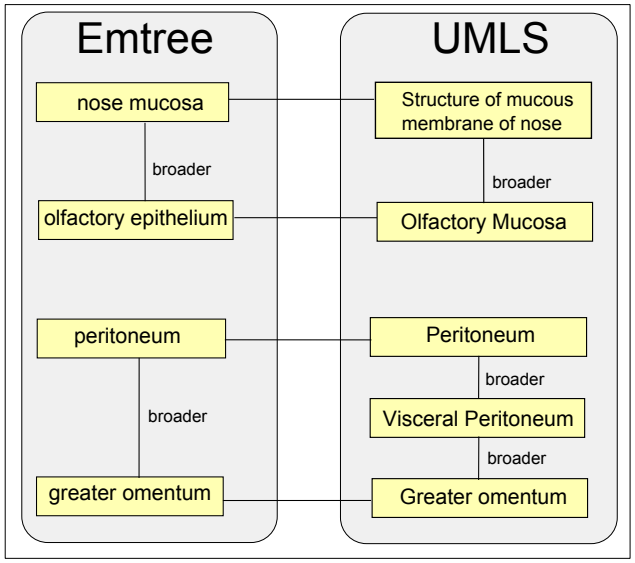

Figura 5.14: Alineamientos recuperados por expansión de broaders

de UMLS. Esto permite completar la información válida para cada concepto Emtree.

Al realizar la expansión por broaders, se recuperan por ejemplo los conceptos mostrados en la figura 5.14. El concepto Emtree *olfactory epithelium* se equipara con el concepto UMLS *Olfactory Mucosa*, aunque este alineamiento fue descartado por el algoritmo al tener menor factor de similitud que el otro elemento del clúster que formaban, *Olfactory Epithelium*. *olfactory epithelium* tiene como broader *nose mucosa* que se equipara con *Structure of mucous membrane of nose* que, a su vez, es broader de *Olfactory Mucosa*. Por tanto, se trata de una relación del nivel 1 y se recupera. Otro ejemplo es el del concepto Emtree *greater omentum* que se equipara con *Greater omentum*, este alineamiento fue descartado inicialmente por tener factor de similitud 0, ya que no tenía narrowers y de los broaders no coincidía directamente ninguno. Tiene como broader *peritoneum* que se equipara con el concepto UMLS *Peritoneum*. Este concepto no es broader directo de *Greater omentum*, sino que es broader de *Visceral Peritoneum* que, a su vez, es broader de él, por tanto es un relación indirecta de nivel 2 y se recupera.

En la figura 5.15, se muestran algunos de los conceptos recuperados por narrowers, como *nose* que se equipara con *External nose structure*, aunque fue descartado por ser el de menor factor de similitud del clúster del que formaba parte con *Nose*. La expansión de vecinos permite detectar que su narrower *nose apex* se equipara a *Tip of nose* que es narrower de *External nose structure*. Por tanto, es un ejemplo de nivel 1. El alineamiento entre

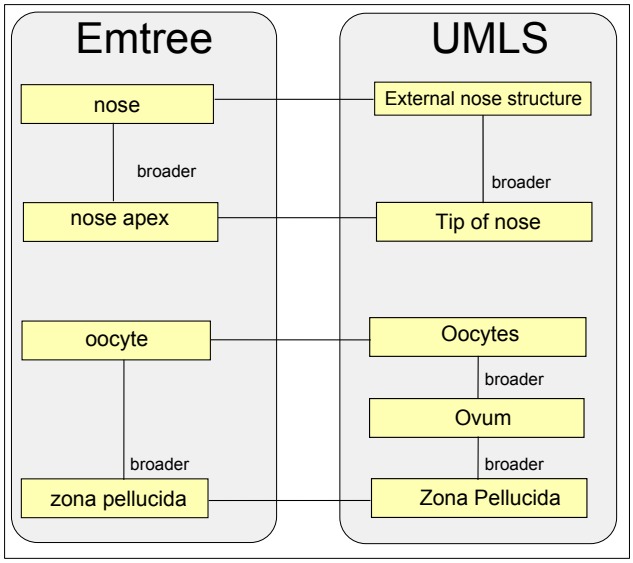

Figura 5.15: Alineamientos recuperados por expansión de narrowers

el concepto Emtree *oocyte* y el concepto UMLS *Oocytes* fue descartado por la misma razón que el anterior, tener menor factor de similitud que el otro concepto en el clúster, *Ovum*. En este caso, *oocyte* tiene como narrower *zona pellucida* que se equipara a *Zona Pellucida* es narrower de *Ovum* que a su vez es narrower de *Oocytes*. Por tanto, es una relación de nivel 2 y se recupera.

De esta forma, exploramos las relaciones de los conceptos UMLS para encontrar más similitudes con el concepto Emtree. Así, podemos validar aquellos alineamientos entre un concepto Emtree y un concepto UMLS que no tienen relación directa, pero sí indirecta. Esta relación será igual de importante que la directa debido al mayor nivel de detalle de UMLS.

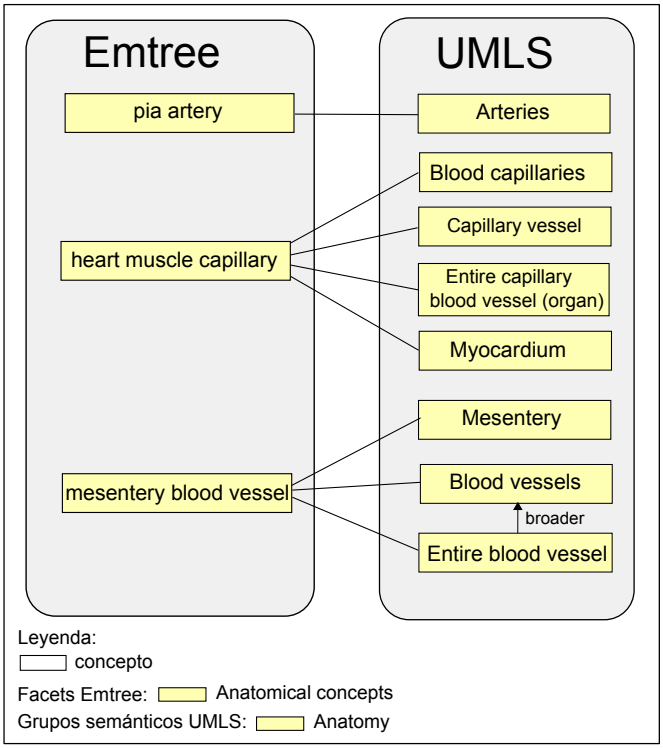

Figura 5.16: Ejemplos tras validar el GS

5.6. Validación del análisis basado en técnicas de procesado de lenguaje natural

En el apartado 3.8, para aquellos conceptos fuente sin alineamiento léxico, obtuvimos alineamientos léxicos a sus partes constituyentes. En la figuras 4.18, 4.19 y 4.20, se mostraron varios ejemplos de ellos. Como puede observarse, los conceptos UMLS obtenidos son de grupos semánticos variados. Por tanto, la validación de estos alineamientos consistirá en descartar aquellos de grupos semánticos no compatibles.

De esta forma, algunos de los ejemplos explicados en el apartado 3.8 quedan como se indica en la figura 5.16. El concepto Emtree *pia artery* queda equiparado al concepto UMLS *Arteries*, que se obtuvo tras buscar *artery*. De la parte *pia*, se descartan todos los conceptos UMLS a los que se equiparaba al no ser del grupo semántico compatible. El concepto Emtree *heart muscle capillary* queda equiparado a 4 conceptos UMLS, *Blood capillaries*, *Capillary vessel* y *Entire capillary blood vessel (organ)* de la parte *capillary* y *Myocardium* de la subfrase *heart muscle*. Sólo se descarta un concepto perteneciente al grupo semántico *Disorders*. El concepto Emtree *mesentery blood vessel*

conserva todos los alineamientos obtenidos al pertenecer al grupo semántico *Anatomy*.

5.7. Alineamiento complejo

Como resultado de este proceso, un concepto fuente queda equiparado semánticamente a uno o más conceptos destino para los que nuestro método ha podido validar que su significado conjunto contiene la mayor parte posible del significado del concepto fuente.

En la mayor parte de casos, nuestro método ha podido elegir un único alineamiento como el mejor de todos. En este caso, decimos que es un ExactMatch. Cuando no es posible indicar un único concepto UMLS, se trata de un alineamiento complejo. Sin embargo, esto no quiere decir que el concepto Emtree sea idéntico a todos los conceptos UMLS del alineamiento complejo, sino que hay 2 casos posibles:

- **BroadMatch**: cuando el concepto de la terminología fuente queda equiparado a conceptos de significado más general en la terminología destino.
- **UnionMatch**: cuando el concepto de la terminología fuente queda equiparado a varios conceptos de la terminología destino cada uno de lo cuales tiene parte de su significado, incluso pudiendo solaparse.

Como ya vimos, un alineamiento complejo BroadMatch aparece cuando no existe en la terminología destino un concepto que se equipare con el buscado y se realiza la separación en subfrases, bien basada en broaders o en unidades léxicas, o si no los hay, directa. Ejemplos de estos pueden verse en la figura 5.16. Un alineamiento complejo UnionMatch aparece en 2 casos, el primero de ellos es cuando, aún como resultado del proceso de validación y desambiguación, un concepto puede quedar equiparado a varios conceptos UMLS. En este caso, cada concepto aporta parte del significado del término, como se vio en los ejemplos de la figura 5.12. El otro caso es en los conceptos de tipo enumeración que no obtuvieron equiparación léxica directa con UMLS, al buscar cada una de sus partes, se obtienen varios conceptos que incluyen el significado del concepto original. Pueden verse ejemplos en la figura 5.17.

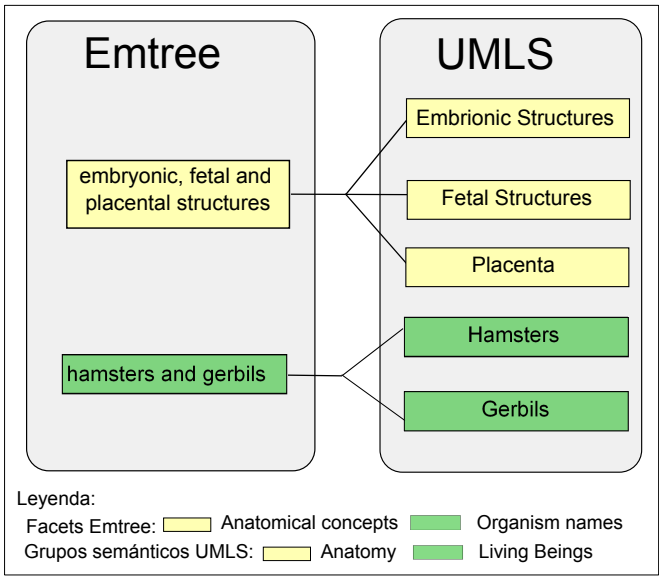

Figura 5.17: Alineamientos complejos UnionMatch

Capítulo 6

Resultados

6.1. Introducción

En este capítulo, realizaremos un análisis de los resultados obtenidos al alinear EMTREE y UMLS, siguiendo el proceso de equiparación y validación de terminologías visto en los capítulos anteriores. Para ello, inicialmente explicaremos el procedimiento de evaluación de resultados propuesto. A continuación, ampliaremos la información sobre las terminologías usadas en el experimento, Emtree y UMLS, detallando las características estructurales que han sido utilizadas en nuestro método, entre ellas, las categorías de alto nivel y las relaciones broader-narrower entre conceptos. En los siguientes apartados, mostraremos los resultados cuantitativos obtenidos en cada fase del método, centrándonos en el número de alineamientos obtenidos y su calidad. Por último, analizaremos los resultados desde el punto de vista cualitativo poniendo el énfasis en las conclusiones que extraemos en cada fase.

6.2. Procedimiento de evaluación

La evaluación de las equiparaciones entre terminologías es la parte más crítica en la investigación de alineamiento de terminologías. La técnica ideal de evaluación automática consiste en comparar los resultados del método frente a un alineamiento de referencia ("gold standard") elaborado por un grupo de expertos en el dominio. Sin embargo, especialmente para los tesauros de gran tamaño, la construcción de tal alineamiento de referencia es tediosa y requiere recursos muy costosos. En su ausencia, las evaluaciones son complejas, difíciles y principalmente manuales. Como alternativa, en otros ámbitos, como la minería de datos y el descubrimiento de conocimiento o

de ontologías, se idearon competiciones (KDDCup[1], OAEI[2]) entre diferentes equipos orientadas a alcanzar un consenso para la evaluación de los diferentes métodos. En particular, para el alineamiento de ontologías, los informes de competiciones pasadas han reflejado algunas dificultades encontradas en el alineamiento, tales como las debidas al gran tamaño de las fuentes o la transformación de éstas desde su formato nativo al alineamiento de referencia.

Una segunda opción consiste en comparar la validación frente a otro método de alineamiento [CSG+03]. Para ello, se alinean dos fuentes dadas con diferentes métodos y se comparan los resultados. Aunque esta técnica no es ideal, la validación cruzada ha proporcionado algunas ideas sobre las fortalezas y debilidades de cada enfoque.

Una tercera opción es usar un alineamiento indirecto a través de una ontología de referencia [ZB06]. En este caso, se realiza el alineamiento de cada ontología con la de referencia y el alineamiento final se deriva de éstos. El análisis de las diferencias entre los alineamientos directos e indirectos ha revelado que la presencia de sinónimos y relaciones en la ontología de referencia es responsable de los alineamientos que sólo identifica el método indirecto. Por su parte, aquellos alineamientos que sólo identificaba el alineamiento directo entre las terminologías se debe a las diferencias entre ellas en la cobertura del dominio. Por tanto, el uso de alineamientos indirectos a través de una ontología de referencia es razonablemente eficiente.

Resumiendo, siempre es recomendable realizar algún tipo de validación aunque ésta sea manual, ya que, a pesar de ser imperfecta y no ideal, puede sacar a la luz las fortalezas y limitaciones de un enfoque particular.

6.2.1. Procedimiento de evaluación propuesto

El procedimiento de evaluación que hemos seguido en esta tesis se ha realizado manualmente, al no disponer de un alineamiento de referencia ni de suficientes recursos para crearlo. Realizamos la evaluación tras cada fase del método, tras la validación por categoría de nivel superior, tras el algoritmo de desambiguación y tras la construcción de los alineamientos complejos. Evaluaremos nuestro método con las medidas de cobertura, precisión y recall, que definimos a continuación.

> Cobertura es el porcentaje de términos (o conceptos) de la terminología fuente para los el método encuentra un alineamiento correcto en la terminología destino.

[1] http://www.kdd.org/
[2] http://oaei.ontologymatching.org/

Precisión es el porcentaje de alineamientos correctos validados por el método con respecto al total de alineamientos validados.

Recall es el porcentaje de alineamientos correctos identificados por el método con respecto al total de alineamientos correctos.

Nuestro procedimiento de revisión para cada alineamiento procesado por nuestro método en cada fase del método consistió en:

1. Comprobamos el significado del concepto fuente (Emtree) involucrado. Como muchos conceptos Emtree no incluyen una definición, sabemos cuál es el significado del concepto revisando su ubicación en la estructura broader-narrower así como el significado de los conceptos más generales y más específicos.

2. Comprobamos el significado del concepto destino (UMLS) revisando la definición (si la hay), los conceptos más generales, más específicos y los tipos semánticos a los que pertenece.

3. En los casos ambiguos, revisamos el diccionario Medline Plus.

4. Comparamos los significados de los conceptos en el alineamiento como sigue.

   - Si el significado del concepto Emtree es igual al significado del concepto Metathesaurus, marcamos el alineamiento validado como correcto.
   - Si los significados son ambiguos o no podemos garantizar que los conceptos sean diferentes, entonces marcamos el alineamiento validado por nuestro método como incorrecto.

Por lo tanto, el método de evaluación sólo considera que nuestro método valida bien alineamientos correctos si hay completa seguridad de que los significados de los conceptos en el alineamiento son iguales.

6.3. Descripción de las terminologías equiparadas

En este apartado, completamos la información ya aportada en capítulos anteriores sobre las terminologías de nuestro estudio, detallando la información necesaria de cara a analizar y comprender los resultados obtenidos.

6.3.1. Emtree

La versión de datos EMTREE utilizada en nuestros experimentos está en formato XML, se está usando en la actualidad, es de gran tamaño y es muy rica en sinónimos. La versión utilizada contiene 46.427 conceptos y más de 190.000 términos (incluidos PT más sinónimos), lo cual da una media de 4,18 sinónimos por concepto. Están distribuidos en 15 categorías principales denominadas facets, que representan taxonomías, y que son:

- Facet A *Anatomical concepts* incluye términos referentes a partes del cuerpo humano, sean órganos, músculos, sistemas, genes, células, entre otros. Incluye términos genéricos como *body regions*, *organ* o *chromosome* o términos más particulares como *thorax*, *finger*, *leg blood vessel* o *peroneus nerve*.

- Facet B *Organism names* contiene formas de vida existentes en la naturaleza y que pueden afectar de algún modo a la salud humana como son bacterias (*bacterium*), invertebrados, vertebrados, plantas y virus. De algunas de ellas, se ofrece gran detalle como *Archaebacterium*, *spirochete* y muchos otros tipos de bacterias o *Astrovirus*, *retrovirus* y muchos tipos de virus.

- Facet C *Physical diseases, disorders and abnormalities* contiene los términos relativos a enfermedades y otros problemas de salud, salvo los de tipo psiquiátrico, por ejemplo, *aorta disease*, *mandible fracture*, *hyperuricemia* o *nose cancer*.

- Facet D *Chemicals and drugs* contiene términos para sustancias químicas, como drogas, sustancias tóxicas y medicamentos. Incluye desde nombres comunes de compuestos como *clenbuterol* o sustancias específicas con su fórmula completa como *3 (5 chloro 2 methoxyphenyl) 3 fluoro 2,3 dihydro 6 trifluoromethyl 1h indol 2 one*.

- Facet E *Analytical, diagnostic and therapeutic techniques, equipment and parameters* incluye términos relativos a técnicas, tratamientos, equipamiento y parámetros médicos como *anesthesia*, *nebulizer*, *glucose intake*, *heart left ventricle pressure* o *low calory diet*.

- Facet F *Psychological and psychiatric phenomena* incluye términos relacionados con el ámbito psiquiátrico y psicológico tanto enfermedades como *perception disorder*, *agnosia* o *depression*, como términos genéricos como *emotionality* o *ego*.

- Facet G *Biological phenomena and functions* contiene términos definiendo funciones fisiológicas normales de los organismos, de sistemas de órganos, de partes del cuerpo y de estructuras moleculares y celulares. Ejemplos son *cardiovascular function*, *retina blood flow*,*thrombin inhibitor*, *thrombocyte agglutination*, *immunotoxin*, *carbohydrate antigen*, *intestine parasite* o *prenatal period*.

- Facet H *Chemical, physical and mathematical phenomena* contiene términos definiendo fenómenos químicos, físicos y matemáticos, con algún solapamiento con otras facets orientadas biológicamente como la E y la G. Ejemplos de esta facet son como *chemical reaction*, *photodegradation*, *combustion*, *stochastic model* o *variance*.

- Facet I *Society and environment* incluye una gran variedad de términos del mundo real, salvo los relacionados con la salud que están en la facet N, como *accident*, *smoke*, *moon*, *night*, *risk*, *safety*, *biomass*, *sea surface waters*, *tooth flora*, *wood*, *jurassic* o *computer security*.

- Facet J *Types of article or study* incluye términos sobre el ámbito académico como *article*, *book*, *comparative study* o *human experiment*.

- Facet K *Geographic names* incluye nombres de países, continentes y otros elementos geográficos, entre ellos *Antarctica*, *Madagascar* o *Middle East*.

- Facet L *Groups by age and sex* son términos relacionados con la edad y el género, como *adolescence*, *puberty*, *child* o *male*.

- Facet M *Named groups of persons* incluye términos de grupos étnicos o referentes a condiciones personales como *Asian*, *religious group*, *consumer*, *immigrant* o *nursing staff*.

- Facet N *Health care concepts* incluye conceptos relacionados con el sistema de gestión de la salud, aunque, como hemos visto, alguno está en la facet anterior. Por ejemplo, *cost effectiveness analysis*, *community hospital*, *laundry*, *ambulatory monitoring*, *rehabilitation* o *health care planning*.

- Facet Q *Biomedical disciplines, science and art* incluye términos generales de medicina como *anthropology*, *biomechanics*, *medical illustration*, *gerontology* o *proctology*.

En la figura 6.1, se representa mediante un gráfico de barras el número de conceptos en cada una de esas facets. Este número es muy variable oscilando

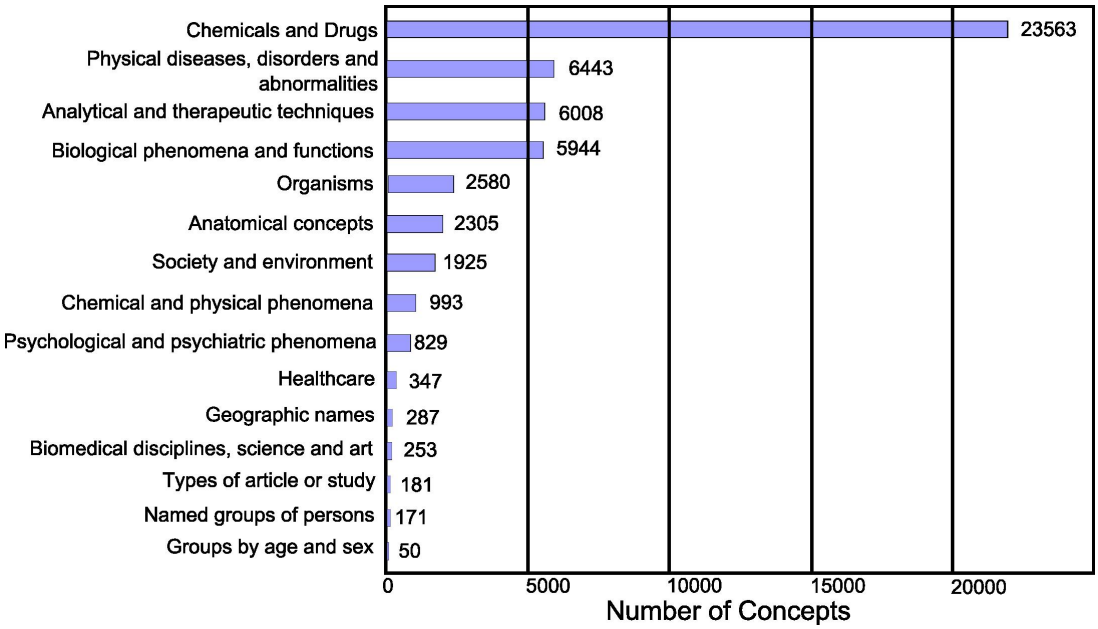

Figura 6.1: Distribución de conceptos en Emtree

entre los 50 para *Groups by age and sex*, que es la más pequeña, a los 23.563 para *Chemicals and Drugs*, que es la mayor. Como algunos de los conceptos aparecen en múltiples facets, el número total de los conceptos de la figura 6.1 supera el número real de los conceptos en Emtree.

6.3.2. UMLS

Los recursos proporcionados por el NLM, junto con el Metathesaurus UMLS, han sido descritos con detalle en la apartado 2.3.3. La versión utilizada del UMLS contiene alrededor de 1,3 millones de conceptos. Aunque el UMLS Metathesaurus integra más de 130 vocabularios, también puede verse como un vocabulario único con los conceptos y las relaciones (narrower, broader y otras) entre estos conceptos. Los conceptos del Metathesaurus se clasifican utilizando un conjunto de categorías semánticas básicas llamadas Tipos Semánticos (TS), como *Anatomical Structure* o *Chemical*. En total, incluye 134 TSs, lo cual para algunos propósitos, resultan excesivos. Para ello, se realizó un estudio y se agruparon, a su vez, en 15 categorías de alto nivel, denominadas Grupos Semánticos (GS), tales como *Anatomy* o *Disorders*. La distribución de conceptos en estos grupos semánticos puede verse en la figura 6.2. Aquí las cantidades se muestran en unidades de millar, lo que da idea de las dimensiones y granularidad que puede aportar el Metathesau-

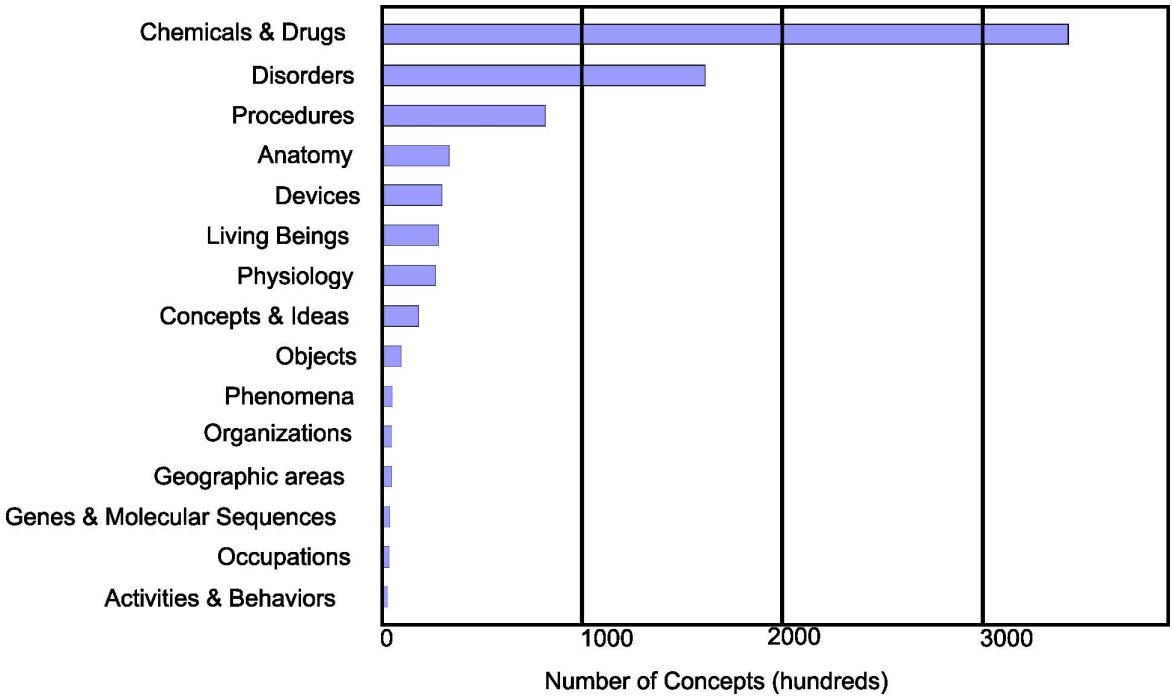

Figura 6.2: Distribución de conceptos en UMLS [MBB01]

rus. El grupo semántico de mayor tamaño es *Chemicals & Drugs* con más de 300.000 conceptos y el de menor tamaño *Activities & Behaviors* con unos 3000.

Los grupos semánticos son:

- *Activities & Behaviors* son actividades del mundo real, comportamientos o eventos. Incluye comportamientos sociales, diarios, gubernamentales, como *Cooperative Behavior* o *Domestic activities*.

- *Anatomy* reúne todos los términos referentes a partes del cuerpo, a órganos, a células, como *Entire body region*, *Lumbosacral region*, *Back* o *Heart muscle cell*.

- *Chemicals & Drugs* incluye términos sobre medicamentos, drogas y sustancias usadas en medicina, como *betahistine* o *naratryptan*.

- *Concepts & Ideas* es uno de los grupos genéricos que incluye conceptos de tipo muy variado: conceptuales, funcionales, espaciales, temporales, sobre lenguaje o clasificaciones. Ejemplos son: *Enviroment*, *Books*, *Infancy*, *Cost Effectiveness* o *Computer Systems*.

- *Devices* contiene dispositivos médicos usados en la práctica médica, para detección o tratamiento de enfermedades, como *Fetal monitor*, *Physical model* o *Cardioscope*.

- *Disorders* contiene enfermedades, trastornos y síntomas, como *Tachycardia* o *Thrombosis of renal artery*.

- *Genes & Molecular Sequences* contiene los términos relativos a genes y secuencias moleculares, como *chromosome 17p*.

- *Geographic Areas* contiene términos de zonas geográficas y países.

- *Living Beings* contiene otras formas de vida presentes en medicina como bacterias, hongos o virus.

- *Objects* es otro de los grupos genéricos con objetos del mundo real como comida, sustancias o elementos manufacturados. Por ejemplo, *Fruit*, *Adhesives* o *Deodorants*.

- *Occupations* contiene ocupaciones relacionadas con el ámbito médico, como *Inmunopathology specialty*, *Anesthesiology*, *Endodontics* o *Pediatric surgery speciality*.

- *Organizations* contiene términos relacionados con organizaciones genéricas o concretas en el ámbito médico, como *Prosthodontics Dentist*, *Eastern Cooperative Oncology Group*, *Surgical service* o *Fisheries*.

- *Phenomena* incluye fenómenos, procesos, efectos, resultados de diferentes tipos: naturales, causados por el hombre, ... Ejemplos de ellos son: *Blood concentration results*, *Relapse*, *Deterioration*, *Exposure to radiation* o *Barotrauma mechanism*.

- *Physiology* contiene términos variados relacionados con procesos y conceptos fisiológicos, como funciones de las células, genéticas o moleculares, atributos clínicos o de organismos y procesos mentales. Ejemplos son *Chromosome Pairing*, *Exocytosis* o *Response to antigens*.

- *Procedures* contiene procedimientos médicos de diágnostico, educacionales, preventivos, de laboratorio, ... como *Procedures on nose*, *Protocol Treatment Arm* o *Lymphocyte subset measurement*.

6.3.3. Comparación entre EMTREE y UMLS

Como se puede observar en las figuras 6.1 y 6.2, en la primera el número de términos está en unidades y en la segunda en unidades de millar, lo cual da idea de la diferencia de tamaño entre las terminologías. Emtree es una terminología concreta mientras que UMLS es una metaterminología, surgida de la unión de múltiples terminologías. Por esta razón, UMLS tendrá una mayor granularidad, es decir, un mayor detalle para cada elemento del mundo real, lo que se traduce en diferentes conceptos relacionados entre sí que, en cambio, en Emtree se representa con un único concepto. En la figura 4.8, puede verse el concepto Emtree *body regions* que se equipara léxicamente a varios conceptos UMLS, similares y relacionados entre sí, *Body part*, *Body regions* y *Entire body region*. Por tanto, surge la necesidad de la validación/desambiguación para seleccionar el concepto más similar.

Otro aspecto a estudiar es el de las categorías de alto nivel que ambos tienen, facets en Emtree y grupos semánticos (GS) en UMLS. Éstas han sido definidas por los diseñadores de cada proyecto, siguiendo criterios particulares como, por ejemplo, la utilidad de la terminología o las necesidades concretas del entorno para el que han sido desarrolladas. Por esta razón, dichas categorías pueden llegar a ser muy diferentes.

En primer lugar, tratándose del mismo dominio, una suposición inicial fue que podíamos encontrar categorías similares entre Emtree y UMLS, al menos, parcialmente. Así, en ambas observamos una misma categoría *Chemical and Drugs* que contiene términos de medicamentos y sustancias usadas en medicina. Ésta es, además, en ambas, la de mayor número de términos. También puede verse a simple vista qué determinadas categorías contienen términos similares como *Physical diseases, disorders and abnormalities* y *Disorders* o *Anatomical concepts* y *Anatomy*, las primeras sobre enfermedades y las segundas sobre anatomía, respectivamente.

En segundo lugar, una terminología puede tener diferente alcance que otra y requerir algunas categorías a mayores. Por ejemplo, Emtree está orientada a la recuperación de información bibliográfica, por lo que contiene la facet *Types of article or study*, categoría que, como tal, no existe en UMLS aunque se pueden encontrar algunos de sus términos en otras.

En tercer lugar, las terminologías suelen contener categorías "cajón desastre" donde se acumulan términos no específicos del dominio, genéricos, abtractos, de temática variada y no relacionados entre sí, como los GS *Concepts & Ideas* y *Objects* en UMLS. En el GS, encontramos conceptos funcionales, temporales, espaciales, cuantitativos y cualitativos. En el segundo, entidades del mundo real, objetos hechos por el hombre, comida o sustancias.

Por último, a simple vista, es difícil ubicar ciertas categorías de una ter-

minología en otra; por ejemplo, en los grupos *Occupations* y *Phisiology* de UMLS. El primero puede equipararse con *Biomedical disciplines, science and art*, aunque su objetivo parece diferente. En cuanto al segundo, Emtree coloca conceptos de este GS en la mayoría de las otras facets.

Siguiendo las apreciaciones anteriores, revisamos manualmente las categorías superiores de ambas terminologías y reunimos los resultados en la figura 6.3, donde hay una línea entre las facets y grupos semánticos que, a simple vista, parecen similares. Como puede observarse, algunos de ellos no han sido asignados al no haber una correspondencia clara. En Emtree, las facets *Society and enviroment* y *Groups by age and sex* no se pueden equiparar directamente a ningún grupo semántico de UMLS, aunque es probable que términos similares se encuentren en *Concepts & Ideas*. Más adelante, se analizarán automáticamente los alineamientos obtenidos para verificar o descartar esta similitud a simple vista.

Como ya comentamos en el capítulo 4, un último aspecto a resaltar es la existencia de la información estructural. En ambas terminologías, disponemos de relaciones tipo broader-narrower entre los conceptos que permiten distribuir los conceptos en jerarquías.

6.4. Alineamiento léxico

En esta sección, mostraremos los resultados obtenidos de la equiparación léxica entre las terminología Emtree y UMLS, haciendo uso del servicio web NormalizeString del UMLSKS. El resultado es el conjunto de alineamientos entre cada concepto Emtree y los conceptos UMLS que son léxicamente similares a él. Para cada concepto Emtree, puede haber 0 (sin alineamiento), uno (alineamiento simple) o varios conceptos UMLS (alineamiento ambiguo) que se equiparan con él. A continuación, mostramos los resultados obtenidos.

6.4.1. Alineamiento léxico de términos

Cada concepto Emtree viene definido por un término preferido y un conjunto de sinónimos, con una media de 4.18 sinónimos por concepto. Nuestro método equipara tanto el término preferido como todos los sinónimos con conceptos del UMLS Metathesaurus, obteniendo en muchos casos los mismos conceptos para sinónimos diferentes. En las dos primeras columnas de la tabla 6.1, se muestra el número total de alineamientos léxicos obtenidos para cada término Emtree, así como el porcentaje que suponen con respecto al número total de términos. Los resultados se detallan desglosados en función de si existe o no alineamiento léxico y, en caso de existir, si este es simple

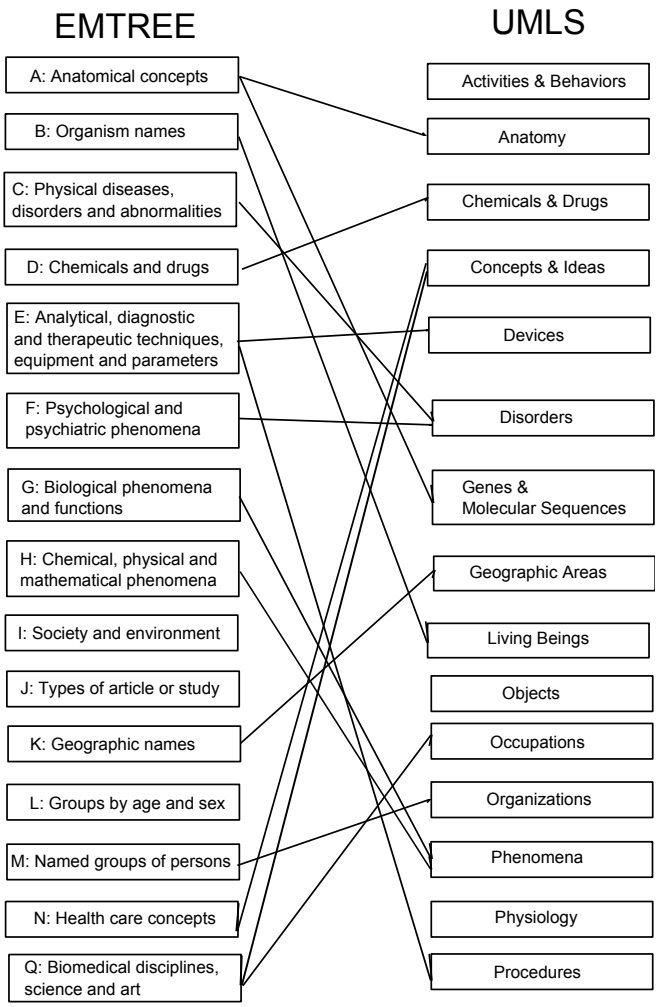

Figura 6.3: Comparativa "a simple vista" entre facets Emtree y Grupos Semánticos UMLS

o ambiguo. Las otras dos columnas de la tabla, que contienen los resultados por concepto, se explican en el apartado siguiente.

De todos los términos de Emtree (en total 264.215), el método no obtiene equiparación para 149.145 términos Emtree, lo que representa el 52,1 % de todos los términos Emtree a equiparar. Por otro lado, hay alineamientos simples para 102.097 términos Emtree (42.5 %), y alineamientos ambiguos para 12.972 términos Emtree (5.4 %). En total, 115.070 términos Emtree coinciden con uno o más conceptos en el Metathesaurus UMLS, lo que supone un 47,9 % de equiparaciones léxicas, de media 1.4 por término Emtree.

En las tablas 6.2 y 6.3, se muestra la misma información desglosada por facet. Ahora analizamos los resultados para las equiparaciones de los términos

	Términos Emtree		**Conceptos Emtree**	
	No	%	**No**	%
Sin alineamiento	149.145	52.1 %	10.345	20 %
Alineamiento simple	102.097	42.5 %	22.644	43.6 %
Alineamiento ambiguo	12.972	5.4 %	18.890	36.4 %
Cobertura total	115.070	47.9 %	41.534	80 %
Alineamientos por término/concepto	1.4	—	3.2	—

Tabla 6.1: Cobertura de términos y conceptos Emtree tras el alineamiento léxico

Emtree y queda para el apartado siguiente los resultados para conceptos.

Sólo en una de las facets, *Geographic areas*, el 85 % de los términos Emtree tiene equiparaciones léxicas a conceptos UMLS. Esto es comprensible dada la temática de la facet, áreas geográficas, ya que países, continentes y otras zonas del mundo se nombran siempre de forma similar y la cantidad de ellos es limitada. Además, en la mayor parte de las facets, 10 de 15, los términos con equiparaciones léxicas superan el 50 %, lo cual supone un índice elevado de resultados. De ellas, en *Organism names* y *Health care concepts*, se obtiene un porcentaje de equiparación del 70 %. Entre 60 % y 70 %, están *Psychological and psychiatric phenomena*, *Groups by age and sex*, *Named groups of persons* y *Biomedical disciplinas, science and art*. Por tanto, el nivel de cobertura de Emtree con UMLS es alto en temas como organismos, conceptos de salud, fenómenos del ámbito psiquiátrico y nombres de agrupaciones de personas. Por último, en *Anatomical concepts*, *Physical diseases, disorders and abnormalities*, *Society and enviroment* y *Types of article or study* entre 50 % y 60 % de los términos obtienen equiparación léxica. En el caso de las 2 primeras, son facets con bastantes términos, que son importantes en el ámbito biomédico ya que abarcan temas fundamentales como anatomía y enfermedades. Es posible que el porcentaje de equiparación no sea más alto debido a las diferencias en el etiquetado de partes del cuerpo y las enfermedades. Para las 2 últimas, *Society and enviroment* y *Types of article or study*, puede deberse a justo lo contrario; son temas más genéricos y no tan relacionados con la medicina por lo que pueden no estar cubiertos de la misma manera en las 2 terminologías.

Las 4 facets restantes tienen menos del 50 %, pero en todo caso más del 30 %, éstas son *Chemicals and Drugs*, *Analytical, diagnostic and therapeutic techniques, equipment and parameters*, *Biological phenomena and functions* y *Chemical, physical and mathematical phenomena*. En las dos primeras,

puede explicarse por el gran tamaño de esas facets y el gran número de sinónimos de los términos. Para las otras 2 facets, puede ser debido al estar formadas por términos generales relativos a fenómenos biológicos, físicos o matemáticos que pueden no estar cubiertos por el dominio biomédico de UMLS.

Por último, comentamos los alineamientos ambiguos obtenidos, es decir, cuando un término Emtree se equipara a más de 1 concepto UMLS. Como ya vimos, UMLS es un Metathesaurus surgido de la unión de muchas terminologías, por lo que puede contener conceptos muy similares de ahí que se obtengan varios. Aún así, el porcentaje no es muy alto en ninguna de las facets, ya que en ningún caso, supera el 20 %, aproximadamente la mitad de las facets, 7 de ellas, supera el 10 % y las restantes es inferior al 10 %. *Anatomical concepts* con un 16.1 % y *Named groups of persons* con 17.2 % son los de mayor número de alineamientos ambiguos, probablemente por tratarse de temáticas donde hay mayor número de conceptos similares. Las de menor número son *Physical diseases, disorders and abnormalities* y *Chemical and Drugs*, con un 4.4 % y un 2.9 %, respectivamente, donde es más difícil encontrar más de un concepto con significado similar. Otra razón para la existencia de la ambigüedad viene dada por la naturaleza de EMTREE, que agrupa diferentes términos muy similares, pero no sinónimos, en los conceptos (recordemos que uno de sus usos principales es indexar EMBASE).

6.4.2. Alineamiento léxico de conceptos

El concepto Emtree se define por los términos que lo componen, tanto el preferido como los sinónimos. Por ello, en esta etapa, agrupamos los conceptos UMLS obtenidos para todos los términos del mismo concepto Emtree, algunos de los cuales pueden coincidir. De esta forma, el concepto Emtree queda equiparado a todos los conceptos UMLS que se han equiparado léxicamente a todos los términos, preferido o sinónimos, del concepto.

En las tablas 6.1, 6.2 y 6.3, se muestran también los totales por concepto y para cada facet. Emtree, al igual que UMLS, tiene una buena colección de sinónimos, 4,18 por cada concepto, lo cual hace el número total de conceptos con alineamiento sea alto. De los 51.879 conceptos en Emtre, 10.345 no se equiparan léxicamente (ni el término preferido ni ninguno de los sinónimos) a ningún concepto del Metathesaurus UMLS, lo que representa sólo el 20 %. Además, hay alineamientos simples (a un solo concepto UMLS) para 22.644 conceptos Emtree, un 43.6 % del total, y alineamientos ambiguos (a más de un concepto UMLS) para 18.890 conceptos Emtree, el 36.4 %. En total, 41.534 conceptos (PTs más sinónimos) en Emtree se equiparan con uno o más conceptos en el Metathesaurus UMLS, es decir, el 80 % de todos los conceptos

Facets Emtree	Tipo de alineamiento	**Emtree Term**		**Emtree Concept**	
		No	%	**No**	%
Anatomical concepts	Sin alineamiento	5.412	49.5 %	392	17 %
	Simple	3.754	34.3 %	773	33.5 %
	Ambiguo	1.763	16.1 %	1.140	49.5 %
	Total	5.517	50.5 %	1.913	83 %
Organism	Sin alineamiento	2.042	28.5 %	81	3.1 %
names	Simple	4.329	60.4 %	1.637	63.4 %
	Ambiguo	797	11.1 %	862	33.4 %
	Total	5.126	71.5 %	2.499	96.8 %
Physical diseases,	Sin alineamiento	15.004	43.3 %	697	10.8 %
disorders and	Simple	18.083	52.2 %	3.238	50.2 %
abnormalities	Ambiguo	1.542	4.4 %	2.508	38.9 %
	Total	19.625	56.7 %	5.746	89.1 %
Chemical	Sin alineamiento	91.312	62.6 %	3.836	16.3 %
and Drugs	Simple	50.187	34.4 %	10.522	44.6 %
	Ambiguo	4.248	2.9 %	9.205	39.1 %
	Total	54.435	37.3 %	19.727	83.7 %
Analytical, diagnostic	Sin alineamiento	13.755	57.7 %	2.073	34.5 %
and therapeutic	Simple	8.514	35.7 %	2.237	37.2 %
techniques, equipment	Ambiguo	1.582	6.6 %	1.698	28.3 %
and parameters	Total	10.096	42.3 %	3.935	65.5 %
Psychological and	Sin alineamiento	1.216	35.2 %	66	8 %
psychiatric phenomena	Simple	2.079	60.2 %	415	50 %
	Ambiguo	157	4.5 %	348	42 %
	Total	2.236	64.8 %	763	92 %
Biological phenomena	Sin alineamiento	14.318	58 %	1.838	31 %
and functions	Simple	8.815	35 %	2.404	40.4 %
	Ambiguo	1.561	6.3 %	1.702	28.6 %
	Total	10.376	42 %	4.106	69 %
Chemical, physical	Sin alineamiento	1.758	61.2 %	409	41.1 %
and mathematical	Simple	966	33.6 %	303	39.6 %
phenomena	Ambiguo	148	5.1 %	191	19.2 %
	Total	1.114	38.8 %	584	58.8 %
Society	Sin alineamiento	2.801	46.2 %	771	40 %
and environment	Simple	2.624	43.3 %	483	25.1 %
	Ambiguo	636	10.5 %	671	34.9 %
	Total	3.260	53.8 %	1.154	60 %
Types of article	Sin alineamiento	248	46.9 %	77	42.5 %
or study	Simple	231	43.7 %	62	34.2 %
	Ambiguo	50	9.4 %	42	23.2 %
	Total	281	53.1 %	104	57.4 %

Tabla 6.2: Resultado por facet del alineamiento léxico con NormalizeString (Continúa en la tabla 6.3)

Facets Emtree	Tipo de alineamiento	**Emtree Term**		**Emtree Concept**	
		No	%	**No**	%
Geographic	Sin alineamiento	83	13.4 %	10	3.5 %
names	Simple	506	82 %	223	77.7 %
	Ambiguo	28	4.5 %	54	18.8 %
	Total	534	86.5 %	277	96.5 %
Groups by age	Sin alineamiento	56	34.8 %	5	10 %
and sex	Simple	77	47.8 %	21	42 %
	Ambiguo	28	17.4 %	24	48 %
	Total	105	65.2 %	45	90 %
Named groups	Sin alineamiento	196	30.7 %	19	11.1 %
of persons	Simple	332	52 %	52	30.4 %
	Ambiguo	110	17.2 %	100	50.5 %
	Total	442	69.3 %	152	88.9 %
Health care	Sin alineamiento	530	29.5 %	33	9.5 %
concepts	Simple	1.073	50.6 %	103	29.7 %
	Ambiguo	196	10.9 %	211	60.8 %
	Total	1.296	70.5 %	314	90.5 %
Biomedical disciplines,	Sin alineamiento	414	38.8 %	38	15 %
science and art	Simple	527	49.3 %	81	32 %
	Ambiguo	127	11.9 %	134	53 %
	Total	654	61.2 %	215	85 %

Tabla 6.3: Resultado por facet del alineamiento léxico con NormalizeString (Viene de la tabla 6.2)

Emtree (Véase tabla 6.1). En promedio, se han encontrado 3,2 alineamientos por concepto Emtree.

El porcentaje de cobertura obtenido es elevado, superior al 50 %, en todas las facets. En 5 de ellas, más del 90 % de los conceptos tienen equiparación léxica a algún concepto UMLS. Prácticamente igualadas están *Organism names* y *Geographic names*, con 96.8 y 96.5 % respectivamente, y las otras son *Psychological and psychiatric phenomena*, *Groups by age and sex* y *Biomedical disciplines, science and art*. Otras 5 tienen entre 80 % y 90 %, que son *Anatomical concepts*, *Physical diseases, disorders and abnormalities*, *Chemical and Drugs*, *Named groups of persons* y *Biomedical disciplines, science and art*. Las 5 facets restantes tienen, en todo caso, más del 50 %. Entre ellas, *Analytical, diagnostic and therapeutic techniques, equipment and parameters*, *Biological phenomena and functions*, *Chemical, physical and mathematical phenomena* y *Society and enviroment*. La facet de menor porcentaje es *Types of article or study* con un 57.4 %. Por tanto, en general, el porcentaje de conceptos Emtree con equiparación léxica a algún concepto UMLS es elevado, sobre todo, en las facets más significativas.

En cuanto al porcentaje de alineamientos ambiguos en el total de alineamientos, éste es mayor que para los alineamientos de términos. De hecho,

en una facet, *Health care concepts*, llega al 60 % y en dos, *Named groups of persons* y *Biomedical disciplines, science and art* supera el 50 %. Además, 10 facets tienen entre un 20 y un 40 % de conceptos con alineamiento ambiguo. Sólo 2 facets tienen menos de 20 %, son *Chemical, physical and mathematical phenomena* con un 19,2 % y *Geographic Names* con un 18.8 %. En el caso de esta última facet, el porcentaje de alineamientos simples es de 77.7 %, lo cual confirma lo explicado en el apartado anterior, ya que al ser una facet de un contenido muy concreto y limitado del dominio, como son los países, tanto el porcentaje de alineamientos por término y por concepto sale muy alto, 86.5 % y 96.5 % respectivamente. Por último, 4 facets tienen un elevado índice de equiparaciones léxicas tanto de términos como de conceptos; son *Organism names*, *Geographic names*, *Named groups of persons* y *Health care concepts*.

6.4.3. Análisis cuantitativo por facet

A continuación, analizaremos facet a facet los resultados obtenidos.

- *Anatomical concepts* obtiene equiparaciones léxicas para el 50.5 % de sus términos y el 83 % de sus conceptos. Esto puede explicarse por el alto número de sinónimos, de los cuales muchos pueden no obtener correspondencias.

- *Organism names* consigue un porcentaje alto tanto de alineamientos de término como de concepto, 71.5 % y 96.8 %, siendo este último el más alto entre las facets. Esto supone que prácticamente todos los conceptos de la facet encuentran correspondencia en UMLS.

- *Physical diseases, disorders and abnormalities* obtiene un 56.7 % de equiparaciones léxicas para términos y de 89.1 % para conceptos, un caso similar al del *Anatomical concepts*. Se observa también que el porcentaje de alineamientos ambiguos para términos es muy bajo, de 4.4 %, aunque para los conceptos asciende a 38.9 %.

- *Chemical and Drugs* es el ejemplo más claro de esa misma tendencia. Se trata de la facet de mayor tamaño pero obtiene el porcentaje de equiparaciones léxicas de los términos más bajo, 37.3 %, mientras que el porcentaje de conceptos es, en cambio, alto, 83.7 %. En cuanto a los alineamientos ambiguos, el porcentaje es también el más bajo de todos para los términos, 2.9 %, y algo más alto para conceptos, 39.1 %.

- *Analytical, diagnostic and therapeutic techniques, equipment and parameters* muestra unos porcentajes de correspondencias medios, el 42.3 % de los términos tiene equiparación y el 65.5 % de los conceptos.

- *Psychological and psychiatric phenomena* obtiene un porcentaje alto para términos, 64.8 %, pero más alto aún para concepto, 92 %.

- *Biological phenomena and functions* es una facet de conceptos genéricos que obtiene correspondencias léxicas para el 42 % de sus términos y para el 69 % de los conceptos.

- *Chemical, physical and mathematical phenomena* tiene un porcentaje bajo de correspondencias léxicas de términos, 38.8 %, y 58.8 % para conceptos, que aunque supera el 50 % no es muy alto comparado con el resto de facets. Esto puede ser debido por el carácter genérico de la facet.

- *Society and enviroment* obtiene alineamientos para el 53.8 % para los términos y el 60 % para conceptos, al tratarse de términos genéricos que UMLS puede no contener.

- *Types of article or study* es un caso similar a otros vistos, donde las correspondencias obtenidas son de nivel medio con un 53.1 % de términos y un 57.4 % de los conceptos, lo que supone además es el porcentaje más bajo de correspondencias léxicas para conceptos.

- *Geographic names* tiene el porcentaje de alineamientos de términos más elevado 86.5 % y el de conceptos, es el segundo más elevado, 96.5 %. El porcentaje de conceptos con alineamientos ambiguos es el más bajo, 18.8 %, lo cual puede deberse a la naturaleza especial de la categoría. Para cada nombre geográfico, es poco probable que haya más de una correspondencia.

- *Groups by age and sex* obtiene equiparaciones léxicas para el 65.2 % de términos y el 90 % de conceptos, lo cual implica un porcentaje bastante alto de cobertura. Es la facet con el porcentaje de términos con alineamiento ambiguo más alto, 17.4 %.

- *Named groups of persons* es un caso similar al anterior, con un 69.3 % de equiparaciones léxicas de términos y un 88.9 % de conceptos.

- *Health care concepts* es la facet con el porcentaje de conceptos con alineamiento ambiguo más alto, 60.8 %, probablemente por tratarse de términos genéricos del ámbito de la salud. El 70.5 % de los términos

Facet	Conceptos	Redundantes	%
Anatomical concepts	4318	108	2.5 %
Organism names	4095	110	2.7 %
Physical, diseases, disorders and abnormalities	9896	498	5.0 %
Chemical and Drugs	37427	1016	2.7 %
Analytical, diagnostic and therapeutic techniques, equipment and parameters	7175	253	3.53 %
Psychological and psychiatric phenomena	1518	39	2.6 %
Biological phenomena and functions	7067	241	3.4 %
Chemical, physical and mathematical phenomena	925	27	2.9 %
Society and enviroment	3027	62	2.0 %
Types of article or study	205	6	2.9 %
Geographic names	453	4	0.9 %
Groups by age and sex	97	6	6.2 %
Named groups of persons	398	3	0.7 %
Health care concepts	1047	13	1.2 %
Biomedical disciplines, science and art	736	3	0.4 %
Total	78384	2389	3.0 %

Tabla 6.4: Resumen de alineamientos redundantes por facet

y el 90.5 % de los conceptos tienen correspondencias léxicas, los cuales son bastantes altos.

- *Biomedical disciplines, science and art* es similar a las anteriores, con porcentajes medios de correspondencias de términos, 61.2 %, y alto de correspondencias de conceptos,85 %.

6.4.4. Alineamientos redundantes

Como resultado del alineamiento, encontramos conceptos UMLS que se equiparan léxicamente con más de un concepto Emtree. En total, suponen el 3 % de los conceptos UMLS. En la tabla 6.4, pueden verse los porcentajes por facet. La facet donde más sucede es en *Groups by age and sex*.

A continuación, explicamos varios ejemplos, que se muestran en la figura 6.4, en la que puede observarse que, entre los conceptos Emtree que tienen correspondencia léxica con el mismo concepto UMLS, puede haber relación broader-narrower directa o indirecta. No hemos localizado ningún caso donde los conceptos Emtree no tuvieran algún ascendiente común.

- El concepto UMLS *Liver* se equipara léxicamente a los conceptos Emtree *liver* y *liver structure*, donde el primero es el broader del segundo.

- El concepto UMLS *Stem cells* se equipara a los conceptos Emtree *stem cell*, *precursor cell* y *blast cell*. El primero de ellos tiene como broaders *cells and cell components* y *hematopoietic system*, precursor cell tiene como broader *cells and cell components* y *blast cell* tiene como broader

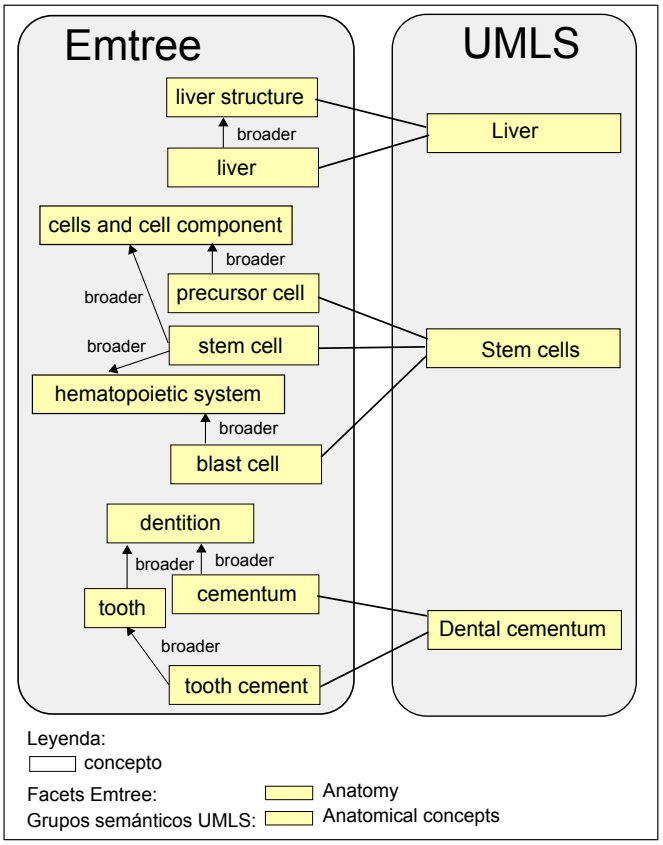

Figura 6.4: Ejemplo de conceptos EMTREE en más de una facet

hematopoietic system. Por tanto, aunque los 3 términos Emtree no están relacionados entre sí, sí tienen los mismos broaders.

- Los conceptos Emtree *cementum* y *tooth cement* se equiparan léxicamente con el concepto UMLS *Dental cementum*. Entre ellos, no hay relación pero sí tienen un broader común, pero no directo, *dentition*, como puede verse en la figura 6.4.

6.5. Validación y desambiguación del alineamiento léxico

En esta sección expondremos los resultados alcanzados tras la validación y desambiguación del alineamiento léxico.

6.5.1. Compatibilidad de categorías de alto nivel

Como se vio en el capítulo anterior, hemos establecido heurísticamente que una facet Emtree y un grupo semántico UMLS son compatibles, si para la facet, al menos, un 60 % de equiparaciones léxicas corresponden a ese grupo semántico. Una vez identificados automáticamente las facets y los grupos semánticos compatibles, validamos los alineamientos léxicos obtenidos, descartando aquellos que pertenecen a grupos semánticos no compatibles con la facet correspondiente.

Identificación automática de grupos compatibles

En la tabla 6.5, se muestra la información de los alineamientos obtenidos por facet y de los grupos semánticos predominantes. Dicha tabla incluye el número de conceptos Emtree, el número de conceptos con alineamiento, el porcentaje que suponen del total, el número de alineamientos obtenidos, desglosando para los 3 grupos semánticos más predominantes.

Analizando esos datos, se ve que algunas de las similitudes entre facets y GS intuidas en el apartado 6.3.3 se confirman. Así, se obtiene más de un 80 % de compatibilidad entre las siguientes categorías de alto nivel: *Anatomical concepts* y *Anatomy*, *Organism names* y *Living Beings*, *Chemicals and Drugs* en ambas, *Physical diseases, disorders and abnormalities* y *Disorders* y, por último, *Geographic areas* en ambas. También *Named groups of persons* tiene un 63 % de conceptos UMLS de un grupo semántico concreto, en este caso, *Living Beings*.

En la figura 6.5, se relaciona cada facet Emtree con el grupo semántico con mayor porcentaje de correspondencia, indicando ese porcentaje al lado. En negrita, están aquellas correspondencias que ya habíamos visto en el análisis a simple vista del apartado 6.3.3.

Consideramos que una facet Emtree y un grupo UMLS son compatibles cuando, al menos, el 60 % de los conceptos UMLS obtenidos de la equiparación léxica de la facet pertenecen al grupo semántico. Como se puede ver en la figura 6.5, esto se cumple para los siguientes pares facet Emtree - grupo semántico UMLS:

- Anatomical concepts - Anatomy
- Organism names - Living Beings
- Physical diseases, disorders and abnormalities - Disorders
- Chemical and drugs - Chemical & Drugs

Facet Emtree	Grupos predominantes Nombre	nº	%
Anatomical concepts	**Anatomy**	3675	81.7 %
	Disorders	378	8.4 %
	Concepts & Ideas	108	2.4 %
Organism names	**Living Begins**	3711	86.7 %
	Chemicals & Drugs	322	7.5 %
	Disorders	163	3.8 %
Physical disorders and abnormalities	**Disorders**	9850	92 %
	Concepts & Ideas	230	2.1 %
	Living Beings	116	1.1 %
Chemicals and drugs	**Chemicals & Drugs**	36419	93.3 %
	Genes & Molecular Seq.	682	1.7 %
	Procedures	572	1.5 %
Analytical, diagnostic and therapeutic techniques, equipment and parameters	Procedures	3669	49 %
	Concepts & Ideas	831	11 %
	Devices	605	8 %
Psychological and psychiatric phenomena	Disorders	664	42 %
	Physiology	293	18.5 %
	Activies & Behaviors	244	15.4 %
Biological phenomena and functions	Chemicals & Drugs	2654	35.7 %
	Physiology	1761	23.7 %
	Disorders	520	7 %
Chemical, physical and mathematical phenomena	Phenomena	262	27.3 %
	Concepts & Ideas	211	22 %
	Physiology	117	12 %
Society and enviroment	Concepts & Ideas	694	22 %
	Activies & Behaviors	393	12.5 %
	Living Beings	372	11.9 %
Types of article or study	Concepts & Ideas	105	49 %
	Procedures	57	26.6 %
	Disorders	11	5.14 %
Geographic areas	**Geographic Areas**	434	92.7 %
	Living Beings	23	4.9 %
	Concepts & Ideas	4	0.9 %
Groups by age and sex	Living Beings	42	39.6 %
	Concepts & Ideas	22	20.7 %
	Physiology	17	16 %
Named groups of persons	**Living Beings**	258	63.4 %
	Concepts & Ideas	90	22.1 %
	Disorders	22	5.4 %
Health care concepts	Procedures	328	30.3 %
	Organizations	216	20 %
	Concepts & Ideas	207	19.2 %
Biomedical disciplines science and art	Occupations	284	38.3 %
	Concepts & Ideas	180	24.3 %
	Living Beings	90	12.1 %

Tabla 6.5: Grupos semánticos predominantes por facet

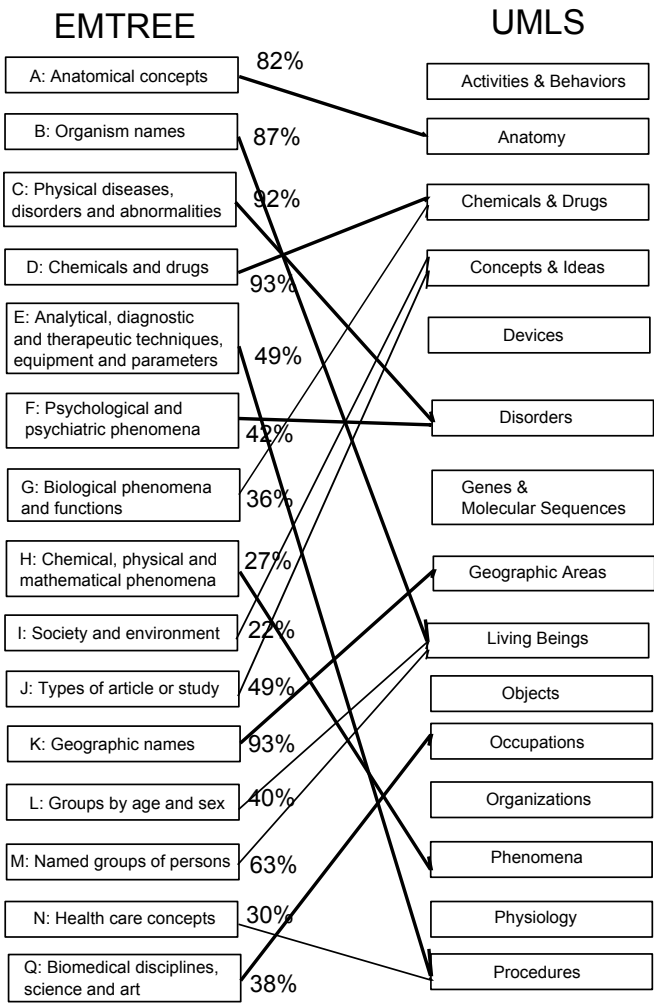

Figura 6.5: Correspondencia entre facets Emtree y Grupos Semánticos UMLS

- Geographic names - Geographic Areas
- Named groups of persons - Living Beings

El resto de facets no se tienen en cuenta en nuestro estudio. Al no tener una correspondencia clara con ningún grupo semántico, no es posible validar las equiparaciones obtenidas. Esto no afecta a la cobertura de nuestro método ya que estas 6 facets suponen el 78 % del total de conceptos Emtree.

Análisis de resultados

Consideramos que un alineamiento es compatible cuando el concepto UMLS pertenece al grupo semántico compatible con la facet del concepto

Facet Emtree	Tipo de alineamiento	Alineamientos de Término Incompatibles	Alineamientos de Concepto Incompatibles	Precisión Total
Anatomical concepts	Simple	290	78	30 %
	Ambiguous	790	16	92 %
	Total	1080 (13.4 %)	94 (4.1 %)	86 %
Organism names	Simple	85	8	38 %
	Ambiguous	687	1	36 %
	Total	772 (12.5 %)	9 (0.4 %)	28 %
Physical diseases,	Simple	527	109	28 %
disorders	Ambiguous	735	49	54 %
and abnormalities	Total	1.262 (5.9 %)	158 (2.5 %)	50 %
Chemicals and drugs	Simple	622	119	56 %
	Ambiguous	2.933	17	95 %
	Total	3.555 (5.9 %)	136 (0.6 %)	93 %
Geographic names	Simple	3	0	0 %
	Ambiguous	31	0	100 %
	Total	34 (6.02 %)	0 (0 %)	100 %
Named groups	Simple	46	5	40 %
of persons	Ambiguous	178	8	81 %
	Total	224 (35.9 %)	13 (7.6 %)	79 %
Totals		6.927 (7.9 %)	410 (1.3 %)	78 %

Tabla 6.6: Número de alineamientos inconsistentes detectados por nuestro método junto con la precisión total

Emtree. A continuación, analizamos las equiparaciones obtenidas para esas 6 facets y elegimos las compatibles.

En la tabla 6.6, se presenta el análisis de los alineamientos considerados incompatibles, distinguiendo entre alineamientos de términos y de concepto. Para cada facet, también hemos separado los resultados en los alineamientos simples y ambiguos. Nuestro método detectó 6.927 (7,9 %) alineamientos inconsistentes sobre el total de alineamientos de términos. Esto dio lugar a 410 alineamientos inconsistentes (1,3 %) del total de los alineamientos de conceptos.

Una mirada más cercana al número de alineamientos incompatibles para los términos y conceptos, distinguiendo entre los alineamientos simples y ambiguos para cada facet, permite destacar los siguientes resultados:

1. Para términos Emtree, el número de alineamientos simples incompatibles es inferior a los ambiguos.

2. Para conceptos Emtree, el número de alineamientos simples incompatibles es superior a los ambiguos, excepto para la facets *Geographic names* y *Named groups of persons*.

La última columna de la tabla 6.6 muestra la precisión global de nuestro método, para cada una de los seis facets Emtree que son compatibles con algún Grupo Semántico de UMLS. Simplemente revisando la tabla 6.6, es evidente que las facets tienen perfiles similares:

- La precisión general de nuestro método en los alineamientos simples es baja: alrededor del 35 %, excepto para las facets *Chemicals and drugs* (56 %) y *Geographic names* (0 %). Este último caso no es significativo, ya que esta facet contiene un número realmente pequeño de alineamientos inconsistentes.

- La precisión general de nuestro método en los alineamientos ambiguos es alta: alrededor del 90 %, excepto para las facets *Organism names* (36 %) y *Physical diseases, disorders and abnormalities* (54 %).

- La precisión general de nuestro método en los alineamientos simples es considerablemente menor que en los alineamientos ambiguos, con la excepción de *Organism names*.

- La precisión total de nuestro método es muy alta: 78 %. Si bien la precisión general en los alineamientos simples es baja, el número de alineamientos ambiguos devueltos por NormalizeString es considerablemente mayor que el número de alineamientos simples. Por lo tanto, nuestro método detecta alineamientos inconsistentes con una alta precisión.

- En algunos casos, hay una correlación entre la precisión global y la puntuación de compatibilidad entre cada facet y el grupo semántico de UMLS correspondiente: cuanto mayor es la compatibilidad entre una facet y un grupo semántico, mayor es la precisión de nuestro método (Tabla 6.6).

Además, las facets (*Organism names* y *Physical diseases, disorders and abnormalities*) presentan las precisiones inferiores de nuestro método (36 % y 50 %).

Con el fin de analizar las causas que producen una disminución en la precisión de nuestro método, la tabla 6.7 presenta los resultados de la evaluación de cada facet particularizada para cada Grupo Semántico inconsistente de UMLS. Es evidente que, con la excepción de *Organism names* y *Physical diseases, disorders and abnormalities*, las facets presentan la menor precisión de nuestro método cuando la cobertura de las facets por GS es el más bajo. Por ejemplo, la facet *Anatomical concepts* presenta la menor precisión (0-20 %) para el GS *Genes and Molecular Sequences*, que sólo representa el 0,4 % de los mappings totales para la facet.

Facet Emtree	**Grupos Semánticos Inconsistentes**	**Cobertura de Facet por GS**	**Precisión**
Anatomical concepts	Activities and Behaviors, Devices, Geographic Areas Phenomena, Procedures Concepts and Ideas Occupations, Disorders , Objects	13.4 %	100-80 %
	Physiology, Living Beings Chemicals and Drugs	4.4 %	80-20 %
	Genes and Molecular Sequences	0.4 %	20-0 %
Organism names	Activities and Behaviors, Genes and Molecular Sequences, Geographic Areas, Disorders, Procedures, Concepts and Ideas	4.7 %	100-80 %
	Phenomena, Anatomy	0.3 %	80-20 %
	Chemicals and Drugs, Objects, Physiology	8.3 %	20-0 %
Physical diseases, disorders and abnormalities	Chemicals and Drugs, Devices, Genes and Molecular Sequences, Geographic Areas, Living Beings, Occupations, Organizations, Procedures	3.6 %	100-80 %
	Objects, Concepts and Ideas	2.2 %	80-20 %
	Activities and Behaviors, Anatomy, Phenomena, Physiology	2.1 %	20-0 %
Chemicals and drugs	Activities and Behaviors, Anatomy, Disorders, Genes and Molecular Sequences, Geographic Areas, Occupations, Organizations, Physiology, Procedures Living Beings, Phenomena	5.9 %	100-80 %
	Concepts and Ideas	0.4 %	80-20 %
	Objects, Devices	0.4 %	20-0 %
Geographic Areas	Chemicals and Drugs, Concepts and Ideas, Disorders Genes and Molecular Sequences, Living Beings, Procedures	7.2 %	100-80 %
Named groups of persons	Anatomy, Devices, Disorders, Geographic Areas, Objects, Procedures Concepts and Ideas	31.4 %	100-80 %
	Activities and Behaviors, Organizations, Physiology	2.9 %	80-20 %
	Occupations	2.2 %	20-0 %

Tabla 6.7: Evaluación de resultados particularizados por los Grupos Semánticos inconsistentes

Facet Emtree	Tipo	Nº Conceptos	Porcentaje
Anatomical concepts	Cluster	583	32.1 %
	Sin Cluster	1236	67.9 %
	Total	1819	
Organism names	Cluster	304	12.2 %
	Sin Cluster	2186	87.8 %
	Total	2490	
Physical diseases, disorders and abnormalities	Cluster	2394	42.8 %
	Sin Cluster	3194	57.2 %
	Total	5588	
Chemicals and drugs	Cluster	6409	32.7 %
	Sin Cluster	13182	67.3 %
	Total	19591	
Geographic names	Cluster	19	6.9 %
	Sin Cluster	258	93.1 %
	Total	277	
Named groups of persons	Cluster	26	18.7 %
	Sin Cluster	113	81.3 %
	Total	139	
Total	Cluster	9735	32.6 %
	Sin Cluster	20169	67.4 %
	Total	29904	

Tabla 6.8: Resultados de Clusters

6.5.2. Desambiguación basada en información estructural

Identificación de Clusters

En la tabla 6.8, se muestra la información sobre los clusters identificados, indicando el número de conceptos Emtree que tienen clusters y el porcentaje que suponen del total. Como se puede verse, un 32.6 % de los conceptos Emtree tienen 1 o más clusters.

El mayor porcentaje de conceptos Emtree con clusters se da en la facet *Physical diseases, disorders and abnormalities* con un 42.8 %, seguido de *Chemical and drugs* con un 32.7 % y de *Anatomical concepts* con un 32.1 %. Esto indica un mayor número de conceptos similares procedentes probablemente de las terminologías fuente de UMLS y que, por tanto, estarán relacionados entre sí. La facet con menor porcentaje es *Geographic names* con un 6.9 % que puede deberse a que, para el ámbito geográfico, zonas del mundo, países, etc., cada elemento del dominio suele tener un solo concepto que lo representa.

Facet Emtree	Similitud	Nº Alineamientos	Porcentaje
Anatomical concepts	Sin sim.	1471	55.9 %
	Baja	167	6.3 %
	Alta	357	13.6 %
	Muy alta	640	24.3 %
Organism names	Sin sim.	2251	74.2 %
	Baja	93	3.1 %
	Alta	205	6.8 %
	Muy alta	485	16.0 %
Physical diseases, disorders	Sin sim.	3751	51.5 %
and abnormalities	Baja	673	9.2 %
	Alta	1327	18.2 %
	Muy alta	1529	21.0 %
Chemicals and drugs	Sin sim.	15344	66.5 %
	Baja	2029	8.8 %
	Alta	2823	12.2 %
	Muy alta	2886	12.5 %
Geographic names	Sin sim.	158	51.5 %
	Baja	13	4.2 %
	Alta	23	7.5 %
	Muy alta	115	37.5 %
Named groups of persons	Sin sim.	165	83.8 %
	Baja	6	3.0 %
	Alta	8	4.1 %
	Muy alta	20	10.2 %
Total	Sin sim.	23140	63.4 %
	Baja	2983	8.2 %
	Alta	4743	13.0 %
	Muy alta	5675	15.5 %

Tabla 6.9: Resultados de Factores de similitud

Cálculo del factor de similitud

En la tabla 6.9, se muestran los porcentajes de factores de similitud obtenidos para los alineamientos de cada facet. Para ello, se han definido 4 rangos de similitud que son:

- Sin similitud, cuando el factor de similitud es 0
- Similitud baja cuando es mayor que 0 y menor que 50
- Similitud alta cuando es mayor o igual que 50 y menor que 80
- Similitud muy alta si es mayor o igual que 80.

Un 63.4 % de los alineamientos no tienen similitud estructural directa, mientras que el 8 % la tiene baja, el 13 % la tienen alta y el 15.5 % la tienen muy alta. Como puede verse, el porcentaje de alineamientos sin similitud es

mayor del 50 % para todas las facets, siendo la facet con mayor porcentaje *Named groups of persons* con un 83.8 %. Esto puede deberse a una diferente estructuración de los conceptos a equiparar en Emtree y UMLS, sobre todo teniendo en cuenta que ésta era la facet con un porcentaje más bajo de compatibilidad con el grupo semántico *Living Beings*, con un 63 %. Le siguen *Organism names* con un 74.5 % y *Chemical and Drugs* con un 66.5 %, lo cual puede deberse a lo mismo, a las diferencias en las relaciones de los conceptos. *Anatomical concepts*, *Physical diseases, disorders and abnormalities* y *Geographic names* rondan el 50 %.

El porcentaje de alineamientos con similitud baja oscila entre el 3 % y el 9 %, siendo *Named groups of persons* de nuevo la de menor porcentaje y *Physical diseases, disorders and abnormalities* el de mayor porcentaje. Esto coincide con las facets con menor y mayor de porcentaje de alineamientos con similitud alta, que en este caso oscila entre el 4 % y el 18 %.

Por último, la facet con mayor porcentaje de alineamientos con similitud muy alta es *Geographic names* con un 37 % seguida por *Anatomical concepts* con un 24 %. Esto puede deberse a una mayor similitud en la estructura jerárquica entre Emtree y UMLS, lo cual tiene sentido dado que su contenido incluye términos geográficos como países, continentes o regiones la primera y partes anatómicas la segunda, que son áreas de conocimiento con una estructura más estandarizada que otras. Con menor porcentaje, están *Named groups of persons* con un 10.2 % y *Chemical and Drugs* con un 12.5 %, que serán las facets con la menor similitud al grupo semántico compatible.

Validación usando información estructural

En la tabla 6.10, se muestran los resultados del algoritmo de validación estructural por cada facet Emtree y los datos totales. Para ello, se indican en las 2 primeras columnas el número y el porcentaje de conceptos Emtree con algún alineamiento considerado válido. A continuación, se indica el número de alineamientos considerados válidos y el porcentaje que supone con respecto al número de alineamientos obtenidos a UMLS.

Como puede observarse, el porcentaje total de alineamientos considerados válidos no es muy alto, un 24.05 %, lo que supone una cobertura del 41.22 % de los conceptos Emtree. Este porcentaje puede parecer bajo, pero hay que tener en cuenta que nuestro método no sólo hace una validación, sino también una desambiguación entre conceptos muy similares. Por tanto, tenemos la seguridad de que éstos son los más similares de los disponibles y su validez está confirmada semánticamente por el uso de las relaciones broader-narrower.

Por facets, *Geographic names* y *Physical diseases, disorders and abnormalities* son las facets con mayor porcentaje de alineamientos validados, un

Facets Emtree	Conceptos Emtree		Alineamientos	
	No	**Cobertura(%)**	**No**	%
Anatomical concepts	953	52.39 %	1106	30.1 %
Organism names	739	29.68 %	767	20.67 %
Physical diseases, disorders and abnormalities	3159	56.53 %	3356	34.07 %
Chemicals and drugs	7294	37.23 %	7659	21.03 %
Geographic names	148	53.43 %	148	34.1 %
Named groups of persons	32	23.02 %	32	12.4 %
Total	12325	41.22 %	13068	24.05 %

Tabla 6.10: Conceptos con alineamiento que superan el algoritmo de desambiguación

34.1 % y 34.07 % respectivamente, posiblemente debido una mayor similitud en la estructuración entre las facets y su grupo semántico compatible, *Geographic names* y *Disorders*, respectivamente. *Named groups of persons* es la de menor porcentaje de alineamientos validados, 12.4 %, posiblemente por la razón contraria, una muy diferente estructuración. El orden se mantiene en el porcentaje de cobertura, superando *Physical diseases, disorders and abnormalities*, *Geographic names* y *Anatomical concepts* el 50 % de conceptos Emtree con alineamientos validados. La de menor cobertura es de nuevo *Named groups of persons* con un 23.02 %.

Podemos dar 2 razones para estos resultados. Por un lado, de aquellos conceptos de los que hay información estructural, no se consideran validados todos los alineamientos sino sólo aquellos de mayor puntuación. Por otro lado, como se veía en la tabla 6.9, hay un porcentaje alto, 63.4 %, de conceptos sin factores de similitud, debido fundamentalmente a las diferencias en la construcción de las relaciones estructurales en las terminologías.

Expansión de vecinos

Como se explicó en el apartado 5.5.3, debido a la mayor granularidad de UMLS junto a la agrupación de terminología similar en los conceptos EMTREE, dos conceptos que en Emtree están directamente relacionados, en UMLS pueden estarlo indirectamente con 1 o más conceptos intermedios. Por esta razón, en esta fase el proceso recupera los alineamientos cuando se cumple que el broader de un concepto Emtree está relacionado con el broader de un concepto UMLS, aunque no sea directamente, sino indirectamente. También se aplica lo mismo para los narrowers, recuperando los alineamientos cuando algún narrower del concepto Emtree está relacionado con el narrower

Facets Emtree	Nivel 1		Nivel 2		Nivel 3	
	No	%	**No**	%	**No**	%
Anatomical concepts	213	13.95 %	261	17.09 %	276	18.07 %
Organism names	71	3.13 %	79	3.49 %	79	3.49 %
Physical diseases, disorders and abnormalities	519	13.23 %	771	19.66 %	852	21.72 %
Chemicals and drugs	3403	22.07 %	4020	26.07 %	4141	26.85 %
Geographic names	28	17.61 %	31	19.5 %	31	19.5 %
Named groups of persons	6	3.64 %	6	3.64 %	6	3.64 %
Total	4240	18.07 %	5168	22.03 %	5385	22.95 %

Tabla 6.11: Conceptos recuperados por similitud en varios niveles de broader

del concepto UMLS.

Los resultados de la exploración de los broaders se muestran en la tabla 6.11, indicando el número de alineamientos léxicos recuperados y el porcentaje que suponen de los conceptos Emtree que habían sido descartados en la fase anterior. Hay que tener en cuenta que los resultados se suman en cada nivel, es decir, al explorar el nivel 2 se recuperan todos los de nivel 1 y lo mismo, los del nivel 3 incluyen los de nivel 2.

Como puede observarse, de esta forma, en el nivel 3 se recupera el 22.95 % de los alineamientos descartados. *Chemical and Drugs* es la facet para la que más alineamientos se recuperan, con un 26.85 %, seguida de *Physical diseases, disorders and abnormalities* con un 21.72 %. La mayoría de ellos ya se recuperan en el nivel 1 de exploración, es decir, teniendo en cuenta las relaciones directas. En *Chemical and Drugs*, en el nivel 2, se recuperan unos 600 alineamientos más, un 4 %, y en el nivel 3 unos 120, menos de 1 %. En *Physical diseases, disorders and abnormalities*, se recuperan más de 6 % en el nivel 2 y un 2 % en el nivel 3. Esto indica también que, aunque UMLS tiene más granularidad que Emtree, se mantiene cierta similitud estructural.

Organism names recupera un 3.49 % y es la facet que recupera menor porcentaje de alineamientos, lo cual parece indicar que la facet y el grupo semántico compatible *Living Beings* pueden tener diferencias en la jerarquización de contenidos. Eso se confirma al ver que, entre el nivel 2 y el 3, no se recuperan alineamientos nuevos.

El caso más llamativo es *Named groups of persons*, que recupera 6 alineamientos en el nivel 1 pero en los niveles 2 y 3 no recupera ninguno más. Además, son el 3.64 % de los alineamientos de la facet. Esto parece indicar que hay poca similitud entre la estructura de la facet y del grupo semántico compatible, *Living Beings*, lo cual puede verse confirmado por su diferente objetivo. Por un lado, *Living Beings* contiene formas de vida y organismos vi-

Facets Emtree	Nivel 1		Nivel 2		Nivel 3	
	No	%	**No**	%	**No**	%
Anatomical concepts	48	3.14 %	49	3.21 %	49	3.21 %
Organism names	16	0.71 %	16	0.71 %	16	0.71 %
Physical diseases, disorders and abnormalities	144	3.67 %	147	3.75 %	147	3.75 %
Chemicals and drugs	43	0.28 %	44	0.29 %	46	0.3 %
Geographic names	1	0.63 %	1	0.63 %	1	0.63 %
Named groups of persons	0	0.0 %	0	0.0 %	0	0.0 %
Total	252	1.07 %	257	1.1 %	259	1.1 %

Tabla 6.12: Conceptos recuperados por similitud en varios niveles de narrower

vos y *Named groups of persons* son conceptos de oficios, estados de personas y grupos étnicos.

Tampoco se recuperan alineamientos nuevos en el nivel 3 con respecto al nivel 2 en *Geographic names*, que es la segunda facet en recuperación de conceptos en el nivel 1 con un 17.61 % y la tercera en el nivel 2 con un 19.5 %. Esto parece indicar que a nivel estructural la facet y el grupo semántico compatible son similares, también en su granularidad.

Por último, puede observarse que para las otras facets, *Anatomical concepts*, *Physical diseases, disorders and abnormalities* y *Chemicals and Drugs*, entre el nivel 2 y el nivel 1 se recuperan mayor porcentaje de alineamientos que entre el nivel 3 y el nivel 2. Esto indica que es más probable encontrar similitudes con una sola relación intermedia que con 2.

A continuación, mostramos los resultados para los narrowers en la tabla 6.12. Como puede observarse, el número de alineamientos recuperados es mucho menor a los recuperados por broader, 1.1 % frente a 22.95 %, lo cual indica que la similitud entre las terminologías es mucho mayor en los niveles superiores. Esto puede ser debido, entre otras causas, a la gran cantidad de narrowers que tienen habitualmente los conceptos UMLS y a la mayor granularidad de UMLS, siendo más probable encontrar más elementos intermedios (aumenta el nivel de detalle) a medida que se baja en las jerarquías.

En este caso, dos de las facets superan el 3 % de conceptos recuperados, *Physical diseases, disorders and abnormalities* y *Anatomical concepts*, mientras que las demás no llegan al 1 %. En la exploración de broaders, eran la segunda y la tercera respectivamente que más recuperaban lo cual confirma la similitud entre la facet y el grupo semántico compatible.

Puede observarse que 2 de las facets, *Geographic names* y *Named groups of persons*, recuperan 1 y 0 alineamientos respectivamente, lo cual indica que la estructura inferior para los conceptos de estas facets es muy diferente en

Facets Emtree	Nivel 1		Nivel 2		Nivel 3	
	No	%	**No**	%	**No**	%
Anatomical concepts	232	15.19 %	279	18.27 %	293	19.19 %
Organism names	79	3.49 %	87	3.84 %	87	3.84 %
Physical diseases, disorders and abnormalities	594	15.15 %	828	21.11 %	906	23.1 %
Chemicals and drugs	3429	22.24 %	4046	26.24 %	4169	27.03 %
Geographic names	29	18.24 %	32	20.13 %	32	20.13 %
Named groups of persons	6	3.64 %	6	3.64 %	6	3.64 %
Total	4369	18.62 %	5278	22.5 %	5493	23.42 %

Tabla 6.13: Conceptos con alineamientos válidos por expansión de vecinos

Emtree y en UMLS o en alguna de ellas no existe, al tener menos nivel de detalle.

Sólo una de las facets, *Chemical and Drugs*, recupera alineamientos nuevos en el nivel 3, lo cual confirma que esta facet tiene una estructura similar al grupo semántico compatible, aunque algunos de ellos tienen relación indirecta. Las dos facets que más alineamientos recuperan, *Anatomical concepts* y *Physical diseases, disorders and abnormalities*, no recuperan ninguno en el nivel 3. Por último, en las otras 3 facets, *Organism names*, *Geographic names* y *Named groups of persons* no se recuperan alineamientos en los niveles 2 y 3.

Cabe resaltar que la facet *Chemical and Drugs* era la facet que más alineamientos recuperaba por broader, un 26.85 %, mientras que por narrower es la segunda que menos recupera, un 0.3 %, tras *Named groups of persons*. Es decir, en esta facet los narrowers casi no tienen similitud con el grupo semántico, mientras que los broaders sí.

Por último, unimos los resultados de ambas exploraciones y mostramos en la tabla 6.13 el resumen de los alineamientos recuperados por esta técnica. Hay que tener en cuenta que si un concepto es recuperado por broader, ya no se tiene en cuenta en la exploración por narrower.

Los porcentajes de alineamientos recuperados se elevan un poco pero se mantienen similares a la exploración de niveles con broaders. *Named groups of persons* es la única que se mantiene igual al no recuperar ningún mapping por broader. De los demás, *Chemical and Drugs* sigue siendo la facet que más recupera, un 27.03 %, seguida de *Physical diseases, disorders and abnormalities* con un 23.01 %. Le siguen Geographic names con un 20.13 % y Anatomical concepts con un 19.19 %. Por último, *Named groups of persons* es la que menos recupera con un 3.64 % seguida de Organism names con 3.84 %.

Facets Emtree	Conceptos Emtree		Alineamientos		% Dif
	No	Cobertura(%)	No	%	
Anatomical concepts	1229	53.32 %	1399	31.12 %	20.94 %
Organism names	821	31.82 %	854	19.94 %	10.19 %
Physical diseases, disorders and abnormalities	3990	61.93 %	4262	39.82 %	21.26 %
Chemicals and drugs	10996	46.67 %	11829	30.32 %	35.25 %
Geographic names	176	61.32 %	180	38.46 %	17.78 %
Named groups of persons	38	22.22 %	38	9.34 %	15.79 %
Total	17250	48.8 %	18562	31.26 %	29.6 %

Tabla 6.14: Conceptos con alineamientos válidos

Alineamientos válidos totales

Tras el proceso indicado, las equiparaciones obtenidas inicialmente de UMLS se han clasificado en válidas o inválidas, según la información estructural existente permita afirmar la similitud entre el concepto Emtree y el concepto UMLS. Los resultados totales se muestran en la tabla 6.14, en la que se muestran el número de conceptos Emtree con alineamientos válidos y la cobertura, porcentaje que suponen del total de conceptos Emtree buscados. También se indican el número de alineamientos válidos y el porcentaje que suponen del número total de alineamientos recuperados inicialmente de UMLS.

El porcentaje total de alineamientos que nuestro método identifica como válidos es del 31.26 % sobre el total de alineamientos recuperados de UMLS. Esto supone el 48.8 % de los conceptos Emtree, lo cual consideramos una buena cobertura, ya que para casi el 50 % de los conceptos buscados nuestro método es capaz de indicar uno o varios alineamientos validados semánticamente. Este porcentaje es mayor del 60 % para 2 facets, *Physical diseases, disorders and abnormalities* y *Geographic names*. *Anatomical concepts* tiene alineamientos válidos para el 53 % de sus conceptos y *Chemicals and Drugs* para el 46.67 %. Las facets con menor cobertura son *Named groups of persons* con un 22.22 % y *Geographic names* con un 31.82 %.

En cuanto al porcentaje de alineamientos válidos sobre el total de alineamientos obtenidos, los porcentajes no superan el 40 %, siendo *Physical diseases, disorders and abnormalities* y *Geographic names* las de mayor porcentaje y *Named groups of persons* la de menor porcentaje. Esto se debe, como ya indicamos, a que el algoritmo no solamente valida a partir de la información semántica, sino también desambigua entre varios similares eligiendo el de mejor factor de similitud. Por tanto, se descartan tanto aquellos conceptos para los que no tenemos información como aquellos que no son los

Facets Emtree	Alineamientos simples		Alineamientos ambiguos	
	No	%	**No**	%
Anatomical concepts	1048	85.27 %	181	14.73 %
Organism names	783	95.37 %	38	4.63 %
Physical diseases, disorders and abnormalities	3707	92.91 %	283	7.09 %
Chemicals and drugs	10314	93.8 %	682	6.2 %
Geographic names	174	98.86 %	2	1.14 %
Named groups of persons	38	100.0 %	0	0.0 %
Total	16064	93.12 %	1186	6.88 %

Tabla 6.15: Conceptos con alineamientos válidos simples y ambiguos

más similares.

El último paso del proceso, la exploración de similitud en varios niveles, supone un aumento de los alineamientos validados semánticamente del 29.6 %, siendo la facet *Chemical and Drugs* que más aumenta con un 35.25 % y *Geographic names* la que menos con un 19.94 %.

Por último, en la tabla 6.15, analizamos este resultado indicando el porcentaje de conceptos que tienen alineamientos simples y alineamientos ambiguos. Como vemos, el porcentaje de alineamientos simples es muy alto, 93.12 %, lo cual indica que nuestro método consigue desambiguar casi siempre. Sólo el 6.88 % de los conceptos tienen alineamientos ambiguos en los casos en que nuestro algoritmo no ha conseguido elegir sólo uno.

Por facets, *Named groups of persons* no tiene ningún alineamiento ambiguo. Le sigue la facet Geographic names con un 1.14 % de alineamientos ambiguos. De las demás, sólo *Anatomical concepts* supera el 10 % con un 14.73 %. Esto puede ser debido a que en esta facet es frecuente encontrar conceptos muy similares, procedentes de diferentes terminologías fuente, para un mismo concepto pero que, aunque se aprecia directamente esa similitud, es difícil determinar automáticamente dadas las diferencias en las relaciones, también heredadas de las terminologías fuente. En menor proporción, puede pasar esto en *Physical diseases, disorders and abnormalities* y *Chemicals and drugs*, que tiene un 7.09 % y un 6.2 % de alineamientos ambiguos respectivamente.

6.5.3. Análisis de los alineamientos complejos obtenidos a partir de técnicas de procesado de lenguaje natural

En la tabla 6.1, puede observarse que un 20 % de los conceptos Emtree (10.345 conceptos) no tienen ningún alineamiento léxico con UMLS. De estos

	Subfrase Emtree		Concepto Emtree	
	No	%	**No**	%
Sin alineamiento	814	2.8 %	7	0.3 %
Alineamiento simple	3328	35.4 %	231	1.7 %
Alineamiento ambiguo	8127	61.8 %	4060	98 %
Cobertura general	11455	97.2 %	4291	99.7 %
Alineamientos por subfrase/concepto	1	—	1	—

Tabla 6.16: Resultado del alineamiento léxico para las partes constituyentes

conceptos, 5.035 pertenecen a categorías de alto nivel compatibles. Nuestro método intenta generar alineamientos complejos para ellos.

De los 5.035 conceptos, 738 son conceptos cuyo término preferido viene descrito por una sola palabra. De los 4.297 conceptos restantes, se obtienen 9.194 partes constituyentes, lo cual da una media de 2.15 partes por concepto. Este número es tan alto debido a la facet *Chemical and Drugs*, que es la de mayor número de conceptos, donde muchos de ellos son nombres de compuestos químicos que están formados por muchas palabras y números.

Aplicamos el mismo proceso de alineamiento léxico que en las etapas iniciales para obtener los conceptos UMLS que se equiparan a las partes constituyentes. En la tabla 6.16, se muestran los resultados de la equiparación, el número de términos y conceptos sin alineamiento y el número total de alineamientos obtenidos, simples y ambiguos, por términos y por conceptos.

Como puede observarse, el porcentaje de alineamientos ambiguos es muy alto por subfrase, 61.8 %, mientras que por concepto son prácticamente todos, un 98 %. También un 35.4 % de los alineamientos de partes son simples, mientras que sólo el 1.7 % de los conceptos, lo cual es debido a la agrupación de los alineamientos obtenidos por las subfrases en el concepto.

El porcentaje de conceptos sin alineamientos de sus subfrases es muy bajo dada la gran probabilidad de encontrar equiparación léxica en UMLS para ellas al ser normalmente unidades con significado propio.

Análisis cuantitativo por facet

A continuación, mostramos los alineamientos por facet en la tabla 6.17, indicando el número de alineamientos y los porcentajes de subfrases y conceptos que no han obtenido equiparación (sin alineamiento), que han obtenido una equiparación (alineamiento simple) y que han obtenido más de 1 equiparación (alineamiento ambiguo). También se indica el porcentaje total de subfrases y conceptos que tienen equiparación.

Facets Emtree	Tipo	Subfrase Emtree		Concepto Emtree	
		No	%	**No**	%
Anatomical concepts	Sin alineamiento	22	2.93 %	0	0.0 %
	Simple	237	31.6 %	6	1.74 %
	Ambiguo	491	65.47 %	339	98.26 %
	Total	728	97.07 %	345	100.0 %
Organism names	Sin alineamiento	6	4.84 %	0	0.0 %
	Simple	48	38.71 %	1	1.67 %
	Ambiguo	70	56.45 %	59	98.33 %
	Total	118	95.16 %	60	100.0 %
Physical diseases, disorders	Sin alineamiento	35	2.36 %	1	0.15 %
and abnormalities	Simple	562	37.82 %	9	1.38 %
	Ambiguo	889	59.83 %	639	98.31 %
	Total	1451	97.64 %	648	99.69 %
Chemicals and drugs	Sin alineamiento	750	7.61 %	6	0.2 %
	Simple	2464	25.0 %	214	6.45 %
	Ambiguo	6642	67.39 %	3000	90.32 %
	Total	9106	92.39 %	3214	96.77 %
Geographic names	Sin alineamiento	1	7.14 %	0	0.0 %
	Simple	4	28.57 %	1	14.29 %
	Ambiguo	9	64.29 %	6	85.71 %
	Total	13	92.86 %	7	100.0 %
Named groups of persons	Sin alineamiento	0	0.0 %	0	0.0 %
	Simple	13	33.33 %	0	0.0 %
	Ambiguo	26	66.67 %	17	100.0 %
	Total	39	100.0 %	17	100.0 %

Tabla 6.17: Resultado del alineamiento léxico por facet

El mayor porcentaje de subfrases sin alineamiento se produce en la facet *Chemical and Drugs*, con un 7.61 %, seguido de *Geographic names* con un 7.14 %, mientras que *Named groups of persons* tiene un 0 %, es decir, en ésta todas las subfrases obtienen algún alineamiento. Si tenemos en cuenta los conceptos, en cambio, se encuentran alineamientos para todos salvo para las facets *Physical diseases, disorders and abnormalities*y *Chemicals and drugs*, que no tienen alineamiento para 1 y para 6 conceptos respectivamente.

Por tanto, tanto para las subfrases como los conceptos, el porcentaje de ellos que obtienen alineamiento es muy alto, superior siempre al 90 % y en algunas facets, el 100 %. *Chemical and Drugs* es la de menor porcentaje en ambos casos, 92.39 % por partes y 96.77 % por concepto. *Named groups of persons* tiene el 100 % de partes con alineamiento, mientras que tienen un 100 % conceptos con alineamiento esa misma facet y Anatomical conceptos, *Organism names* y *Geographic names*.

Entre un 25 y un 40 % de las subfrases, tienen un alineamiento simple, siendo *Organism names* la de mayor cantidad con un 38.71 % y *Chemical and Drugs* el menor con un 25 %. En cambio en los conceptos, la mayor parte de ellos tienen entre 0 y 2 %, salvo la facet *Geographic names*, que muestra un 14.29 %. De los demás, *Organism names* tiene un 1.67 % y *Physical disea-*

Facets Emtree	Conceptos Emtree		Alineamientos	
	No	**Cobertura(%)**	**No**	%
Anatomical concepts	302	87.54 %	2630	42.33 %
Organism names	57	95.0 %	897	53.91 %
Physical diseases, disorders and abnormalities	580	89.23 %	4705	40.27 %
Chemicals and drugs	2995	90.32 %	36567	97.99 %
Geographic names	6	85.71 %	624	55.27 %
Named groups of persons	16	94.12 %	302	42.6 %
Total	3956	89.1 %	45725	77.87 %

Tabla 6.18: Validación de alineamientos obtenidos tras separación de subfrases

ses, disorders and abnormalities tiene un 0.15 %. Las otras 3 facets no tiene conceptos con alineamientos simples, sino que todos son ambiguos.

Por último, analizamos los resultados de alineamientos ambiguos. Los conceptos de la facet *Named groups of persons* son equiparados siempre a más de 1 concepto UMLS. Para el resto de facets, el porcentaje es muy alto, rondando el 96 %, salvo para *Geographic names* que es de un 85.71 %. Esto puede ser debido a que al hablar de conceptos geográficos es más probable encontrar una sola equiparación que para el resto de las facets. Para las subfrases, el porcentaje también es alto para todas las facets entre el 55 y el 67 %, siendo *Chemical and Drugs* la facet con mayor porcentaje con 67.39 % y *Organism names* la menor con 56.45 % .

Por tanto, puede concluirse que el porcentaje de alineamientos ambiguos es alto para las subfrases, que al reunirse los datos, pasa a ser un porcentaje muy alto para los conceptos, cercano al 100 % en todos los casos. El porcentaje de conceptos que quedan sin equiparación es muy bajo.

Validación semántica de los alineamientos

Como resultado del proceso anterior, un concepto Emtree, que fue particionado en sus subfrases, queda equiparado a 0, 1 o más conceptos UMLS de diferentes grupos semánticos, tal como se vio en el apartado 5.6. A continuación, se realiza la validación de estos alineamientos descartando los conceptos UMLS que pertenecen a grupos semánticos no compatibles a la facet.

En la tabla 6.18, se muestran los resultados de esta validación. Se indican el número de conceptos Emtree con alineamientos válidos y el porcentaje que eso supone de los conceptos Emtre iniciales, la cobertura. También se incluyen el número de alineamientos considerados válidos y el porcentaje que eso supone de los alineamientos inicialmente obtenidos de UMLS.

Para todas las facets, el porcentaje de conceptos con alineamientos válidos es muy alto, superior al 85 %, lo cual consideramos que es un buen resultado, ya que demuestra el alineamiento léxico con UMLS ha obtenido equiparaciones relevantes en la mayor parte de conceptos. Comparando estos resultados con los de la tabla 6.10, donde se obtenía un porcentaje del 41.22 % de conceptos con alineamientos válidos, hay que tener en cuenta que ahora sólo se realiza la validación por grupo semántico, no la de desambiguación. En resumen, el 89.1 % de los conceptos Emtree tiene un alineamiento válido por estar en un grupo semántico compatible.

El porcentaje de alineamientos válidos frente a los alineamientos recuperados inicialmente de UMLS también es elevado, un 77.87 %. Al analizar por facets, vemos que *Chemical and Drugs* tiene el porcentaje más elevado, 97.99 %, lo que indica que la mayor parte de alineamientos conseguidos para los conceptos de esta facet pertenecían al grupo semántico compatible, por tanto, la equiparación con UMLS devuelve menos conceptos potencialmente no válidos. En cambio, para las otras facets, ronda entre el 40 y el 55 %, siendo la menor *Physical diseases, disorders and abnormalities* con un 40.17 % y la mayor *Geographic names* con un 55.27 %.

6.5.4. Análisis de los alineamientos complejos

Como se ve en la tabla 6.18, de los 4.291 conceptos que obtuvieron equiparación léxica de alguna de sus subfrases, 3.956 ha resultado tener alineamientos válidos semánticamente. Estos alineamientos complejos pueden ser de dos tipos, en función de la información que aportan sobre el concepto Emtree, que puede ser:

- **UnionMatch**: cuando el concepto de la terminología fuente queda equiparado a varios conceptos de la terminología destino cada uno de lo cuales tiene parte de su significado, incluso pudiendo solaparse.
- **BroadMatch**: cuando el concepto de la terminología fuente queda equiparado a conceptos de significado más general en la terminología destino.

En la tabla 6.19, se muestran los alineamientos de estos dos tipos por facet, indicando el número y porcentaje que suponen en el total.

Geographic names es la facet con mayor porcentaje de alineamientos complejos UnionMatch, con un 33 %. Las demás facets oscilan entre el 8.77 % de *Organisms names* y el 2.65 % de *Anatomical concepts*. Por otro lado, esta es la facet con mayor porcentaje de alineamientos BroadMatch, un 97.35 %, siendo *Geographic names* la de menor porcentaje con un 66 %.

Facets Emtree	Conceptos Validados	UnionMatch		BroadMatch	
		No	%	No	%
Anatomical concepts	302	8	2.65 %	294	97.35 %
Organism names	57	5	8.77 %	52	91.23 %
Physical diseases, disorders and abnormalities	580	23	3.97 %	557	96.03 %
Chemicals and drugs	2995	214	7.14 %	2781	92.86 %
Geographic names	6	2	33.33 %	4	66.67 %
Named groups of persons	16	1	6.25 %	15	93.75 %
Total	3956	253	4.15 %	3703	95.85 %

Tabla 6.19: Conceptos Emtree con alineamientos complejos

Como puede observarse, el porcentaje total de alineamientos complejos UnionMatch encontrados es muy bajo, 4.15 %. Esto viene dado por la propia naturaleza de los conceptos Emtree a equiparar. Cuando el concepto Emtree contiene una conjunción, las subfrases generadas muestran los elementos unidos por ella, como por ejemplo, los casos mostrados en la figura 4.21. En el resto de casos, la separación en subfrases se realiza en función de los broaders o las unidades lógicas y, en caso de que no las haya, las subfrases serán todos los nombres. Por tanto, serán conceptos más generales que el buscado, como los vistos en la figura 4.20.

Resultados totales de alineamientos UnionMatch

Como puede verse en la tabla 6.15, como resultado del proceso de validación y desambiguación, un concepto Emtree puede quedar también equiparado a más de un concepto UMLS, que aunque son similares aportan información ligeramente diferente sobre el concepto. Estos casos también son alineamientos UnionMatch ya que el significado del concepto Emtree parece repartido en varios conceptos UMLS. Por esta razón, la tabla 6.20 muestra el total de alineamientos complejos UnionMatch con respecto al número total de alineamientos validados y reúne tanto los mostrados en el apartado anterior (Tipo 1) como los obtenidos de esa forma (Tipo 2).

Como puede observarse, la facet con mayor número de alineamientos complejos UnionMatch es *Anatomical concepts* con 8 alineamientos del tipo 1 y 181 del tipo 2, que suponen el 7.85 % del total de alineamientos para esta facet. Le sigue *Physical diseases, disorders and abnormalities* con 23 alineamientos del tipo 1 y 283 del tipo 2, que suponen el 4.75 % del total. Las otras facets tienen menos de un 5 % de alineamientos complejos, siendo *Geographic names* y *Named groups of persons* las de menor porcentaje con 1.14 % y

Facets Emtree	Tipo 1		Tipo 2	
	No	%	**No**	%
Anatomical concepts	8	0.35 %	181	7.85 %
Organism names	5	0.19 %	38	1.47 %
Physical diseases, disorders and abnormalities	23	0.36 %	283	4.39 %
Chemicals and drugs	214	0.11 %	682	2.89 %
Geographic names	2	0.7 %	2	0.7 %
Named groups of persons	1	0.58 %	0	0.0 %
Total	253	0.18 %	1186	3.36 %

Tabla 6.20: Alineamientos complejos UnionMatch

Facets Emtree	ExactMatch		UnionMatch		BroadMatch	
	No	%	**No**	%	**No**	%
Anatomical concepts	1048	68.45 %	189	12.34 %	294	19.2 %
Organism names	783	89.18 %	43	4.9 %	52	5.92 %
Physical diseases, disorders and abnormalities	3707	81.12 %	306	6.7 %	557	12.19 %
Chemicals and drugs	10314	73.72 %	3651	26.1 %	26	0.19 %
Geographic names	174	95.6 %	4	2.2 %	4	2.2 %
Named groups of persons	38	70.37 %	1	1.85 %	15	27.78 %
Total	16064	75.75 %	4194	19.78 %	948	4.47 %

Tabla 6.21: Conceptos Emtree con alineamientos ExactMatch, UnionMatch y BroadMatch totales

0.58 %.

6.5.5. Resultados totales

Los resultados mostrados en las tablas 6.14, 6.15 y 6.18 se resumen en la tabla 6.21, donde se muestra el porcentaje de conceptos Emtree con alineamientos de tipo ExactMatch, UnionMatch y BroadMatch.

Como puede observarse, el porcentaje de ExactMatch es muy alto en todas las facets, mayor del 65 %, siendo *Geographic names* la facet con mayor porcentaje, con un 95.6 % y *Anatomical concepts* con un 68.45 %. Estos resultados coinciden con lo visto hasta ahora. Por un lado, los términos de tipo geográfico, como países, regiones, ... tienen mayor probabilidad de encontrar un solo concepto equiparado en la terminología fuente. Por otro, *Anatomical concepts* es la facet donde hemos encontrado más casos totales de alineamientos complejos, el 31 %, procedentes de diferentes fuentes de UMLS, ya que muchos de los vocabularios integrados eran de Anatomía.

Con respecto a los alineamientos complejos, hay mayor porcentaje total

Facets Emtree	**Conceptos Emtree**		**Alineamientos**	
	No	**Cobertura(%)**	**No**	%
Anatomical concepts	1531	57.77 %	4029	37.63 %
Organism names	878	33.26 %	1750	29.43 %
Physical diseases, disorders and abnormalities	4570	64.43 %	8911	39.81 %
Chemicals and drugs	13991	59.3 %	47811	62.63 %
Geographic names	182	61.9 %	804	50.34 %
Named groups of persons	54	28.72 %	340	30.47 %
Total	21206	58.16 %	63645	53.89 %

Tabla 6.22: Conceptos Emtree con alineamientos válidos

de UnionMatch, 19.78 %, frente a un 4.47 % de BroadMatch, lo cual es muy positivo, ya que indica que sólo para el 4.47 % de los conceptos Emtree no hemos podido encontrar correspondencias directas y la información que se aporta es más general.

Chemicals and Drugs es la facet con el porcentaje de alineamientos UnionMatch más alto, un 26.1 % y la que menos porcentaje tiene de alineamientos BroadMatch, un 0.19 %. Esto indica que para la mayor parte de los conceptos con alineamientos complejos, los conceptos UMLS obtenidos reúnen el significado el concepto Emtree. *Named groups of persons* es la facet con mayor porcentaje de alineamientos tipo BroadMatch, un 27.78 %, seguido de Anatomical concepts con un 19.2 %.

Por último, en la tabla 6.22, indicamos la cobertura de nuestro método, dando el número de conceptos Emtree con alineamientos válidos y el porcentaje de alineamientos válidos con respecto al número inicial de alineamientos obtenidos de UMLS.

Nuestro método encuentra alineamientos válidos para el 58.16 % de los conceptos Emtree a equiparar, lo que supone el 53.89 % de los alineamientos obtenidos inicialmente. Este porcentaje puede no parecer alto pero indica que UMLS devolvió inicialmente muchos conceptos UMLS que tenían sólo similitud léxica con los buscados.

La facet con mayor porcentaje de conceptos con alineamientos válidos es *Physical diseases, disorders and abnormalities* con un 64.43 %. Tienen también un porcentaje alto, superior al 50 %, *Anatomical concepts* y *Chemical and Drugs*. Dado que la información usada en la validación y desambiguación es semántica, esto parece indicar una mayor similitud entre la facet y el grupo elegido como compatible. La facet de menor porcentaje es *Named groups of persons* con un 28.72 %.

El mayor porcentaje de alineamientos válidos se encuentra, en cambio,

Facets Emtree	Conceptos Analizados	Alineamientos Obtenidos	Alineamientos Correctos	Precisión
Anatomical concepts	151	165	165	100.0 %
Organism names	825	214	210	98.13 %
Physical diseases, disorders and abnormalities	743	502	494	98.41 %
Chemicals and drugs	623	393	385	97.96 %
Geographic names	56	15	15	100.0 %
Named groups of persons	97	11	11	100.0 %
Total	2592	1300	1280	98.46 %

Tabla 6.23: Precisión del método

en la facet *Chemical and Drugs* con un 62.63 %, lo cual puede tener que ver con la cantidad de alineamientos que fueron recuperados, donde *Chemical and Drugs* era la facet con mayor número. La facet con menor porcentaje es *Organism names* con un 29.43 %, lo cual puede indicar que un porcentaje alto de los conceptos de la facet estaban en otros grupos semánticos.

Esto lo analizamos con detalle en el próximo apartado, el análisis cualitativo de los resultados.

6.5.6. Precisión del método

Para analizar la precisión del método, tenemos en cuenta en qué fase se han obtenido los alineamientos validados, ya que eso permite valorar su idoneidad. Además, hay que tener en cuenta que analizar manualmente los alineamientos obtenidos para los más de 46000 conceptos de la terminología Emtree sería una tarea titánica. Por tanto, hemos decidido elegir aleatoriamente una jerarquía de conceptos en cada una de las facets y revisar manualmente los alineamientos que nuestro método señala como válidos a fin de obtener el porcentaje de ellos que son realmente válidos.

En primer lugar, realizamos el análisis de la precisión de los alineamientos resultantes tras la validación y desambiguación, incluida la fase de expansión de vecinos. Los resultados obtenidos se resumen en la tabla 6.23, donde se indica el número de conceptos Emtree analizados para cada facet, el número de alineamientos considerados válidos por el método, el número de ellos realmente válidos y, por último, el porcentaje que suponen, es decir, la precisión.

Como puede observarse, la precisión total del método es muy alta, cercana al 100 %. La facet con un poco menos de precisión es *Chemicals and drugs*, con un 97.96 %. Las facets *Anatomical concepts*, *Geographic names* y *Named groups of persons* tienen un 100 % de precisión, es decir, todos los alineamientos validados por nuestro método son efectivamente válidos. Esto nos indica la gran utilidad e importancia que tiene la información estructural

Facets Emtree	Conceptos Analizados	Alineamientos Obtenidos	Alineamientos Correctos	Precisión
Anatomical concepts	151	41	41	100.0 %
Organism names	825	11	11	100.0 %
Physical diseases, disorders and abnormalities	743	139	100	71.94 %
Chemicals and drugs	623	3	2	66.67 %
Geographic names	56	3	3	100.0 %
Named groups of persons	97	4	4	100.0 %
Total	2592	201	161	80.1 %

Tabla 6.24: Precisión para mappings complejos UnionMatch

en el proceso de validación y desambiguación.

A continuación, analizamos la precisión para los alineamientos validados por nuestro método para aquellos conceptos Emtree sin equiparación léxica directa con UMLS. En este caso, la búsqueda de las subfrases del concepto proporciona gran cantidad de alineamientos. La validación por grupo semántico permite eliminar aquellos de temática diferente al concepto original, pero aún así el número de alineamientos considerados válidos es alto. De ellos, los alineamientos complejos tipo UnionMatch, donde el concepto original estaba formado por las subfrases vinculadas por una conjunción copulativa, muestran una precisión un poco menor a la del método. Como el número de conceptos en Emtree de este tipo no es muy alto, hemos analizado todos los alineamientos obtenidos y los resultados pueden verse en la tabla 6.24.

El 80.1 % de los alineamientos dados como válidos por nuestro método son efectivamente correctos. *Chemicals and drugs* y *Physical diseases, disorders and abnormalities* son las facets con menor precisión con 66.67 % y 71.94 % respectivamente. Esto se puede justificar porque algunas de las subfrases obtenidas son palabras de significado muy general que obtienen gran cantidad de resultados sin relación con el concepto original. Para el resto de facets, *Anatomical concepts*, *Organism names*, *Geographic names* y *Named groups of persons*, todos los alineamientos son correctos.

Los alineamientos complejos tipo BroadMatch, procedentes de los conceptos Emtree que no obtuvieron equiparación léxica, son, por su naturaleza intrínseca, similares sólo parcialmente, ya que contienen la información relacionada a una subfrase del concepto original que puede ser un broader, una unidad léxica o simplemente una palabra. Por tanto, los alineamientos obtenidos en muchos casos no obtienen información relacionada, por ejemplo, si la subfrase es un concepto de significado general.

Así, hemos considerado correctos aquellos alineamientos donde el concepto UMLS está relacionado, normalmente es más general, al concepto Emtree inicial. El alineamiento complejo BroadMatch da, para un concepto Emtree,

Facets Emtree	Conceptos Analizados	Alineamientos Obtenidos	Alineamientos Correctos	Precisión
Anatomical concepts	151	2589	2389	92.28 %
Organism names	825	885	804	90.85 %
Physical diseases, disorders and abnormalities	743	1733	1559	89.96 %
Chemicals and drugs	623	35979	32198	89.49 %
Geographic names	56	621	617	99.36 %
Named groups of persons	97	298	295	98.99 %
Total	2592	42105	37862	89.92 %

Tabla 6.25: Precisión para mappings complejos BroadMatch

Facets Emtree	Alineamientos correctos	Broader		Unidad léxica		Palabras	
		No	%	No	%	No	%
Anatomical concepts	2389	740	30.98 %	83	3.47 %	1566	65.55 %
Organism names	804	600	74.63 %	0	0.0 %	204	25.37 %
Physical diseases, disorders and abnormalities	1559	331	21.23 %	18	1.15 %	1210	77.61 %
Chemicals and drugs	32198	8	0.02 %	0	0.0 %	32190	99.98 %
Geographic names	617	30	4.86 %	0	0.0 %	587	95.14 %
Named groups of persons	295	5	1.69 %	0	0.0 %	290	98.31 %
Total	37862	1714	4.53 %	101	0.27 %	36047	95.21 %

Tabla 6.26: Estudio de resultados BroadMatch

un conjunto de alineamientos posibles. Los resultados para las mismas jerarquías que los resultados de la tabla 6.23 se muestran en la tabla 6.25.

Como puede observarse, la precisión total de esta parte del método es también alta, un 89.92 %. Todas superan el 85 %, siendo la de mayor precisión *Named groups of persons*, donde casi la totalidad de los conceptos Emtree encuentran alineamientos válidos.

Por último, analizamos estos alineamientos correctos de tipo BroadMatch en función del método en que fue conseguida la división del concepto original en sus subfrases. Como vimos, había 3 formas, la primera era coincidencia de un broader, la segunda era coincidencia de una unidad léxica, y, por último, si no había ninguna de ellas, la separación en sus palabras. En la tabla 6.26, se muestran los porcentajes de cada tipo de subdivisión.

Podemos observar que el caso más habitual es, sin duda, el de la división directa en las palabras, al no encontrar ni broaders ni unidades léxicas, con un 95.21 %. El porcentaje de ocasiones en que se usan las unidades léxicas es muy bajo, 0.27 %, y el restante 4.53 % de casos se usan broaders.

En 3 de las facets, el porcentaje de ocurrencia de las diferentes situaciones es similar, con más de un 90 % de la división en palabras, un 0 % de uso de unidades léxicas y un porcentaje entre el 0.02 % y el 4.86 % de ocurrencia de broaders. Esto sucede en *Chemicals and drugs*, *Geographic names* y *Named*

groups of persons.

Anatomical concepts es la facet con mayor porcentaje de unidades léxicas, un 3.47 %, y también tiene un 30.98 % de uso de broaders. Por último, *Organism names* presenta la ocurrencia más alta de broaders con un 74.63 %, mientras que no tiene ningún caso de unidad léxica.

6.5.7. Recall del método

El recall permite evaluar si el método devuelve todos los alineamientos correctos a la terminología destino. Hay que tener en cuenta que nuestro método no sólo realiza la validación de los alineamientos sino también su desambiguación, eligiendo el más similar estructuralmente entre las terminologías entre todos los alineamientos validados. Por tanto, esto baja el recall, al bajar el número de alineamientos considerados correctos por el método.

También hay que tener en cuenta que, al realizar la consulta a UMLS con el método NormalizeString, nuestro método ya trabaja inicialmente sólo con los conceptos UMLS que superan sus condiciones. Por lo que es posible que otros conceptos correctos ya no entren siquiera en nuestro método. Sin embargo, usando otros métodos de búsqueda, como Word o Normalize Word se obtendrían muchos más alineamientos, a veces más de 100 y en ocasiones más de 1000 para un único concepto Emtree, por lo que realizar la evaluación manual de todos ellos resultaría inviable.

Por tanto, decidimos usar como estándar una de las terminologías incluidas en el Metathesaurus de UMLS, que es MeSH [3]. Las razones de esta elección son dos. Por un lado, MESH es el tesauro del NLM usado para indexar artículos en PubMed, que es el motor de búsqueda en MEDLINE, como ya vimos en el capítulo 2. Como ya sabemos, Emtree es usado por Elsevier para el mismo objetivo. Por otro lado, MeSH organiza los conceptos en 16 categorías de alto nivel, de las cuales muchas coinciden con las de Emtree, como por ejemplo, *Anatomy*, *Organisms*, *Diseases*, *Chemical and Drugs*, *Analytical, Diagnostic and Therapeutic Techniques and Equipment*, *Psychiatry and Psychology*, *Named Groups*, *Publication Characteristics* o *Geographic Locations*.

Hemos considerado válido un alineamiento donde al menos uno de los conceptos UMLS devueltos pertenezca a MeSH. En la tabla 6.27, mostramos los resultados para un subconjunto de casi 400 conceptos elegidos aleatoriamente entre aquellos que obtienen alineamientos mediante el proceso descrito.

Como puede verse, el 44.19 % de los alineamientos dados como correctos por nuestro método pertenecen a MESH. Las facets con mayor recall son

[3] http://www.ncbi.nlm.nih.gov/mesh

Facets Emtree	Conceptos Analizados	Alineamientos Obtenidos	Alineamientos Correctos	Recall
Anatomical concepts	40	54	20	37.04 %
Organism names	23	24	14	58.33 %
Physical diseases, disorders and abnormalities	88	185	39	21.08 %
Chemicals and drugs	234	251	154	61.35 %
Geographic names	7	7	4	57.14 %
Named groups of persons	3	4	1	25.0 %
Total	395	525	232	44.19 %

Tabla 6.27: Recall del método

Chemicals and Drugs, *Organism names* y *Geographic names* con un 62.35 %, un 58.33 % y 57.14 %, respectivamente. Las facets con menos recall son *Physical diseases, disorders and abnormalities* con un 21.08 %, *Named groups of persons* con un 25 % y *Anatomical concepts* con sólo el 37.04 % de los alineamientos dados por correctos en MeSH.

6.6. Análisis cualitativo de los resultados

A continuación, realizamos un análisis de los datos mostrados en el apartado anterior, a fin de valorar la efectividad de nuestro método.

6.6.1. Alineamiento léxico

El alineamiento léxico proporciona aquellos conceptos de la terminología destino que son similares léxicamente a los conceptos de la terminología fuente. Ello implica similitud en las palabras usadas, pero puede no coincidir realmente el significado. A continuación, valoramos estos resultados.

Alineamiento léxico de términos

El servicio web NormalizeString de UMLS, usado para obtener los conceptos UMLS similares a los conceptos Emtree, sólo puede realizar una alineación léxica, ya que se basa en las propiedades léxicas de los términos de las terminologías a ser alineadas. Aún así, el porcentaje de los alineamientos encontrados por término puede ser muy alto. Tal como se muestra en la tabla 6.1, en el alineamiento léxico directo, se alcanza una cobertura de términos Emtree del 47.9 % y de conceptos Emtree de un 80 %, a pesar de que las terminologías fueron desarrolladas independientemente para diferentes propósitos. La razón principal de este alto grado de cobertura es la rica colección de sinónimos en el Metathesaurus. Nuestros resultados sobre la cobertura de

la alineación léxica son compatibles con las dos características propuestas por Burgun [Bur06] para determinar si una terminología (en nuestro caso, el Metathesaurus) sirve como una referencia en la alineación entre terminologías en biomedicina: la amplia cobertura léxica de los conceptos y la inclusión de muchos sinónimos.

Alineamiento léxico de conceptos

En la tabla 6.1, también se incluyen los resultados de los alineamientos obtenidos para los conceptos Emtree, de forma que el 80 % de ellos obtienen alineamientos, un 43.6 % de los cuales son simples y un 36.4 % son ambiguos. Este porcentaje es bastante alto y hemos realizado un análisis de las causas que producen los alineamientos ambiguos y se han encontrado varias:

1. El *amplio uso de los sinónimos* en una terminología, es decir, varios conceptos con significados diferentes pueden contener el mismo sinónimo simultáneamente. Un ejemplo es el término *bacteria*, del concepto *bacterium*, que es sinónimo con 4 conceptos del Metathesaurus (Figura 4.8), que describe: un ser vivo *Bacteria* (*Living Beings*), un procedimiento *Bacterial Count Measurement* (*Procedure*) y dos conceptos funcionales *bacteria aspects* y *Percent of bacteria* (*Concepts and Ideas*). Sin embargo, el término Emtree *bacteria* sólo se utiliza para designar a un organismo.

2. Las *diferencias en la granularidad entre terminologías*. Un ejemplo es el concepto *body regions*, que es descrito en mayor detalle en UMLS que en Emtree: UMLS distingue entre tres conceptos (*Body part*, *Body Regions* y *Entire Body Region*), mientras que Emtree sólo considera uno (*body regions*).

3. *Uso de un término para designar tanto a un concepto y una categoría*. Por ejemplo, el término *fungus* es el PT de un concepto Metathesaurus y el nombre de un tipo semántico, mientras que en Emtree sólo representa un concepto.

4. *Presencia de términos de propósito muy general* en una terminología. Por ejemplo, el término Emtree *axis* se utiliza para designar un concepto de la anatomía. Sin embargo, esta es una cadena de uso general y no designa un concepto con precisión. Por el contrario, el Metathesaurus utiliza cadenas más largas, incluyendo nombres o adjetivos, para identificar el significado más claramente: *Axis vertebra*, *Electrocardiographic axis*, *Genus Axis* y *Entire axis vertebra*.

5. *El uso de algunas categorías de nivel superior como "cajón de sastre"*, que incluyen una gran variedad de conceptos con relaciones difícilmente jerárquicas entre ellos. Un ejemplo es el Grupo Semántico de UMLS *Concepts & Ideas* que contiene una gran cantidad de alineamientos ambiguos.

6.6.2. Validación del alineamiento léxico

Cada concepto Emtree queda equiparado léxicamente a 1 o varios conceptos UMLS, que pueden o no ser similares semánticamente a él. Por tanto, es necesario validar esos alineamientos, para lo cual se usa la información estructural presente en las terminologías, las categorías y las relaciones jerárquicas.

Compatibilidad de categorías de alto nivel

Como se vió en el apartado 4.4, tras analizar los grupos semánticos más comunes en los alineamientos de cada facet Emtree, se determina cúal es el porcentaje mínimo necesario para considerar compatible una facet Emtree y un grupo semántico UMLS y, a continuación, se fijan estas compatibilidades. Ese porcentaje se fija en el 60 % y en la tabla 6.5, se observa que son 6 las facets que superan este porcentaje en el primer grupo semántico predominante.

Aunque el número de categorías encontradas compatibles es pequeño (6 de 15 categorías Emtree de nivel superior), esto corresponde a un número considerable de términos Emtree: el 75,8 % de los alineamientos léxicos totales encontrados por NormalizeString y el 65,3 % del total de conceptos Emtree. Esto es debido a que algunas de esas facets o grupos semánticos son los de mayor número de términos, por ejemplo, *Chemical and Drugs* y *Physical diseases, disorders y abnormalities*.

Esperábamos que nuestro método permitiría detectar más alineamientos inválidos ambiguos que simples. Éste es el caso para los términos, pero no para los conceptos. Para detectar un alineamiento inválido entre conceptos, todos los alineamientos de los términos que lo forman debe ser identificados como inválidos. Por lo tanto, es más probable detectar un alineamiento inválido para términos, ya que sólo un alineamiento de término es necesario para llevar a él, mientras que los conceptos ambiguos requieren la identificación de todos los alineamientos de término correspondientes como inválidos.

Hemos realizado un análisis manual de los errores en la precisión, que se muestran en la tabla 6.7. Evaluando el conjunto de alineamientos inválidos

identificados por nuestro método, hemos detectado varias situaciones que disminuyen la precisión de nuestro método.

- La *similitud estructural en UMLS*. En algunos alineamientos, conceptos Emtree y UMLS que pertenecen a categorías de alto nivel incompatibles tienen proximidad semántica, debido a la similitud estructural en UMLS. Se distinguen tres casos:

    1. El concepto UMLS no pertenece a un GS compatible, sino que pertenece a un Tipo Semántico (TS) con proximidad semántica al GS compatible. En otras palabras, el TS está en la misma parte de la jerarquía de la red semántica que muchos TSs del GS compatible. Por ejemplo, el GS *Genes and Molecular Sequences* incluye un TS, *Gene o genoma*, que pertenece a la misma parte de la red semántica que muchos TSs del GS *Anatomy* (Figura 6.6). Por lo tanto, los conceptos en *Gene o genoma* están semánticamente relacionados con el GS *Anatomy* (hay una relación "es-un" entre *Gene o genoma* y *Fully Formed Anatomical Structure*). Nuestro método ha detectado erróneamente como inválidos todos los alineamientos que incluyen pares de conceptos Emtree-UMLS pertenecientes a *Anatomical concepts* <–> *Genes and Molecular Sequence*. Sin embargo, esto no es una regla general. Hay otros conceptos de UMLS que pertenecen a GSs incompatibles, pero tienen proximidad semántica a través de la jerarquía de la red semántica y que nuestro método detectó como inconsistente correctamente.

    2. El concepto UMLS pertenece a un TS que está relacionado, a través de una relación no jerárquica, a muchos TSs en el GS compatible. Por ejemplo, el GS *Physiology* incluye el TS *Organism attribute* que está relacionado con muchos TSs de *Living Beings* (*Plant*, *Fungus*, *Virus*, etc.) a través de la relación 'propiedad de'. Una vez más, nuestro método erróneamente detectó como inválido todos los alineamientos que incluyen pares de conceptos Emtree-UMLS pertenecientes *Organisms* <–> *Physiology*. De nuevo, esta no es una regla general.

    3. El concepto UMLS está relacionado, a través de una relación de 'broader' o de 'narrower', a otro concepto perteneciente al GS compatible. Por ejemplo, en el alineamiento *body surface* <–> *Body surface*, el concepto UMLS *Body surface*, perteneciente al GS *Physiology*, es una especialización de *Anatomical surface*, el cual es una especialización de *Non-material physical anatomical entity*,

que pertenece al GS *Anatomy* (Figura 6.6). Por lo tanto, este alineamiento fue erróneamente detectado como inválido con nuestro método. Hemos verificado que esta situación es una regla general, lo que sugiere que las relaciones del Metathesaurus (broader y narrower) son semánticamente más fuertes que las de la Red Semántica.

- *Categorías de alto nivel de tipo 'cajón desastre' en UMLS.* UMLS incluye dos categorías de alto nivel, *Concepts and Ideas* y *Objects*, cuyos conceptos que participan en los alineamientos son difíciles de evaluar correctamente debido a las razones siguientes:

  1. Los conceptos son muy abstractos y muy diferentes unos de otros.
  2. Los conceptos no tienen relaciones jerárquicas entre ellas.
  3. Los conceptos son muy similares a los otros clasificados en GSs diferentes, pero no tienen relaciones con ellos.

- *Categorías de nivel superior organizadas como un cluster.* En Emtree, *Physical diseases, disorders and abnormalities* incluye no sólo los conceptos relacionados con los trastornos (disorders), sino también los conceptos más generales relacionados con la vida, la salud o el estado de la enfermedad. Estos conceptos se clasifican en UMLS como Tipos Semánticos *Activities and Behaviour*, *Anatomy*, *Phenomena* o *Physiology*. Por lo tanto, los conceptos de estas categorías son difíciles de evaluar correctamente.

- *Simples diferencias en la clasificación.* Las excepciones restantes corresponden a diferencias en la clasificación en ambas terminologías. Esto ocurre principalmente con la facet *Organism*, que incluye muchos crustáceos y otros seres vivos, a los que UMLS clasifica como *Food (Objects)*, y microorganismos que producen enfermedades, que UMLS clasifica como *Chemicals and drugs*.

Como resultado de la etapa anterior, hemos clasificado las equiparaciones obtenidas inicialmente entre conceptos Emtree y UMLS en compatibles e incompatibles en función del grupo semántico al que pertenecen los conceptos UMLS obtenidos. Las equiparaciones incompatibles se descartan ya que se considera que se han obtenido por similitud léxica pero que el concepto UMLS tiene otros significados. Estos son algunos ejemplos de equiparaciones incompatibles:

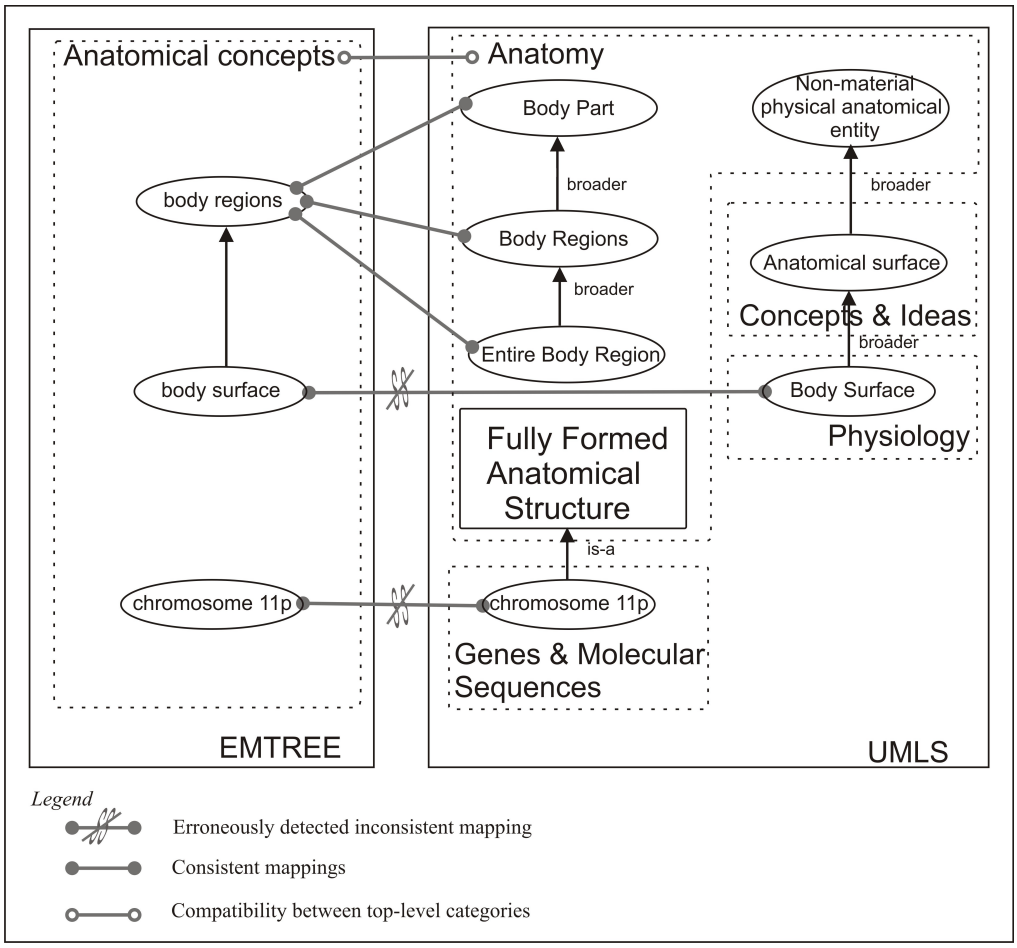

Figura 6.6: Ejemplo de alineamientos consistentes e inconsistentes entre Emtree y UMLS

- el término Emtree *spine* de la facet *Anatomical concepts* se equiparaba léxicamente, entre otros, a los conceptos UMLS *Spine problem* y *Malignant neoplasm of vertebral column* que pertenecen al grupo semántico *Disorders*. Semánticamente, se aprecia que mientras el término Emtree hace referencia a una parte del cuerpo humano, los conceptos UMLS se refieren a una enfermedad, por tanto, aunque están relacionados, no son equiparables léxicamente.

- el término Emtree *head* de la facet *Anatomical concepts* se equiparaba léxicamente al concepto UMLS *Procedures on Head*, del grupo semántico *Procedures* que hace referencia a procedimientos sobre la cabeza, pero no información anatómica sobre ella.

- el término Emtree *bacterium* de la facet *Organism names* se equiparaba léxicamente al concepto UMLS *bacteria aspects*, del grupo semántico *Concepts and Ideas*, del tipo semántico *Functional concept* lo cual sugiere que puede hacer referencia a aspectos funcionales pero no a una descripción de la bacteria en sí misma.

- el término Emtree *Legionella longbeachae* de la facet *Organism names* se equiparaba léxicamente al concepto UMLS *Legionella longbeachae antigen*, del grupo semántico *Chemicals and Drugs*, es decir, UMLS ha equiparado léxicamente una droga o medicamento con el término que es un organismo vivo.

Desambiguación basada en similitud directa

Como ya hemos visto, para realizar la desambiguación, es decir, la selección del mejor alineamiento, usamos la información de los clusters (agrupaciones de los conceptos UMLS que están relacionados directamente) y los factores de similitud, que comparan las relaciones del concepto de Emtree con sus vecinos con las de los conceptos UMLS.

En la tabla 6.8, se muestran los resultados de los clusters obtenidos por facet y en total. El porcentaje es bastante alto, un 32.6 % de los conceptos tienen uno o más clusters entre sus conceptos UMLS, es decir, obtienen conceptos relacionados directamente entre sí y, por tanto, muy similares. Esto es debido a que UMLS es la unión de muchos terminologías por lo que muchos conceptos se repiten. La facet con mayor porcentaje es *Physical diseases, disorders and abnormalities* con un 42.6 % y la de menor *Geographic names* es con un 6.9 %.

A su vez, en la tabla 6.9, se muestran los porcentajes de alineamientos sin similitud y con similitud baja, alta o muy alta, obtenidos por facet y en total. El porcentaje de alineamientos con alguna similitud no es alto, un 36.7 %, donde un 15.5 % tienen similitud muy alta, un 13 % alta y 8.2 % baja. Por tanto, nuestro método de cálculo del factor de similitud no encuentra similitud directa en el 63.4 % de los alineamientos. La facet con mayor porcentaje de alineamientos sin similitud es *Named group of persons* con un 83.8 % y las de menor *Physical diseases, disorders and abnormalities* y *Geographic names* es con un 51.5 %. De esto se puede deducir que la similitud de la estructuración directa de conceptos en Emtree y UMLS no es alta, aunque esto puede deberse a la mayor granularidad de UMLS.

Tras analizar los resultados, vemos que estos resultados se deben a varios factores:

- Términos aparentemente similares no tienen relaciones entre sí por pro-

venir de diferentes fuentes de UMLS. Un ejemplo se ve en la figura 4.9 para el concepto *spine*, donde los conceptos *Vertebral column* y *Entire Vertebral column*, aún pareciendo similares no tienen relación directa.

- La existencia de relaciones indirectas entre los conceptos UMLS, debidos a la mayor granularidad de UMLS. Un ejemplo de esto se vió en la figura 4.9, para el concepto Emtree *back*, donde *Entire back (surface region)* no está relacionado directamente con el concepto *Back*, por lo que no forman cluster, pero tiene como broader *Entire trunk* que a su vez tiene como broader *Back*. Por, hay un relación indirecta.

- La diferente estructuración de los conceptos en la jerarquía en ambas terminologías. Esto se aprecia claramente para la facet *Named groups of persons* y su grupo semántico compatible *Living Beings*. El porcentaje de coincidencia entre ellos (tabla 6.5) era alto, de 63.4 %, pero no se encuentra ninguna similitud estructural en el 83.8 % de los mappings. En menor medida, en la facet *Organism names*, compatible con el mismo grupo semántico, el porcentaje de alineamientos sin similitud es del 74.2 %. Esto indica que UMLS estructura los conceptos referentes a formas de vida de forma diferente a Emtree.

Como vemos, estos resultados nos aportan cierta información sobre las diferencias entre cada facet y el grupo semántico compatible. Así, la facet *Named groups of persons* tiene un porcentaje alto de alineamientos sin similitud, 83.8 % y el porcentaje de clusters es bajo, 18.7 %, lo cual indica que la facet y el grupo semántico *Living Beings* tiene una estructura muy diferente. Lo contrario que lo que sucede entre la facet *Disorders* y el grupo semántico *Physical diseases, disorders and abnormalities* y la facet *Anatomical concepts* y el grupo *Anatomy* en las que hay un porcentaje alto, del 30 y 40 % conceptos con cluster y cercano al 50 % de alineamientos con similitud. Por tanto, algunas de las partes en ambos están estructuradas de forma similar. Un caso ligeramente diferente es *Geographic names* que tiene un porcentaje alto de alineamientos con similitud, 48.5 %, 37.5 % de ellos muy alta, pero un porcentaje bajo de clusters, 6.9 %. Esto es debido a que en esta área del dominio, lo más habitual es que haya un solo término para representar un país, región o zona geográfica.

En la tabla 6.10, se muestran los resultados del algoritmo que usa la información anterior para elegir, si es posible, un único alineamiento como el mejor de ellos. Se busca descartar aquellos que son prácticamente iguales a otros y a aquellos que no tienen ninguna similitud estructural. Así, muchos alineamientos que fueron consideramos compatibles se descartan. Este

es el objetivo de la parte final de nuestro método, desambigüar entre los alineamientos obtenidos y considerados compatibles para un mismo concepto Emtree.

El porcentaje de alineamientos considerados los mejores no es muy alto, 24.05 %, para el 41.22 % de los conceptos Emtree. Esto indica que solo ese porcentaje de alineamientos muestran una similitud estructural directa entre Emtree y UMLS. Realmente, entonces el resultado no está mal, ya que la estructuración, como vimos, depende de los objetivos y necesidades de los constructores de las terminologías.

A continuación, mostramos varios ejemplos de alineamientos que han sido descartados:

- Para el concepto Emtree *abdomen*, se descarta *Entire abdomen* ya que forma cluster con el concepto *Abdomen*, pero tiene un factor de similitud (41.5) menor que él (75).
- Para el concepto Emtree *mesentery*, se considera válido solamente *Mesentery*, con un factor de similitud de 50, mientras que *Connective Tissue - Mesentery (MMHCC)* y *Structure of mesentery of small intestine* son descartados al no tener ninguna similitud
- Para el concepto Emtree *greater omentum*, se descarta el único concepto al que fue equiparado léxicamente *Greater omentum*, ya que no hay similitud estructural con él.
- Para el concepto *phlebitis*, se descarta correctamente *Postphlebitic Syndrome* aunque forma cluster con *Phlebitis*, al tener factor de similitud 0 y se elige a *Phlebitis* que tiene 75.
- En la facet *Named groups of persons*, el concepto Emtree *Indian* se equipara al concepto UMLS *Indian ethnic group* pero se descarta ya que no es posible confirmar la similitud a través de las relaciones broader-narrower. Este caso se da en muchos otros conceptos.

Expansión de vecinos

Analizando los alineamientos que consideramos válidos o mejores por el algoritmo anterior, se vio que se descartaban alineamientos igualmente válidos. Eso se debe a que esa fase del método usa la información estructural directa, comparando las relaciones directas del concepto Emtree (broaders y narrowers) con las relaciones directas de los conceptos UMLS a los que se equipara. Sin embargo, como ya se vio, UMlS tiene mayor granularidad que Emtree, que se traduce en un mayor número de conceptos y un mayor nivel

de detalle en las jerarquías. Por tanto, en muchas ocasiones, conceptos relacionados directamente en Emtree lo estaban indirectamente en UMLS. Por eso, se ha realizado la expansión de vecinos para explorar esas relaciones. Con ellos, se recuperan conceptos con coincidencias en relaciones directas e indirectas.

En las tablas 6.11 y 6.12, se muestran los alineamientos recuperados por la expansión de vecinos, para broaders y narrowers respectivamente, para los niveles 1, 2 y 3. Como puede observarse, se recuperan muchos más conceptos por broaders que por narrowers, 22.95 % frente al 1.1 %. Esto es debido a la diferencia de tamaño entre las terminologías. UMLS es la fusión de más de 130 vocabularios y terminologías, por lo que la cantidad de broaders y narrowers que se encuentran para cada concepto es, en muchas ocasiones, muy alto. Emtree ha sido desarrollada para un objetivo preciso y por tanto, su tamaño es mucho menor y no llega al nivel de detalle de UMLS. Eso hace que el número de niveles de la jerarquía sea menor y por tanto es más frecuente encontrar coincidencias hacia los niveles superiores que hacia los inferiores.

Por niveles, se recuperan más en el primer nivel, el de las relaciones directas descartadas en el paso anterior. En el nivel 2, se recuperan algunos, un 3.98 % por broader y un 0.03 % por narrower. En el nivel 3, ya se recuperan muchos menos, solo un 0.98 % por broader mientras que por narrower se recuperan 2 conceptos más, con lo cual el porcentaje final no varía.

En la tabla 6.13, se pueden observar los resultados totales de los conceptos recuperados por expansión de vecinos, un 23.42 % más. Hay que tener en cuenta que un mapping a veces es recuperado tanto por broader como por narrower, de ahí que el total sea un poco inferior. De los ejemplos vistos en el apartado anterior, se recupera *Greater omentum* para *greater omentum* y *Postphlebitic Syndrome* para *phebitis*

Como puede observarse, el uso de las relaciones indirectas en UMLS, al ser de mayor granularidad, nos permite recuperar alineamientos descartados en un primer momento pero que igualmente se pueden considerar buenos, al haber podido comprobar que ocupan un lugar similar en la jerarquía en Emtree y en UMLS.

En la tabla 6.14, se reúnen los resultados validados por el algoritmo y los recuperados por expansión de vecinos, que consigue añadir un 29.6 % de alineamientos. De esta forma, el 31.26 % de los alineamientos que han sido validados por el grupo semántico, es decir, el concepto Emtree y el concepto UMLS pertenecen al mismo área de conocimiento, han sido localizados en lugares similares de la jerarquía, por tanto, son similares estructuralmente y se puede afirmar que son alineamientos válidos. Esto supone el 48.8 % de los conceptos, es decir, para casi la mitad de los conceptos Emtree podemos

proporcionar un alineamiento validado semánticamente. En algunos casos, también puede darse que el método detecta 2 conceptos son similares pero que también presentan diferencias, en cuyo caso, los 2 son validados y forman un alineamiento complejo, como veremos más adelante.

Por facet, 4 de ellas tienen están entre el 30 y el 40 % de alineamientos validados, *Physical diseases, disorders and abnormalities*, *Geographic names*, *Anatomical concepts* y *Chemicals and drugs*, siendo la primera la de mayor porcentaje con 39.82 %. La facet de menor porcentaje es *Named groups of persons* con un 9.34 %. Este porcentaje bajo indica que la facet y el grupo semántico compatible *Living Beings* no tienen una estructura jerárquica similar.

6.6.3. Análisis de los alineamientos derivados de las técnicas de procesado de lenguaje natural

Como ya comentamos en el capítulo 4, los conceptos Emtree que no han obtenido equiparación léxica directa de UMLS se particionan en sus subfrases constituyentes que serán de nuevo buscados en UMLS. Como resultado, se obtienen las equiparaciones léxicas para estas subfrases.

Alineamiento con UMLS

La mayor parte de los conceptos cuando se particionan en sus partes constituyentes obtienen alineamientos, casi el 100 %, la mayor parte de las cuales son ambiguos. Esto es debido a que siempre existe algún alineamiento para alguna parte constituyente de cada término.

Validación del alineamiento

Realizamos la validación considerando alineamientos válidos aquellos que pertenecen al grupo semántico compatible a la facet. Los resultados se ven en la tabla 6.18. El porcentaje de alineamientos válidos es alto, de un 77.87 %, en el 89.1 % de los conceptos. Es decir, sólo el 10.9 % de los conceptos quedan sin equiparaciones léxicas válidas.

En este caso, no realizamos la desambiguación ya que los conceptos UMLS no son obtenidos directamente y, por tanto. no pueden ser comparados sus broaders y narrowers con los de concepto original. Aún así, consideramos que la validación por grupo semántico permite descartar aquellos conceptos UMLS de significado completamente diferente al concepto Emtree al que se equiparan y que los conceptos que quedan aportan algún tipo de información relevante sobre él.

Así, cada concepto Emtree queda equiparado a varios conceptos UMLS que, al proceder de subfrases, contienen un parte del significado original.

A continuación, explicamos algunos ejemplos vistos en las figuras 4.18 y 4.20:

- *pia artery* queda equiparado a un único concepto UMLS *Arteries*, ya que ninguno de los conceptos que se equiparaba a *pia* supera la validación por grupo semántico.

- *heart muscle capillary* queda equiparado a 4 conceptos: *Myocardium* procedente de la subfrase *heart muscle* y *Blood capillaries*, *Capillary vessel* y *Entire capillary blood vessel(organ)* procedentes de *capillary*.

- En *embryonic, fetal and placental structures*, al buscar cada subfrase, se obtiene una equiparación léxica, luego validada semánticamente, para cada parte. Son *Embrionic Structures* para la subfrase *embrionic structures*, *Fetal Structures* para *fetal structures* y *Placenta* para *placental structures*.

6.6.4. Alineamiento complejo

Los alineamientos complejos BroadMatch suponen casi el 96 % del total, tal como se muestra en la tabla 6.19. Estos alineamientos aportan información relacionada al concepto Emtree de sus broaders, aunque la correspondencia no es exacta.

Los alineamientos complejos UnionMatch tienen, entre todos, una correspondencia exacta con el concepto Emtree, es decir, entre todos ellos, reúnen su significado. Así, para los enumeraciones, al buscar cada término por separado, los alineamientos obtenidos juntos componen el significado del concepto. Por otro lado, como resultado del proceso de validación y desambiguación, puede dar como resultado que conceptos Emtree queden equiparados a más de un concepto UMLS, ya que nuestro algoritmo determina que no son exactamente iguales sino que su significado se solapa, por tanto, todos se consideran válidos y necesarios. En la tabla 6.20, se resumen los porcentajes que hay de estos casos con respecto al número total de conceptos Emtree buscados.

A continuación, vemos cómo quedan algunos de los ejemplos que hemos visto a lo largo de esta memoria:

- En la figura 4.7, se vieron varios ejemplos de alineamientos simples. De ellos, el alineamiento entre *cornea vascularization* y *Vascularization of cornea* no se valida por grupo semántico ya que el primero es de

Anatomy y el segundo de *Disorders*. Los otros dos, *animal anatomy - animal structures* y *beak - Beak* aunque sí validan no pasan el proceso de desambiguación al no coincidir sus jerarquías.

- Para *body regions*, visto en la figura 4.8, el concepto *Body Regions* es el único que supera el algoritmo de desambiguación, al ser el miembro del cluster con mayor puntuación y donde el Índice Comparativo entre él y *Body Part* no era menor al 75 %. Es decir, *Body Regions* incluía toda la información.

- Para *bacterium*, tanto *Bacteria* como *Schizomycetes* se validan por grupo semántico, pero sólo el primero supera el algoritmo de desambiguación al tener un factor de similitud mayor que 0. *Schizomycetes* no lo supera por tener ese factor igual a 0 y tampoco es recuperado al revisar sus vecinos al no coincidir la estructura jerárquica.

- Para el concepto *back*, mostrado en la figura 4.9, de los 8 conceptos UMLS a los que queda validado léxicamente, 6 validan por grupo semántico (los descartados son de los GS *Disorders* y *Concepts and Ideas*). Entre ellos, hay dos clusters de 2 conceptos cada uno y 2 conceptos quedan libres. Los dos conceptos *Back structure, excluding neck* y *Back structure, including back of neck* se descartan por tener factor de similitud 0. *Back* supera el algoritmo de desambiguación al ser el concepto con mayor factor de similitud en los clusters. En la expansión de vecinos, se recupera *Entire back (surface region)* tras encontrar una relación indirecta, a través de *Entire trunk*, con el concepto *Back*. Por tanto, se trata de un mapping complejo UnionMatch.

- El concepto *spine* se equiparaba léxicamente con 7 conceptos de los cuales 3 son descartados por pertenecer al GS *Disorders*. Había un clúster entre *Entire spine* y *Entire vertebral column*. El primero supera el algoritmo al ser el de mayor factor de similitud y el segundo es recuperado en la expansión de vecinos. Los otros dos conceptos UMLS, *Vertebral column* y *Dorsal column*, se descartan por tener factor de similitud 0 y en la expansión de vecinos tampoco se recuperan.

- El concepto *greater omentum* se equiparaba léxicamente al concepto UMLS *Greater omentum* y al pertenecer a grupo semántico compatible, era válido pero no supera el algoritmo de desambiguación al tener factor de similitud 0. Sin embargo, al hacer la expansión de vecinos, si se encuentra una relación indirecta, y por tanto se considera válido, formando un alineamiento ExactMatch.

- El concepto *lumbar disk* se equipara léxicamente a un único concepto UMLS *Entire lumbar disc*, que es validado semánticamente al pertenecer al GS compatible. Sin embargo, el algoritmo de desambiguación lo descarta al tener factor de similitud 0, pero la expansión de vecinos lo recupera al encontrar una coincidencia en los broaders. Por tanto, es otro ejemplo de ExactMatch.

- Los ejemplos vistos en el apartado anterior, *pia artery* y *heart muscle capillary*, forman alineamientos BroadMatch. De esta forma, el primero queda equiparado a un único concepto que da información genérica de arterias, mientras que el segundo queda equiparado a conceptos sobre músculos del corazón y vasos sanguíneos. Otro caso visto es el de *mesentery blood vessel* que forma un alineamiento BroadMatch con *Mesentery*, *Blood vessels* y *Entire blood vessel*.

- El ejemplo *embryonic, fetal and placental structures* forma un alineamiento UnionMatch. Es el mismo caso que *hamters and gerbils*, que queda equiparado a los conceptos UMLS *Hamters* y *Gerbils*, donde cada concepto a los que quedan equiparados tiene una parte del significado global del concepto.

6.6.5. Resultados totales

Tal como se muestra en las tablas 6.21 y 6.22, la cobertura total de nuestro método es del 58.16 %, los cuales el 75.75 % de los conceptos Emtree obtienen alineamientos simples y el resto complejos, lo cual consideramos que es un buen porcentaje.

Tal como se ha visto en el proceso, para los alineamientos simples, podemos afirmar que el alineamiento aportado es el mejor de entre los disponibles. Para los alineamientos complejos, nuestro proceso recupera toda la información disponible que considera relevante. En el caso de los alineamientos BroadMatch, la correspondencia no es exacta pero consideramos que la información es relevante, ya que hace referencia a partes o broaders del concepto original.

Para casi el 42 % de los conceptos Emtree, nuestro método no ha podido elegir, entre las equiparaciones léxicas obtenidas inicialmente de UMLS, una o varias similares semánticamente a ellos. Las causas de esto son varias:

- Algunos conceptos Emtree solo obtuvieron equiparaciones léxicas con conceptos UMLS de grupos semánticos incompatibles y, por tanto, no se han podido validar.

- La revisión manual de los alineamientos descartados por el algoritmo de desambiguación y la expansión de vecinos nos permite ver que la mayoría de ellos lo fueron por no coincidir la información semántica en Emtree y en UMLS, que es lo que nos permite hacer la validación. Así, aunque el concepto Emtree y el UMLS pertenecen a facets y grupo semántico respectivamente compatibles, estos están estructurados de diferente manera por lo que la posición en las jerarquías no es similar.

6.6.6. Valoración de la precisión

Como vimos en las tablas 6.23, para aquellos conceptos que han obtenido equiparación léxica y han sido considerados válidos por nuestro método, la precisión es muy alta, la mayoría de los alineamientos que el método da como correctos efectivamente son correctos. La razón principal es el uso de toda la información estructural disponible, lo que nos permite descartar aquellos alineamientos con conceptos UMLS de otras temáticas y elegir entre los restantes los que ocupan un lugar en la jerarquía similar al concepto Emtree.

De las jerarquías estudiadas, pocos son los alineamientos que al revisar manualmente no consideramos correctos. Hay que tener en cuenta que, al carecer de un experto en el proceso de revisión, alguna de las observaciones pueden no ser correctas. Algunos ejemplos de alineamientos considerados incorrectos en la revisión son:

- El concepto Emtree *Gram negative anaerobic bacteria* se equipara a *Sulfur-Reducing Bacteria*. No hemos encontrado información que lo confirme.

- El alineamiento del concepto Emtree *Yersinia pseudotuberculosis* con el concepto UMLS *Pasteurella pseudotuberculosis* lo consideramos incorrecto ya que parecen tipos diferentes de bacterias y no hemos encontrado información que indique lo contrario.

- El concepto Emtree *physical disease by body function* es una categoría dentro de la facet *Physical diseases, disorders and abnormalities* queda equiparado a un concepto UMLS también general *Signs and Symptoms*, pero que no está relacionado.

- En los casos de los alineamientos entre *hyperammonemia* y *Ornithine carbamoyltransferase deficiency* y entre *3,3,14,14 tetramethylhexadecanedioic acid* y *MEDICA 16* no disponemos de información suficiente para confirmar su similitud por lo que los consideramos incorrectos.

Para los conceptos que no obtuvieron equiparación léxica directa con UMLS y que dieron lugar a los alineamientos complejos UnionMatch, la precisión también es alta, como puede verse en la tabla 6.24. Esto indica, en estos casos, que la validación por grupo semántico consigue descartar bien los alineamientos incorrectos, aunque tendremos probablemente más de un concepto UMLS por cada subfrase.

La casi totalidad de alineamientos considerados incorrectos en la validación manual son de la facet *Physical diseases, disorders and abnormalities*, en conceptos Emtree generales de nombre largo que recuperan muchos conceptos UMLS, como por ejemplo, *growth, development and aging disorders* y *physical disease by composition of body fluids, excreta and secretions* que obtienen algunas enfermedades no relacionadas. En *Chemical and drugs*, solo consideramos incorrecto un alineamiento, el existente entre *agents acting on the auditory and vestibular systems* y *Agent*, ya que recupera un concepto de significado general y no incluirá información relevante al concepto Emtree. En el resto de facets, todos los alineamientos revisados son correctos.

Para los restantes, que son alineamientos complejos BroadMatch, la precisión también es alta, como se ve en la tabla 6.25, ya que las subfrases suelen obtener resultados más generales que el concepto Emtree original pero aún así son relevantes. A continuación, vemos algunos ejemplos de alineamientos:

- El concepto *leg blood vessel* queda equiparado a varios conceptos relacionados con *leg* (*Lower extermity*, *Leg*, *Entire lower limb* y *Entire lower leg*) y varios con *blood vessel* (*Blood vessels* y *Entire blood vessel*).

- Para el concepto *extravascular space*, que se separa *extravascular* y *space*, se obtienen 445 conceptos UMLS, la mayoría procedentes de la segunda subfrase, que tiene un significado muy general y es usado en muchos conceptos. Todos ellos hacen referencia a otras zonas del cuerpo, donde se usa esa palabra, como *Retroperitoneal Space* o *Entire second intercostal space*.

- El concepto *sweat gland cancer* forma un alineamiento complejo con los conceptos *Sweating*, *Malignant Neoplasm* y *Primary malignant neoplasm*.

- En *Named groups of persons*, el concepto *chain saw operator* tiene como subfrase *operator* que obtiene de UMLS decenas de operadores.

En todos los alineamientos que hemos considerado incorrectos, ninguno de los conceptos UMLS obtenidos por el método tenía relación con el concepto Emtree original.

6.6.7. Estudio del recall del método

Como ya hemos visto en el apartado 6.5.7, decidimos usar como estándar una de las terminologías incluidas en el Metathesaurus de UMLS, MeSH, al tener un objetivo y jerarquización similares a Emtree y hemos considerado válido un alineamiento donde, al menos, uno de los conceptos UMLS devueltos pertenezca a MeSH.

Debido a que la revisión manual de todos los alineamientos es una tarea ingente, hemos optado por realizar una selección aleatoria de 400 alineamientos. Los resultados se muestran en la tabla 6.27. Como vimos, están en MeSH el 44.19 % de los alineamientos consideramos correctos por el método, siendo *Organism names* la facet de mayor porcentaje con un 58.33 % y *Physical diseases, disorders and abnormalities* la de menor con un 21.08 %.

Este recall no es alto y las causas son:

- Emtree incluye todos los términos de MeSH, pero además es de mayor tamaño. Emtree tiene unos 56000 términos preferidos, mientras que MeSH tiene menos de la mitad. Teniendo en cuenta los sinónimos, Emtree llega casi a los 300.000 términos, mientras que MeSH tiene unos 160.000.

- Nuestro método usa el tipo de búsqueda NormalizeString de UMLS que devuelve menos resultados que otros tipos a fin de facilitar el tratamiento de los resultados. Por tanto, algunos alineamientos correctos ya no se obtienen en el alineamiento léxico. Sin embargo, esto también permite obtener una precisión muy elevada, al tiempo que se reduce el coste computacional y se simplifica el proceso de desambiguación.

- Nuestro método es muy restrictivo, ya que realiza una desambiguación eligiendo el más similar de ellos en el 75 % de los casos.

- La elección del alineamiento más similar entre los disponibles se realiza en función del factor de similitud, según la mayor similitud de relaciones broader-narrower entre el concepto Emtree y el concepto UMLS. Por lo tanto, la forma en que estas relaciones están construidas produce que, en el 55.81 % de los casos, el concepto de mayor factor de similitud no sea de MeSH.

Por tanto, la evaluación del recall da lugar a unos valores muy por debajo de los valores reales. Pero ante la falta de un procedimiento de evaluación manual mejor, creemos que el método debe contemplar los resultados para el caso peor. Teniendo en cuenta esto y que el método depende del alineamiento léxico inicial, creemos que los resultados globales obtenidos son buenos.

Capítulo 7

Conclusiones

Como hemos visto a lo largo de esta memoria, el alineamiento de terminologías presenta varias dificultades importantes que impiden la creación de métodos automáticos para realizar la tarea. El primero de ellos es el gran tamaño de algunas terminologías, lo que hace que los recursos computacionales necesarios para hacer el alineamiento sean grandes. Otro problema es cierta informalidad en la creación de las propias terminologías, tanto a nivel léxico debido a la ambigüedad del lenguaje natural como a nivel estructural en la creación poco sistemática de relaciones entre los conceptos. Por último, el tercer gran problema es que la necesidad de la participación de expertos para la validación de los alineamientos, lo cual es una tarea imposible en la práctica para terminologías de gran tamaño. De ahí, la necesidad de un método automático.

En esta tesis doctoral, hemos presentado un método semiautomático para la equiparación de terminologías de gran tamaño, y la validación y desambiguación de los alineamientos resultantes. El método incluye la combinación de diferentes técnicas (léxicas, estructurales y de procesamiento de lenguaje natural), con el fin de incrementar la automatización del proceso, a la vez que la precisión global.

Las principales conclusiones extraídas de esta tesis se pueden resumir en los siguientes puntos:

- La combinación de diversas técnicas para el alineamiento de terminologías a gran escala incrementa la precisión, al mismo tiempo que facilita la validación y desambiguación automática del alineamiento.

- El establecimiento inicial de principios bien fundamentados asegura la calidad de la validación y la desambiguación, con la consiguiente mejora en la calidad del alineamiento final.

- El uso de técnicas léxicas permite obtener alineamientos iniciales entre la terminología fuente y la terminología destino, basados en la similitud léxica de los términos que dan nombre a los conceptos. Aunque hoy en día las técnicas léxicas han alcanzado un alto grado de madurez, aún producen alineamientos incorrectos, tanto por homonimia de los términos a equiparar como por pertenencia de éstos a diferente subdominio de conocimiento.

- El empleo de técnicas estructurales permite validar automática y semánticamente los alineamientos léxicos obtenidos inicialmente, descartando aquellos de temáticas diferentes y valorando la similitud de los demás.

- La importancia de la verificación del principio de compatibilidad de las terminologías con el alineamiento resultante, como un procedimiento automático de validación.

- La relevancia de la verificación de la similitud estructural de las terminologías como un procedimiento automático de validación y desambiguación.

- El uso de técnicas de procesado de lenguaje natural en combinación con técnicas léxicas y estructurales permite obtener alineamientos parciales para aquellos conceptos que no obtienen alineamiento directo.

Principales aportaciones

En esta tesis, hemos intentado proporcionar una solución para resolver la falta de métodos automáticos para interpretar y evaluar los alineamientos léxicos entre terminologías de gran tamaño, de forma más eficaz que la revisión manual. Así, las principales aportaciones del método propuesto son:

- Un proceso automático donde, dadas dos terminologías, se obtiene un elevado número de alineamientos correctos sin intervención humana en ninguna fase del proceso.

- Una arquitectura genérica para el alineamiento de terminologías a gran escala. La arquitectura está basada en los recursos disponibles en el momento, tales como las técnicas de equiparación léxicas proporcionadas por el servidor de conocimiento del UMLS o técnicas de procesamiento de lenguaje natural. La arquitectura ha integrado todas estas facilidades para el alineamiento de las terminologías fuentes, con la mínima intervención del usuario final. Además, la arquitectura planteada se diseñó para permitir la adaptación de algunos de los elementos que la

integran con el fin de obtener una mayor precisión y recall del alineamiento resultante.

- Un procedimiento automático de validación de los resultados alcanzados durante la equiparación de las terminologías usando la combinación de técnicas propuestas. El procedimiento permite conocer el nivel de similitud semántica que presentan los alineamientos resultantes, seleccionando los más adecuados. Para ello, usa información semántica presente en las terminologías y parámetros calculables automáticamente como son la similitud de las categorías de alto nivel y los factores de similitud de los conceptos de cada alineamiento.

- Integración de la meta-terminología más importante en el ámbito biomédico, el Metathesaurus de UMLS, permite tener disponible una cantidad ingente de información, que redunda en unos mejores resultados del método.

- Un método para generar alineamientos parciales para conceptos sin alineamiento directo también haciendo uso de la similitud entre el conocimiento estructural de las terminologías, de técnicas de procesamiento de lenguaje natural y del conocimiento sobre las unidades léxicas más frecuentes en las terminologías a equiparar.

- Un método basado en la definición del proceso en bloques computacionales independientes, tal como se explica en el apéndice C, permite adaptar cualquiera de ellos en función de las necesidades particulares. Para el alineamiento con otra terminología diferente, incluso con otro tipo de acceso (no en XML, sino vía Internet, por ejemplo), basta sustituir los módulos iniciales del proceso.

Trabajo futuro

El método propuesto define un entorno mínimo que cumplen la mayor parte de terminologías existentes como puede ser la estructuración en conceptos, la categorización de los mismos y la existencia de relaciones jerárquicas. Estos son los elementos utilizados para validar y desambiguar los alineamientos léxicos iniciales. Por tanto, puede ser aplicado directamente para otras terminologías y en otros ámbitos de conocimiento. Además, si el caso concreto donde nos interesa aplicarlo dispone de características especiales puede ser adaptado.

Diferentes aplicaciones de la solución presentada pueden ser:

- Recuperación de información. El método puede integrarse en una interfaz web y/o a través de una API, a fin de que dada una consulta,

ésta pueda ser procesada semánticamente a fin de buscar con más éxito en las fuentes de información accesibles. El método puede dar resultados clasificados por su similitud y su fiabilidad. Esto puede ser de gran utilidad en el campo de los buscadores semánticos y en la integración de sistemas.

- Extracción de información. Actualmente, existe mucha información disponible en Internet en documentos de texto plano o casi plano (con etiquetas HTML) cuya integración puede ser interesante. A través de técnicas de procesado de lenguaje natural, el método puede ser adaptado para reconocer, analizar e integrar esa información.

- Interoperabilidad entre sistemas de información. Éste es el objetivo general a conseguir. En la práctica, se puede conseguir definiendo servicios web que, usando el método, permitan a sistemas de información diferentes compartir información.

En línea con estas aplicaciones, nos parece de gran interés el uso de SKOS para la representación de los alineamientos, a fin de facilitar su integración posterior en otros sistemas de información globales.

Además, en el caso concreto del ámbito biomédico, se pueden aplicar algunas mejoras en el método, aprovechando particularidades que presentan Emtree y UMLS. Entre ellas, está el uso de otras relaciones entre conceptos disponibles además de las jerárquicas, lo cual puede ampliar los alineamientos validados. También pretendemos valorar los alineamientos complejos BroadMacth indicando cuales de ellos son mejores. Por último, esperamos contar con expertos en el futuro para evaluar el método.

Por tanto, tomando como partida los objetivos alcanzados en esta tesis, se pueden desarrollar aplicaciones y líneas de investigación diferentes a fin de avanzar en un tema de tanta actualidad como es la Web semántica.

Apéndice A

Diccionario

A continuación, relacionamos la descripción de los términos que usamos a lo largo de esta tesis.

- **Alineamiento** es el proceso de búsqueda en la terminología destino de los conceptos que se equiparan al concepto de la terminología fuente que estamos evaluando.
- **Categoría** es una de las nociones abstractas y generales por las cuales los conceptos son reconocidos, diferenciados y clasificados. Con ellos, se pretende una clasificación jerárquica de los conceptos del dominio. Conceptos similares y con características comunes formarán una categoría, y a su vez varias categorías con características afines formarán una categoría superior.
- **Concepto** representa una entidad existente en el ámbito que se quiere modelizar. Un concepto puede ser un objeto real, físico, pero también abstracto o imaginario.
- **Correspondencia** es la relación de similitud léxica o semántica entre un concepto de la terminología origen y un concepto de la terminología fuente.
- **Desambiguación** es el proceso de seleccionar entre varios alineamientos aquel que se considere mejor en función de los parámetros fijados.
- **Equiparación(Matching)** es el proceso de encontrar relaciones o correspondencias entre entidades de diferentes ontologías.
- **Facet** es el nombre que Emtree da a las categorías con las que se clasifican sus conceptos.

- **Grupo semántico** es el nombre que UMLS da a las categorías de nivel superior con las que se clasifican sus conceptos.
- **Alineamiento (mapping)** es el vínculo que se establece, tras el proceso de equiparación, entre un concepto de la terminología fuente y uno o varios conceptos de la terminología destino al que se equipara.
- **Alineamiento complejo** es el vínculo entre un concepto de la terminología fuente y varios de la terminología destino, que entre todos éstos, dan toda la información disponible en la terminología destino sobre el concepto de la terminología fuente.
- **Alineamiento complejo BroadMatch** es un caso particular del anterior, cuando no se encontró ningún concepto en la terminología destino que se equipare léxicamente con el concepto de la terminología fuente y queda equiparado a conceptos más generales derivados de los broaders o de unidades léxicas dentro del término preferido del concepto.
- **Alineamiento complejo UnionMatch** es un caso particular del alineamiento complejo, cuando no es posible encontrar un único concepto en la terminología destino que contenga todo el significado del concepto de la terminología fuente y se dan varios.
- **Alineamiento simple** es el par formado por un concepto de la terminología fuente y un concepto de la terminología destino al que se equipara. Se denomina ExactMatch cuando el proceso de validación y desambiguación ha determinado que es el mejor de los alineamientos disponibles para ese concepto.
- **OpenNLP** es un conjunto de herramientas para el procesado de textos en lenguaje natural, tanto en inglés como en castellano, desarrollados bajo la supervisión de la The Apache Software Foundation.
- **Procesado de lenguaje natural** es una subdisciplina de la Inteligencia Artificial que se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas o entre personas y máquinas por medio de lenguajes naturales.
- **Relación** es el vínculo entre 2 conceptos de una terminología entre los que hay una conexión. En una terminología, las relaciones más frecuentes son las jerárquicas, que definen qué un concepto es superior o padre de otro.

- **Sinónimo** es una relación semántica de identidad o semejanza de significados entre 2 términos. Por tanto, sinónimos son palabras que tienen un significado similar o idéntico entre sí y pertenecen a la misma categoría gramatical.

- **SKOS** (Simple Knowlegde Organization System) es una iniciativa del W3C con el objetivo de mejorar la funcionalidad e interoperabilidad de la Web, que 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, folksonomías y otros vocabularios controlados similares.

- **Subfrase** es cada parte en que se divide un concepto inicial, tras aplicar técnicas de procesado de lenguaje natural y técnicas estructurales.

- **Término** es la representación de un concepto en una terminología. En alguna terminología, se marca como término preferido aquel usado para referenciar el concepto y los demás son sinónimos.

- **Terminología** es un vocabulario especial de una disciplina o un ámbito de conocimiento desarrollado para un determinado objetivo.

- **Unidad léxica** es la agrupación de dos o más palabras, de las que componen un término, que forman una unidad con significado propio, diferente a cada término por separado.

- **URI** o Uniform Resource Identifier (en español, identificador uniforme de recurso) es una cadena de caracteres corta que identifica inequívocamente un recurso (servicio, página, documento, dirección de correo electrónico, enciclopedia, etc.). Normalmente estos recursos son accesibles en una red o sistema.

- **Validación** es el proceso de análisis de un conjunto de mappings para determinar cuales de ellos no son válidos, en nuestro caso, cuales no son similares al concepto al que están equiparados.

Apéndice B

UMLS

En este apéndice, incluidos toda la estructura de la Semantic Network de UMLS, tanto tipos semánticos como relaciones.

B.1. Tipos Semánticos de la Semantic Network

B.1.1. Entity

- Physical Object
  - Organism
    - Plant
      - Alga
    - Fungus
    - Virus
    - Rickettsia or Chlamydia
    - Bacterium
    - Archaeon
    - Animal
      - Invertebrate
      - Vertebrate
        - Amphibian
        - Bird
        - Fish
        - Reptile
        - Mammal
          - Human
  - Anatomical Structure
    - Embryonic Structure

- Anatomical Abnormality
  - Congenital Abnormality
  - Acquired Abnormality
- Fully Formed Anatomical Structure
  - Body Part, Organ, or Organ Component
  - Tissue
  - Cell
  - Cell Component
  - Gene or Genome
- Manufactured Object
  - Medical Device
  - Research Device
  - Clinical Drug
- Substance
  - Chemical
    - Chemical Viewed Functionally
      - Pharmacologic Substance
        - Antibiotic
      - Biomedical or Dental Material
      - Biologically Active Substance
        - Neuroreactive Substance or Biogenic Amine
        - Hormone
        - Enzyme
        - Vitamin
        - Immunologic Factor
        - Receptor
      - Indecator, Reagent, or Diagnostic Acid
      - Hazardous or Poisonous Substance
    - Chemical Viewed Structurally
      - Organic Chemical
        - Nucleic Acid, Nucleoside, or Nucleotide
        - Organophosphorus Compound
        - Amino Acid, Peptide, or Protein
        - Carbohydrate
        - Lipid
          - Steroid
          - Eicosanoid
      - Inorganic Chemical
      - Element, Ion, or Isotope
    - Body Substance

- Food
- Conceptual Entity
  - Idea or Concept
    - Temporal Concept
    - Qualitative Concept
    - Quantitative Concept
    - Functional Concept
      - Body System
    - Spatial Concept
      - Body Space or Junction
      - Body Location or Region
      - Molecular Sequence
        - Nucleotide Sequence
        - Amino Acid Sequence
        - Carbohydrate Sequence
      - Geographic Area
  - Finding
    - Laboratory or Test Result
    - Sign or Symptom
  - Organism Attribute
    - Clinical Attribute
  - Intellectual Product
    - Classification
    - Regulation or Law
  - Language
  - Occupation or Discipline
    - Biomedical Occupation or Discipline
  - Organization
    - Health Care Related Organization
    - Professional Society
    - Self-help or Relief Organization
  - Group Attribute
  - Group
    - Professional or Occupational Group
    - Population Group
    - Family Group
    - Age Group
    - Patient or Disabled Group

B.1.2. Event

- Activity
  - Behavior
    - Social Behavior
    - Individual Behavior
  - Daily or Recreational Activity
  - Occupational Activity
    - Health Care Activity
      - Laboratory Procedure
      - Diagnostic Procedure
      - Therapeutic or Preventive Procedure
    - Research Activity
      - Molecular Biology Research Technique
    - Governmental or Regulatory Activity
    - Educational Activity
  - Machine Activity
- Phenomenon or Process
  - Human-caused Phenomenon or Process
    - Environmental Effect of Humans
  - Natural Phenomenon or Process
    - Biologic Function
      - Physiologic Function
        - Organism Function
          - Mental Process
        - Organ or Tissue Function
        - Cell Function
        - Molecular Function
          - Genetic Function
      - Pathologic Function
        - Disease or Syndrome
          - Mental or Behavioral Dysfunction
          - Neoplastic Process
        - Cell or Molecular Dysfunction
        - Experimental Model or Disease
  - Injury or Poisoning

B.2. Relaciones de la Semantic Network

- isa
  - associated_with
    - physically_related_to
      - part_of
      - consists_of
      - contains
      - connected_to
      - interconnects
      - branch_of
      - tributary_of
      - ingredient_of
    - spatially_related_to
      - location_of
      - adjacent_to
      - surrounds
      - traverses
    - functionally_related_to
      - affects
        - manages
        - treats
        - disrupts
        - complicates
        - interacts_with
        - prevents
      - brings_about
        - produces
        - causes
      - performs
        - exhibits
        - practices
      - occurs_in
        - process_of
      - users
      - manifestation_of
      - indicates
      - result_of
    - temporally_related_to
      - co occurs_with
      - precedes

conceptually_related_to
- evaluation_of
- degree_of
- analyzes
  - assesses_effect_of
- measurement_of
- measures
- diagnoses
- property_of
- derivative_of
- developmental_form_of
- method_of
- conceptual_part_of
- issue_in

Apéndice C

Implementación

Nuestro método ha sido programado en Java, se ejecuta en un ordenador personal sobre Linux, en algunas ocasiones, y Microsoft Windows XP y Vista, en otras ocasiones. Para ello, se han implementado una serie de clases Java que realizan cada paso de este método, tomando de entrada los ficheros XML, uno por facet, con los datos necesarios y devuelven uno o varios ficheros XML, también uno por facet, con el resultado del proceso correspondiente. A continuación, se explican con detalle estos módulos y el modelo de datos que subyace en esta tesis.

C.1. Módulos implementados

Los módulos (clases Java) desarrollados han sido:

- **SeparateFacets**: recibe como entrada el fichero XML con la terminología Emtree, separa los conceptos por la facet a la que pertenecen. Como salida, produce un fichero XML por cada facet.
- **UMLSQuery**: recibe como entrada los ficheros anteriores, se conecta a UMLS usando la API UMLSKS y realiza las consultas para cada término y sinónimo usando el método NormalizeString. Como resultado, genera un fichero XML por cada facet, con todos los alineamientos obtenidos para término y sinónimo, recuperando para cada concepto UMLS su nombre, su identificador (CUI), la definición, el tipo semántico y la lista de narrowers y broaders.
- **ProcessMapping**: recibe como entrada los ficheros con los alineamientos por todos los términos y los agrupa por concepto. Da como resultado un conjunto de ficheros XML, uno por facet, con los alineamientos obtenidos para cada concepto Emtree.

- **PSGFilter**: teniendo como entrada los ficheros anteriores, los explora obteniendo los grupos semánticos más comunes y genera dos ficheros XML por facet, uno con los alineamientos que pertenecen a ese grupo y otro con los que no, que ya son descartados.

- **CompareRelations**: recibe como entrada los ficheros con los alineamientos validados por grupo semántico. Realiza el análisis de las relaciones del concepto Emtree y de los conceptos UMLS obteniendo el factor de similitud de cada alineamiento. También forma los clusters para aquellos conceptos UMLS que los contengan. Tiene como salida un fichero xml con facet con toda esta información.

- **GenerateResult**: Recibe como entrada los ficheros anteriores y aplica el algoritmo de validación. Como resultado, genera dos ficheros por facet, uno con los alineamientos que el algoritmo valida y otro con los que descarta.

- **ComparingStructure**: Recibe como entrada los ficheros con los alineamientos descartados, analiza las relaciones de los conceptos UMLS explorando varios niveles en los broaders y los narrowers y buscando coincidencias con las relaciones del concepto Emtree. De esta forma, recupera los alineamientos para los cuales se encuentran coincidencias. Como salida, genera para cada facet 4 ficheros, con los conceptos que se recuperan y se descartan por broader y por narrower, que luego se unifican en otros dos.

- **NotFound**: recibe como entrada, por cada facet, el fichero con los conceptos Emtree y el fichero con los alineamientos obtenidos y agrupados. Revisa los conceptos Emtree iniciales para determinar cuales de ellos no han obtenido alineamientos y genera como salida los ficheros con esos conceptos.

- **ParsingNF**: analiza los conceptos anteriores, aplicando el método que hemos desarrollado para separar en subfrases. Da como resultado los ficheros con las partes por cada concepto.

- **NotFoundQuery**: Adaptación de la clase UMLSQuery que recibe como entrada los ficheros anteriores y tiene como salida los ficheros con los alineamientos obtenidos para las subfrases.

- **PSGFilterNF**: Adaptación de la clase PSGFilter para filtrar por grupo semántico los alineamientos obtenidos a partir de las subfrases.

- **AnalizeValidMappings**: recibe como entrada, por facet, todos los ficheros con los alineamientos validados en el proceso, que son los resultantes de los módulos **GenerateResult**, **ComparingStructure** y **PSGFilterNF**. Se analizan los alineamientos estudiando su procedencia y construyendo los alineamientos complejos. Genera un fichero por facet con todos los alineamientos validados, incluyendo un campo para indicar su tipo.

Tras la realización de la consulta a UMLS, todos los ficheros tienen una estructura similar, incluyendo para cada mapping el nombre del concepto, su identificador (CUI), el grupo y el tipo semántico y la lista de narrowers y broaders. Como puede verse, al realizar cada módulo una parte del proceso, es posible modificar cualquiera de ellos y ejecutar igualmente el proceso.

C.2. Modelo de datos

El modelo de datos puede verse en la figura C.1. Incluye las siguientes clases (con sus atributos y métodos relevantes), relaciones y restricciones:

- Una **Terminología** tiene una o varias **Categorías** que, a su vez, tienen uno o varios **Conceptos**.
- Un **Concepto** pertenece a una o más **Categorías**
- Un **Concepto** tiene 1 o más **Términos**, uno de los cuales es el término preferido.
- Un **Concepto** tiene 1 o más **Relaciones** con otros **Conceptos**. Una *Relación* siempre se establece entre 2 *Conceptos*.
- Un **Concepto** forma 0 o más **Alineamientos** con **Conceptos** de otras **Terminologías**.
- Un **Alineamiento** relaciona a 2 o más **Conceptos**, donde un **Concepto** de la **Terminología** fuente se equipara a 1 o más **Conceptos** de la **Terminología** destino
- Un **Alineamiento** es simple cuando vincula a un **Concepto** de la **Terminología** fuente con uno y sólo un **Concepto** de la **Terminología** destino.

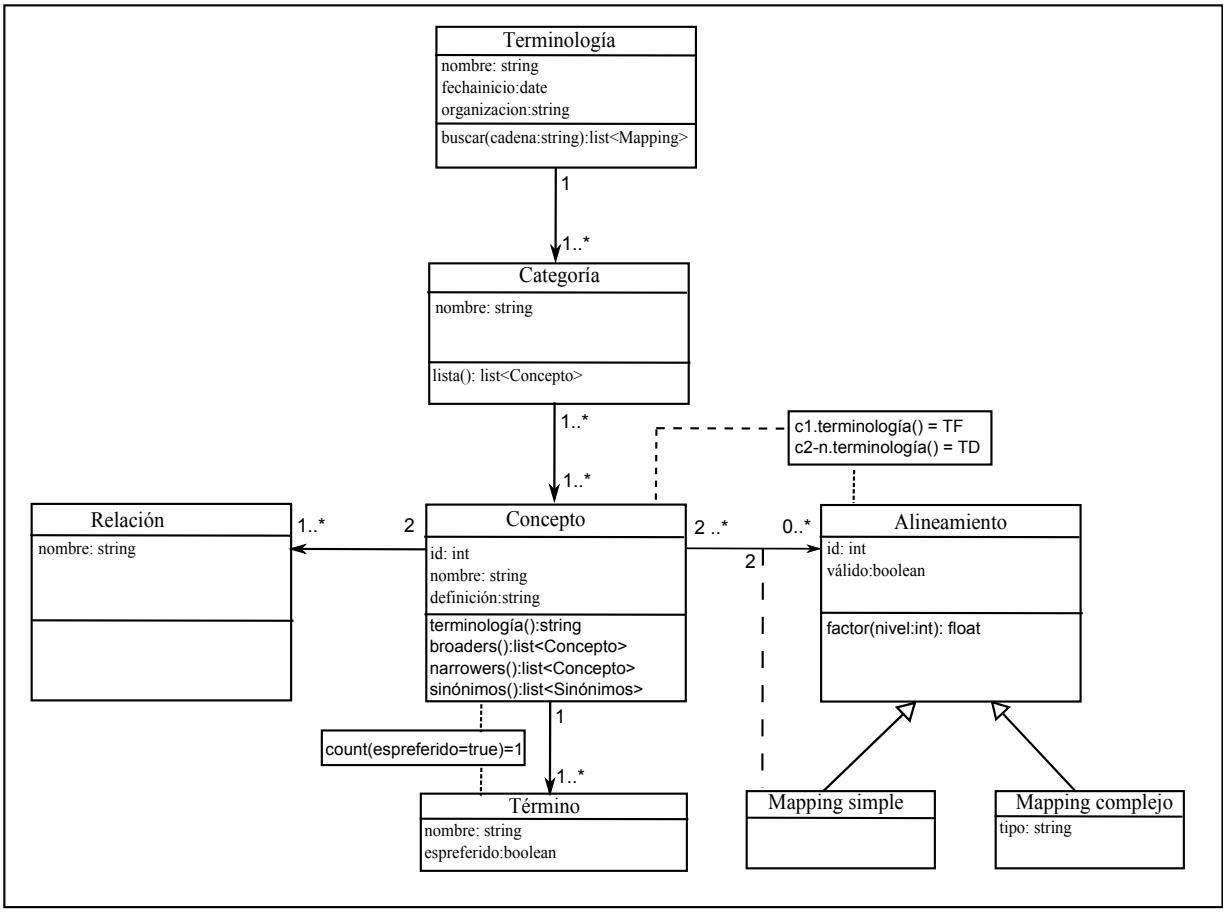

Figura C.1: Modelo de datos

Bibliografía

[aHBM93] D.A. Lindberg ahd Humphreys BL and A.T. McCray. The unified medical language system. *Methods Inf Med*, 4(32):281–291, August 1993.

[Bet05] National Library of Medicine Bethesda. Unified medical language system umls. 2005. `http://umlsinfo.nlm.nih.gov/`.

[Bod04] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, (32):267–270, January 2004.

[Bur06] Anita Burgun. Desiderata for domain reference ontologies in biomedicine. *Journal of Biomedical Informatics*, 39:307–313, June 2006.

[Cim98] J. J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, 37(4-5):394–403, November 1998.

[CL03] M. N. Cantor and Y. A. Lussier. Putting data integration into practice: using biomedical terminologies to add structure to existing data sources. *AMIA Annu Symp Proc.*, pages 125–129, 2003.

[CSG+03] M.N. Cantor, I.N. Sarkar, R. Gelman, F. Hartel, O. Bodenreider, Y.A. Lussier, and Y. A. Lussier. An evaluation of hybrid methods for matching biomedical terminologies: Mapping the gene ontology to the umls. *Stud Health Technol Inform*, (95):62–67, 2003.

[DNH04] AnHai Doan, Natalya F. Noy, and Alon Y. Halevy. Introduction to the special issue on semantic integration. *SIGMOD Rec.*, 33(4):11–13, 2004.

[Doe01] M. Doerr. Semantic problems of thesaurus mapping. *Journal of Digital Information*, 1(8), 2001.

[ECH+94] D. A. Evans, J. J. Cimino, W. R. Hersh, S. M. Huff, and D. S. Hell. Toward a medical-concept representation language. the canon group. *Methods Informatical Medical*, (1):207–217, November 1994.

[ES07] Jerome Euzenat and Pavel Shvaiko. *Ontology Matching*. Secaucus, NJ, USA, 2007.

[FBA+07] K. W. Fung, O. Bodenreider, A. R. Aronson, W. T. Hole, and S. Srinivasan. Combining lexical and semantic methods of inter-terminology mapping using the umls. *Stud Health Technol Inform*, 19:605–609, 2007.

[FCJ94] C. Friedman, J. J. Cimino, and S. Johnson. A schema for representing medical language applied to clinical radiology. *JAMIA*, 1(3):233–248, 1994.

[HGH+09] K. Huang, J. Geller, M. Halper, Y. Perl, and J. Xu. Using wordnet synonym substitution to enhance umls source integration. *Artificial Intelligence Medical*, 46(2):97–109, 2009.

[HOTO04] T. Hishiki, O. Ogasawara, Y. Tsuruoka, and K. Okubo. Indexing anatomical concepts to omim clinical synopsis using the umls metathesaurus. *Silico Biol.*, 4(1):31–54, 2004.

[IHT] IHTSDO. Snomed ct. `http://www.ihtsdo.org/`.

[Int85] International. *ISO 5964: Guidelines for the establishment and development of multilingual thesauri*. British Standards Institution, London, 1985.

[Iso86] Iso. *ISO 2788:1986, Documentation – Guidelines for the establishment and development of monolingual thesauri*. Multiple. Distributed through American National Standards Institute (ANSI), 1986.

[KS03] Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18(1):1–31, 2003.

[MBB01] A. T. McGray, A. Burgun, and O. Bodenreider. Aggregating umls semantic types for reducing conceptual complexity. *Proceedings of Medinfo 2001*, 10(1):216–220, 2001. `http://www.ncbi.nlm.nih.gov/pubmed/11604736`.

[Nat05] National. ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. page 184, 2005.

[Noy04] Natalya F. Noy. Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, 33:204, 2004.

[oMa] National Library of Medicine. Blast basic local alignment search tool. `http://blast.ncbi.nlm.nih.gov/`.

[oMb] National Library of Medicine. Umls reference manual: Metathesaurus. `http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls&part=ch02`.

[oMc] National Library of Medicine. Umls reference manual: Semantic network. `http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls&part=ch05`.

[PS] Juan Antonio Pastor Sánchez. Diseño de un sistema colaborativo para la creación y gestión de tesauros en internet basado en skos. Tesis doctoral presentada en la Universidad de Murcia el 30-03-2009.

[Rec98] A. L. Rector. Thesauri and formal classifications: terminologies for people and machines. *Methods Inf Med*, (37):4–5, November 1998.

[RMJ+08] S. T. Rosenbloom, R. A. Miller, K. B. Johnson, P. L. Elkin, and S. H. Brown. A method for evaluating interface terminologies. *JAMIA*, 15(1):65–76, November 2008.

[SCG+03] I. N. Sarkar, M. N. Cantor, R. Gelman, F. Hartel, and Y. A. Lussier. Linking biomedical language information and knowledge resources. *GO and UMLS. PSB 2003*, pages 439–450, 2003.

[SCK+05] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Kohler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46, 2005.

[SR06] Julian Seidenberg and Alan Rector. Web ontology segmentation: analysis, classification and use. *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 13–22, 2006.

[SS06] J. Y. Sun and Y. Sun. A system for automated lexical mapping. *J Am Med Inform Assoc.*, 13(1):334–343, 2006.

[VGHHT04] D. Vizine-Goetz, C. Hickey, A. Houghton, and R. Thompson. Vocabulary mapping for terminology services. *Journal of Digital Information*, 2004.

[W3C] W3C. Skos simple knowledge organization system reference. `http://www.w3.org/2004/02/skos/`.

[WPJ+] Jan Walker, Eric Pan, Douglas Johnston, Julia Adler-Milstein, David W. Bates, and Blackford Middleton. The value of health care information exchange and interoperability. `http://content.healthaffairs.org/cgi/content/full/hlthaff.w5.10/DC1`.

[Yu06] Alexander C. Yu. Methods in biomedical ontology. *Journal of biomedical informatics*, 39(3):252–266, 2006.

[ZB06] Songmao Zhang and Olivier Bodenreider. Aligning multiple anatomical ontologies through a reference. 2006.

[ZB07] S. Zhang and O. Bodenreider. Experience in aligning anatomical ontologies. *International Journal on Semantic Web and Information System*, 3(2):1–26, 2007.

[ZC04] Marcia Lei Zeng and Lois Mai Chan. Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology*, 55(5):377–395, 2004.

[ZMBB07] Songmao Zhang, Peter Mork, Olivier Bodenreider, and Philip A. Bernstein. Comparing two approaches for aligning representations of anatomy. *Artificial Intelligence in Medicine*, 39:227–236, March 2007.